University of California
Santa Barbara

Imperfect Label Information in Multimodal Human-Centric Machine Learning

A Dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Yi Ding

Committee in charge:

Professor Tobias H. Höllerer, Chair
Professor Matthew Turk
Professor William W. Wang
Professor Elizabeth Belding
Professor Barry Giesbrecht

September 2022

The Dissertation of Yi Ding is approved.

Professor Matthew Turk

Professor William W. Wang

Professor Elizabeth Belding

Professor Barry Giesbrecht

Professor Tobias H. Höllerer, Committee Chair

September 2022

Imperfect Label Information in Multimodal Human-Centric Machine Learning

by

Yi Ding

To family, friends, and all the people in my life.

Acknowledgements

I usually have no trouble thinking of how to thank people for all of their help. However, when writing the acknowledgements section of this thesis, I find that I am at a loss for how I can express this gratitude. This Ph.D. has been seven years in the making, during which time I have met so many wonderful peers, colleagues, mentors, mentees, and friends. Each person played a pivotal role in helping me get to today and has helped me in difficult to explain ways. So please excuse my lack of eloquence when expressing these acknowledgements.

First and foremost, I would like to thank my mentor and advisor Tobias. Without his guidance, encouragement, and leadership, this Ph.D. would have happened more than two years later, if at all. He taught me how to question and analyze the fundamental processes of science and to dive into the uncertainty with confidence and optimism. I would also like to extend my appreciation to my committee members, Matthew, Barry, Elizabeth, and William for their advice, support, and encouragement right when I needed them. Additionally, I am deeply grateful to Tonja, Pradeep, and Jennifer for their mentorship.

I would also like to express gratitude to all my collaborators and members of the FourEyes Lab. I feel so fortunate to have found such a wonderful lab filled with supportive colleagues and friends. To Brandon, Alex, Noah, Radha and Aiwen: thank you for your support in making many ideas come to life. Thank you Ehsan, You-Jin, and CY for your insightful questions and contributions in all our discussions. Thank you to all the students I've worked with through the research mentorship program, Kento, Mason, Jacob, Vrish, and many others! I hope that I taught you as much as you taught me. Also, a big thank you Lina for supporting me through the RMP program.

Friends are among the most important things in my life. One of the many reasons is that they help you through the challenging times (like during a Ph.D.). Thank you

Dennis, Jeemin, and Eddie for our discussions off the golf course. Thank you Andrew, Pranthik, Ben, and Jess for visiting and supporting me. Thank you Xin, Jinjin, Naveen, and countless others for our shared experiences over the last many years.

Lastly, I would like to thank my family. Thank you mom and dad for always challenging my decisions, supporting me through my difficult times, and celebrating my successes. To my brother Nick, thank you for making me be a good brother. To Nicole, my wife, your support, companionship, and love is invaluable.

Curriculum Vitæ

Yi Ding

Education

2022	M.S./Ph.D. in Computer Science, University of California, Santa Barbara.
2011	B.S. in Computer Science, University of Massachusetts, Amherst.
2011	B.S. in Mathematics, University of Massachusetts, Amherst.

Publications

- **Yi Ding**, Jacob You, Jennifer Jacobs, Tonja Machulla, and Tobias Höllerer. "Impact of Demographics on Sentiment Labeling." To appear at Computer-Supported Work and Social Computing (2022).
- **Yi Ding***, Alex Rich*, Mason Wang, Noah Stier, Matthew Turk, Pradeep Sen, and Tobias Höllerer. "Sparse Fusion for Multimodal Transformers." Under Review at the Winter Conference on Applications of Computer Vision (2023, *Equal Contribution).
- **Yi Ding***, Kento Nishi*, Alex Rich, and Tobias Höllerer. "Augmentation Strategies for Learning with Noisy Labels." 2021 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR 2021, *Equal Contribution).
- Kento Nishi, **Yi Ding**, Alex Rich, and Tobias Höllerer. "Improving Label Noise Robustness with Data Augmentation and Semi-Supervised Learning (Student Abstract)." (2021).
- **Yi Ding**, Radha Kumaran, Tianjiao Yang, and Tobias Höllerer. "Predicting Video Affect via Induced Affection in the Wild." In Proceedings of the 2020 International Conference on Multimodal Interaction, (ICMI 2020).
- Anish Kachinthaya, **Yi Ding**, and Tobias Hollerer. "Exploring the Benefits of Depth Information in Object Pixel Masking (Student Abstract)." Proceedings of the AAAI Conference on Artificial Intelligence, (AAAI Student Abstracts 2020).
- Virshab Krishna, **Yi Ding**, Aiwen Xu, and Tobias Höllerer. "Multimodal biometric authentication for VR/AR using EEG and eye tracking." In Adjunct of the 2019 International Conference on Multimodal Interaction (ICMI LBR 2019).
- **Yi Ding**, Brandon Huynh, Aiwen Xu, Tom Bullock, Hubert Cecotti, Matthew Turk, Barry Giesbrecht, and Tobias Höllerer. "Multimodal Classification of EEG During Physical Activity." In 2019 International Conference on Multimodal Interaction (ICMI 2019).

Patents

- Pillai, Biju Balakrishna, Kenneth Mark Karakotsios, Peter Cheng, David Wayne Stafford, Stephen Vincent Mangiat, and **Yi Ding**. "Wirelessly preparing device for high speed data transfer." U.S. Patent 10,237,329, issued March 19, 2019.
- Karakotsios, Kenneth Mark, Stephen Vincent Mangiat, Peter Cheng, and **Yi Ding**. "Controlling content zoom level based on user head movement." U.S. Patent 10,585,485, issued March 10, 2020.
- Pillai, Biju Balakrishna, Kenneth Mark Karakotsios, Peter Cheng, David Wayne Stafford, Stephen Vincent Mangiat, and **Yi Ding**. "Transmitting content to kiosk after determining future location of user." U.S. Patent 9,940,583, issued April 10, 2018.
- Bell, Matthew Paul, Peter Cheng, Stephen Michael Polansky, Amber Nalu, Alexander Li Honda, **Yi Ding**, David Wayne Stafford, and Kenneth Mark Karakotsios. "One-handed zoom." U.S. Patent 10,019,140, issued July 10, 2018.
- **Ding, Yi**, Stephen Vincent Mangiat, Peter Cheng, Kenneth Mark Karakotsios, Steven Michael Sommer, and Peter Andrew Schiller. "Approaches for controlling a computing device based on head movement." U.S. Patent 9,665,249, issued May 30, 2017.
- Cheng, Peter, Steven Scott Noble, Matthew Paul Bell, **Yi Ding**, Stephen Michael Polansky, and Alexander Li Honda. "Machine-learning based tap detection." U.S. Patent 9,235,278, issued January 12, 2016.
- Pillai, Biju Balakrishna, Kenneth Mark Karakotsios, Peter Cheng, David Wayne Stafford, Stephen Vincent Mangiat, and **Yi Ding**. "Kiosk Providing High Speed Data Transfer." U.S. Patent Application 14/498,270, filed March 31, 2016.

Professional Experience

- **Virtual Power Systems**, 2014-2015
- **Amazon**, 2013-2014
- **Cisco Systems**, 2011-2013

Abstract

Imperfect Label Information in Multimodal Human-Centric Machine Learning

by

Yi Ding

Multimodal machine learning studies the ability to take multiple streams of input data to make predictions on an output. The classic notion is that by using multiple streams of input, we can make better predictions by accounting for multiple contexts. Such applications include audio-visual speech recognition, emotion prediction, and much more. While this research has enabled novel and effective ways to fuse the data for improved modeling performance, few works have examined how highly uncertain and varied human opinions and behavior can impact model performance.

Accounting the variability or differences in human opinions is important for multimodal machine learning because in many human-centric applications the labels contain high degrees of uncertainty. One notable example of this is in predicting human sentiment or emotions. In current datasets, we do not get a complete picture for the variability of human opinions. This is further complicated by the fact that the inclusion of additional modalities leads to an increase in discriminating features, causing models to fit to imperfect data faster.

This thesis lays a foundation for examining the effect of label variability on multimodal algorithms and datasets. We propose and develop novel techniques for unimodal label tolerance and strive to bring this to a multimodal domain. The goal is that by explicitly accounting for ambiguities in the output, we can improve the effectiveness and understanding of label noise in a multimodal domain.

Contents

Chapter 1

Introduction

This thesis is primarily motivated by the desire to understand human emotions and behavior in relationship to machine learning. In contrast to simply pursuing Artificial Intelligence (AI) techniques which attempt to replicate human capabilities such as sight, sound, or speech, we are interested in how these techniques can be incorporated into understanding how humans feel, think and act. The reason for this is two-fold: 1) it can enable a better understanding of ourselves, and 2) it can enable far richer interaction experiences.

In the last few years, there has been an incredible development in the capabilities of AI to replicate human behavior. And, on some benchmarks such as image recognition or natural language understanding, AI can actually surpass human capability. A well-known application is AlphaGo, an AI who is arguably the strongest Go player in history. However, despite all these developments, most AI applications have yet to become effectively integrated into society. And even with services that are becoming increasingly available, such as Chatbot assistants, something in the technology *feels* amiss. We believe that at least some of the current limitations are due to the lack of understanding of the behavioral and psychological properties of humans, both by us and by AI.

> Human-sourced data that depends on human interpretation and opinion is inherently imperfect; understanding this pattern and improving the handling of this uncertainty, first in unimodal scenarios and then extended to multimodal human signals, are critical for effective and natural integration of AI into society.
>
> THESIS STATEMENT

To tackle these issues, we investigate two questions that we believe are critical components for solving this problem:

1. How do we improve multimodal modeling of human-centered data? (Part I)
2. How do we address the variability and imperfection of human-centered data? (Part II)

We believe that these issues are core components towards enabling better understanding of human behavior by an AI because 1) human interaction is naturally multimodal, and 2) humans are inherently different from each other. By gaining a better understanding of these issues, as well as how they inter-relate, we can make critical strides towards enabling AI to help humans more effectively.

1.1 Multimodality and Variability

We experience and interact with the world through our five senses: sight, sound, taste, touch, and smell. The human brain is incredibly good at processing all of this information, paying attention only to the few things that matter. Imbuing a computer with the ability to process multimodal data effectively is highly desirable because it would enable a vast array of multi-sensory applications. We begin this thesis by investigating how multimodal

models can be applied on human-centric data and improved. In particular, we focus on applications that contain elements of motion and mobility. This is because we would like interaction with an intelligent multimodal agent to occur anywhere and at any time.

The second issue that lies at the heart of this thesis is regarding variability. While groups of humans often behave according to some broad average pattern, there are large individual variances that can make it difficult to model expected behavior and emotions, as well as interaction processes. As such, we should account for the inaccuracies in human perceptions when building multimodal systems. A classic example for the importance of studying multimodal machine learning is presented by the McGurk effect [1]. In this situation, a slight incongruence in auditory and visual stimuli can cause a human to perceive drastically different things. Once such example, is the perception of /da:/ when hearing the syllable /ba:/ but watching the lips of a person saying /ga:/.

In re-examining the McGurk effect, we pose the question: should the label in such a situation be /da:/, /ba:/ or /ga:/? In such situations, we argue, the true label would depend on how the labeling process was conducted and the specifics of what the data collection parameters were. However, in reality, datasets typically are not collected with such clear distinctions as we neither fully understand the labeling effects nor ambiguities and differences in opinion. Sometimes, it might even be infeasible to do so. For example, more complex datasets involving an application such as emotion recognition can depend on body movements, tonal changes, speech cadence, audio, visual, language cues and many more as input for prediction. These inputs are used as a whole and assigned an emotion score by humans which may have different interpretations of the input based on a range of factors such as culture, experiences, gender, environment, and much more. Under such a scenario, providing a comprehensive way to address the labeling uncertainties appears to be an impossible task. In this thesis we attempt to dissect these issues of multimodality and variability, and attempt to examine the impact of these effects.

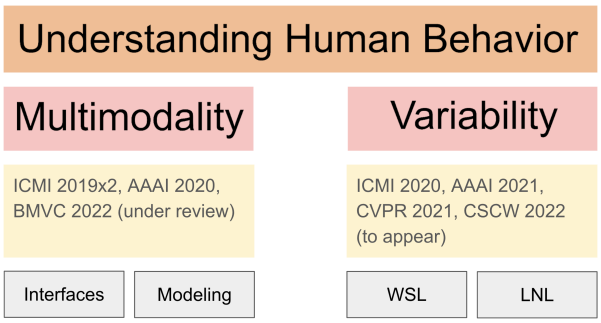


Figure 1.1: Diagram of the contributions made by this thesis. Part I discusses the contributions for multimodality which covers topics on novel interfaces and multimodal modeling. Part II discusses the contributions on variability, covering topics on weakly-supervised learning and learning with noisy labels.

It should be noted that we consider noise and differences of opinion to be synonymous when in fact they are not. In other words, we consider a reliable annotation to be valid when modeling this type of data. There are two reasons that we do this. First, quantifying the validity [2] of results for subjective data is difficult. While current datasets account for reliability via agreement scores, it is much more difficult to assess validity as there is no quantitative metric for it. Secondly, to us, it is the first logical step towards enabling the modeling of human variability due to existing datasets and models. We anticipate future work built on this thesis will separate these two concepts.

1.2 Contributions

This thesis contributes research on modeling techniques aimed at tackling uncertainties on labels in a unimodal setting and on incorporating this into multimodal settings. We first examine problems involving multimodal inputs (Part 1). We find that there is a lack of clarity on how to incorporate techniques which tolerate high degrees of uncertainty into multimodal settings. This topic is highly important, as many problems that deal with multimodal inputs, such as modeling human behavior, are affected by high degrees

of uncertainty. Motivated by this, we examine problems in multimodal settings that expose high degrees of uncertainty, and how uncertainty is typically addressed. We then explore the role of human uncertainty in the setting of multimodal sentiment analysis bring the two lines of work together. We present this thesis to gain an understanding for how we can build smarter systems while relying on less human input. A conceptual diagram of these contributions is provided in Figure 1.1. The following summarizes the contributions in each chapter:

1. We start by exploring a few applications of multimodal machine learning in Chapter 2. We discuss our explorations using EEG as a modality as well as RGB depth prediction. We present a novel multimodal approach for classifying the P300 event related potential (ERP) component by coupling EEG signals with nonscalp electrodes (NSE) that measure ocular and muscle artifacts.

2. In Chapter 3, we present Sparse Fusion Transformers (SFT), a novel multimodal fusion method for transformers that performs comparably to existing state-of-the-art methods while having greatly reduced memory footprint and computation cost.

3. We begin our examination of uncertainty in multimodal learning in Chapter 4 by examining the potential to use *unlabeled* public reactions in the form of textual comments to aid in classifying video affect. We examine two popular datasets used for affect recognition and mine public reactions for these videos. We learn a representation of these reactions by using the video ratings as a weakly supervised signal.

4. In Chapter 5, we present an algorithm that obtains state-of-the-art performance on multiple benchmarks on the learning with noisy labels problem. We achieve this by proposing a novel way to inject augmentations into the learning pipeline. We

find that using one set of augmentations for loss modeling tasks and another set for learning is most effective. We use this as a grounded method to examine methods to improve uncertainty handling in human-centric data.

5. Lastly, in Chapter 6 we examine how label uncertainty finds its way into human-centric datasets and its impact. In particular, we ask >1000 crowdworkers to provide their demographic information and annotations for multimodal sentiment data and its component modalities. We show that demographic differences among annotators impute a significant effect on their ratings, and that these effects also occur in each component modality.

Part I

Multimodality

Chapter 2

Applications of Multimodal Learning

Parts of the contents of this chapter were published with multiple collaborators at the International Conference on Multimodal Interaction (ICMI) 2019 and ICMI Late-Breaking Results 2019.

This chapter discusses some common applications of multimodal machine learning and how they can be improved. It explores some of the existing works and motivates our work in label noise tolerance. We examine the ERP classification problem as it relates to Brain-Computer Interfaces in a grounded evaluation. We then examine a way in which such a BCI scheme can be used as an easy way for user authentication. This chapter aims to provide a few examples of how data coming directly from a human can be modeled.

2.1 Multimodal EEG Classification Under Motion

Brain Computer Interface (BCI) systems enable the control of a computer through brain signals [3]. Traditionally, BCIs have been utilized as an assistive technology for people with mobility impairments [4]. There is however a growing interest in general purpose, non-invasive BCI technologies to improve the computing experience of perfectly

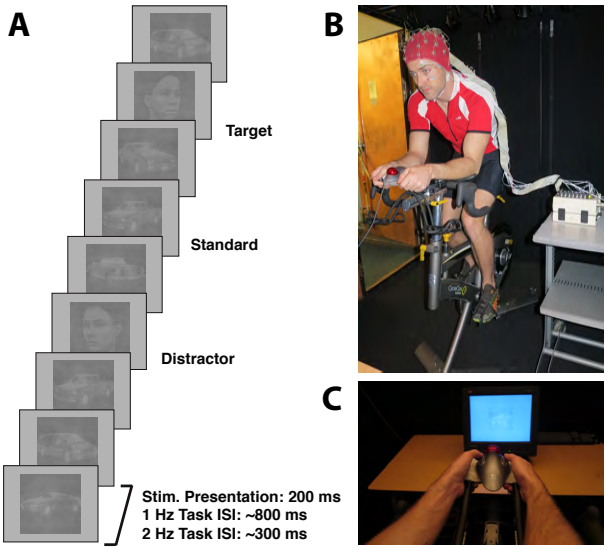


Figure 2.1: Methods and Tasks. (A) Example of the oddball task. Participants were required to detect targets (right oriented faces) in a stream of distractors (left oriented faces) and standards (cars oriented left or right). (B) The participant was fitted with an EEG cap and positioned on a stationary bike. (C) The participant rested their elbows on a pair of "aero bars" attached to the bike handlebars and used their right thumb to respond to targets.

healthy people. A number of consumer facing products have been developed such as EEG headsets by Emotiv, the Muse meditation headband, and an EEG-integrated virtual reality headset by Looxid Labs. These products promise to enhance the computing experience by enabling intelligent interfaces that sense and react to changes in the user's cognition.

Recent work from neuroadaptive systems have explored the use of these brain signals (EEG) as an additional input modality for a wide variety of interaction tasks. A central idea is to use EEG information for user modeling to build adaptive interfaces [5, 6] that can implicitly and quickly react to user state [7, 8]. This information can be used to quantify user states such as cognitive load [9, 10], emotion [11, 12], and attention [13, 14] to inform better interaction experiences for education [15], entertainment [16], equipment operation [17], and others [18, 19, 20, 21]. Nonetheless, an open research question remains as to how these signals should be integrated and how reliable they are

for non-trivial computing applications.

The key advantage of these technologies is that they opens the door to real Ubiquitous Computing, where computing may occur in any time or place [22]. In ubiquitous computing, users may be interacting with the system while moving around in their environment and engaging in physical activity (e.g. an augmented reality task). However, EEG signals are commonly known to be severely impacted by a wide range of biophysiological artifacts associated with movement. To ensure that the system remains usable, it is crucial to understand how classification performance of EEG signals changes under these adverse conditions, so that we can develop techniques for robust classification and analysis.

Typical BCI solutions are based on laboratory studies and rarely replicate the conditions outside the lab in which the system should be deployed. There are three main ways to tackle such a problem: 1) to extract features that are invariant to the expected noise, 2) to denoise the signal, and 3) to be robust to the noise. An extensive amount of data collection, manual feature extraction, and domain knowledge is typically necessary to identify, classify, and correlate these signals to a particular application [23]. A number of techniques have been used to alleviate the need for manual feature extraction, including spatial and temporal filtering, and neural networks [24].

Due to the success of Deep Learning models in other fields such as Computer Vision and Speech Recognition in which the performance reaches human-like levels of performance, there has been a resurgence in the use of deep neural networks for feature extraction and classification of EEG signals [25]. However, deep learning methods require a large amount of training examples to be successful in order to model the variability that exists across examples [26], which is not typically the case for BCI data. First, the EEG datasets have a low number of examples per class compared to typical computer vision problem, and they have unbalanced datasets, in particular for event-related potential

(ERP) based BCI in which the target class has a low probability. Second, EEG is highly non-stationary and characteristics of the signal can change depending on the behavioral state of the wearer (e.g. fatigue/arousal). These effects can be partially remedied through the use of data augmentation [27, 28, 29], but the applications of these techniques have not been well studied for EEG signals.

To help address these issues, we introduce a dataset in which participants were positioned on a stationary bike and engaged in a visual three-stimulus oddball task [30] while at rest and during bouts of low- and high- intensity cycling exercise [31]. We open source[1] this dataset with the goal of encouraging further research. We investigate whether classification performance of a state-of-the-art deep learning model suffers under different intensity levels of physical activity, and discover that it does, suggesting room for improvement of feature representation. We propose a model to improve performance by incorporating the use of a denoising autoencoder. Furthermore, we consider the addition of signals from nonscalp electrodes and user state data, to provide supplementary information.

2.2 Related Work

2.2.1 Classification Methods

Lotte et al. [32, 24] has provided a review of classification algorithms for EEG-based BCI. They concluded that the current state-of-the-art is Riemannian Geometry (RG) classifiers, and suggested it is time to move away from classical approaches which usually use Linear Discriminant Analysis (LDA) with Common Spatial Pattern (CSP) filters. In fact, the winning approach for the Kaggle BCI competition at NER 2015 used xDAWN

[1] https://github.com/yding37/mcann

spatial filters with RG [33].

Deep learning models have been applied to EEG classification since at least 2008 [34] and there has been a sharp increase in activity thanks to the recent success of these models in the natural language processing and computer vision domains. The best performing deep learning approach is currently Convolutional Neural Networks (CNN), which are able to borrow technical advancements from the computer vision community. CNNs were first used for EEG classification by Cecotti et al. in 2011 [35, 36] for P300 ERP classification. Schirrmeister et al. [25] conducted an in-depth survey of CNN architectures for EEG classification and provided an open source software library for evaluating them.

Recently, the US Army Research Lab released EEGNet, a CNN architecture that reached performance comparable to the state-of-the-art on 4 different BCI tasks [37]. The authors used depthwise and separable convolutions which helped to reduce the amount of trainable parameters in the model [38]. Our model uses a variant of EEGNet as the basis for our encoder. By applying EEGNet within an autoencoder paradigm, we are able to learn a more robust representation of the EEG data.

Although not as common, autoencoder methods have been explored in a few EEG classification studies. In Yin and Zhang[39], the authors utilized a Stacked Denoising Autoencoder (SDEA) in order to classify mental workload. We also utilize a denoising autoencoder, but we target the P300 signal which is better characterized and understood [40]. The authors also compared computation time and concluded that their SDEA model could be used for online classification, which supports our intended use case of ubiquitous BCI.

Said et al. [41] used a multimodal fusion approach with stacked autoencoders coupling EEG and EMG data. However, their approach utilized two separate pathways for EEG and EMG and learned a latent vector representation of the data. However, their network did not show improvements over CNN based approaches. In our approach, we share the

weights of the encoder for both EEG and NSE and fuse their latent representations.

2.2.2 EEG During Physical Activity

EEG recordings are known to suffer from motion artifacts, as they simply measure the electrical signals in the brain. Moreover, the activation of muscles produces large electrical signals and is the main source of noise in many studies. Participants are usually trained to stay completely still in order to minimize contamination of the dataset. For this reason, there are relatively few datasets where EEG is actually recorded while under motion.

Nathan and Contreras-Vidal [42] collected EEG recordings of subjects walking at 3 different speeds on a treadmill. Contrary to expectations, they did not observe significant contamination of the EEG signal by motion artifacts. However, it should be noted that the fastest speed investigated (4.5 km/h) is still less than the preferred walking speed of an average person [43], making it difficult to extrapolate to Ubiquitous BCI settings.

A handful of studies have collected EEG data during acute bouts of exercise. Yagi et al. [44] and Grego et al. [45] both measured EEG with a P300 task while cycling. More recently, other studies have investigated the impact of acute exercise on other types of brain responses, such as orientation-selective responses in visual cortex [46] and neural oscillatory activity associated with inhibitory control [47]. For a comprehensive summary of sport and exercise related EEG studies, see Cheron et al. [48]. However, to the best of our knowledge, the present study is the first to apply classification methods to EEG collected during acute bouts of physical activity, with the goal of improving the efficacy of ubiquitous BCI.

Artifact removal is another common practice for removing sfig/ources of interference such as muscle activity. Gwin et al. [49] compare artifact removal methods for EEG

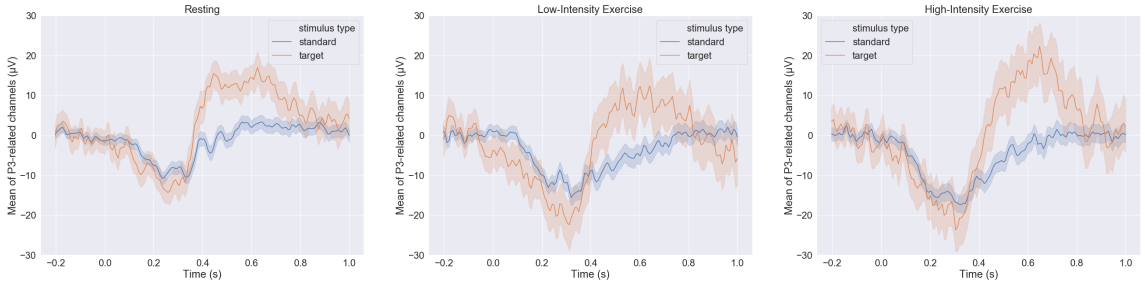


Figure 2.2: Single subject (sj04) ERPs are shown for each of the physical activity conditions. To avoid visual clutter, only ERPs generated from the standard and target conditions are shown. Error bars represent ± standard error of the mean (SEM)

Dataset	# Samples	Ratio
BCI-2a [51]	2.5K	1:1
Kaggle [52]	8.8K	1:1
P300 [37]	30K	1:5.6
Bike (ours) [31]	72K	1:1:8

Table 2.1: Our dataset is the largest among current publicly available datasets for BCI tasks of similar purpose.

collected while walking or running. General guidelines and good practice for artifact removal can be found in Urigüen and Garcia-Zapirain [50]. Ultimately, given that we are interested in online classification, we chose not to perform any motion artifact removal. Instead, we perform minimal pre-processing on the datasets, which we describe in the next section.

2.3 Description of Dataset

2.3.1 Task and Exercise Protocol

The EEG dataset used in this work was previously described in Bullock et al. [31]. Twelve adult student volunteers took part in the study in exchange for course credit or financial compensation. Figure 2.1 provides an overview of the methodology used for collection. Each participant performed two different versions of a three-stimulus oddball

task [30] while seated on a stationary bike. Participants were required to respond to target stimuli (left-oriented faces) and ignore the distractor stimuli (right-oriented faces) and the standard stimuli (cars oriented either to the left or right). The ratio of targets to distractors and standards was 1:1:8, respectively. In the two different versions of the task the stimuli were presented at different rates. Stimuli were either presented at 1 Hz (200 ms stimulus presentation with 800 ms inter-stimulus intereval (ISI) or 2 Hz (200 ms stimulus presentation with 300 ms ISI). The 2 Hz data were collected for the purpose of a BCI study and were not reported in the original paper.

Participants completed the 1 Hz and 2 Hz tasks at rest (sat on the bike but not pedaling), during low-intensity exercise (pedaling at at a very light resistance level of 40W) and during high-intensity exercise (pedaling at a resistance level which the participant reported to be "somewhat hard" according to their Rating of Perceived Exertion (RPE; Borg 1970 [53].)). The order of completion was counterbalanced between participants.

EEG data were recorded continuously during each task using a BioSemi Active Two System consisting of 32 scalp electrodes arranged in an elastic cap (Electro-Cap, OH, USA) and placed in accordance with the 10-20 system. Additional non-scalp electrodes (NSE) were fixed to the right and left mastoids, 1 cm lateral to the left and right canthi (horizontal EOG), above and below each eye (vertical EOG) and on the right and left trapezius muscles (EMG).

2.3.2 Classification Goals and Challenges

Here, the goal of the classifier was to determine which stimulus (target, distractor or standard) the participant viewed for each trial. The inclusion of two physical activity conditions sets this dataset apart from typical P300 datasets. This dataset fits with our goal of building BCI paradigms for ubiquitous computing, because compared to

other P300 datasets, the conditions in this task are more similar to those that might be encountered in real life. In traditional P300 EEG data collection, participants typically sit in a comfortable position and are told to minimize non-task related physical motion. However, for BCIs to be useful in real life, the actions of the user should not be controlled. In contrast, this P300 dataset incorporates physical exercise, an indispensable part of day-to-day life. Therefore, achieving a good classification accuracy on this dataset is a first step to building a useable BCI for everyday activities.

The inclusion of physical exercise introduced extra noise, which makes the classification task more difficult. The presence of extra noise under physical exercise is visualized in the error bands in Figure 2.2. Here, we identify at least three sources of noise which may not be present in other existing P300 datasets. The first is EMG noise. EMG activity can be present in the range 10 - 250 Hz, which overlaps with the useful frequency band of EEG signals at 1 - 40 Hz. The second is that sweating can cause low-frequency noise [54]. The third is physical motion itself. During the task, the participants were biking at 50 RPM, which corresponds to 100 pedal downstrokes per minute (left and right). This introduced a noise at 1.33 Hz, which also overlaps with the 1 - 40 Hz range. Due to the overlap, naive filtering methods cannot eliminate those sources of noise completely.

2.4 Method

The ERP classification task is defined as follows: A set of EEG channels C and its signal over time $\mathbf{x} \in \mathbb{R}^{C \times T}$ is given where T depends on the sampling rate and duration of an epoched trial. The task is to take each epoched trial x and output a 3-class probability distribution $\mathbf{y}$. We are additionally given $\mathbf{x}_{nse} \in \mathbb{R}^{C_n \times T}$ for NSE information and $\mathbf{s} \in [0.1, 0.5, .9]$ for resting, low, and high exercise states.

In this chapter, we propose an end-to-end deep learning architecture, Multimodal

Layer	Parameters
Conv2d	1x10x10
Batch Norm	f: 10, eps: 1e-3, m:0.1
Conv2d, Elu	Cx10x10
AvgPool	k: 1x4 s: 1x4
Renorm	p: 2, mn: 1
Conv2d	10x1x20

(a) Temporal Encoding Network

Layer	Parameters
Dropout	p: .25
Conv2d	1x16x20, g: 20
Conv2d	1x1x10
AvgPool	k: 1x8, s: 1x8
Dropout	p: .25
Fully Connected, Elu	(T/1.6)x64

(b) Fusion Network

Layer	Parameters
Fully Connected, Elu	64x(20*T//32)
BatchNorm	
Deconv2d	1x1x20
Fully Connected, Elu	(T/4 + 1)x(T/4 + 1)
Deconv2d	Cx1x10, g: 10
Deconv2d, Elu	1x5x1
Deconv2d	1x10x1

(c) Decoder Network

Table 2.2: *Network parameters* are abbreviated as follows: **eps** for epsilon, **m** for momentum, **f** for number of filters, **k** for kernel size, **s** for stride, **p** for power, **mn** for max norm, and **g** for groups. Convolutions filter sizes are expressed in channel by time by number of filters.

Context-Aware Neural Network (MCANN), for modeling the ERP prediction problem. Our model (Figure 2.3) is in part motivated by the ability for unsupervised techniques to build a good representation of data. We break our model into 4 components: 1) a temporal feature extraction module in section , 2) a fusion component which combines these features, 3) a decoder which to reconstruct the signal for unsupervised learning and 4) a classification network for predicting the final class distribution.

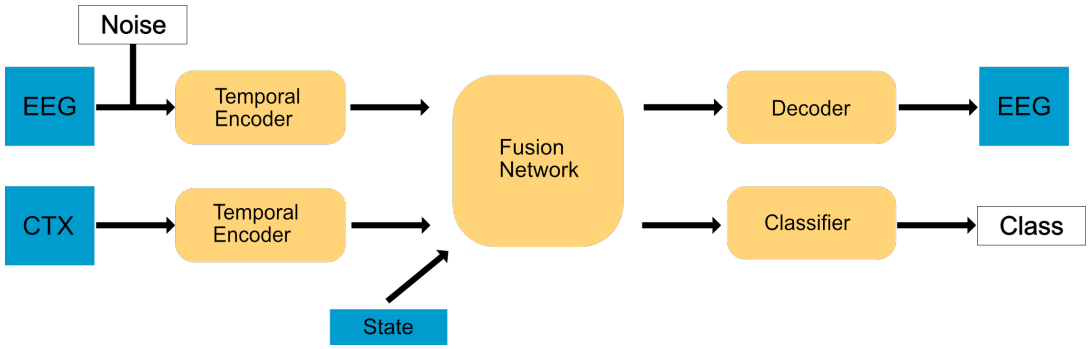


Figure 2.3: Our proposed model for evaluating EEG data with additional input modalities.

2.4.1 Unsupervised Representation Learning

Autoencoders first map an input $\mathbf{x}$ into a latent representation by a deterministic mapping: $\mathbf{z} = f_\theta(x)$. The latent representation $\mathbf{z}$ is then remapped back to $\mathbf{x}' = g_{\theta'}(\mathbf{z})$. A denoising autoencoder additionally takes a corrupted version of the original input $\tilde{\mathbf{x}}$ to reconstruct $\mathbf{x}$. Autoencoders of this sort have been shown to be robust to partial destruction of input for a wide range of tasks. Here $f_\theta(x)$ and $g_\theta(x)$ are modeled by multiple neural networks.

Noisy input $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{Wn}$ is formed by the addition of a noise vector whose values are sampled independently from a normal distribution $\mathbf{Wn}_{ij} \sim \mathcal{N}(0, \sigma^2)$ for all $i \in \{1, \ldots, C\}$ channels and all $j \in \{1, \ldots, T\}$ time samples. Here we use the standard normal distribution ($\sigma^2 = 1$), however other values could be considered depending on the dataset. Additionally, alternative methods of corruption could be explored which can be informed via user context or state information.

The noisy input $\widetilde{\mathbf{x}}$ is concatenated with prior noise information $\mathbf{x}_{nse}$ and $\mathbf{s}$ and fed into our network. The reconstructed signal $\mathbf{x}' = g_{\theta'}(f_\theta(\tilde{\mathbf{x}}, \mathbf{x}_{nse}, \mathbf{s}))$ is obtained.

2.4.2 Temporal Encoder

Table 2.2a describes the temporal feature extraction network. Weights from the first convolutional layer are shared to extract common temporal signal properties. The output of our temporal extraction process is denoted by: $\mathbf{v}_{eeg} = h_{\phi}(\widetilde{\mathbf{x}})$ and $\mathbf{v}_{nse} = h_{\phi'}(\mathbf{x}_{nse})$ for encoded eeg and nse features.

2.4.3 Fusion Network

The outputs of the temporal encoder are concatenated channel-wise with state and NSE information $\mathbf{v} = [\mathbf{v}_{eeg}; \mathbf{v}_{nse}; s]$, where the state is broadcast for each temporal feature value. A single layer, fully connected, network $\mathbf{u} = MLP(\mathbf{v})$ is used to fuse channel information over time for each time step. The output $\mathbf{u}$ is passed through our multimodal fusion encoding network (Table 2.2b). Note that we can adjust, add, or remove other modalities of input by modifying the representations concatenated to $\mathbf{v}$ and fused through $\mathbf{u}$ with ease.

For all evaluations 64 dimensions are used as the latent representation output by the last layer of the fusion network along the temporal dimension. For datasets with a smaller number of temporal time steps, the $min(64, T/1.6)$ is used.

2.4.4 Decoder Network

The network is parameterized by $\mathbf{x}' = g_{\theta'}(\mathbf{z})$ using the network given in Table 2.2c. At a very high level it approximates the opposite order of layers presented by the encoder network to obtain the non-corrupt signal. The final output $\mathbf{x}'$ is mapped to the noise-free input via the reconstruction loss:

$$\mathcal{L}_r = -\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}_i - \mathbf{x'}_i\|_1^2. \tag{2.1}$$

2.4.5 Classification Network

The classification network is a single layer fully connected network which maps the latent vector z to a softmax distribution of 3 classes $\mathbf{y}' = softmax(\ MLP_\rho(z))$. The negative log likelihood is utilized to maximize the correct class distribution:

$$\mathcal{L}_c = -\mathbb{E}[\log p(\mathbf{y}|\mathbf{x})] \tag{2.2}$$

2.4.6 Joint Loss Function

The final optimization function is a linear combination of the reconstruction and classification loss:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_r, \tag{2.3}$$

Where λ is an adjustable weight that can be annealed. For all experiments, we set λ to be linearly annealed from 1e$-$2 to 1e$-$5 over 5 training epochs. While other functions for annealing are possible, we did not evaluate them for this study.

2.5 Experiment Setup

Two methods of **preprocessing** were used to examine the performance as well as noise tolerance properties of our model. For each of 1Hz-Bike and 2Hz-Bike data (the two conditions of data collection previously described), the mastoid electrodes were used as the reference electrodes. The original data were also band-pass filtered between .1Hz-255Hz. We analyze a "denoised" version of the dataset, 1Hz-Bike-Filtered and 2Hz-Bike-Filtered, by applying a band-pass filter from 1-40Hz and down sampling to 128Hz. This step removes the high (40Hz+) and low (.1Hz-1Hz) frequency noise as the P300 ERP signal is known, a priori, to be within the 1Hz-40Hz frequency band. Com-

paring these different preprocessing methods allows us to examine algorithmic behavior under different conditions of noise and prior information.

To evaluate the accuracy and robustness of our algorithm, we split the data in two ways. For **subject-independent** splitting, the post-processed data is split into 80% training and 20% testing instances. We do this for each user stratifying by class distribution. The training data for each user is concatenated and shuffled to create a large training dataset. The same is done for the test dataset. Model and parameter tuning is conducted by randomly splitting 10% of the data from the training set for validation. This method for data splitting was used because this is common practice for current machine learning methodologies, as well as its more challenging condition over single-subject within-subject classification.

Cross-subject splitting allows us to analyze the generalizability of a technique to novel users. We follow the procedure from EEGNet [37] for subject splitting. Due to our smaller subject pool, we choose 1 subject iteratively and select an additional subject randomly. The remaining 10 subjects are used for training. This process is repeated 12 times so that each subject's test data is tested at least once. We set the training epochs to be 150 when validation accuracy appears to have converged for all models.

2.5.1 Algorithm Comparison

The MCANN model is compared against a traditional non-deep learning approach as well as a previous state-of-the-art deep learning model for ERP classification. For the traditional approach, xDAWN with 5 spatial filters was trained on the EEG data for each class, estimate covariance matrices, and project them into tangent space. Classification is performed using logistic regression with Riemannian distance [33]. This is similar to the technique used to win the Kaggle BCI challenge.

For the deep-learning model, MCANN is compared against EEGNet [37], a CNN architecture which performs comparably to state-of-the-art methods on a number of BCI tasks. For a fair comparison and to study the effects of multimodal signals on existing architectures, two versions of EEGNet are used. The EEGNet (UM) is a unimodal model which is only trained on EEG data. EEGNet (MM) is a multimodal model where we concatenate state information and non-scalp electrodes to the input.

2.5.2 Training and Setup

Training was conducted using a dropout of .25, the adam optimizer with an L2 weight decay of 1e−8, and a learning rate of 1e−3. All hyperparameters were tuned on the validation set of subject independent splitting and kept same throughout evaluation. Early stopping was used during tuning.

Training and evaluation was conducted on a single AMD 2700X with a single NVidia RTX 2070. We measure the performance of running a classification on the test set with a mini-batch size of 1 to simulate how samples would be received during a real-time scenario. Running a single end-to-end evaluation of a single sample takes 39 ms.

We examine the macro-averaged precision, recall and F1-scores of all algorithms. Table 2.3 shows the classification results averaged across all physical activity conditions. We compare all algorithms using both methods of pre-processing for two different data collection parameters (1 Hz and 2 Hz).

2.5.3 Subject-Independent Evaluation

While xDAWN+RG provided the best performance in recall for one of the four conditions, the MCANN model exhibited the best performance in F1-score and in all metrics for all other conditions. In this study, our overall F1-score for subject-independent clas-

	1Hz-Bike			1Hz-Bike-Filtered		
Method	R	P	F1	R	P	F1
xDAWN+RG	**74.16**	54.40	62.76	70.52	52.32	60.07
EEGNet (UM)	64.98	55.99	60.16	68.18	58.29	62.85
EEGNet (MM)	64.98	57.86	61.21	71.72	57.42	63.78
MCANN (Ours)	69.62	**61.92**	**65.55**	**75.33**	**62.31**	**68.20**

(a) 1Hz sampling results.

	2Hz-Bike			2Hz-Bike-Filtered		
Method	R	P	F1	R	P	F1
xDAWN+RG	64.7	46.93	54.40	67.59	49.33	57.03
EEGNet (UM)	58.82	48.63	53.24	66.94	52.00	58.53
EEGNet (MM)	62.04	50.59	55.73	67.70	54.33	60.28
MCANN (Ours)	**67.09**	**56.33**	**61.24**	**71.92**	**58.27**	**64.38**

(b) 2Hz sampling results.

Table 2.3: Summary results for all conditions under subject independent splitting. Average percentage metrics for (R)ecall, (P)recision, and (F1) score reported. Bold signifies best performance.

sification improved 6.28 points or approximately 10% in performance.

Figure 2.7 provides a confusion matrix for the 1Hz-Bike-Filtered condition. Precision and recall values for recognizing the distractor signal improved with MCANN for the high-intensity exercise condition. Additionally, the decrease in performance between resting and high-intensity exercise (higher noise) for target recognition is noted in this and other experiments. For the four cases we analyzed, the true positive rate for distractors showed the greatest increase.

Noise Robustness

We study the effects of noise on algorithms by changing the sampling and filtering parameters. Our dataset contains noise both within the known ERP frequency (1-40hz) and outside. Future applications of algorithms to EEG might make use of these additional

data ranges, potentially preventing the bulk filtering of large frequency bands.

All algorithms (with the exception of the 1 Hz case for xDAWN+RG) typically demonstrated approximately 3% increase in performance metrics over unfiltered data (Table 2.3). However, when given noisy data, MCANN scores higher on the overall F1-score than other methods under filtered conditions. This suggests MCANN has high tolerance to noise.

Effect of Exercise Intensity

All algorithms for all evaluated conditions and preprocessing methods experienced a drop in performance from resting to low- or high- intensity activity. Figure 2.7 provides an example comparison for the user context versus performance. While MCANN also experiences a drop in performance, it more than doubles the precision for prediction of the distractor in the 1Hz-Bike-Filtered scenario.

Table 2.4 examines the algorithmic performance over the three conditions. We see that our model and EEGNet (MM) demonstrate an improvement in performance, especially in the high-intensity exercise (higher noise) condition. Our algorithms produced a greater increase in performance under noisy conditions over previous state-of-the-art classifiers, indicating an improved tolerance to noise. EEGNet (MM) also showed improvements over it's UM variant.

Effect of Multimodality

There is some evidence to suggest that the addition of multimodal information can improve classification performance. When comparing the EEGNet (UM) model to EEGNet (MM) performance, we see on average a 1.5 point improvement in F1 score in Table 2.3. Our proposed method, which also uses the additional modalities performs best on precision and recall scores and leads to an average 6 point improvement in F1

User Context	Resting	Low	High
Precision			
xDAWN+RG	52.65	43.7	44.46
EEGNet (UM)	49.89 (-5%)	49.44 (+13%)	46.55 (+5%)
EEGNet (MM)	52.08 (-1%)	49.22 (+13%)	50.47 (+14%)
MCANN (Ours)	**59.93** (+14%)	**54.75** (+25%)	**54.34**(+22%)
Recall			
xDAWN+RG	**70.67**	60.75	59.43
EEGNet (UM)	64.73 (-8%)	60.35(-1%)	53.12 (-11%)
EEGNet (MM)	68.21 (-3%)	59.62(-2%)	59.54 (+0%)
MCANN (Ours)	69.8 (-1%)	**67.8** (+12%)	**63.38** (+7%)

Table 2.4: 2Hz-Bike, subject independent condition with percentage difference from xDAWN+RG.

over EEGNet (UM).

Looking at the confusion matrix in Figure 2.7, we see that EEGNet (UM) target prediction true positives drops by 10 percentage points between the resting and high-intensity exercise conditions. However, for our multimodal approach, we maintain reasonable performance for target predictions and only drop by about 2 points. This suggests that the extra modalities may enhance target signal detection.

2.5.4 Cross-Subject Evaluation

Cross subject evaluation is conducted on the best overall algorithm performance case (1Hz-Bike-Filtered) and the worst overall algorithmic performance case (2 Hz-Bike) from subject-independent evaluation.

A two-sample t-test assuming equal variances is used. We report p-values and effect size using Cohen's d. Our method performs significantly better on metrics against xDAWN+RG, the traditional approach, with greater accuracy (p=0.005, d=1.28), precision (p=0.008, d=1.19), recall (p=0.031, d=0.94), and F1-score (p=0.006, d=1.25).

When compared to EEGNet (UM), our model performs better on accuracy (p=0.014,

Method	Resting	Low	High
Precision			
xDAWN+RG	49.44 ± 1.34	46.95 ± 1.24	46.08 ± 1.02
EEGNet (UM)	52.01 ± 1.26	51.34 ± 1.16	48.34 ± 1.17
EEGNet (MM)	52.42 ± 1.37	51.81 ± 1.23	48.51 ± 1.29
MCANN (Ours)	**56.01 ± 1.85**	**54.65 ± 1.65**	**49.33 ± 1.56**
Recall			
xDAWN+RG	68.52 ± 1.86	63.71 ± 2.04	59.31 ± 1.96
EEGNet (UM)	68.38 ± 1.42	61.78 ± 1.36	56.18 ± 1.81
EEGNet (MM)	68.35 ± 1.16	62.46 ± 1.56	56.41 ± 1.79
MCANN (Ours)	**72.02 ± .89**	**68.52 ± 1.41**	**62.29 ± 1.41**

Table 2.5: Cross subject evaluation on the 1Hz-Bike-Filtered condition. ± standard error is reported.

d=1.09), recall (p=0.002, d=1.44), and F1-score (p=0.032, d=0.93). We did not find any significant difference for precision. Likewise, when compared to EEGNet (MM), we perform significantly better on accuracy (p=0.027, d=0.97), recall (p=0.005, d=1.26), and F1-score (p=0.05, d=0.85), but there were no significant differences on the precision metric.

Additional significance tests were computed to compare across biking condition. When comparing to EEGNet (UM) we see significant increases for metrics on accuracy (p=0.013, d=1.10), recall (p=0.003, d=1.34), and F1 (p=0.028, d=0.96) for the low condition. For the high activity condition when compared to EEGNet (UM) we perform significantly better on accruacy (p=0.021, d=1.01) and recall (p=0.021, d=1.02). No other significant differences were found when compared to EEGNet (UM). Under resting conditions, no significant difference was found between EEGNet (UM) and our model.

Similar tests are conducted between MCANN and xDAWN+RG. We generally see significantly better performance on almost all metrics for our model. On the high intensity condition, we see accuracy (p<0.005, d=.64), precision (p<0.005, d=.70), recall (p<0.005, d=.49), and F1-Score (p<0.005, d=.68). In the low condition, we see improve-

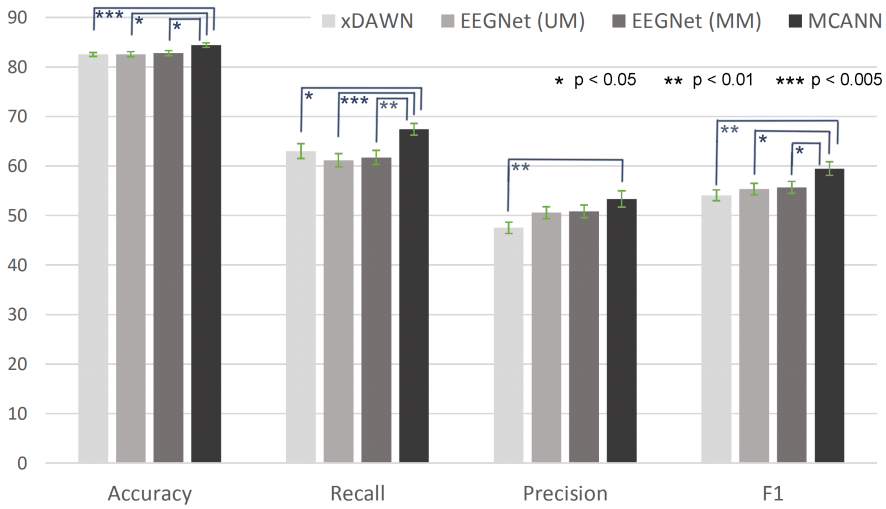


Figure 2.4: 1Hz-Bike-Filtered performance measures for each algorithm: Accuracy, Recall, Precision, and F1-metric. Error bars show standard error. Brackets indicate p-value significance groups from paired t-tests.

ments in accuracy ($p<0.005$, $d=1.6$), precision ($p<0.005$, $d=1.4$), and F1-Score ($p<0.005$, $d=1.4$). In the resting condition, we found significant improvements in accuracy ($p=0.004$, $d=1.32$), precision ($p=0.01$, $d=1.16$), and F1-score ($p=0.007$, $d=1.22$). Both tests of significance on EEGNet(UM) and xDAWN+RG indicate MCANN's stronger tolerance to noise.

When looking at the worst case scenario with the 2Hz-Bike dataset, we did not find any significant differences among any of the methods. We believe that, due to the minimal amount of processing and higher presentation frequency, the 2Hz-Bike dataset has a very large amount of individual subject variance outside the traditional 1-40Hz range, making it difficult for any model to generalize to new subjects. Some of these differences may be due to differences in individual motion patterns such as riding posture and cadence. These factors introduce unique noise patterns that compound the already challenging task of performing classification across users. These results highlight the need for more robust multimodal sensors to be used in conjunction with EEG sensors. We intend to investigate these inter-user differences in future research.

2.6 An example application

To demonstrate how ERPs can be incorporated into a real-world use case, we provide a demonstration for how it can be used for VR/AR authentication. As VR/AR headsets become pervasive, alternative methods for fast, secure, and non-intrusive authentication systems such as face and fingerprint recognition on modern mobile devices must be considered. This is especially important as private information stored in these headsets, such as eye and facial movement as well as financial and geo-tracking information, is an important security risk.

A potential answer to this problem lies in the use of Brain-Computer Interface (BCI) technology. BCIs enable interaction with computing devices via electroencephalogram (EEG) information with applications in education, marketing, security, medicine, and entertainment [55, 56]. Head-mounted devices such as VR/AR headsets offer a natural, non-intrusive way for widespread deployment of this technology. However, the generalizability of BCI algorithms across the EEG data of users is a major challenge. What if we take advantage of these inter-user differences for biometric authentication? Additionally, can we make authentication more accurate using additional modalities/biometrics from head-mounted devices?

Due to its morphological, anatomical, and functional plasticity, EEG based biometrics have been found to have potential discriminating capability [57] enabling it to be a reliable, convenient and universal biometric [58]. As a behavioral biometric, EEG signals are harder to imitate compared to physiological biometrics such as face and iris due to their temporal variations [59]. Another non-intrusive behavioral biometric, eye tracking, which is commonly used in VR/AR headsets, depends on the subtly different reactions of the eyes to stimuli [60] and can thus be easily applied to such systems.

Combinations of biometrics/modalities such as EEG and face [61], EEG and ECG

EEG	AUC	99%
	Accuracy	98%
	EER	3.4%
	FRR	8.4%
	FAR	1.8%

Table 2.6: Results for EEG Authentication: Note that this system is very accurate and the low FAR and FRR values are very valuable in a biometric security system.

Eye Tracking	Accuracy	79%
	Log Loss	0.82
	FRR	36.7%
	FAR	7.4%

Table 2.7: Results for Eye Tracking Authentication: Note that the FAR is low but the FRR is high indicating the high rejection of subjects from the system.

[62], and Eye Tracking and facial recognition [63] have been shown to achieve high levels of accuracy through multimodal fusion in biometric authentication. The authors did not find previous works combining EEG and eye tracking for the particular use case of biometric authentication.

2.7 Evaluating a Multimodal Authentication Scheme

The proposed method consists of three major steps: EEG authentication, Eye Tracking authentication, and Multimodal Fusion.

2.7.1 EEG Authentication

Task and Dataset: The EEG data used for processing were ERPs generated in a motor imagery task [64, 65, 66]. The left and right fist movements from the EEG Motor Movement/Imagery (EEG MMI) dataset from the Physionet bank [67, 68] were chosen

due to the simplicity of such motions in a potential practical application of such a system and the abundant use of such tasks in BCIs.

Preprocessing: The EEG signals are epoched and preprocessed using the MNE package [69]. Each individual epoch was band-pass filtered using a Finite Impulse Response windowed filter between 0.5 Hz to 42 Hz and normalized to zero mean and z-scores from zero.

Classification: The unnormalized cross-correlation is used to measure the similarity between two signals and is applied in a template matching procedure between the 64 electrode signal pairs from the samples being compared. The maximal value of the cross-correlation is used to create a 64×1 feature vector. Support Vector Machines (SVM) with linear and radial basis function (RBF) kernels are applied to this feature vector.

2.7.2 Eye Tracking Authentication

Task and Dataset: The dataset A of the EMVIC 2012 competition, containing positional data of the eye fixations across time, was utilized [70]. The dataset contains eye tracking data from a "jumping dot stimulus" task. Samples were disproportionately grouped towards certain subjects. To reduce dataset bias, we pool subjects with fewer than 40 samples into a separate group called the "unauthorized users group". This procedure has the positive side effect of allowing for more variety and fewer samples per subject for unauthorized users which better reflects real-world conditions.

Classification: A random forest classifier with 100 trees was trained on feature vectors composed of the concatenated eye-tracking signals. The model predicts an array of posterior probabilities that the given sample belongs to each of the possible labels consisting of $n = 5$ authorized users and the unauthorized group (totally $n+1 = 6$ bins).

Models	FAR	FRR
SVM Fusion	23.6%	29.2%
Weighted Mean (Eye tracking + EEG)	60.5%	23.6%
EEG only	42.1%	27.8%

Table 2.8: Comparison of fusion methods and EEG baseline in cases of low confidence

2.7.3 Multimodal Fusion

Each subject in the EMVIC dataset was matched to a participant in the EEG MMI dataset to create a fused dataset of hypothetical subjects with Motor imagery and Eye Tracking data. We thus have 5 authorized subjects and 32 unauthorized subjects in our newly composed dataset. We conduct match-score level fusion as it preserves adequate discriminating information and is modular in its execution. Here, two fusion methods have been implemented: weighted mean and fusion by SVM with linear kernels, each providing a normalized match-score from the individual predictions.

Weighted Mean: A weighted linear combination of the EEG and eye tracking scores gives a normalized match-score by the following formula:

$$z = \alpha u + (1 - \alpha)v \tag{2.4}$$

where u, v are the scores of the individual modalities and α is a parameter tuned on the training set.

SVM: An SVM with a linear kernel is applied to the predicted EEG score and the eye tracking distribution to obtain the fused score.

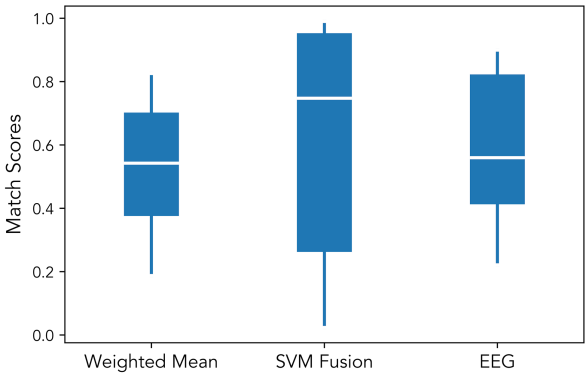


Figure 2.5: Analysis of positive samples. Normalized match scores are presented for each type of input. Higher scores indicate that the model is predicting a match with high confidence while lower scores indicate that the model is predicting a mismatch with high confidence. Values between 0.2 and 0.8 are considered low-confidence predictions with .5 being a neutral prediction. SVM fusion performs the best overall, however it is more likely to give a negative prediction when the expected label is positive.

2.8 Results and Discussion

2.8.1 Individual Modalities:

The results for the individual modalities are provided in Table 2.6 and Table 2.7. We tested the EEG system using an 80:20 train-test split. We report only the metrics of the linear kernel SVM due to its significantly better performance compared to the RBF kernel SVM. For the Eye Tracking system, we assume the label with maximum score in the distribution as the match-score. An 80:20 train-test split was applied and the various metrics for the 6 way classification are described.

2.8.2 Multimodal Evaluation

We did not observe appreciable improvement in ROC curves, EER values, and AUC values when using the fusion system versus the EEG system. In fact, the weighted average method performed worse compared to the EEG baseline and the SVM fusion method provided marginal improvements. We try to understand this by examining the confidence

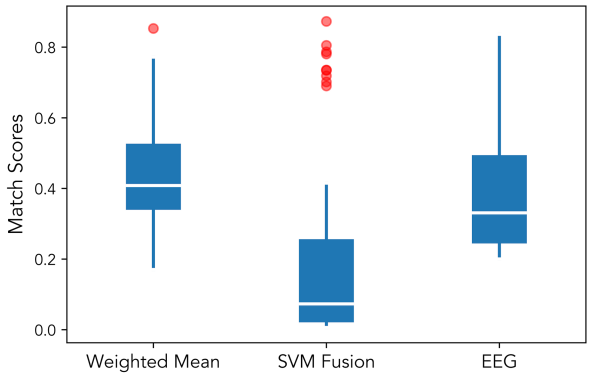


Figure 2.6: Analysis of negative samples. Normalized match scores are presented for each type of input. Higher scores indicate that the model is predicting a match with high confidence while lower scores indicate that the model is predicting a mismatch with high confidence. Values between 0.2 and 0.8 are considered low-confidence predictions with .5 being a neutral prediction. SVM fusion also appear to perform the best when negative predictions are expected.

of the predictions. From the median match-score of the SVM fusion method in positive and negative ground truths (Figure 2.6, Figure 2.5), the system shows higher confidence towards the correct label. However, when examining low confidence predictions of the EEG system (0.2 to 0.8), SVM fusion is more likely to give a negative prediction. This results in the FAR values being significantly lower for the SVM Fusion as compared to other methods without the FRR being affected as can be seen in Table 2.8. Here, we see eye tracking providing a benefit when EEG confidences are low without affecting the values of high confidence.

2.9 Conclusion

In this chapter we presented a challenging dataset for developing BCI classification algorithms. We provided a novel method for classifying EEG signals under conditions that varied dramatically with regard to noise. We observed significant improvements in our test set during cross-subject evaluation when compared to previous state-of-the-

art techniques. Additionally, our new algorithm is capable of incorporating additional modalities for improved classification of brain data. We additionally provide a feasibility study towards using EEG and eye tracking for multimodal biometric authentication. Future work will include (1) the application of our classifier to online BCI scenarios that involve motion, such as navigation of real-life or virtual environments, and (2) testing classifier performance during other types of physical activity that may involve more extreme head and body movements.

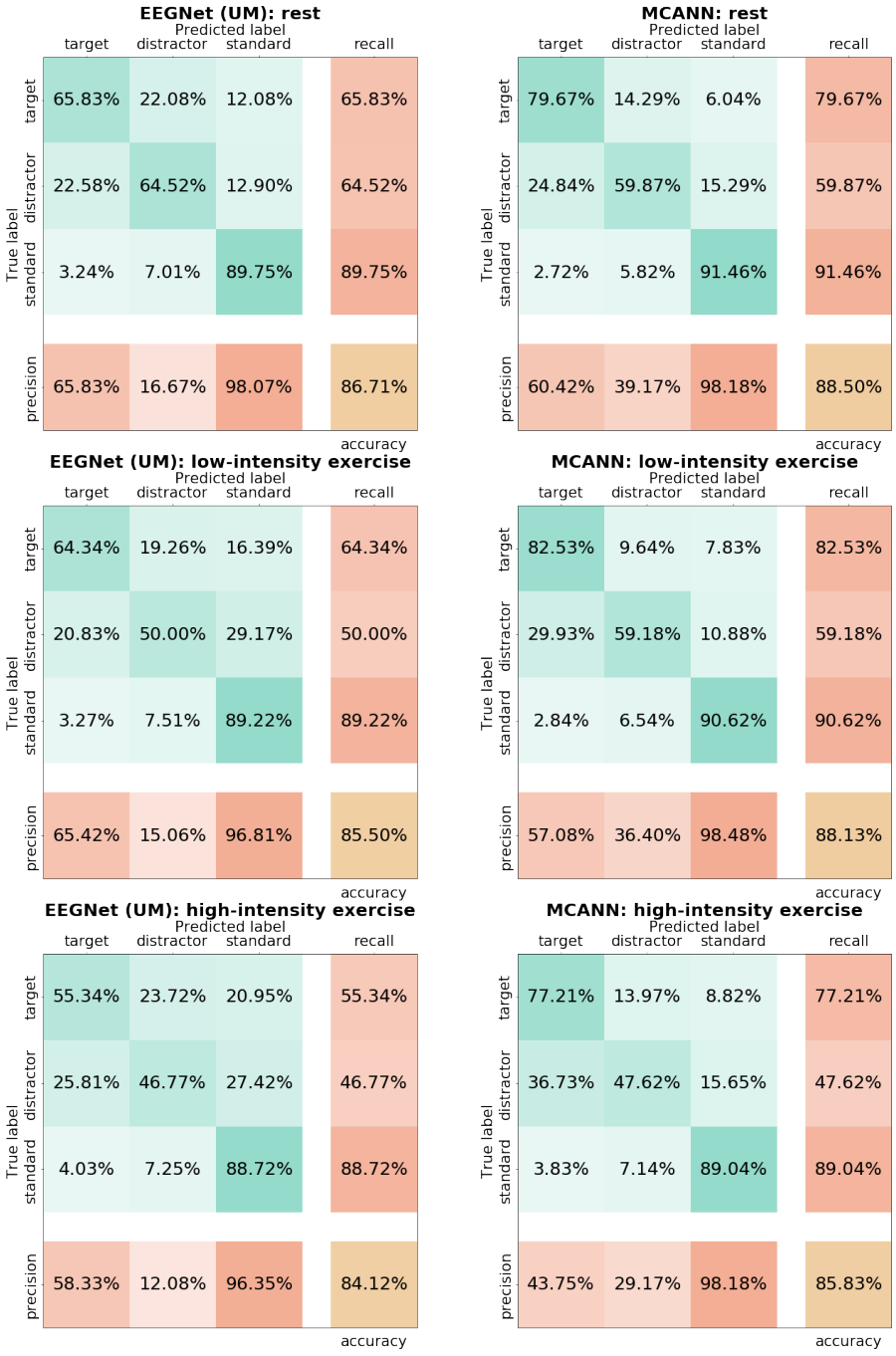


Figure 2.7: Detailed breakdown of performance for the 1Hz-Bike-Filtered, subject-independent condition.

Chapter 3

Fusing Multimodal Data with Computation Constraints

Parts of this chapter are in preparation with collaborators for the Winter Conference on Applications in Computer Vision 2023

This chapter looks at how we can reduce the computational complexity of multimodal machine learning algorithms. Because we would like these models to work in a real-time environment, high-powered computing devices may not be always available. We perform the experiments in this chapter with this purpose in mind and try to examine ways in which we can improve model performance while minimizing computational cost.

3.1 Motivation

Processing multiple data streams increases computational cost, and it is therefore a high priority to develop efficient algorithms in the multimodal domain. Additionally, many multimodal applications, such as the detection of instances of domestic abuse, or detection of prolonged emotional and psychological struggles, are particularly well-

suited for mobile or low-resource devices. In these resource-constrained settings, the computation cost and memory footprint become critical factors that must be considered for practical use.

Current multimodal algorithms involve some level of modality-independent feature processing followed by a fusion process which then jointly models the dependencies and cross-dependencies between the modalities. In particular, deep-learning transformer models have been used in this way to achieve state-of-the-art performance on numerous tasks [71, 72]. However training and processing such data remains prohibitively expensive in many cases, in terms of time, computational resources, and energy consumption. For example, a single layer of a vision transformer [73] requires approximately 1.35 billion floating-point operations (GFlops) for a 224×224 image for a single forward pass. If we represent a sequence of 30 frames in a similar manner for video data, this explodes to 88.24 GFlops. Although recent advancements have been made to sparsify transformers, these efforts have primarily approached the problem from a unimodal perspective [74, 75, 76, 77, 78].

Motivated by these concerns, we propose a sparse fusion method for multimodal transformers called Sparse Fusion Transformers (SFTs) that drastically reduces training time and memory consumption while maintaining the quality of existing fusion methods. Our approach is based on the hypothesis that the large amount of complementary information across different modalities allows us to sparsify unimodal information prior to multimodal fusion without the loss of accuracy. In particular, approaching a problem from a multimodal perspective enables us to sparsify the unimodal information far more aggressively. With our sparse-fusion method, we achieve faster performance with less memory use while attending to features that are most important.

Our proposed fusion process is agnostic to input modality and makes a full multimodal classification network robust to sparsification of input representations. It is composed

of three parts: a block-sparse within-modality attention to learn strong local representations, a pooling method for extracting them, and dense self-attention for cross-modal feature fusion. Furthermore, we propose to use a customized mixup to apply spatio-temporal regularization to the learned representations in a modality agnostic manner. Fusing features in this way demonstrates comparable or better performance than existing methods while requiring significantly less computation and memory. In summary, our contributions are:

- We propose a novel fusion method that maintains or exceeds the performance of previous fusion methods while demonstrating up to a six-fold reduction in computation and memory requirements.
- We demonstrate that multimodal algorithms can tolerate far more token reduction than unimodal algorithms due to complementary cross-modal information. We show that by accounting for multimodal information during sparsification, more information can be removed without loss of performance.
- We perform extensive ablation studies on fusion components using real-world datasets to determine the efficacy of each model component. We further experiment with multiple pooling methods to demonstrate model robustness under different pooling requirements.

3.2 Related Work

The problem of modality fusion has been explored in numerous problem spaces for a long time [79]. The primary challenge is to find an effective way to combine representations of data from disparate modalities into a single representation for more accurate modeling. While the first methods for multimodal fusion were proposed to address signal

inadequacies in individual modalities [80], we are now at a time when the resolution in each modality is much higher, making some computation costly and intractable. Therefore, we wish to purposely trade off some of the signal bandwidth to improve performance.

Many methods have been proposed to tackle the task of fusion. A way to categorize all these techniques is by the time of fusion occurrence. Early fusion typically refers to combining base level representations or even input values, while late fusion primarily refers to its application near the output. Early deep-learning methods typically make use of linear layers and cross products to combine modalities [81, 82, 83]. More rudimentary forms of fusion simply involve adding the logits of individual modality predictions together. As transformer-based architectures have become very popular recently, some recent techniques have also explored their use in multimodal settings. Originally proposed in [84] for neural machine translation (NMT) tasks, they have demonstrated superior performance on multiple benchmark problems such as image classification [73], action recognition [72] and 3D reconstruction [85, 86]. The basic functionality is to apply layers of self-attention, on sequential representations. To classify a discrete output, transformers typically rely on the use of a special token (`CLS`) that is prepended to the sequence for classification.

The most natural form of transformer fusion is simply to concatenate the sequence of tokens and rely on self-attention to learn their inter-dependencies. Works such as [87, 88] that do this learn better cross-modal representations and have shown benefits relative to naive fusion methods. Very recently, multimodal bottleneck transformers [72] have demonstrated a way for early fusion to occur without the use of costly cross-modal operations. However, the process of fusing multimodal information with some form of concatenation and dense attention remains costly due to the $O(N^2)$ complexity of transformers for input sequences of length N. It is this cost we seek to address with our sparsification approach.

Recent efforts have focused on reducing computational complexity for transformers

and large-scale deep learning [89, 90, 91, 74]. An effective method for this is to exploit the representation of features within a small sliding window of tokens [92] on a long sequence. However, these methods require significant engineering efforts and are hard to train [93]. Other works approach the problem via sparsification of the attention mechanism, such as random or local attention [94, 95, 96]. Sparsification methods have also been applied successfully for some computer vision tasks [77].

Training optimizations for transformers have also been explored. Regularization techniques such as dropout [97], weight decay [98], and mixup [99] have all been applied. While weight decay and dropout can be applied in a modality-agnostic manner directly onto the weights, the use of mixup has primarily been used to tackle problems in the vision domain, as its application is easily interpretable and offers large benefits to the algorithms [100, 101]. Although some recent efforts have been made to enable the application of mixup on domains in a modality agnostic manner [102], its application in a fundamentally multimodal domain remains under-explored. Its use in the mixing of fused features across modalities spatially and across time demonstrates large benefits for our application.

3.3 Method

In this section, we describe our proposed Sparse Fusion Transformers (SFT). See Fig. 3.1 for a visualization of our algorithm. As input, our method takes token sets from M different modalities, $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$, with each modality consisting of N tokens of dimension D, $\mathbf{Z}_i = [\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iN}] \in \mathbb{R}^{N \times D}$. Note the number of tokens N can vary from modality to modality but for simplicity of notation, we keep it fixed in our description. Additionally, if the token dimension D varies from modality to modality, we apply a per-token projection to keep the token dimension constant across all modalities. Following

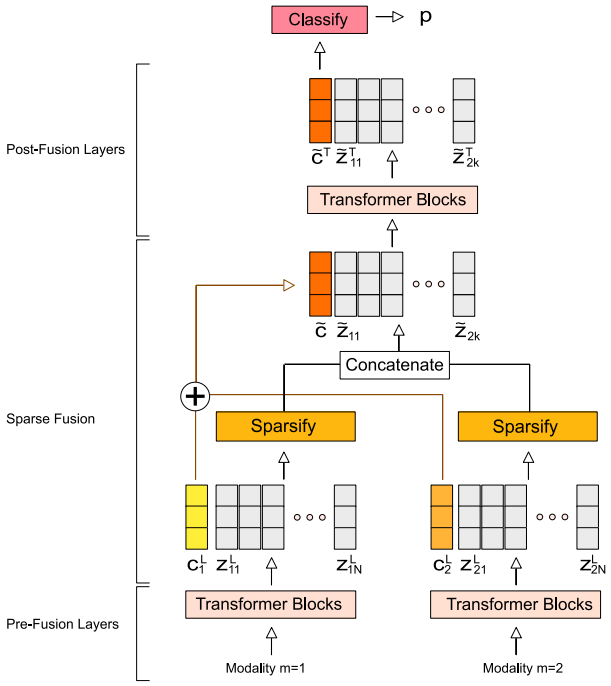


Figure 3.1: Visualization of our fusion method with two modalities. Following existing work, a special `CLS` token is appended to each unimodal token set prior to unimodal transformers. After unimodal transformers, the `CLS` token ($\mathbf{c}_1^L$ and $\mathbf{c}_2^L$) from each modality is summed. A pooled block-sparse attention is applied to local regions of each modality. The `CLS` token and pooled representations are then combined, and dense self-attention is applied to model global and cross-modal dependencies.

existing work, we prepend a special `CLS` token $\mathbf{c}$ with learnable parameters to each token set for each modality for the purpose of classification: $\hat{\mathbf{Z}}_i = [\mathbf{c}_i|\mathbf{Z}_i] = [\mathbf{c}_i, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iN}] \in \mathbb{R}^{(N+1)\times D}$. The goal of our method is classification, i.e., we want to learn a function $f_\theta : \mathbb{R}^{M\times(N+1)\times D} \rightarrow \mathbb{R}^C$:

$$f_\theta(\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_M) = \mathbf{p}, \tag{3.1}$$

such that $\mathbf{p}$ is the probability distribution over C classes.

Our method consists of three main parts. First, we model relationships between tokens within modalities using a standard transformer that is applied unimodally (Sec. 3.3.1). Second, we aggregate information within local regions of each sequence using block-sparse attention and then apply local subsequence pooling to sparsify the to-

ken set for each modality (Sec. 3.3.2). Third, we concatenate the sparsified features from each modality and run dense self-attention to predict a final class (Sec. 3.3.3). During training, we apply a novel multimodal variation of manifold mixup [102] for regularization of intermediate latent representations (Sec. 3.3.4).

3.3.1 Unimodal Modeling

In this stage, we apply a separate transformer to the token set from each modality. Following Vaswani *et al.* [84], we use a standard L-layer transformer encoder to model relationships between tokens in each modality. Each layer of the encoder consists of layer normalization (LN), Multi-head Self-Attention (MSA), and a Multi-Layer Perceptron (MLP). Given token set $\hat{\mathbf{Z}}^l$ after l transformer layers, the output of layer $l+1$ is:

$$\mathbf{Y}^l = \text{MSA}(\text{LN}(\hat{\mathbf{Z}}^l)) + \hat{\mathbf{Z}}^l \tag{3.2}$$

$$\hat{\mathbf{Z}}^{l+1} = \text{MLP}(\text{LN}(\mathbf{Y}^l)) + \mathbf{Y}^l \tag{3.3}$$

We apply a separate L-layer transformer per modality to get token sets $\hat{\mathbf{Z}}_1^L, \ldots, \hat{\mathbf{Z}}_M^L$.

3.3.2 Sparse Multimodal Fusion

In this stage, we apply local pooling blocks to each token set $\mathbf{Z}_i^L$ to extract k descriptive tokens per modality $\tilde{\mathbf{Z}}_i = [\tilde{\mathbf{z}}_{i1}, \ldots, \tilde{\mathbf{z}}_{ik}] \in \mathbb{R}^{k \times D}$, as represented by the "Sparsify" blocks in Fig. 3.1. As shown in our experiments in Sec. 3.5.3, information is quite redundant within and across each modality, and we hypothesize simple sub-sequence pooling to be a cheap and effective method for capturing important information while removing redundancies. Prior to pooling, we first apply a single bi-directional strided sparse attention layer [103] to enforce aggregation of dense local context and sparse global con-

text to every token in the sequence to each modality. We then apply non-overlapping per-channel pooling blocks of stride s for each token set:

$$\tilde{\mathbf{z}}_{ij} = \text{pool}\left(\mathbf{z}^L_{i(js+1)}, \ldots, \mathbf{z}^L_{i(js+s)}\right). \tag{3.4}$$

A natural choice for pooling is either per-channel max pool or average pool. We explored several options in ablation studies and found our method to be robust to the choice of pooling (see Table 3.4). However, for our main experiments we use average pooling.

We additionally form a multimodal classification token $\tilde{\mathbf{c}}$ by summing the unimodal classification tokens:

$$\tilde{\mathbf{c}} = \sum_{i=1}^{M} \mathbf{c}^L_i \tag{3.5}$$

The final, fused token set $\mathbf{F}$ is formed using this classification token and the union of the unimodal pooled token sets $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_M$:

$$\mathbf{F} = [\tilde{\mathbf{c}}, \tilde{\mathbf{z}}_{11}, \ldots, \tilde{\mathbf{z}}_{Mk}] \tag{3.6}$$

3.3.3 Dense Cross-modal Modeling and Prediction

To model cross-modal relationships, we apply a dense, T-layer transformer on the token set $\mathbf{F}$. Note the tokens of $\mathbf{F}$ are aggregated from all modalities. We adopt the same architecture used in the unimodal modeling task, denoting the token set after t transformer layers as $\mathbf{F}^t$, with the final output denoted $\mathbf{F}^T = [\tilde{\mathbf{c}}^T, \tilde{\mathbf{z}}^T_{11}, \ldots, \tilde{\mathbf{z}}^T_{Mk}]$. Finally, a small MLP followed by softmax is applied to $\tilde{\mathbf{c}}^T$ to produce a C-way class prediction $\mathbf{p}$.

3.3.4 Multimodal Manifold Mixup

We apply a novel variation of manifold mixup [102] for improved generalization. In the originally proposed mixup [99], given two random training inputs $\mathbf{x}_i$ and $\mathbf{x}_j$, their corresponding ground-truth labels $\mathbf{y}_i$, $\mathbf{y}_j$, and an interpolation weight $\lambda \in [0, 1]$, a classifier is trained using the following virtual training examples:

$$\tilde{\mathbf{x}} = \lambda \mathbf{x}_i + (1 - \lambda)\mathbf{x}_j \tag{3.7}$$

$$\tilde{\mathbf{y}} = \lambda \mathbf{y}_i + (1 - \lambda)\mathbf{y}_j \tag{3.8}$$

Generally, the interpolation term λ is sampled from a Beta distribution $\text{Beta}(\alpha, \alpha)$, where α is a hyperparameter. Manifold mixup extends this by also selecting a random layer l in an L layer network f and interpolating the latent representations $\mathbf{v}_i^l, \mathbf{v}_j^l$ of that layer instead of the input example:

$$\tilde{\mathbf{v}}^l = \lambda \mathbf{v}_i^l + (1 - \lambda)\mathbf{v}_j^l \tag{3.9}$$

Layers $l+1, \ldots, L$ of f are then applied to $\tilde{\mathbf{v}}^l$ and the output is supervised using Eq. 3.8. Manifold mixup has been shown to be more effective for regularization than input mixup.

We extend manifold mixup to the multimodal case for use with our model. Given our $(L+T)$-layer network, with the first L layers involving separate, unimodal transformers and the last T layers involving a single, multimodal transformer, we sample a single layer $l \in [1, L+T]$ for manifold mixup. If $l > L$, we use standard manifold mixup using Eqs. 3.8 and 3.9. If $l \leq L$, we sample a different interpolation term for each of the M modalities, $\lambda_1, \ldots, \lambda_M \sim \text{Beta}(\alpha, \alpha)$. Given latent representation $\mathbf{v}_{mi}^l, \mathbf{v}_{mj}^l$ of layer l for

modality m, the new latent representation is given as:

$$\tilde{\mathbf{v}}_m^l = \lambda_m \mathbf{v}_{mi}^l + (1 - \lambda_m)\mathbf{v}_{mj}^l \tag{3.10}$$

This is applied to every latent representation of layer l for every modality $1, \ldots, M$. After running the remaining $L + T - l$ layers, the output of the network is supervised using:

$$\tilde{\mathbf{y}} = \overline{\lambda_*}\mathbf{y}_i + (1 - \overline{\lambda_*})\mathbf{y}_j \tag{3.11}$$

where $\overline{\lambda_*}$ is the average of the M sampled λ values.

3.4 Experimental Setup

We now describe the datasets used for training and evaluation (Sec. 3.4.1), dataset pre-processing (Sec. 3.4.2), baseline network architectures used for comparison (Sec. 3.4.3), and training hyper-parameters we used (Sec. 3.4.4).

3.4.1 Datasets

We perform extensive experiments on two benchmark multimodal datasets: VGG-Sound [104] and CMU-MOSEI [105] The datasets tackle popular and broadly applicable tasks in multimodal machine learning for audio-visual classification and multimodal sentiment classification. The modalities evaluated include video, audio, and text data. Additionally, these datasets have differences in modality characteristics such as cross-modality alignment and information content.

VGG-Sound

VGG-Sound [104] consists of over 200,000 YouTube videos and their associated audio streams, each annotated with one of over 310 class labels. The audio spans a large range of challenging acoustic environments and noise characteristics of real applications. All videos are captured "in the wild." There are clear audio-visual correspondences, i.e., the sound source is visually evident. Each segment is 10 seconds long. To aid in evaluation, we select two subsets of data from VGG-Sound containing 10 classes and 100 classes each. We call these VGGS10 and VGGS100, respectively. We select VGGS10 by choosing pairs of easily confused classes, such as "baby babbling" and "baby laughing". We then build VGGS100 using these ten classes and additionally include 90 randomly chosen classes. The total training and testing set sizes for VGGS10 are 6,051 and 459. For VGGS100, the training set size is 66,180 and the test set size is 4,549. A validation set is extracted by taking 20 percent of the training set.

We summarize the classes we used in VGGS10 and provide the label distributions in VGGS10 and VGGS100. VGGS10 is a manually curated dataset built by selecting pairs of difficult to separate classes from the full VGGSound dataset as well as for differences between video and audio modalities. We chose the following ten classes: airplane, baby babbling, baby crying, baby laughter, cat meowing, cat purring, people marching, people running, playing bass guitar, playing electric guitar. The final training set distribution for VS10 in Fig. 3.2, and the final VGGS100 dataset distributions are show in Fig. 3.3.

CMU-MOSEI

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [105] dataset is one of the largest multimodal sentiment analysis and emotion recognition datasets to date. The dataset contains more than 23,500 sentence utterance videos from

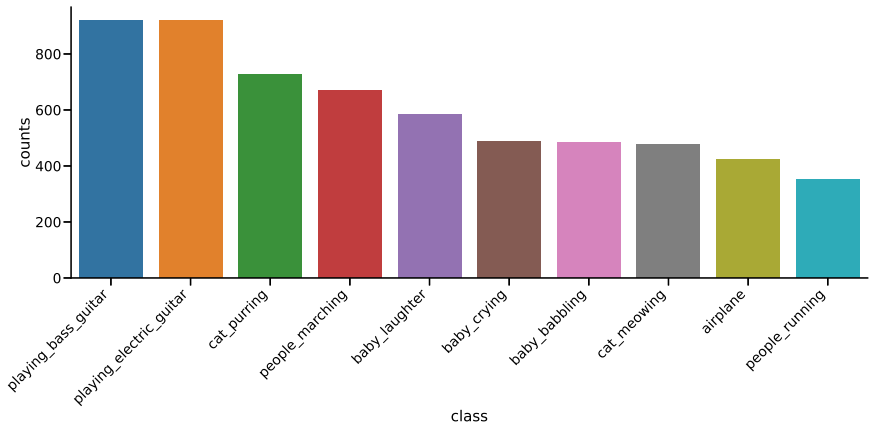


Figure 3.2: Distribution of samples in the VGGS10 dataset by class.

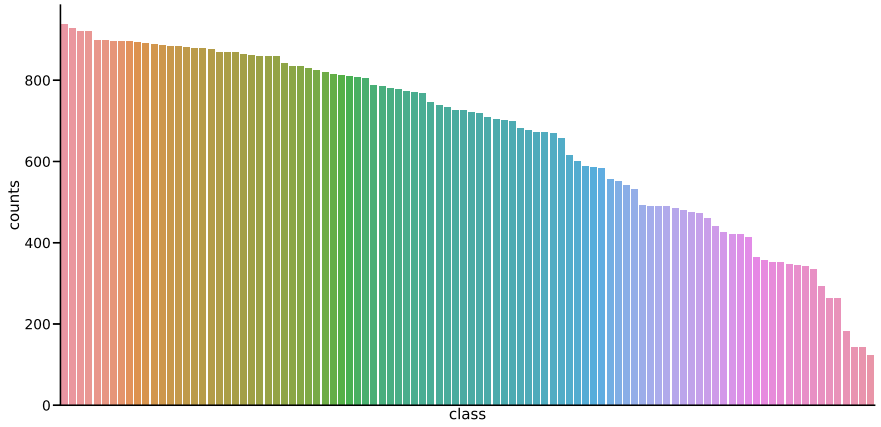


Figure 3.3: Distribution of samples in the VGGS100 dataset by class.

more than 1000 online YouTube speakers. The dataset is gender-balanced. All utterances are randomly chosen from various topics and monologue videos. The task is to predict a 7-class sentiment score of a particular multimodal video sample. Each sample contains audio, video, and text modalities. This dataset is frequently used to explore the unaligned nature of multimodal sequences between text and video.

3.4.2 Pre-processing

Each modality is pre-processed with a feature extraction pipeline in order to generate the input token sequence. For the MOSEI dataset, we use the pre-processed data provided by the authors. The pre-processing pipeline that was used assumes that each video depicts a "talking head": a single human talking, whose face is visible and whose voice is clearly audible. This assumption is valid for the MOSEI dataset, and the pre-processing pipeline therefore extracts visual features such as facial landmark positions and audio features such as estimated vocal parameters. We refer the reader to Zadeh *et al.* [105] for the full details. To pre-process VGGSound, we employ a feature extraction pipeline that can be applied to videos more generally, without assuming human faces or voices are present.

For the VGGS10 and VGGS100 datasets, we extract *visual features* using I3D [106], a spatio-temporal video feature extraction model that was pre-trained on the Kinetics human action recognition dataset [106]. This is a two-stream model, which processes optical flow and raw RGB independently as two separate modalities. We also extract TV-L^1 optical flow from the VGGSound videos. For *Audio* pre-processing we follow Nagrani *et al.* [72]: we resample all audio at 16Hz and convert to mono, then compute log mel spectrograms with 128 frequency bins, using a Hamming window with size 25ms and stride 10ms.

3.4.3 Baseline Network Architectures

We compare against the following transformer-based fusion methods:

Self-Attention Fusion (Concat): A baseline method of fusion is to concatenate the individual modality representations prior to input to any network and rely exclusively on dense self-attention. This is a form of early fusion.

Late Fusion (LF): This method works by applying transformer blocks on individual modalities only. The final prediction is obtained via a summation of logits derived from individual class tokens. This helps us compare the benefit of modeling cross-modal interactions.

Multimodal Transformer (MulT): [87] MulT is a hybrid early-late attention-based fusion method using a unique cross-modal attention mechanism. The data is first fused via an attention mechanism by using one modality each for key, query, and value. Transformer blocks are then stacked on top. At the very end, the features are concatenated and a prediction is obtained after an FC layer.

Bottleneck Fusion (MBT): [72] This is a form of fusion in which special tokens called bottleneck tokens are introduced. These tokens are shared among all modalities, and transformers alternate operating on each modality independently. The final `CLS` token is summed from each modality and used for prediction. We additionally evaluate MBT using manifold mixup (MBT+MM) as the original paper used input mixup, and our inputs are features.

3.4.4 Implementation details

Our model is implemented in PyTorch. For all experiments on the smaller datasets VGGS10 and MOSEI we use a learning rate of 10^{-4}. For the larger dataset VGGS100 we use a learning rate of 10^{-3}. Learning rate is decayed by factor of 10 every 10 epochs based on minimum validation loss. We use a batch size of 24 for all experiments. For all datasets, we report results based on averaging performance training from 5 different seeds for generalization purposes and to minimize tuning effects. We use a standard 12-layer network and 5 attention heads for all evaluations. We project embeddings from each modality to 40 to minimize the effects of over-parameterization. For experiments

involving latent mixup, we used a strength of $\alpha = 0.3$. We use an initial warm-up of 5 epochs in which no mixup is applied. For all other experiments we applied dropout $p = 0.2$ for regularization. For baselines, we follow descriptions in original papers and publicly available code for comparison. All experiments were conducted on consumer-grade graphics cards. We make our code and preprocessed data publicly available.

3.4.5 Metrics

We report results using commonly used metrics. **Top1** represents the accuracy of the most likely class. **mAP** represents the mean of per-class average precision scores. We also report the computational cost in Giga floating-point operations (GFlops) which is estimated similar to previous methods [77]. We present all flop estimates in the chapter using the following equations. We primarily follow the flop estimation from [77] with some minor changes due to layer differences. Each transformer layer consists of a multi-head attention and multilayer perceptro block. A multi-head attention (MHA) block has cost of:

$$\begin{aligned} \phi_{MHA} &= \phi_{qkv} + \phi_A + \phi_O + \phi_{proj} \\ &= 3nd^2 + n^2d + n^2d + nd^2 \\ &= 4nd^2 + 2n^2d \end{aligned} \tag{3.12}$$

where n, d represent the length and embedding dimension, ϕ_{qkv} is the cost of projecting to the query, key, and values. ϕ_A is the cost of the attention map, ϕ_O is the cost of the self attention, and ϕ_{proj} is the cost of projection for self-attention outputs.

A MLP block includes two linear layers as well as a normalization layer for a cost of:

$$\begin{aligned}\phi_{MLP} &= \phi_{proj1} + \phi_{norm} + \phi_a + \phi_{proj2} \\ &= nd^2 + 3nd + nd + nd^2 \\ &= 2nd^2 + 4nd \end{aligned} \tag{3.13}$$

where ϕ_{proj1} and ϕ_{proj2} are cost of projecting into and out of latent space for transformer block, ϕ_{norm} represents cost of applying layer normalization, and ϕ_a represents the cost of an activation function.

Many experiments examine the effect of a reduction factor, which refers to reducing the number of tokens in the sequence dimension for transformer architectures. We report most results as a mean and standard deviation of experiments run with five different seeds.

3.5 Results

We first report our results against state of the art (Sec. 3.5.1) showcasing our performance on multiple datasets from different domains. We then perform a series of ablation studies to explore the effects of sparsification (Sec. 3.5.2), and the benefits of addressing within-modality redundancies during fusion (Sec. 3.5.3). We also study the effect of pooling choice (Sec. 3.5.4) and the effect of our proposed multimodal manifold mixup (Sec. 3.5.5).

3.5.1 Comparison against state of the art

We present our summary benchmark performance on real-world datasets VGGS10, VGGS100, and MOSEI in Tables 3.1 and 3.2. For each dataset, our model keeps a subset

	VGGS10		VGGS100		MOSEI	
	Top1	**mAP**	**Top1**	**mAP**	**Top1**	**mAP**
Concat	<u>67.62</u> ± 1.3	**71.46** ± .63	51.72 ± .26	51.64 ± .13	48.47 ± .23	<u>32.40</u> ± .83
LF	67.10 ± .79	70.46 ± .79	52.00 ± .73	46.92 ± .28	49.10 ± .33	31.75 ± .78
MulT	65.49 ± .40	69.73 ± 1.1	51.35 ± .43	49.25 ± .43	<u>49.36</u> ± .34	31.92 ± .79
MBT	66.84 ± .61	70.98 ± .78	51.67 ± .66	51.29 ± .37	49.12 ± .27	32.15 ± .47
MBT+MM	66.80 ± 1.8	70.56 ± .61	**55.97** ± .42	**57.29** ± .37	48.77 ± .37	32.03 ± 1.2
Ours	**67.71** ± 1.3	<u>71.06</u> ± .81	<u>55.61</u> ± .61	<u>57.18</u> ± .39	**49.67** ± .23	**33.66** ± .85

Table 3.1: Accuracy comparison for each dataset and model. For all benchmarks we report the mean and standard deviation performance over 5 seeds to minimize tuning effects. Bold indicates best, underline second best. We are either best or close to best in all metrics.

of tokens from each modality during pruning. For VGGSound data after pooling we have 12 tokens of RGB and flow information and 20 tokens spectrogram data. For MOSEI, we keep 10 tokens of visual and audio information and 25 tokens of text information. These numbers were chosen according to experiments described in Sec. 3.5.3.

We maintain the performance of existing fusion methods and exceed them in some situations while significantly reducing the amount of computation required. For MOSEI we report more than a five-fold reduction in computational cost while achieving the best performance in terms of both Top1 accuracy and mAP. For VGGS10 and VGGS100, we observe approximately a six-fold reduction in computational cost. Our method also exceeds the performance of multiple fusion methods on the VGGS100 dataset.

3.5.2 Effect of Sparsification

In this section, we explore the effect of how naively applying pooling can affect multimodal models. In particular, we are interested in how pooling affects fused versus modality-independent features. We answer this question by comparing the performance of late fusion, concatenation fusion, and our fusion method. For concatenation fusion, we concatenate all the input tokens prior to input into the model. From here, we apply

	Mem (GB)	Eval (ms)	Train (ms)	GFlops
Concat	1.52 (3.16×)	3.59 (2.46×)	10.93 (2.56×)	1.68 (6.72×)
LF	1.35 (2.80×)	3.54 (2.42×)	12.32 (2.88×)	1.51 (6.04×)
MulT	1.18 (2.45×)	3.53 (2.41×)	16.73 (3.91×)	2.64 (10.56×)
MBT	1.35 (2.82×)	3.59 (2.45×)	12.02 (2.81×)	1.52 (6.08×)
Ours	**0.48**	**1.46**	**4.27**	**0.25**

(a) VGGS10/VGGS100 comparison

	Mem (GB)	Eval (ms)	Train (ms)	GFlops
Concat	1.04 (11.95×)	2.71 (2.39×)	8.02 (2.01×)	1.16 (11.60×)
LF	0.49 (5.66×)	2.00 (1.77×)	7.66 (1.92×)	0.59 (5.90×)
MulT	0.62 (7.10×)	2.84 (2.50×)	11.49 (2.88×)	1.03 (10.30×)
MBT	0.50 (5.72×)	2.14 (1.89×)	7.42 (1.86×)	0.59 (5.90×)
Ours	**0.09**	**1.13**	**3.99**	**0.10**

(b) MOSEI comparison

Table 3.2: Computational cost comparison for each dataset and model. For all metrics we obtain results with a single RTX 3090. Metrics are normalized by the batch size. Our method has the lowest cost. GFlops is estimated based on number of transformer blocks and token operations and represents a theoretical cost for a single forward pass through the network.

a single transformer block as if the number of modalities is $M = 1$. We then apply max pool with a kernel and stride of 64. Afterwards, we apply eleven more transformer layers to obtain the result. For late fusion and our method, we also apply pooling on the representations after the first layer. However, the pooling is conducted on unimodal representations. In late fusion, transformer layers are applied independently for each modality and the final result is obtained via a summation of logits obtained from the `CLS` token. In experiments described in Sec. 3.5.3, we observe a drop in performance for our method with strides larger than 32 for some datasets and 128 for others, thus we assume a stride of 64 will provide meaningful comparisons between fusion methods.

The results shown in Table 3.3 demonstrates that our method for sparsification is more robust than naive methods. We see that in both naive methods of pooling, the reduction in the sequence dimension causes a significant drop in performance. Our method does

		Token Reduction Factor		
		None	**64×**	**Diff.**
Concat	*Top1*	$51.72 \pm .26$	$46.29 \pm .73$	-5.45
	mAP	$51.64 \pm .13$	$47.49 \pm .63$	-4.49
LF	*Top1*	$52.00 \pm .73$	$49.76 \pm .71$	-2.24
	mAP	$46.92 \pm .28$	$45.96 \pm .62$	-0.96
Ours	*Top1*	$\mathbf{55.57} \pm .23$	$\mathbf{55.98} \pm .28$	$+0.41$
	mAP	$\mathbf{56.54} \pm .49$	$\mathbf{56.91} \pm .60$	$+0.37$

Table 3.3: Comparison of our method for sparsification versus application of only pooling in baseline methods on VGGS100. `Diff` column shows difference between no reduction of tokens and taking 1/64ths of the tokens, where the minimum is one token per modality. Our method is more robust than naive methods of pooling. Pooling has a large effect when training with fused features (Concat) which we solve using our method. Difference for the same reduction factors between Top1 and mAP shows that late fusion (LF) tends to fit some samples better than others and suggests the advantages of an early-fusion method.

not see any reduction, instead experiencing a small boost in performance. Furthermore, we see that concatenation fusion tends to have a higher mAP metric, whereas late fusion has a higher Top1. Overall, our method is robust, and pooling has no detrimental effect even when removing over 98% of tokens.

3.5.3 Within-Modality Information Redundancy

We provide experiments to analyze why it is advantageous to address the within-modality redundancy problem during fusion. In particular, we wish to show that pooling when accounting for multimodal information is more robust than pooling without this information. We set up the experiment so that a max-pooling layer is applied after the first layer of transformers to simulate modality-independent feature sparsification for each method. We then compare pruning by an equal factor for each modality to observe the effect on overall performance, referred to as "sequence reduction factor." We set the minimum allowed sequence length to one to avoid removing all tokens. We compare

	VGGS10		VGGS100		MOSEI	
Pooling Method	*Top1*	*mAP*	*Top1*	*mAP*	*Top1*	*mAP*
Max	67.0 ± 1.1	70.7 ± 0.7	55.7 ± 0.4	57.3 ± 0.4	49.4 ± 0.2	33.2 ± 0.3
Average	67.7 ± 1.2	71.1 ± 0.7	55.6 ± 0.5	57.2 ± 0.3	49.7 ± 0.2	33.7 ± 0.9
Attn Average	67.5 ± 1.0	71.1 ± 0.7	55.4 ± 0.3	56.9 ± 0.4	49.4 ± 0.2	33.7 ± 0.8

Table 3.4: Comparison of pooling method on VGGS10, VGGS100, and MOSEI datasets. Based on Top 1 accuracy and mean average precision metrics, we find our method robust to pooling type.

against unimodal transformers for each modality. We also evaluate two versions of our method: SFT which is our full pipeline, and SFT-PO which removes the strided sparse attention layer and multimodal manifold mixup and includes only the strided pooling.

In the first column of Fig. 3.4, we present Top1 accuracy as a function of sequence reduction factor. In the second column, we present the relative change in Top1 accuracy when compared with no sequence reduction. Lower indicates a performance degradation from sequence reduction. Multimodal models exceed unimodal performance in all reduction factors. We generally see a performance decrease for each unimodal model as the reduction factor increases. However, some modalities do not decrease due to two likely reasons: 1) from redundancies in information and 2) that all useful information was extracted after just a single layer of transformers. We also see that some modalities experience an increase in performance as we reduce the number of tokens, signifying better feature extraction for those. However, in general, the performance of unimodal models with less redundant information all decrease, while our model (SFT) is more robust. In particular, SFT is better than using just pooling (SFT-PO) as is evident from it maintaining higher performance with greater reduction factors.

We see that up to a factor of 50 for evaluations conducted on MOSEI, there is very minimal drop in performance in the multimodal model. However, the performance of the text-only transformer drops observably larger than our multimodal model. The perfor-

mance of the RGB and Audio transformers remains the same throughout the experiment. This signifies two things: that the information for label present in the text classifier is less redundant than in RGB and Audio features for this dataset, and that application of sparse fusion can compensate for the loss of information necessary for classification by exploiting the other modalities. The effect of unimodal models experiencing a decrease in performance is also evident for the optical flow modality on the VGGS10 dataset at 8× reduction, and at 64× reduction for spectrogram data. On VGGS100, we see the same, where both the RGB and flow modalities experience decreases in performance with a pruning factor of just 2× while our model's performance remains relatively flat. Furthermore, our multimodal model with one token per-modality after pruning still achieves better performance than a unimodal model which uses all tokens.

These observations signify that certain modalities contain information that is more redundant than others and that even if we filter out more than what a model with redundant information is able to predict, the multimodal model is able to make up for that. The same is not true for unimodal models, which cannot filter out unnecessary information as well, and is not robust to this reduction. Even under extreme circumstances where information is reduced to the length of a single token, performance of the multimodal degrades but still remains the overall top performer.

3.5.4 Effect of Pooling

In this section, we explore the effects of using various pooling choices in the network. See Table 3.4 for results on the VGGS10, VGGS100, and MOSEI datasets. We use max pooling, average pooling, and attention-weighted average pooling, denoted "Max," "Average," and "Attn Average" respectively. For attention-weighted averaging, we weight using a simple, attention-based per-token significance metric proposed by Goyal *et al.* [107].

mixup?	**Top1**	**mAP**
✓	$55.61 \pm .61$	$57.18 \pm .39$
×	$51.30 \pm .80$	$51.80 \pm .44$

Table 3.5: Comparison of model performance on VGGS100 when trained with and without our multimodal manifold mixup.

Given the attention weights $\mathbf{W}^h \in \mathbb{R}^{N \times N}$ calculated from layer L head $h \in \{1, \ldots, H\}$ of the pre-fusion network, the significance (sig) for token i is:

$$\text{sig}(i) = \sum_{h=1}^{H} \sum_{n=1}^{N} W_{in}^{h} \tag{3.14}$$

Interestingly, all metrics are within 1 percentage point of each other across the three pooling types. This indicates our model is quite robust to the choice of pooling type. Average pooling appears better, but this is well within the std. dev.

3.5.5 Effect of Multimodal Manifold Mixup

See Table 3.5 for results from SFT trained on VGGS100 with and without the use of our multimodal manifold mixup during training. Without mixup, we observe over a 4% reductin in Top1 and over a 5% reduction in mAP. This drop in performance is quite significant, indicating the effectiveness of training with our proposed multimodal manifold mixup.

3.6 Limitations

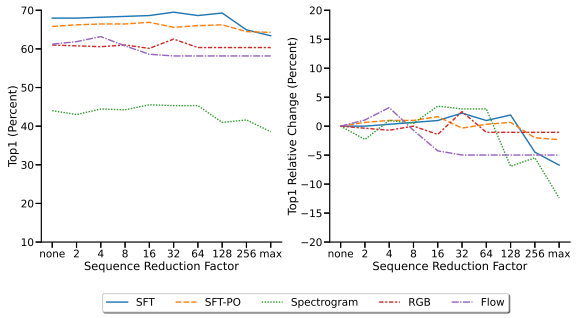
We provide an effective method for quickly ingesting and classifying large quantities of multimodal sequential data with high levels of accuracy. However, we do not provide evaluations on how this fusion method might behave as part of a generative network and

we leave this for future work. Secondly, our methods operate on extracted features such as I3D and spectrogram data. While we follow popular and common settings for feature extraction, improved unimodal modeling might be able to condense the representations and reduce within-modality redundancy. This would lead to slightly reduced complexity benefits. However, the large differences between our results and unimodal approaches as well as maintaining performance under extreme sparsification support our conclusions.

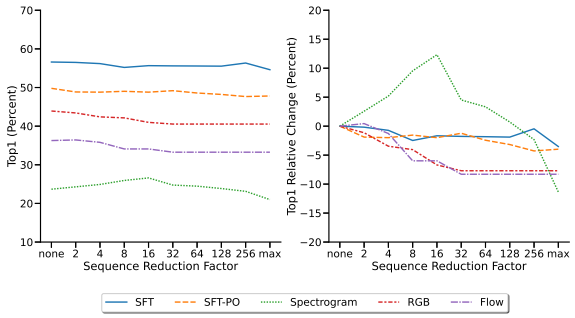
3.7 Conclusion

We present an effective technique that offers more than a five-fold reduction in computational cost while maintaining the performance of state-of-the-art fusion techniques. Different fusion methods exhibit improved performance under varying conditions when all input conditions are equal. However, when optimizing for speed, there are drastic improvements that can be made to feature selection during cross-modal modeling that can improve performance.

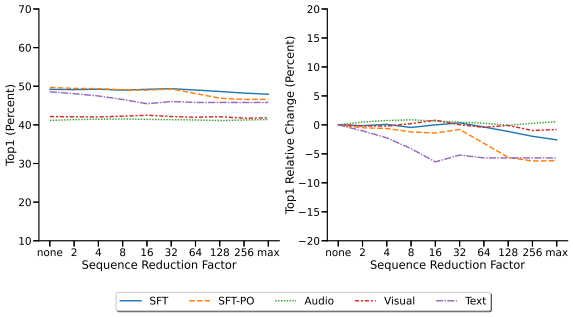
Broader Impacts: We propose sparse fusion for multimodal transformers as a method to reduce computational costs. This translates to energy savings and is beneficial for numerous applications including on mobile devices. Namely, it has the potential to train and fine-tune a network for use to a specific user without needing to offload the training to a server. This preserves the privacy of the user while providing benefits of performance and energy savings. Furthermore, we hope to spur democratization of learning on large datasets by enabling rapid development and evaluation on consumer-level hardware. However, we hope that by enabling this technology on mobile devices it is not applied to tasks such as unlawful surveillance.



(a) Top 1 absolute score and relative change from no pooling for VGGS10. Multimodal performance degredation occurs after a 64-fold reduction in sequence length. Compared to flow at 4, and spectrogram at 64. We outperform all all methods at all reduction levels. SFT exceeds the pooling only variant (SFT-PO).



(b) Top 1 metrics for VGGS100. SFT degradation occurs at 256-fold reduction compared to 64 for SFT-PO and 2 for Flow and RGB. Audio representations might benefit from better feature extraction, however there is dramatic loss of performance with very few tokens, while we remain tolerant.



(c) Top 1 metrics for MOSEI. SFT degrades minimally until max while SFT-PO degrades at 32. Text modality degrades immediately. Information appears highly redundant in Audio-Visual modalities.

Figure 3.4: Comparison of reduction factor effect on performance difference against no reduction for unimodal and multimodal models. Reduction in total length of fused features reported in the x-axis. In cases where the reduction factor is greater than sequence length of a particular modality, a single token along the sequence dimension is passed through. Sequence lengths for VGGS10 and VGGS100 are 38 for RGB and Flow, 1200 for Spectrogram. For MOSEI, Audio and Visual is 500 while text is 50. Top1 absolute score and relative change from using no pruning is reported. For all experiments we used a batch size of 24. Multimodal models will tolerate more pruning over unimodal models by making up for the lost information through fusion. Notably, SFT exceeds performance of SFT without sparse attention or mixup (SFT-PO) in all cases and tolerates more reduction. Pooling offers some benefits for feature extraction in some cases for longer sequences.

Part II

Variability

Chapter 4

A Weakly Supervised Application to Induced Affect Prediction

Parts of the contents of this chapter were published with collaborators at the International Conference on Multimodal Interaction 2020.

We start our examination variability by exploring an application in Affective Computing. Affective Computing encapsulates many forms of human variability and is a natural place to start. It uses computational techniques to model human psycho-physiological states [108]. Researchers have tackled the recognition of these states uni- and multimodally. By recognizing these states, we can enable richer human-computer interaction by encoding human state beyond what is explicitly expressed. Application opportunities are broad and include the ability to automatically determine user opinion, empower individuals with better social cues, automated detection of misbehavior and many more.

Several high quality datasets have recently been developed to study this important problem [12, 109, 110]. These recent works provide both categorical measures and also incorporate continuous ratings of affect to capture subtle emotional differences. However, creating such high quality datasets is very resource intensive.

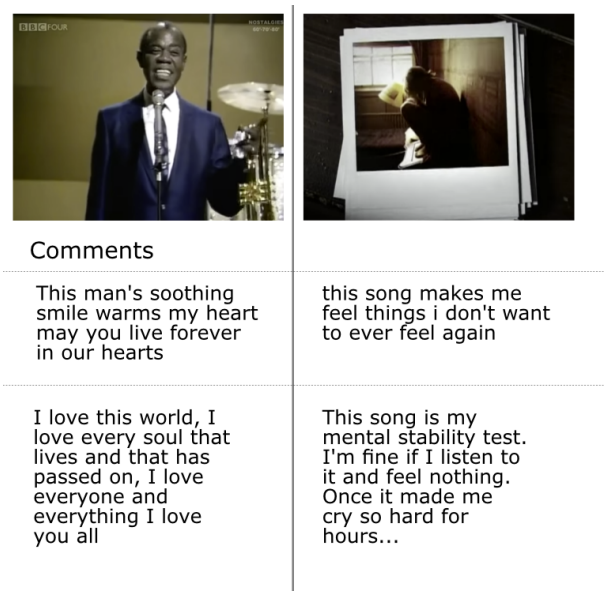


Figure 4.1: Example of comments for Youtube music videos. The left video has a low arousal and high valence rating. The right video has a low arousal and low valence rating. Corresponding comments from YouTube are shown below the figure and demonstrate correlation with the affect contents of the video.

To alleviate this issue, we investigate whether we can use unlabeled public reactions to aid in determining the affect of a corresponding video. Since many emotional reactions are gathered in the lab by presenting a stimuli to induce an emotional response, then we would expect some similar reactions in the wild. For example if the average rating by test subjects in a lab for a particular video is 4 out of 5 for a happiness rating, then we should expect similar reactions and statistical measures by people in the public. Furthermore, we examine whether these responses can then be used to help determine the affect rating of the video.

We study these questions by examining comments found in the wild of videos used to induce emotions in a laboratory setting. We attempt to learn a language model respective of the affect dimensions when given only the laboratory affect ratings of the video. The learned language features are then incorporated into a multimodal affect prediction model to determine video affect. Since each video has potentially millions of

comments, designing an effective way to model this data can drastically reduce the video annotation burden.

Mathematically, our problem can be construed as learning a language representation over a mixture of Gaussians. We assume that each video occupies a region in affect space, and elicits emotional responses according to its distribution (an expected value as determined in a lab). We hypothesize that these emotional responses are reflected in the comments posted to a video. We take each video rating to induce an emotional response which can be used to "mold" the multiple comments associated with each video to region occupied by the video. That is, when given a weak prior in the form of expectations to video reactions, we would like to embed language in this affect space to fit its affective properties.

To learn this, we define a custom variational objective, an approach with demonstrated effectiveness in learning unsupervised sentence representations [111, 112]. By taking advantage of the smooth distribution learned by a VAE as well as weakly supervised information offered by the videos, we can mold the latent distribution of the comments such that they conform closer to the defined affect dimensions.

In summary, we provide the following contributions:

1. We propose a novel problem for learning language representations from induced affect in the wild when given weakly supervised signals.

2. We formalize the problem within the context of a Gaussian mixture and design an effective optimization method.

3. We augment existing datasets with public reactions and make these augmentations publicly available.

4. Experiments that indicate the potential to use induced signals in the wild for affect

prediction tasks.

4.1 Related Work

Affect representation has primarily been studied in two ways: as a categorical selection [113] or via dimensional representations such as the well known arousal-valence model [114, 115]. The research field of sentiment analysis often focuses on measures the valence dimension of affect: positive, negative or neutral [116]. We use the arousal-valence model, which measures affect on two orthogonal dimensions: arousal (the level of alertness/involvement) and valence (a pleasure-displeasure continuum) [114]. However, it should be noted that whether the dimensions are truly orthogonal can be deemed controversial [117].

Early works in affect computation primarily attempted to group people's emotional state into distinct categories. Over the years, researchers have expanded this capability to include more continuous affect representations [110, 109], by capture emotional state in addition to intensity which enable the modelling of modelling subtle difference. Multiple recent works have attempted to learn improved multimodal representations of affect-based data to improve downstream tasks such as video affect classification [118].

Emotion elicitation can by achieved by having subjects watch music videos [119, 12]. Visual [109] and audio stimuli [120] are among the most common modalities for inducing emotions. There has also been an increasing interest in using data collected in response to multimodal stimuli for the task of emotion recognition. Audio-visual stimuli have been studied in the form of monologues [121], conversations [122], and music videos [123]. The DEAP Database [12] contains records of EEG and peripheral physiological signals of participants who watched selected music videos, as well as the participants' self assessment of their emotional state after each trial. This has been used in emotion

recognition and classification tasks [124, 125]. We use the arousal and valence values that have been provided for each video in the database as weak signals to supervise the learned representations in our model.

Language models seek to learn a representation and have been studied actively [126] [127] [128]. More recently, work has been done on using additional modalities in language modelling [129], incorporating symbolic knowledge into language models to allow for generation of rare words [130], and learning generalized representations of data for use in multiple language understanding tasks [131]. Frameworks to learn sentence representations by unsupervised learning methods have also been widely studied, such as asymmetric encoder-decoder structures [132], improvements to the VAE that learn semantics better [133], and the use of discourse relations to learn accurate sentence representations [134] [135]. Our approach, however, is to use a weakly supervised learning method, to embed language with explainable dimensions.

Some techniques used for affect recognition tasks include transfer learning [136] [137], attention modelling [138], and Tree-LSTMs [139]. Ghosh et al. [140] extend the LSTM model for text generation in conversations, allowing for control of the emotional content of the sentences generated. Another approach to controlling the emotion of generated sentences assume that the emojis in Twitter messages indicate the emotion of the conversation, and accordingly generates responses with appropriate emotion [141]. Song et al. [142] have explored affect-based text generation using not only explicitly emotional words, but also neutral words which express an emotion when combined in a specific pattern.

Traditional variational autoencoders (VAEs) usually incorporate a single Gaussian for regularizing latent variables, and Gaussian for the output as well. The output of VAE has been extended with mixture models and it has performed in unsupervised clustering [143] where the clusters are modelled by GMM, and an uniform distribution to model

major clusters and the remaining data [144]. Mixture models in the latent space for semi-supervised learning in classification problems have also been adopted [145], where different mixture components share parameters.

4.2 Problem Formulation

We formalize the problem as follows: We denote V as a given set of videos that have been assigned an emotional rating, where a rating is a two-dimensional vector consisting of valence and arousal scores. Our task is to learn a mapping of comments $f : C \to \mathbb{R}^2$ such that $\mu_c := f(c)$ and μ_c reflects the true valence and arousal of the comments.

Each video has multiple ratings, and the mean $\mu_v \in \mathbb{R}^2$ and variance $\Sigma_v \in \mathbb{R}^{2\times 2}$ of the ratings are given. When the number of raters is large enough, we can reasonably assume that all reactions towards a video v follows a normal distribution $\mathcal{N}(\mu_v, \Sigma_v)$. Explicitly, μ_v is the mean valence $\mu_{(1)}$ and mean arousal score $\mu_{(2)}$, while the covariance matrix Σ_v is diagonal, since we assume that valence and arousal are two uncorrelated [114], orthogonal criterion, i.e.,

$$\mu = (\mu_{(1)}, \mu_{(2)})^T,$$

$$\Sigma = \begin{pmatrix} \sigma_{(1)} & 0 \\ 0 & \sigma_{(2)} \end{pmatrix}.$$

For simplicity, we use $\text{diag}\{\sigma_{(1)}, \sigma_{(2)}\}$ to denote covariance.

The set of comments is denoted by C and comments associated with the video v are represented by C_v. Each video is given a rating μ_v which codes its effect on viewers. We are interested in exploiting the potential emotional influence of the video on any commenters. That is, the learned distribution of comments for a particular video should occupy the region described by the mean and variance of the video. Intuitively, while

there may be a few deviating comments, a large proportion of the comments C_v in a video should agree roughly with the rating. Once we obtain the learned language model, we can then use the average comment scores for affect as an indicator for video affect.

4.3 Learning an Affect Embedding

A variational autoencoder (VAE) is an unsupervised architecture with demonstrated ability to produce quality representations of text [111]. They work by optimizing the parameter θ and maximizing the probability of each c such that:

$$P(c) = \int_{\mathcal{Z}} P_\theta(c|z)P_\theta(z)dz,$$

where $z \in \mathcal{Z}$ is a latent variable sampled by another function $Q(z|c)$ in order to reproduce c. VAEs are an extension of the standard autoencoder which imposes a prior distribution on z. It assumes that samples of z can be first drawn from a standard Gaussian distribution $p(z) \sim \mathcal{N}(0, I)$, where I is the identity matrix of the same column dimension with z. This has been empirically shown to learn smooth regions and enable better continuity in language representations. [111]

Hence we expect the distance between $Q_\phi(z|c)$ and $P_\theta(z|c)$ to be small. Mathematically, the standard VAE objective can be defined as [146] :

$$L_{\theta;\phi}(c) = E_{q_\phi(z|c)}[\log p_\theta(c|z)] - D_{KL}(q_\phi(z|c)||p(z)) \tag{4.1}$$

However, since the posterior distribution learned by the VAE is arbitrary, we cannot guarantee that a representation is learned in the dimensions that we want. Here we propose a simple tweak to use the valence and arousal ratings as a prior to shape the distribution. As a result, the ratings of comments from a video can be modelled as a

two-dimensional uncorrelated Gaussian distribution $\mathcal{N}(\mu_p, \Sigma_p)$, where $\mu_p = (\mu_p^{(1)}, \mu_p^{(2)})$, $\Sigma_p = \text{diag}\{\sigma_p^{(1)}, \sigma_p^{(2)}\}$.

Giving us the following KL term instead of the KL term in equation 4.1 by the property of a diagonal matrix:

$$\begin{aligned} D_{KL}(Q||P) = \frac{1}{2}(&\text{tr}(\log \Sigma_p) - \text{tr}(\log \Sigma_q) - n \\ &+ tr(\Sigma_p^{-1}\Sigma_q) \\ &+ (\mu_p - \mu_q)^T \Sigma_p^{-1}(\mu_p - \mu_q)), \end{aligned}$$

where $\log \Sigma_p := \text{diag}\{\log \sigma_p^{(1)}, \log \sigma_p^{(2)}\}$.

4.3.1 Centered VAE (C-VAE)

Since it is known that stronger stimuli tends to produce a stronger emotional response, we introduce a second KL divergence term. To explain the reasoning, we introduce the definition of *uncertain response*: a response without a specific appropriate stimulus class [147]. Since stimuli with an uncertain response – close to the origin $(0, 0)$ – do not provide additional information regarding the prior, these videos should still contain comments that can vary wildly depending on personal preference which could "cover" the latent space.

Since the comments should match the center of the ratings with weight $1 - \lambda$, we construct the second KL term with the prior $p(x)$ sampled from the distribution of the entire dataset, i.e., for all $v \in V$. $p(x) \sim \mathcal{N}(\mu_x, \Sigma_x)$. For a total of N videos in V,

$\mu_x = \frac{1}{N}\sum_{i=1}^{N}\mu_i$; and we compute $\sigma_x^{(r)},\ r = 1, 2$ as follows:

$$\begin{aligned}(\sigma_x^{(r)})^2 &= E[(x^{(r)})^2] - (\mu_x^{(r)})^2 \\ &\approx \frac{1}{N}\sum_{i=1}^{N} E[(x_i^{(r)})^2] - (\mu_x^{(r)})^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\big\{\big(\mu_i^{(r)}\big)^2 + \big(\sigma_i^{(r)}\big)^2 - \big(\mu_x^{(r)}\big)^2\big\}.\end{aligned}$$

Using this mean and variance, we can create our new term $D_{KL}\big(Q||\mathcal{N}(\mu_x, \sigma_X)\big)$. We use a λ term to weigh the potential variance in emotional responses based on its Euclidean distance to the original. This gives us our final loss function:

$$\begin{aligned}L_{\theta;\phi}(c) = E_{q_\phi(z|c)}[\log p_\theta(c|z)] \\ - \lambda D_{KL}\big(q_\phi(z|c)||p(z)\big) \\ - (1-\lambda) D_{KL}\big(Q||\mathcal{N}(\mu_x, \Sigma_x)\big)\end{aligned} \tag{4.2}$$

4.3.2 Centered Gaussian Mixture VAE (CGM-VAE)

As it is a strong assumption that all comments to videos are normally distributed, we propose to use a Gaussian mixture prior. This allows us to be more accurate with respect to the center of the video ratings given by the prior $p(z)$. We extend the work of Hershey and Olsen [148] which demonstrates numerous ways to approximate the KL divergence of Gaussian mixtures.

Recall that a Gaussian mixture consists of multiple Gaussian distributions and the proportion of each mixture component which is represented as a latent variable that yields the multinomial distribution. So we use unlabeled samples $\{y_i\}_{i=1,\ldots,n}$ from n

multi-dimensional Gaussians with known covariance matrices. This yeilds the mixture:

$$f_\theta(y) = \sum_{i=1}^{n} \pi_i \phi(y; \mu_i, \sigma_i^2 I_d)$$

where π_k is the mixing proportion of the k-th Gaussian distribution satisfying $\sum_{i=1}^{n} \pi_i = 1$, and $\phi(\cdot; \mu, \Sigma)$ denotes the density of a $\mathcal{N}(\mu, \Sigma)$ random vector in $\mathbb{R}^d$:

$$\phi(w; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(w-\mu)^T \Sigma^{-1}(w-\mu)\}$$

Specifically, for the total $n = N$ number of videos, we give equal weight to each part of the mixture, i.e., $\pi_k = \frac{1}{N}$. Since we don't distinguish the weight of each Gaussian distribution if there is no further information on the importance of the videos, the dimension of each Gaussian is $d = 2$ for the valence and arousal ratings.

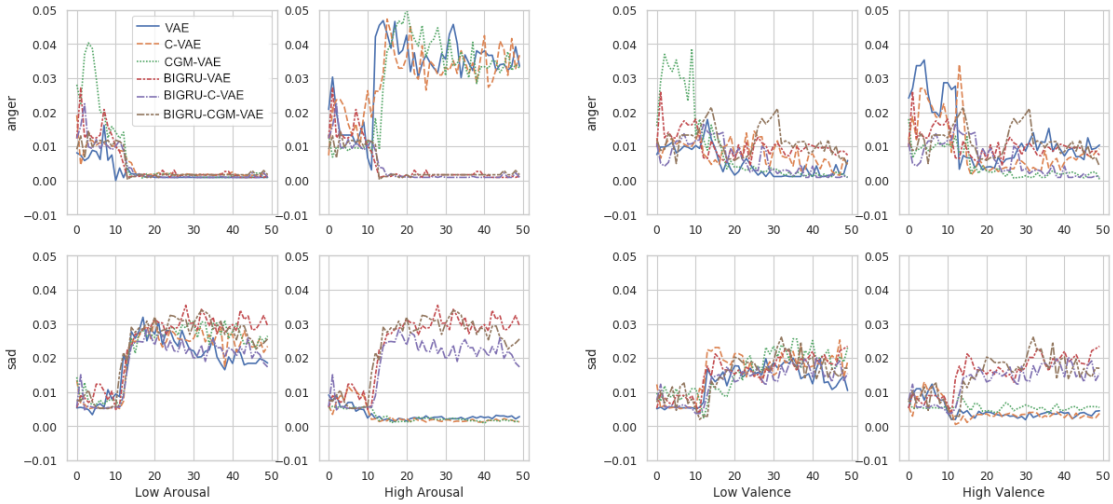
Thus the second KL divergence follows

$$\begin{aligned} D_{KL}(Q||P) = & -\frac{1}{2}\mathrm{tr}(\log \Sigma_q) - 1 \\ & - \frac{1}{n}\mathbb{E}_Q\left[\log \sum_{i=1}^{N} \frac{1}{|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}A_k}\right], \end{aligned} \tag{4.3}$$

where $A_k := (x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)$.

One possible way to make equation 4.3 computationally tractable is through Monte Carlo sampling. We draw K i.i.d samples $\{x_j\}_{j=1}^{n}$ from the distribution of Q, $\mathcal{N}(\mu_q, \Sigma_q)$:

$$\begin{aligned} D_{KL}(Q||P) = & -\frac{1}{2}\mathrm{tr}(\log \Sigma_q) - 1 \\ & + \frac{1}{2N}\sum_{k=1}^{n}\Big\{\mathrm{tr}\big(\log \Sigma_k\big) + \mathrm{tr}\big(\Sigma_k^{-1}\mu_k\mu_k^T\big) \\ & + \frac{1}{K}\sum_{j=1}^{K}\mathrm{tr}\big(\Sigma_k^{-1}x_j(x_j^T - 2\mu_k^T)\big)\Big\}, \end{aligned}$$



(a) Arousal comparison over testing epochs epochs

(b) Valence comparison over testing epochs

Figure 4.2: Average of fine grained scores provided by LIWC for top and bottom 500 comments of learned embeddings. Anger is associated with high arousal but neutral valence, while sadness is associated with low arousal and low valence. A large inflection can be seen at approximately 12 epochs of training time due to the delay of kl-annealing. As can be seen, there is a correlation with some tested models

where $A_k^j := (x_j - \mu_k)^T \Sigma_k^{-1} (x_j - \mu_k)$.

4.4 Experiment Setup

We conduct two primary experiments to validate our optimization methods presented in equations 4.2 and 4.3. We examine the predictive power of our technique and its ability to learn embeddings through induced emotion signals. We then compare the fusion of our learned embeddings into a multimodal model to examine results compared on a state of the art benchmark. Our optimization methods are referenced as VAE (the standard VAE objective but with a modified prior), C-VAE (Centered VAE), and CGM-VAE (Centered Gaussian Mixture VAE).

Statistic	DEAP	MOSEI
# Videos	120	3228
# Videos with comments	82	764
# Comments	31481	17217

Table 4.1: Summary statistics of dataset

4.4.1 Data

We apply our approach to use text comments on videos as a signal provider for instilled affect to augment two different datasets for affect prediction:

DEAP [12] provides affect annotations for music videos available on Youtube. We used the online subjective annotations video list containing 120 Youtube videos each with 14 to 16 ratings. A 9-point rating for valence, arousal, and dominance were collected, although we only examine the valence and arousal dimensions. We ask readers to refer to the original paper for detailed analysis [12]. For our use case, we used DEAP's valence and arousal ratings to embed comment language in a 2-dimensional space.

MOSEI [110] is a large multimodal sentiment and emotional dataset containing 23453 segments of videos by 1000 distinct speakers. Each video is an opinion video clip which is annotated in segments by 14 expert judges. Sentiment annotations on a Likert scale from -3 to 3 and Ekman emotions are annotated on a Likert scale of [0,3] from no evident emotion to high presence of emotion. For our use case, MOSEI's emotional space provided an additional 6-dimensional embedding vector for each comment. Additionally, since no video rating was provided, we took the mean of all segment-level ratings for each video as the overall video rating as input to our model.

For all videos we crawled the available comments. The maximum number of comments per video was limited to 1000 and we exclude videos with no comments. Additionally, some videos were no longer available at the time of data collection. This resulted in 82 videos with comments for DEAP and 764 usable videos for MOSEI. Summary statistics

are available in Table 4.1.

DEAP expresses instilled emotion (i.e. emotion of the viewer of a music video), while MOSEI characterizes the emotional state of the speaker in a video. User comments on a video may be more directly indicative of the user's emotion than the speaker's emotion, and we will evaluate this use case in Sections 4.5.1-4.5.3. Regarding the second case, it is our hypothesis that a causal connection sufficient for a distinctive signal likely exists as well, i.e., if I see a video of a happy/angry/sad person, I'm more likely to write a happy/angry/sad comment myself. We will show results on the MOSEI dataset in Section 4.5.4.

4.4.2 Preprocessing

The crawled comments are preprocessed to keep only the top-most level comment to remove any unrelated discussion using the `@user` expression. We also removed non-english comments and discarded sentences longer than 50 words, and shorter than 2 words for ease of language modeling. GloVE [149] word embeddings are used and kept fixed during training.

The dataset is split into an 80%-20% training-testing by randomly selecting 80% of the videos and their associated comments for training and the rest for testing. Validation is split from the training set during model tuning and for cross validation experiments in an 80%-20% fashion. For MOSEI evaluations, we observed the training, validation and testing splits provided by the MOSEI sentiment classification dataset.

4.4.3 Network Architecture

We train our comments embedding network using a Recurrent Neural Networks (RNNs) connected in an end-to-end fashion [150] as the foundation for our modeling.

We follow the the work from [111] closely in learning and optimization procedures, but use our learning objective.

Multiple network architectures were evaluated for our experiments. We used a Gated Recurrent Unit (GRU) as the base recurrent architecture. Single layer and 2 layer GRUs were used to evaluate our results. A 2-layer MLP is added to the output of the GRU to predict output distributions. Decoder architectures were varied with single, and 2-layer bidirectional GRU variants. The BiGRU tag is used to indicate the 2-layer bidirectional variant.

4.4.4 Hyperparameter Tuning

A the encoder and decoder hidden vector size was set to 100. Glove embeddings of 100 dimensions were used and kept fixed during model training. A two layer feed-forward network is attached to the output of the decoder GRU to predict word tokens. Monte Carlo samples used to approximate the gaussian mixture prior was set to 200. Although we experimented with different λ values, no large differences were noticed and were set at .5 for the entire experiment.

Hyperparameters were tuned on the validation set. The standard AdamW optimizer with all default options. A batch size of 128. Both Dropout and word dropout are used and is set to 0.2. Sigmoid KL annealing as used to train the evaluated VAEs offset by 15 epochs.

4.5 Results & Discussion

1) We provide empirical evaluation of the predicted video emotion ratings (Sections 6.1 & 6.2). 2) We show that our metric approximations agree with crowd-sourced user rated affect scores (Section 6.3). 3) We apply our technique to a large-scale public

benchmark dataset for multimodal emotion analysis (MOSEI) and show that we can learn fine-grained emotion ratings for individual user comments while matching overall video emotions as the aggregate emotion of all user comments. We also demonstrate the ability for our embeddings to successfully extend an existing model with user comments as an additional dimension (Section 6.4).

4.5.1 Analysis on predictive power of comments

We perform an empirical evaluation of the learned language representations using the augmented DEAP dataset, as music performance is known for its ability to induce emotions, as demonstrated in lab studies. DEAP provides annotations for each video by multiple users in a valence, arousal and dominance space. We examine whether individual comments can be placed in a space close to their valence and arousal rating without the supervised information from the videos. We perform multiple experiments to correlate our predictions of valence and arousal scores with supervised tools which analyze language affect.

Supervised language analysis tools

The evaluations of the learned language representations are compared with two well-known tools for analyzing affect content in text. These tools are often used to provide distantly supervised information for related machine learning tasks [140] and are built on supervised knowledge. It is our expectation that if our learned representation *without* supervised information, demonstrates a positive correlation with existing supervised techniques, then we can expect that the video has 1) induced an expected emotion in the user and 2) our model can extract this information. This supervised information is *not* provided during training or testing time.

Two popular language analysis tools are used to analyze the video coments:

LIWC2015 [151] is a proprietary tool which produces scores for various dimension of language use. The typical output measures the fraction of words which fall under some variable.

The tone score is used to analyze our predicted valence score. It is a variable that measures the positive or negative tone of a text. Additionally, measures for anxiety, anger and sadness are provided which are typically associated with high, high and low arousal emotions [152].

VADER [153] is a lexicon and rule-based sentiment analysis tool designed for social media contexts. provides ratings for proportion of text which fall under categories of positive, neural, or negative as well as a normalized compound score. The compound score ranges from -1 to 1 and represents a summarizing of the overall positiveness or negativeness of the input text sequence. It is the expectation that the compound score most closely related to valence.

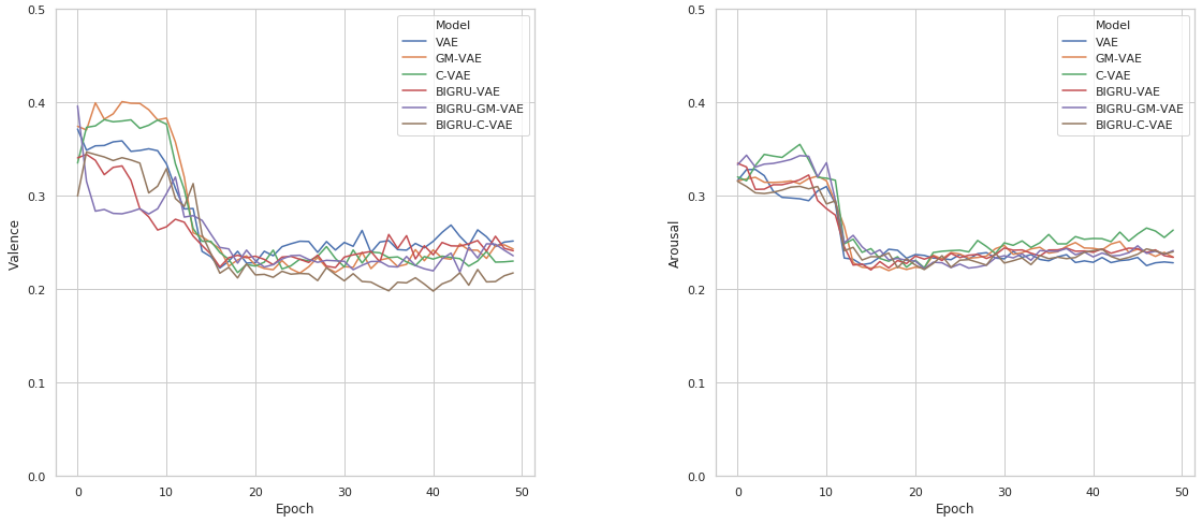
Metrics

As we do not have ground truth for sentence-level affect scores, we define an approximation based on expected emotion correlations:

$$V = \frac{S_c + S_t}{2} - .5$$
$$A = S_a + S_x - S_s$$

Where the S_c represents the normalized (between 0 and 1) in VADER compound score, and $S_t, S_a, S_x S_s$ represents the normalized in LIWC tone, anger, anxiety, and sadness scores respectively.

The tone and compound scores reflect the positive versus negative emotions present



(a) MAE of predicted valence by epoch

(b) MAE of predicted arousal by epoch

Figure 4.3: 10-fold Cross-validation prediction of video rating with the DEAP dataset. MAE valence and arousal of our predicted comments ratings with defined metrics per epoch is shown.

in the sentence. As there is no direct measure for arousal, we correlate with the LIWC measure for anger and sadness which are respectively positively and negatively correlated with arousal. We found that these metrics typically produced affect scores between -.5 and .5.

4.5.2 Video Affect Regression

We perform a 10-fold cross-validation evaluation with random initialization. The training set is split into 80% training and 20% validation. The MAE distance of predicted affect scores from our model with our valence and arousal metrics are shown in Figure 4.3. As can be seen, while the initial training shows large variations, all models eventually converge to a value closer to scores given by supervised approaches. Our Gaussian mixture optimization technique also shows the best average overall performance at epoch 50.

Model	Valence MAE	Arousal MAE
Random	.33	.33
VAE	.222	.225
C-VAE	.217	.220
GM-VAE	.217	.220
BIGRU-VAE	.223	.221
BIGRU-C-VAE	.198	.221
BIGRU-GM-VAE	.218	.222

Table 4.2: Minimum MAE for valence and arousal ratings of video. Random shows the average MAE from randomly choosing scores and represents a baseline. Minimum possible value and best possible score is 0. As can be seen, our optimization method can learn embeddings that enable predictions close to the valence and arousal rating of the video.

Table 4.2 shows the epochs with the minimum valence and arousal MAE values. As can be seen, the learned embeddings are moving away from a randomly embedded space into one which correlates with valence and arousal. Additionally, we see in figure 4.2, emotion scores from LIWC for each comment correlates with the expected embedding within valence and arousal space. We also see that comments are slowly conforming to the mold given by the prior distribution.

4.5.3 Perception Study

We conducted a user study via Amazon Mechanical Turk asking users to rate the valence and arousal properties of learned comments representations. We compare the top and bottom ranked 100 comments for each dimension (valence and arousal) for each algorithm. All participating workers were from the US, with an approval rating greater than 98%. Workers provides ratings on a 5 point Likert scale, for valence as well as arousal of each comment.

Workers worked in batches of 16 comments with each sentence being rated by two unique workers. Inter-rater agreement was measured using Krippendorff's α as an ordinal

Method	Bottom 100	Top 100
VAE	0.58 ± 0.02	0.60 ± 0.02
C-VAE	0.50 ± 0.03	0.67 ± 0.02
GM-VAE	0.56 ± 0.02	0.66 ± 0.02
BiGRU-VAE	0.54 ± 0.01	0.65 ± 0.02
BiGRU-CGM-VAE	0.54 ± 0.01	0.62 ± 0.02

Table 4.3: User valence scores for the comments with model-estimated valence scores ranked in the top 100 and bottom 100. Scores range from 0 to 1. As we can see based on user ratings, the comments scored lower by our algorithm exhibit lower valence ratings and higher ranked comments received overall higher ratings from human raters.

Method	Bottom 100	Top 100
VAE	0.47 ± 0.02	0.61 ± 0.02
C-VAE	0.45 ± 0.02	0.57 ± 0.02
GM-VAE	0.50 ± 0.03	0.59 ± 0.02
BiGRU-VAE	0.51 ± 0.01	0.59 ± 0.02
BiGRU-CGM-VAE	0.51 ± 0.01	0.57 ± 0.01

Table 4.4: User arousal scores for the comments with model-estimated arousal scores ranked in the top 100 and bottom 100. Scores range from 0 to 1. A similar correlating trend is seen here with arousal scores.

metric, with $\alpha = 0$ representing perfect disagreement and $\alpha = 1$ representing perfect agreement. For this study $\alpha = 0.475$ for the *arousal* scale, and $\alpha = 0.686$ for the *valence* scale.

Table 4.3 shows the mean valence scores, as rated by the workers, of the top and bottom 100 comments. We perform a two population means t-test to compare the models, with a significance level $\alpha = 1\%$. C-VAE is significantly better than GM - VAE (p-value $= 0.0444$) and VAE (p-value $= 0.0126$) when identifying low-valence comments, and also outperforms VAE (p-value $= 0.0135$) in identifying high-valence comments.

The mean arousal scores are displayed in Table 4.5.3. The empirical analysis suggested that no model significantly outperforms another, and the results of this perception study indicate the same - no model performs significantly better than the others, both for low

Label	H	Sa	A	F	D	Su
Video Ground Truth	0.13	0.07	0.17	0.17	0.1	0
Video Prediction	**0.18**	0.07	**0.07**	0.01	0.05	0
Comments Prediction	**0.3**	0.08	**0.04**	0	0.03	0.01

(a) Ratings by method for sentence: "I liked lord of the flies in High School . It was really good i thought ."

Label	H	Sa	A	F	D	Su
Video Ground Truth	0.33	0	0.17	0	0	0
Video Prediction	**0.23**	**0.07**	0.06	0	**0.04**	0.01
Comments Prediction	**-0.05**	**0.09**	0.17	-0.02	**0.08**	0

(b) Ratings by method for sentence: "I hate how my English teacher just makes us write and give a public speech without even teaching how to do that ! ! ! ! ! ! ! "

Label	H	Sa	A	F	D	Su
Video Ground Truth	0.17	0.17	0	0.06	0.06	0
Video Prediction	**0.18**	**0.09**	0.08	0.01	0.05	0
Comments Prediction	**0.41**	**0.05**	0.04	0.01	0.02	0

(c) Ratings by method for sentence: "Very interesting video, Melody, thank you!"

Table 4.5: Selected results, comparing the ground truth and predicted emotion of a video to that of a single comment on the same video. This demonstrates that the network learns better fine-grained embeddings than are provided by the overall video ratings. The six dimensions of emotion are happiness(H), sadness(Sa), anger(A), fear(F), disgust(D) and surprise(Su). A higher score indicates a stronger presence of the emotion.

as well as high arousal sentences.

4.5.4 MOSEI benchmark

We evaluate the performance of our learned comment embeddings to aid in a multimodal affect classification task. We used the MOSEI sentiment unaligned dataset for this task. We compare against two state of the art techniques MulT [87] and Raven [81]. We follow the experiment setup from [87] and use their CTC augmentation of RAVEN.

We learn comment-level embeddings on the training set and accordingly predict the

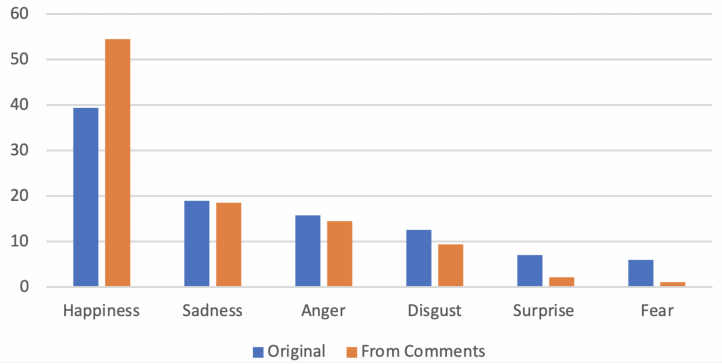


Figure 4.4: Percentage of emotions predicted by our learned comments compared to the original dataset.

6-dimensional emotion vector for each comment in the test set. Figure 4.4 shows that our video predictions capture the overall relative distribution of emotions from the original segment ratings.

Video-level emotional embeddings are generated by averaging the segment level predictions.

We concatenate the predicted video emotion ratings for each video onto the text embedding to fuse the comments context vector with segment-level text information. The resulting text representation is fed through the the network from [118] to perform the prediction.

Table 4.6 shows results from the MOSEI sentiment classification task which predicts sentiment classes for video segments. Our augmented affect predictions incurs but a slight effect on the final predicted segment sentiment scores. One limitation of our approach is that the user comments refer to the entire video, whereas MOSEI sentiment classification occurred on the level of shorter video segments.

Note also that the YouTube dataset makes up a small portion of the overall dataset and thus of the original 3228 videos, only 764 had comments (cf. Table 4.1). With additional comment information performance could potentially improve.

Additionally, we provide a case study on the predicted emotion ratings of individual

Model	Acc7	Acc2	F1	MAE	Corr
CTC+RAVEN	45.5	75.4	75.7	0.664	0.599
MulT	50.1	81.0	81.2	.610	.681
MulT + C-VAE (ours)	49.1	81.2	81.5	.618	.681

Table 4.6: MOSEI Sentiment classification results on unaligned data. We see the augmentation of existing state of the art techniques improves its performance in a few situations.

comments in table 4.5. As can be seen in multiple examples, despite the overall video not providing detailed emotion representations, we can still provide an effective prediction of the comment's emotions. For example, looking at the last row of table 4.5, we notice that the video has a positive emotion overall (H is larger than all the negative emotions). The comment however (which is clearly sad), is predicted to have Sa larger than H, which is accurate.

4.6 Conclusion

In this chapter we examined the problem of learning sentence representations in affect space when given a weak prior in the form of a video affect rating. We introduced a novel problem and proposed and evaluated an effective optimization technique.

Our empirical evaluation of the predicted video emotion ratings show that it is possible to deduce affect from video content alone and that our approximation metrics agree with crowd-sourced user rated affect scores.

When applying our technique to a large-scale public benchmark dataset for multimodal emotion analysis (MOSEI), we show that we can learn fine-grained emotion ratings for individual user comments while matching overall video emotions as the aggregate emotion of all user comments. This demonstrates that our embeddings can successfully extend an existing multimodal model with user comments as an additional dimension.

While we did not achieve the best performance here likely due to the differences in the way MOSEI and DEAP obtained affect labels.

Overall, we provided new augmentations of multimodal video datasets and demonstrated the potential for reactive signals in the wild, in the form of user comments, to predict the affect induced by the videos, through modeling effective language representations in affect space.

Chapter 5

Improving Learning under Imperfect Data Conditions

Portions of this chapter were published with collaborators at the Conference on Computer Vision and Pattern Recognition 2021

We seek to find a method to address imperfections in the data more directly. This leads us to examine how we can perform effective classification in the presence of label noise. In particular, we are interested in how we might improve such techniques. In this chapter, explore how data augmentation can be effectively applied for this type of problem.

Data augmentation is a common method used to expand datasets and has been applied successfully in many computer vision problems such as image classification [154] and object detection [155], among many others. In particular, there has been much success using learned augmentations such as AutoAugment [156] and RandAugment [157] which do not require an expert who knows the dataset to curate augmentation policies. It has been shown that incorporating augmentation policies during training can improve generalization and robustness [158, 159]. However, few works have explored their efficacy

for the domain of learning with noisy labels (LNL) [160].

Many techniques which tackle the LNL problem make use of the network memorization effect, where correctly labeled data fit before incorrectly labeled data as discovered by Arpit et al. [161]. This phenomenon was successfully explored in Deep Neural Networks (DNNs) through modeling the loss function and the training process, leading to the development of approaches such as loss correction [162] and sample selection [163]. Recently, the incorporation of MixUp augmentation [99] has dramatically improved the ability for algorithms to tolerate higher noise levels [164, 165].

While many existing works use the common random flip and crop image augmentation which we refer to as *weak augmentation*, to the best of our knowledge, no work at the time of writing has explored using more aggressive augmentation from learned policies such as AutoAugment during training for LNL algorithms. These stronger augmentation policies include transformations such as rotate, invert, sheer, etc. We propose to incorporate these stronger augmentation policies into existing architectures in a strategic way to improve performance. Our intuition is that for any augmentation technique to succeed, they must *(1)* improve the generalization of the dataset and *(2)* not negatively impact the loss modeling and loss convergence behavior that LNL techniques rely on.

With this in mind, we propose an augmentation strategy we call Augmented Descent (AugDesc) to benefit from data augmentation without negatively impacting the network memorization effect. Our idea for AugDesc is to use two different augmentations: a weak augmentation for any loss modeling and pseudo-labeling task, and a strong augmentation for the back-propagation step to improve generalization.

In this chapter, we propose and examine how we can incorporate stronger augmentation into existing LNL algorithms to yield improved results. We provide some answers to this problem through the following contributions:

- We propose an augmentation strategy, Augmented Descent (AugDesc), which demonstrates state-of-the-art performance on synthetic and real-world datasets under noisy label scenarios. We show empirically that this can increase performance across all evaluated noise levels (Section 5.3.4). In particular, we improve accuracy on the CIFAR-10 benchmark at 90% symmetric noise by more than 15% in absolute accuracy, and we also improve performance on the real-world dataset Clothing1M (Section 5.3.5).

- We show that there is a large effect on performance depending on how augmentation is incorporated into the training process (Section 5.3.2). We empirically determine that it is best to use weaker augmentation during earlier epochs followed by stronger augmentations to not adversely affect the memorization effect. We analyze the behavior of loss distribution to yield insight to guide effective incorporation of augmentation in future work (Section 5.3.3).

- We evaluate the effectiveness of our augmentation methodology by performing generalization studies on existing techniques (Section 5.3.7). Without tuning any hyperparameters, we were able to improve existing techniques with only the addition of our proposed augmentation strategy by up to 5% in absolute accuracy.

5.1 Related Work

Learning with Noisy Labels The most recent advances in training with noisy labels use varying strategies of *(1)* selecting or heavily weighting a subset of clean labels during training [166, 167, 163, 168], or *(2)* using the output predictions of the DNN or an additional network to correct the loss [169, 170, 171, 162, 172].

Many methods use varying strategies of training two networks, using the output of

one or both networks to guide selection of inputs with clean labels. Decoupling [166] maintains two networks during training, updating their parameters using only inputs which the two networks disagree on. MentorNet [167] pre-trains an extra network and uses the pre-trained network to apply weights to cleanly labeled inputs more heavily during training of a student network. Co-teaching [163] maintains two networks, and feeds the low-loss inputs of each network to its peer for parameter updating. The low-loss inputs are expected to be clean, following the finding that DNNs fit to the underlying clean distribution before overfitting to noisy labels [161]. INCV [168] trains two networks on mutually exclusive partitions of the training dataset, then uses cross-validation to select clean inputs. INCV uses the Co-teaching architecture for its networks. The main drawback of these strategies is they only utilize a subset of the information available for training.

The second category of techniques attempts to use the model's output prediction to correct the loss at training time. One such common method is to estimate the noise transition matrix and use it to correct the loss, as in forward and backward correction [170] and S-Model [171]. Another common method is to linearly combine the output of the network and the noisy label for calculating loss. Bootstrap [169] replaces labels with a combination of the label and the prediction from the DNN. Joint Optimization [162] uses a similar approach to the work in [169], but adds a term to the loss to optimize the correction of noisy labels. D2L [172] monitors the dimensionality of subspaces during training, using it to guide weighting of a linear combination of output prediction and noisy label during loss calculation.

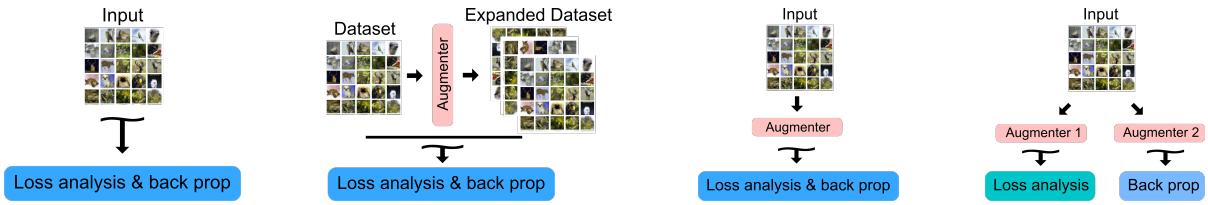
Optimized Augmentation Augmentation of training data is a widely used method for improving generalization of machine learning models. Recent works such as AutoAugment [156] and RandAugment [157] have focused on studying which augmentation policies are optimal. AutoAugment uses reinforcement learning to determine the selection

and ordering of a set of augmentation functions in order to optimize validation loss. To remove the search phase of AutoAugment and therefore reduce training complexity, RandAugment drastically reduces the search space for optimal augmentations and uses grid search to determine the optimal set. Both techniques are widely used in semi-supervised settings.

In semi-supervised learning settings, augmentation has been successfully applied to consistency regularization [173, 174, 175, 176]. In consistency regularization, a loss is applied to minimize the difference in network prediction between two versions of the same input during training. [173] uses a mixture of augmentation, random dropout, and random max-pooling to produce these two versions. More recently, unsupervised data augmentation [174] and ReMixMatch [175] minimize the network predictions between a strongly augmented and weakly augmented version of the input. All of these findings motivate us to incorporate strong augmentation within the realm of LNL to improve performance.

The semi-supervised learning problem itself is similar to the LNL problem with the subtle difference that some labels are unknown rather than corrupt. As techniques in semi-supervised learning have been able to make predictions on a larger dataset from a smaller clean dataset, it would be logical that LNL techniques would benefit from the generalization effects of augmentation. In fact, the recent semi-supervised techniques MixUp [99], and Luo et al. [177] all exhibit strong robustness to label noise.

Most recently, FixMatch [176] successfully combines strong vs. weak augmentation in consistency regularization with pseudo-labeling to achieve state-of-the-art results in semi-supervised classification tasks. While we similarly employ two separate pools of augmentation functions for use in downstream tasks, there are key important differences. Most notably, our key idea is separating augmentations used during loss analysis from augmentations used during back-propagation, rather than focusing on pseudo-labeling



(a) Raw (b) Dataset Expansion (c) Runtime (d) AugDesc

Figure 5.1: Visualization of training methods when incorporating different augmentation strategies. Raw takes the input directly and feeds it into the model for loss analysis and back-propagation. Dataset expansion first creates an expanded dataset which is then sampled by batches and fed into the network. Runtime Augmentation applies a random augmentation policy during runtime for each sampled batch. Augmented Descent produces two sets of random augmentations at the batch level: one is used for all loss analysis tasks, and the other is used for gradient descent.

and consistency regularization. Additionally, we apply this idea to LNL, a separate domain with different considerations. We experimentally show improvements for a wide variety of LNL algorithms and demonstrate improvements on both synthetic and real-world datasets.

5.2 Method

We first describe how various algorithms operate within the context of the network memorization effect [161]. We then propose the Augmented Descent strategy for filtering and generating pseudo-labels for high confidence samples based on one set of augmentations, then performing gradient descent on a different set of augmentations. Lastly, we provide an example for how to retrofit existing techniques.

5.2.1 Loss Modeling Under Noisy Label Scenarios

For some training data $D = (x_i, y_i)_{i=1}^{N}$, a classifier can be trained to make predictions using the cross entropy loss:

$$l(\theta) = -\sum_{x,y \in D} y^T \log(h_\theta(x)),$$

where h_θ is the function approximated by a neural network. Fundamentally, many algorithms are exploiting the behavior outlined in Arpit et al. [161] which finds that correctly labeled data tends to converge before incorrectly label data when training neural networks.

Many existing algorithms are then employing some degree of "pseudo-labeling", where the network is using its own guesses to approximate the labels for the remainder of the dataset. This is done by encouraging the learning of high confidence (or lower initial loss) samples via filtering or modifications to the loss function.

For example, in the sample selection technique Co-teaching [163], this is accomplished by feeding low-loss samples to a sister network, training the networks on data which it believes is correct. Abstractly, this would create two datasets from the input for each training epoch of what is believed to be correctly labeled $C = arg\,min_{D:|D| \geq R(T)|D|} l(f, D)$, where $R(T)$ is a threshold for the number of samples to place into the clean set determined empirically by the loss behavior, and incorrectly labeled $I = D \setminus C$. Using these sets, we obtain the loss:

$$l(\theta) = -\sum_{x,y \in C} y^T \log(h_\theta(x)) - 0 * \sum_{x,y \in I} y^T \log(h_\theta(x)).$$

Here, the learning process is ignoring samples which are believed to be incorrectly labeled as the training progresses. This is represented by the 0 term multiplied into what the model believes to be incorrect samples.

By contrast, Arazo et al. [164] accomplishes noise tolerance by incorporating the network's own prediction into its loss as a weighted sum based on the confidence determined by a mixture model fit to the previous epoch's losses, enabling a softer incorporation of the labels:

$$l(\theta) = -\sum_{x,y \in D, w \in W} (1-w) y^T \log(h_\theta(x)) - \sum_{x \in D, w \in W} w z^T \log(h_\theta(x)),$$

where W is a set of weights learned using a beta mixture model and z is the model's prediction for input x. More recently, DivideMix [165] combines these ideas and assigns weights to inputs to incorporate network guesses, separates the input into two sets, and trains with the resulting data in a semi-supervised manner using MixMatch [101].

With this understanding, we propose Augmented Descent (AugDesc) for LNL techniques that employ loss modeling to separate correctly labeled from incorrectly labeled data. We propose to use one augmentation of the input for sample loss modeling and categorization to create the hypothetical sets C and I or to determine the pseudo label z, while utilizing another different augmentation as input to the network h_θ for purposes of back-propagation. This would require twice the number of forward passes during training for each input. The goal of this is so that we do not adversely affect any loss modeling but also be able to inject more generalization during the learning process. We provide an example in section 5.2.4 for how we can incorporate AugDesc into DivideMix.

5.2.2 Augmentation Strategies

We examine the following strategies for incorporating augmentation into existing algorithms. Figure 5.1 presents a conceptual representation for incorporating our augmentation strategy into existing techniques.

Raw: Original image is used without any modifications.

Dataset Expansion: A dataset is created that is twice the original size of the dataset. This is then fed directly into the model without further augmentation.

Runtime Augmentation: Images are transformed before being fed into network at

Algorithm 1

Input: θ^1, θ^2, training batch possibly labeled x, possibly unlabeled u, dataset labels y, gmm probabilities w, number of augmentations M, augmentation policies Augment_1 and Augment_2

$x^{desc} = \text{Augment}_2(x)$
$u^{desc} = \text{Augment}_2(u)$

for $m = 1$ to M
  $x = \text{Augment}_1(x)$
  $u = \text{Augment}_1(u)$
end // co-guessing and sharpening
$p = \frac{1}{M} \sum_m p_{model}(x; \theta^{(k)})$
$\bar{y} = wy + (1 - w)p$
$\hat{y} = Sharpen(y, T)$
$\bar{q} = \frac{1}{2M} \sum_m (p_{model}(\hat{u}; \theta^{(1)})$
$+p_{model}(\hat{u}; \theta^{(2)}))$
$\hat{q} = Sharpen(\bar{q}, T)$
// train using a different augmentation
$\hat{\mathcal{X}} = \{(x, y) | x \in x^{desc}, y \in \hat{y}\}$
$\hat{\mathcal{U}} = \{(u, q) | u \in u^{desc}, q \in \hat{q}\}$
$\mathcal{L}_x, \mathcal{L}_u = \text{MixMatch}(\hat{\mathcal{X}}, \hat{\mathcal{U}})$
$\mathcal{L} = \mathcal{L}_x + \lambda_u \mathcal{L}_u + \lambda_r \mathcal{L}_{reg}$
$\theta^{(k)} = \text{SGD}(\mathcal{L}, \theta^{(k)})$

Figure 5.2: Batch level training modifications to DivideMix for Augmented Descent. Full implementation provided in the supplemental.

runtime.

Augmented Descent (AugDesc): Two sets of augmented images are created. One set is used for any loss analysis tasks, while the other is used for gradient descent. The motivation is that we can learn a better representation for each image while not compromising the sample filtering and pseudo-labeling process.

5.2.3 Augmentation Policy

We evaluate three different augmentation policies, classified into "weak" and "strong". Many algorithms make use of the standard random crop and flip for augmentation [178].

We call this process *weak augmentation*. We experiment with *strong augmentations* using automatically learned policies from AutoAugment [156] and RandAugment [157]. AutoAugment and RandAugment both provide a way to apply augmentations without hand-tuning the particular policy. Our strong augmentation policy first applies a random crop and flip, followed by an AutoAugment or RandAugment transformation, and lastly normalization. For dataset expansion and runtime augmentation, we experiment with both weak and strong augmentations.

We examine three variants of Augmented Descent. AugDesc-WW means we perform loss analysis using a weakly-augmented input, then use this label to train a different weakly augmented version of the same input. Similarly, AugDesc-SS represents strongly-augmented loss analysis, coupled with strongly augmented gradient descent. Finally, AugDesc-WS corresponds to weakly-augmented loss analysis with strongly augmented optimization.

Because AutoAugment is learned on a small subset of the actual data, it is easy to incorporate into existing architectures. We further perform an ablation study using RandAugment to show that our augmentation strategy is agnostic to augmentation policy, as well as the fact that no dataset-specific or pre-trained augmentations are necessary. We use AutoAugment for most of our experiments as it prescribes a pre-trained set of policies, while RandAugment requires tuning that can depend on the networks used as well as the training set size.

5.2.4 Application to State of the Art

While many techniques beyond those above have similar characteristics that we can analyze in a similar manner, we examine this augmentation strategy within the context of the current state-of-the-art DivideMix [165] in this chapter. We then extend our

augmentation strategy to other techniques and report results in the experiments section.

DivideMix incorporates aspects of warm-up, co-training[167, 163], and MixUp [99]. The original DivideMix algorithm works by first warming up using normal cross-entropy loss with a penalty for confident predictions by adding a negative cross entropy term from Pereyra et al. [179]. Afterwards, for each training epoch, the algorithm first uses a GMM to model the per-sample loss with each of the two networks. Using this and a clean probability threshold, the network then categorizes samples into a labeled set x and an unlabeled set u. Batches are pulled from from each of these two sets and are first augmented. Predictions using the augmented samples are made and a sharpening function is applied to the output [101] to reduce the entropy of the label distribution. This produces sharpened guesses for the labeled and unlabeled inputs which is used for optimization.

We outline the application of our augmentation strategy in Algorithm 5.2. We require two different sets of augmentations: one for the original DivideMix pipeline, and one to augment the original input for training with MixMatch losses. Additional examples of implementation in previous techniques are included in the supplemental.

5.3 Experiments

We first perform evaluations on synthetically generated noise to determine an effective augmentation strategy. We then conduct generalization experiments on real-world datasets, apply our strategies to previous techniques, and experiment with alternative augmentation policies.

5.3.1 Experimental Setup

We perform extensive validation of each augmentation technique on CIFAR-10 and CIFAR-100, two well-known synthetic image classification datasets frequently used for this task. CIFAR-10 contains 10 categories of images and CIFAR-100 contains 100 categories for classification. Each dataset has 50K color images for training and 10K test images of size 32x32. Symmetric and asymmetric noise injection methods [162, 180] are evaluated. We perform most of the ablation studies within the DivideMix framework as this is the state-of-the-art technique. We then extend the augmentation strategies we found to other techniques.

We use an 18-layer PreAct Resnet [181] as the network backbone and train it using SGD with a batch size of 128. Some experiments are conducted using a batch size of 64 due to hardware constraints but consistency is maintained in the comparisons. We conduct the experiments using the method outlined in DivideMix [165] with all the same hyperparameters: a momentum of 0.9, weight decay of 0.0005, and trained for roughly 300 epochs depending on the speed of convergence. The initial learning rate is set to 0.02 and reduced by a factor of 10 after roughly 150 epochs. Warm-up periods where applicable are set to 10 epochs for CIFAR-10 and to 30 epochs for CIFAR-100. We keep the number of augmentations parameter $M = 2$ fixed for a fair comparison.

5.3.2 Comparison of Augmentation Strategies

We examine the performance of each proposed augmentation strategy outlined in Section 5.2.2 using DivideMix as our baseline model. We investigate the performance impact on lower label noise (20%) and very high label noise (90%) for some performance bounds. We report results in Table 5.1.

As shown in the table, there is a large effect on algorithm performance based on how

		CIFAR-10		CIFAR-100	
Method/Noise		20%	90%	20%	90%
Raw	Best	85.94	27.58	52.24	7.99
	Last	83.23	23.92	39.18	2.98
Expansion-W	Best	90.86	31.22	57.11	7.30
	Last	89.95	10.00	53.29	2.23
Expansion-S	Best	90.56	35.10	55.15	7.54
	Last	89.51	34.23	54.37	3.24
Runtime-W [165]	Best	96.10	76.00	77.30	31.50
	Last	95.70	75.40	76.90	31.00
Runtime-S	Best	**96.54**	70.47	**79.89**	40.52
	Last	**96.33**	70.22	79.40	40.34
AugDesc-WW	Best	96.27	36.05	78.90	30.33
	Last	96.08	23.50	78.44	29.88
AugDesc-SS	Best	96.47	81.77	79.79	38.85
	Last	96.19	81.54	**79.51**	38.55
AugDesc-WS	Best	96.33	**91.88**	79.50	**41.20**
	Last	96.17	**91.76**	79.22	**40.90**

Table 5.1: Performance differences for each augmentation strategy. The best performance in each category is highlighted in bold. Removing all augmentation is highly detrimental to performance, while more augmentation seemingly improves performance. However, too much augmentation is also detrimental to performance (AugDesc-SS). Strategically adding augmentation by exploiting the loss properties (AugDesc-WS) yields the best results in general.

augmentations are included. While in some aspects this is unsurprising, what is surprising is the huge effect augmentation can have with regards to higher noise datasets. In the best case, we see AugDesc-WS at 90% noise achieve results on CIFAR-10 close to accuracies reported on augmentation techniques with 20% label noise. For CIFAR-100, we also witness a large effect with higher noise rates but it remains a challenging benchmark for noisy datasets. Overall, we find that AugDesc-WS achieves the strongest result across the board.

It should be noted that a vast number of image-based machine learning algorithms incorporate some level of weak augmentation (flip, crop, and normalization) during training

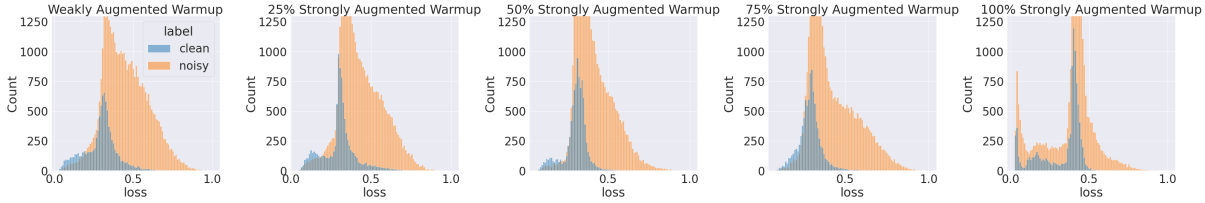


Figure 5.3: Effect of augmentation strength on the distribution of normalized loss for noisy versus clean segments of the dataset during warm-up for 90% label noise. Too much augmentation can cause samples in the clean dataset to be have higher loss, causing lower loss in samples from the noisy dataset.

time. For completeness, we retrospectively examine the effect of removing these augmentations to tease out the effect of augmentation, i.e. the raw input method. We see that including some very small amount of augmentation is hugely beneficial, particularly evident when examining the transition from raw to weak augmentation at runtime.

5.3.3 Effect of Augmentation During Warm-up

LNL algorithms generally rely on fact that clean samples are fit before noisy ones. To take advantage of such a property, many algorithms create scheduled learning or tune the loss function, explicitly designating warm-up period to exploit the label noise learning property [164, 165, 182]. We test the effect of introducing augmentation before and after this period by comparing the performance of models injected with augmentations from the first epoch and models trained with augmentations after the designated warm-up period.

We report performance metrics in Table 5.2 for various noise levels. We find that injecting strong augmentations during the warm-up period in low noise datasets benefit performance, but is detrimental when the dataset becomes increasingly noisy. This is particularly evident when examining the 90% noise rate. Conversely, weakly augmented warm-up greatly increases performance at higher noise levels.

Model	Noise	20%	50%	80%	90%	40% Asym
DivideMix (baseline) [165]	Best	96.1	94.6	92.3	76.0	93.4
	Last	95.7	94.4	92.9	75.4	92.1
DM-AugDesc-WS-SAW	Best	**96.3**	**95.6**	93.7	35.3	94.4
	Last	**96.2**	**95.4**	**93.6**	10.0	94.1
DM-AugDesc-WS-WAW	Best	**96.3**	95.4	**93.8**	**91.9**	**94.6**
	Last	**96.2**	95.1	**93.6**	**91.8**	**94.3**

(a) Results on CIFAR-10

Model	Noise	20%	50%	80%	90%
DivideMix (baseline) [165]	Best	77.3	74.6	60.2	31.5
	Last	76.9	74.2	59.6	31.0
DM-AugDesc-WS-SAW	Best	**79.6**	**77.6**	61.8	17.3
	Last	**79.5**	**77.5**	61.6	15.1
DM-AugDesc-WS-WAW	Best	79.5	77.2	**66.4**	**41.2**
	Last	79.2	77.0	**66.1**	**40.9**

(b) Results on CIFAR-100

Table 5.2: Application of strong versus weak augmentation during the warm-up period of DivideMix, in comparison to the baseline model. WAW signifies weakly augmented warm-up, SAW represents strongly augmented warm-up. Weak warm-up appears to benefit datasets with higher noise while strong warm-up benefits datasets with lower noise.

To better understand why this is, we perform an experiment by stochastically applying strong augmentation to each batch with increasing chance to observe its distribution at epoch 20. Figure 5.3 shows the loss distribution for samples in the training set associated with the clean versus the noisy dataset. We find that applying too much augmentation too soon can encourage lower noise data to have too high of a loss and noisy data to have lower loss.

5.3.4 Synthetic Dataset Summary Results

We report the summary results in Table 5.3. The results show that augmenting the state-of-the-art algorithm using our best augmentation strategy increases accuracy

		CIFAR-10				CIFAR-100			
Model	Noise	20%	50%	80%	90%	20%	50%	80%	90%
Cross-Entropy	Best	86.8	79.4	62.9	42.7	62.0	46.7	19.9	10.1
	Last	82.7	57.9	26.1	16.8	61.8	37.3	8.8	3.5
Reed et. al. [183]	Best	86.8	79.8	63.3	42.9	62.1	46.6	19.9	10.2
	Last	82.9	58.4	26.8	17.0	62.0	37.9	8.9	3.8
Yu et al. [182]	Best	89.5	85.7	67.4	47.9	65.6	51.8	27.9	13.7
	Last	88.2	84.1	45.5	30.1	64.1	45.3	15.5	8.8
Zhang et al. [99]	Best	95.6	87.1	71.6	52.2	67.8	57.3	30.8	14.6
	Last	92.3	77.6	46.7	43.9	66.0	46.6	17.6	8.1
Yi & Wu [184]	Best	92.4	89.1	77.5	58.9	69.4	57.5	31.1	15.3
	Last	92.0	88.7	76.5	58.2	68.1	56.4	20.7	8.8
Li et al. [180]	Best	92.9	89.3	77.4	58.7	68.5	59.2	42.4	19.5
	Last	92.0	88.8	76.1	58.3	67.7	58.0	40.1	14.3
Arazo et al. [164]	Best	94.0	92.0	86.8	69.1	73.9	66.1	48.2	24.3
	Last	93.8	91.9	86.6	68.7	73.4	65.4	47.6	20.5
Li et al. [165]	Best	96.1	94.6	92.9	76.0	77.3	74.6	60.2	31.5
	Last	95.7	94.4	92.3	75.4	76.9	74.2	59.6	31.0
DM-AugDesc-WS-SAW	Best	**96.3**	**95.6**	93.7	35.3	**79.6**	**77.6**	61.8	17.3
	Last	**96.2**	**95.4**	**93.6**	10.0	**79.5**	**77.5**	61.6	15.1
DM-AugDesc-WS-WAW	Best	**96.3**	95.4	**93.8**	**91.9**	79.5	77.2	**66.4**	**41.2**
	Last	**96.2**	95.1	**93.6**	**91.8**	79.2	77.0	**66.1**	**40.9**

Table 5.3: Performance comparison when incorporating our best augmentation strategy into the current state-of-the-art. Our augmentation strategy improves performance at every noise level. Results for previous techniques were directly copied from their respective papers.

across all noise levels. In particular, the improvement for extremely noisy datasets (90%) is very large, and approaches the best performance of lower noise datasets and represents an error reduction of 65%. For comparison, we achieve 91% accuracy for 90% symmetric noise on the CIFAR-10 dataset while the previous state of the art achieves 96.1% on only 20% label noise. Furthermore, we achieve an over 15% improvement in accuracy over previous state of the art for CIFAR-10 at 90% label noise.

Method	Test Accuracy
Cross Entropy	69.21
M-correction [164]	71.00
Joint Optimization [162]	72.16
MetaCleaner [185]	72.50
MLNT [180]	73.47
PENCIL [184]	73.49
DivideMix [165]	74.76
ELR+ [186]	74.81
DM-AugDesc-WS-WAW (ours)	74.72
DM-AugDesc-WS-SAW (ours)	**75.11**

Table 5.4: Comparison against state-of-the-art methods for accuracy on the Clothing1M dataset.

5.3.5 Clothing1M Performance

Clothing1M [187] is a large-scale real-world dataset containing 1 million images obtained from online shopping websites. Labels are generated by extracting tags from the surrounding texts and keywords, and are thus very noisy. A ResNet-50 with pre-trained ImageNet weights are used following the work of DivideMix [180]. We applied the pre-trained ImageNet AutoAugment augmentation policy for this task.

We report results in table 5.4. Our augmentation strategy obtained state-of-the-art performance when utilizing a strongly augmented warm-up cycle. In addition to obtaining competitive results, this further indicates that the noise level is likely to be below 80% based on our previous experiments, as strong warm-up improves accuracy. This is in concordance with the estimates of the noise level of Clothing1M, said to be approximately 61.54% [187].

5.3.6 Automatic Augmentation Policies

In our evaluation benchmarks, we primarily used AutoAugment pre-trained policies. These policies are trained on a small subset of the original dataset with regards to CIFAR-

		CIFAR-10		CIFAR-100	
Method/Noise		20%	90%	20%	90%
Baseline [165]	Best	96.1	76.0	77.3	31.5
	Last	95.7	75.4	76.9	31.0
AutoAugment	Best	**96.3**	**91.9**	**79.5**	**41.2**
	Last	**96.2**	**91.8**	**79.2**	**40.9**
RandAugment	Best	96.1	89.6	78.1	36.8
	Last	96.0	89.4	77.8	36.7

Table 5.5: Comparison of different automated augmentation policy algorithms. We compare performance of each policy using the AugDesc-WS approach. Adjusting the augmentation policy has minimal effect but still handily outperforms the runtime augmentation used in the baseline. The improved performance is still large with a noise ratio of 90%.

10 and CIFAR-100 (5000 samples). We do this due to the simplistic nature of integrating pre-trained AutoAugment policies. For completeness, we evaluate whether we can achieve similar performance with an untrained set of augmentations, as theoretically we could then tune policies based on validation accuracy. To do this, we examine whether we can achieve performance on-par with AutoAugment using RandAugment [157], which can be tuned by adjusting 2 parameters. For these experiments, we used $N = 1$ and $M = 6$ for RandAugment hyperparameters.

We report results in Table 5.5. As shown in the table, RandAugment can achieve performance on-par with AutoAugment with minimal tuning and demonstrates the validity of our approach. Furthermore, since we were able to outperform the state-of-the-art on Clothing1M while using a pre-trained ImageNet AutoAugment policy for the task, optimizing an AutoAugment policy on Clothing1M could potentially yield better results.

5.3.7 Generalization to Previous Techniques

Based on our evaluations, we find that a weakly augmented warm-up period followed by the application of strong augmentation works best. Furthermore, it is beneficial

to perform the loss analysis process on a weakly augmented input, then forwarding a strongly augmented input through the network for training. We apply our most effective augmentation strategy to previous techniques to evaluate generalizability of our approach.

We choose to compare to Cross-Entropy, Co-Teaching+ [182], M-DYR-H [164], and DivideMix [165] due to the range of techniques these algorithms employ. Co-Teaching+ uses two networks and thresholding to exploit the memorization effect and is an updated work based on the popular Co-Teaching [163] technique. M-DYR-H uses mixture models to fit the loss to previous epochs to weight the models predictions using a single network. DivideMix is the current state-of-the-art which combines these and brings in a semi-supervised learning framework.

All source code for each evaluated technique was available publicly published by the original authors. We follow the hyperparameters and models outlined in the original published paper and apply no tuning of our own. This demonstrates the ease at which augmentations can be incorporated without delicate tuning of hyperparameters, highlighting the generalizability of our approach. We detail the exact algorithm modifications for inserting augmentations in the supplemental of this chapter. We perform the evaluation on a lower noise setting (20%) as many previous techniques did not perform well at high noise levels. Table 5.6 shows the performance of our evaluation.

For vanilla cross-entropy, we used RUNTIME-S since as there is no warm-up period. For other techniques, we applied the AUGDESC-WS-WAW strategy. We evaluated our augmentation strategy on these algorithms as they cover a range of general approaches to learning with label noise. Some differences in performance are larger than expected due to the specific implementation of network architecture and synthetic noise generation techniques. We attempted strongly augmented warm-up for Co-teaching and found that there was a very large detrimental impact to performance. This agrees with our earlier observation that too much augmentation during the warm-up period can be detrimental.

		CIFAR-10		CIFAR-100	
		Base	**Aug**	**Base**	**Aug**
Cross Entropy	Best	86.8	**89.9**	60.2	**61.2**
	Last	82.7	**85.1**	59.9	**60.4**
Co-Teaching+ [182]	Best	59.3	**60.6**	**26.2**	25.6
	Last	55.9	**57.4**	23.0	**23.7**
M-DYR-H [164]	Best	**94.0**	93.9	68.2	**73.0**
	Last	93.8	**93.9**	67.5	**72.7**
DivideMix	Best	96.1	**96.3**	77.3	**79.5**
	Last	95.7	**96.2**	76.9	**79.2**

Table 5.6: Performance benefits when applying our augmentation strategy to previous techniques at 20% noise level. Baseline and augmented accuracy scores are reported.

In particular, it appears to have a strong impact on the way noisy and clean data converge during the warm-up period, which these algorithms typically rely on.

The AugDesc-WS-WAW strategy and even augmentation in general benefits performance in multiple categories (Table 5.6). As the experiments conducted were with no tuning of hyperparameters, we expect that further improvements can be seen when tuning with augmentation in mind due to the ways in which these algorithms exploit the loss distributions. Additionally, we see that across the board, the average performance of the last few epochs with augmentation is better than performance without. This indicates that using our augmentation strategy aids in learning a better distribution.

5.4 Conclusion

In this chapter, we propose and examine the effect of various augmentation strategies within the domain of learning with label noise. We find that it is advantageous to add additional augmentation, particularly for higher noise ratios. Furthermore, copious amounts of augmentation during warm-up periods should be avoided if the noise rate is high, as this can have detrimental effects on the property that neural networks fit

clean data before noisy data [161]. We performed extensive studies and found that the AugDesc-WS strategy is capable of producing improvements across all noise levels and in multiple datasets. We further show its generalization capabilities by applying it to previous techniques with demonstrated success. This is additional evidence for how using two separate pools of augmentation operations for two separate tasks in these machine learning algorithms can be beneficial. This idea has previously been demonstrated to be effective in SSL settings [176], and we now show this for LNL settings.

In summary, we examined where it is advantageous to incorporate varying degrees of augmentation, and were able to demonstrate a strategy to advance the state-of-the-art as well as improve the performance of previous techniques. We hope the insights regarding the strength and amount of augmentation will be beneficial for future applications of augmentation when developing LNL algorithms.

Chapter 6

Impact of Demographics on Multimodal Dataset Labels

Portions of this chapter are slated for publication at the Conference On Computer- Supported Cooperative Work And Social Computing '22

In the previous chapter we presented a method for improving machine learning algorithms in the presence of label noise in the unimodal setting. In this chapter, we look at how label noise and label variability manifests in a multimodal setting.

Multimodal sentiment analysis presents the challenge of computationally determining how humans would emotionally interpret a given input. For example, what is the sentiment expressed by a speaker in a video? This problem is critical for building richer human-computer interaction experiences and providing automated assistance to people, for example. Leveraging the power of deep learning, researchers have made progress modeling sentiment in any modality of input including text, video, audio, or some combination of multiple modalities. To train complex, non-linear deep learning models, researchers have created datasets consisting of thousands of video examples labeled according to sentiment [188, 189, 190, 105, 191, 192]. Most of these datasets consist of

example videos of "talking head" speakers. Each example is labeled independently by a set of human annotators who are asked to gauge the emotion of the speaker (whether they are saying something positive, negative, or neutral, for instance). As research in sentiment analysis has progressed, AI models that classify sentiment have been applied to a range of decision making pipelines and applications, including emotional chat support [193] and determining hate speech [194]. Furthermore, as these technologies advance and become more pervasive, it can drastically alter how we live our daily lives from seeking out medical help [195] to and changing our work environment [196].

To create the datasets to train the models in these systems, researchers often rely on hiring crowd workers through services such as Amazon's Mechanical Turk to label the data. There has been extensive work examining how to leverage crowd workers to obtain quality labels at scale [197, 198], and sentiment is frequently labeled in the same manner. However, sentiment is highly subjective, and when appraising sentiment of others, psychologists have found that our experiences and opinions play an important role [199, 200, 201, 202, 203, 204, 205, 206, 207, 208]. Since these studies have shown that demographics can be used to capture differences in sentiment appraisal, it therefore suggests that differences in annotator demographics can lead to differences in their interpretation of sentiment.

To attempt to control for the impacts of demographics, researchers have developed previous datasets that balance some variables. For example, balancing for a 50-50 distribution of male and female speakers [105]. Newer datasets also occasionally provide demographic information of the speakers or subjects in the dataset [192, 191] as a means to aid models in making more informed decisions. Other works have also examined the biases inherent to the contents [209, 210, 211]. However, few works have examined biases due to annotators' demographic backgrounds [212, 213, 214]. To the best of our knowledge, no works have examined annotator biases for sentiment from a multimodal

perspective.

If annotator demographics impact sentiment, then any results gleaned from a dataset that does not control for annotator demographics at the time of creation will be biased and skewed in addition to all the other biases that such datasets already exhibit [215, 216, 217] by any "unbalanced" (defined relative to specific application needs and goals) distribution of annotators. Therefore, if the demographics of the annotators did not match the distribution of, say, the general population, then results and analysis using that dataset might not be applicable to the general population. Furthermore, models and evaluations using these datasets would reflect the opinions of those who perform crowdwork versus those who do not. For systems which make decisions based on these models, this would mean lower efficacy for certain groups of users. However, as these technologies become increasingly critical in commonplace technological systems, it would not be far-fetched to notice a disenfranchisement of specific demographics of people [218]

Understanding the potential role of annotator demographic is critical in informing decisions about how we use and trust sentiment analysis technologies going forward. We attempt to provide some answers to this and quantify the impact of demographics on sentiment analysis datasets. We re-labeled the well-established MOSEI dataset [105]: a dataset of "talking head" speakers scraped from YouTube. Using a crowd-sourced labeling process that took the annotators' demographic information into account, we produce a rich annotation that includes 5 times more annotators per video than the original MOSEI dataset. We also collected detailed demographic information for all annotators in our relabeled dataset. Using this dataset, we conduct statistical experiments to establish and begin to quantify the impact of demographic background on sentiment labeling. From this analysis we found that annotator sentiment varies (with statistical significance) based on demographic factors such as age, gender, ethnicity, and educational level. Our results suggests that decisions derived from AI should be used cautiously and a strong need for

interdisciplinary collaboration for more inclusive AI development.

Our work provides the following contributions:

- We present a large set of annotations for multimodal sentiment analysis containing rich demographic information. Additionally, we provide annotations for independent modalities (text, audio, visual) in addition to their combined annotations.
- We show that demographics have a significant effect on sentiment labeling, and show that this effect generalizes across all component modalities: text, visual, audio, and their combination. We find noticeable differences in label agreement and ratings for different modalities. These results suggest that decisions derived from AI results should be used cautiously. In particular, they should consider the parameters for data collection that may introduce unintended biases.
- We show that algorithmic claims of sentiment classifier improvement can vary greatly due to demographics, observing up to 4.9% change in absolute algorithmic performance when sampling for various sub-populations. This exceeds the improvement claims over state-of-the-art of most recent sentiment classification machine learning papers. We additionally show the ability for our gathered labels to be used as an improved evaluation metric to account for demographic biases. Data is released for public use.

6.1 Related Work

We examine relevant literature to motivate our work. We first examine modern advancements in multimodal machine learning that is applied to sentiment or emotion classification. We discuss datasets these models are trained with and their annotation process. We then examine how demographics can influence the emotion appraisal process.

Finally, we examine ways in which works have attempted to quantify and mitigate the demographic of annotators.

6.1.1 Multimodal Machine Learning

Enabling machine learning for multimodal data has been explored in many domains over a long period of time [79]. Many machine learning techniques have been applied on the task of sentiment classification [219, 81]. As transformer-based architectures have become very popular recently, some recent techniques have also explored their use in multimodal settings. Originally proposed in [84] for neural machine translation (NMT) tasks, they have demonstrated superior performance on multiple benchmark problems such as in image classification [73] and action recognition [72]. The basic functionality is to apply layers of self-attention, on sequential representations. Recent attempts by researchers to enable multimodal modeling on transformers via cross-modal attention have been successful for sentiment analysis [87, 220]. Inspired by work which showed that shifting one modality (language) using representations from other modalities improves performance, [81], MAG-XLNet [221] incorporates the ability for fine-turning on multimodal data on a transformer-like model built on top of XLNet [222]. XLNet is an extension of transformer based methods that enables learning over longer sequences and the ability to better model the context dependencies.

6.1.2 Multimodal Machine Learning Datasets

There is a long line of work for building large scale datasets for machine learning. There have been numerous works on the development of large scale datasets for the vision, language, and multimodal domains. Many datasets have been gathered over the years to explore sentiment or valence: text based, visual, audio, and via their multimodal

combination [105, 192, 191]. Additional datasets using modalities such as pose [190, 188] and EEG [12] have also been created and analyzed. Datasets built around continuous representations have also been explored [223].

Due to their large size, these datasets are typically labelled via crowdsourcing platforms such as Amazon Mechanical Turk. One frequently used method for obtaining quality labels is the use of multiple annotations and taking the mean or majority class. While many existing datasets simplify annotations to a single ground truth emotion or sentiment label, likely due to a lack of annotators per sample, emotion representation is not necessarily discrete. Representations such as [224] describe emotions in a continuous space. In this work we compare the mean label for ease of comparison with prior art, however, the scope of our data collection enables us to represent labels as a distribution (with mean and variance).

6.1.3 Demographic Effect on Emotional Appraisal

The studies of how emotions are interpreted have a long history in psychology. Demographics such as gender [202, 200, 225], age [206, 204], culture [203], economic background [226], etc., play a particularly large role. There are significant differences in the emotion expression and appraisal as a result of these factors. Combinations between multiple demographic variables have also been considered, such as in age and culture [206]. For example, Plant et al. [200] showed that people typically rate women sadder than men, and that they demonstrated a wider variety of emotions. Fischer et al. [202] found that women's experiences of emotions were modulated by cultural background. Many works have also explored gender stereotypes beyond this [227, 203].

Cultural backgrounds have also played a role. Brody [203] presented data showing that emotion expressiveness across cultures are different. Davis et al. [201] demonstrated

	# samples	Mean annotators per sample	Total annotations	Scale	D	M
Ours	500	15	30,000	7-pt likert	Yes	Yes
CH-SIMS [192]	2281	3	27,372	7-pt likert	No	Yes
MOSEI [105]	23,453	3	70,359	7-pt likert	No	No
SEWA [223]	1990	5	Continous	Continuous	No	Yes

Table 6.1: High level statistics of recent datasets for sentiment or emotion analysis. **D** represents whether the datasets contains annotator demographic information and **M** represents whether the datasets provide per-modality annotations. Our annotation effort produced more ratings per sample and also contains detailed annotator information. We also provide per-modality labels and have a comparable number of total annotations in the entire dataset. This type of annotation enables us to perform in-depth analysis on demographic effects and is comparable to large-scale machine learning datasets by annotation count.

that elicited emotional responses are different between participants of Chinese versus American culture between men and women. Age has also been well studied: Mitchell et al. [204] found that older adults are less accurate at interpreting prosodic emotion cues, and follows numerous previous works studying the age-related decline for identifying emotional cues [228]. Additional differences in age demographics between rater and poser were also discussed by Riediger et al. [205], and that emotional expression by older posers were more difficult to read.

6.1.4 Annotator Bias

The study of annotator demographics and its relationship to machine learning dataset creation is not new. Many techniques have tackled this during data acquisition [229, 230] as well as during model development [231, 100]. Works such as Wauthier and Jordan [232] proposed a framework to mitigate worker biases and downstream effects on model performance. Asking workers to think about other workers responses as demonstrated by Shaw et al. encouraged workers to provide more objective annotations [233]. This

inspired Hube et al. [213] to develop a method for intervention to overcome the strong influence that personal opinions have during annotation of subjective datasets. Chung et al. [234] recently conducted a systematic evaluation of different approaches for obtaining ground truth labels. Most recently, techniques have been proposed by Chen et al. [235] have attempted to capture annotation uncertainty as well as improve consistency by improving the annotation task design.

Furthermore, recent efforts to combat bias have been a topic of focus in the natural language processing domain. In particular, it has been observed that for datasets pertaining to hate-speech, the demographics of annotators play a large role [214, 236, 237, 238, 239, 240]. These results are echoed by Prabhakaran et al. [212] who found that annotators for hate-speech [239], sentiment [241], and emotions [242] for the language modality contained bias due to annotator demographics. That is, aggregated labels did not properly capture the perspectives of annotators from varying demographic groups. In fact, the impact of annotator demographics have also been observed when obtaining credibility ratings for news [243]. Some recent works have begun to quantify and mitigate this effect. Gordon et al. [244] introduced a metric to correct metrics based on the assumption that annotators will provide inaccurate answers with some chance.

We build upon these works by providing more detailed data as well as analysis for the multimodal domain. Additionally, we examine the scale of these effects empirically using state of the art techniques.

6.2 Experimental Design

We examine the impact of how demographics of annotators impact label distribution. In other words, do the demographics of annotators matter when labeling sentiment data? Our hypothesis is that by grouping annotations based on the demographics of annotators

who provided them, we can drastically alter the "ground truth" label distribution, to the point that these differences might even outweigh differences among competitive machine learning classifiers attempting to approximate this ground truth. Furthermore, with the intuition that model performance and evaluation will be strongly affected, we examine how annotations from strategically selected demographic subgroups can be used to create a demographic bound on performance. Naturally, this requires gathering data that can capture nuances of demographic effects and is also capable of being used by recent machine learning architectures for evaluation.

We divide the experiment into two parts: 1) We first conduct a large scale annotation ($>$ 1000 annotators) of videos used for multimodal sentiment classification, and 2) We additionally conduct a set of statistical experiments to determine the significance and impact of annotator demographics on dataset labels.

We choose to evaluate *sentiment*, in this case positive versus negative speaker stance (as evidenced by language, speaker video and speaker audio), due to the simplicity in the annotator decision making (one single axis) compared to more complex emotion measures. Agreement scores among raters are generally much lower for emotion datasets than for sentiment (positive vs negative) only [245]. We perform our investigations in the chosen domain to provide a strong baseline, as more obvious annotations should intuitively be least affected by demographic differences. Furthermore, as we wish to examine whether the effects of modality would modulate any annotator biases, we additionally gather annotations for the individual component modality for each sample. We present our process for data collection in Section 6.3.

We build our dataset by randomly sampling from the test set for MOSEI based on the split from the work of Tsai et al. [87] and Rahman et al. [221] for re-annotation. This is a very suitable dataset for our purposes. The fact that we are extending an existing established dataset means that we have a baseline set of annotations to compare to when

performing analysis of the new annotations. Another benefit is that the machine learning research community actively produced and evaluated classifiers for MOSEI, which can be used in our evaluation [87, 221, 81, 220]. By changing the demographic distribution of the test set, we determine the approximate effect that this would have on machine learning models. We discuss and conduct thorough experiments in Section 6.4. In summary, our goals are to 1) provide a set of annotations large enough for machine learning evaluation and to understand demographic influences and 2) provide empirical evidence for the scale of annotator demographic effects on sentiment dataset labeling across modalities.

6.3 Data Collection

We now describe the data collection process for our large scale study to examine the influences of annotator demographics on unimodal and multimodal sentiment-dataset annotation. We first describe the video samples used for annotation in Section 6.3.1. We then discuss the platform used for recruiting participants (Section 6.3.2), and the collection interface (Section 6.3.3). Lastly, we discuss our way to improve the demographic distribution of the annotators in Section 6.3.4.

6.3.1 Multimodal video samples

We build our dataset using 500 videos segments randomly sampled from the Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) dataset [105]. It is one of the largest multimodal sentiment analysis datasets to date, and is highly regarded in the domain. The dataset is gender-balanced for male/female speakers. All sentences are annotated and randomly selected from various topics and monologues. The dataset contains over 23k video segments of 7.28 seconds long. Each segment was annotated by 3 annotators on a 7-point Likert scale. This resulted in over 70k total annotations.

To answer questions regarding modality effects, we split each video into its component modalities: audio, video, text, and their combination. This results in a total of 2000 different samples. Annotators are randomly assigned 30 of the 2000 samples for annotation. We ask more than 1000 annotators to provide ratings and results in approximately 15 annotations per sample. This enables us to to capture the per sample demographic and population effects on sentiment annotation for all modalities.

6.3.2 Prolific crowdsourcing

We use Prolific to gather annotations for the samples. Prolific is a crowdsourcing tool similar to Amazon's Mechanical Turk available to countries within the Organisation for Economic Co-operation and Development (OECD). In addition to being able to designate tasks to crowdworkers, Prolific gathers demographics of participants and makes this information available to researchers. Researchers can filter for a specific participants based on demographic background. We chose to go with Prolific as our crowd sourcing tool due to this accessibility of diverse participant information and its strict verification process (via government issued identification). We obtained crowdworker ethnicity data, country of birth, employment status, student status, gender identity, fluent languages, highest education level, immigration, whether participants were mono/multi culture, their nationality, and household income.

In addition to filtering for data we need for analysis, we filter participants by parameters to maintain data quality. In particular, we only accept workers with greater than 97 percent approval, who are fluent in English with no language related disorders. It is also required that participants can see video and hear audio.

This work involves reannotating part of an existing dataset consisting of non-offensive video footage of movie and other reviews. We do not collect demographic information

ourselves or have access to any private information of the annotators. The participants have further agreed that some high level demographic information will be shared and used for research purposes when signing up for our study through Prolific. The annotation task is quick, and we experimentors did not interact directly with any annotators. As a result of these factors, our institution's IRB has determined our methodology to be exempt human subjects research. We recognize demographic properties of participants is sensitive information and follow protocols to protect the privacy of the annotators.

The representations of demographic properties are limited by the availability of information provided by Prolific and exclude certain populations from analysis. We encourage the reader to interpret results with these factors in mind.

6.3.3 Collection interface

We follow the same annotation process described in detail by Liang et al. [245] for data collection interface to reduce variability between the experiments. The participants are presented with a series of bullet points explaining the task of sentiment analysis, as we did not have access to the original training videos used by Zadeh et al [105]. We describe sentiment as the speaker's attitude towards the topic of their discussion. We also asked the annotators to rate the sentiment of the speaker, and not their own opinions. As sentiment labeling is a frequent task on crowd working sites, we expected most annotators to be able to accomplish the task with minimal training. We did not provide too much training as we, in accordance with previous annotation goals, wanted to avoid the over-training of subjects and wanted to maintain the in-the-wild goal of the dataset.

After receiving directions, the participants are given 30 random samples to annotate. We asked participants to provide ratings on a 7-point Likert scale for sentiment from

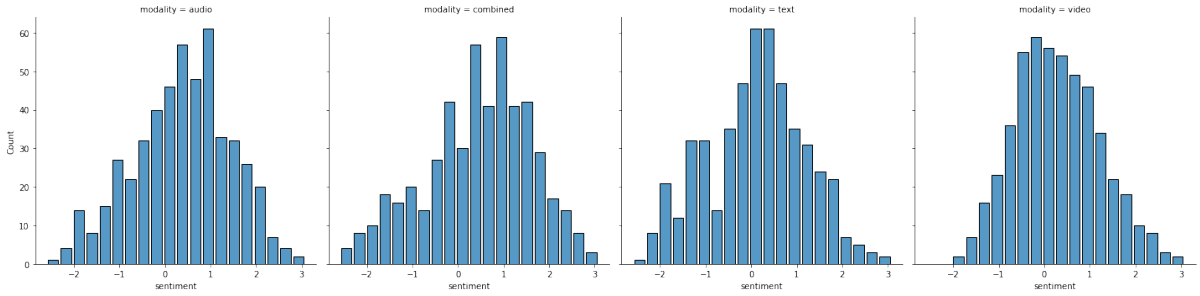


Figure 6.1: Distribution of sentiment ratings by modality. The distributions are similar to the original ratings, however noticeable differences exist when examining different modalities.

highly negative (-3) to highly positive (3). We further ask the participant to rate the gender of the speaker as well as ask if any samples failed to display properly. We match the gender assessment of each sample with the original annotations as an additional way to maintain data quality. While we do not use this assessment of the speaker in our work directly for analysis, we anticipate future work to explore relationships between annotator demographics and data properties.

We estimated the time to complete 30 questions to be approximately 13 minutes or approximately 20 seconds per sample with some extra time for the directions. We targeted payment to be approximately $9.50 per hour as this is the good pay threshold set by Prolific. However, our actual hourly rate ended up being approximately $13 per hour as text was much faster to label and the average taken length for 30 questions ended up being 10 minutes instead. We did not reduce this pay as we did not wish affect the task adoption or completion rate [246].

6.3.4 Improving demographic diversity of annotations

We divide the annotation process into two phases during the summer of 2021. In the first phase, we asked 500 participants from the US to provide annotations without

Age	**Before boosting**	**After boosting**	**Prolific dist.**
$\leq$ 20	27%	21%	17%
21-30	58%	46%	50%
31-40	8%	16%	18%
$>$ 40	6%	17%	14%
Gender			
Female	79%	73%	73%
Other	21%	27%	27%
Ethnicity			
White	73%	67%	74%
Other	27%	33%	26%

Table 6.2: Demographic distribution by age, gender, and ethnicity before and after gathering data from under-represented groups. By gathering more data from under--represented groups we reduced the demographic skew. We report the proportion of labels before and after applying a boosting process to increase the number of under--represented demographic groups of online annotators. Each annotator provided 30 labels and we obtained data from approximately 1000 people. We also provide the approximate active proportion of annotators available to researchers as reported by Prolific. These distributions are in agreement with previous findings that a majority of crowd workers are white, young (under 30), and female. More data is collected from under-represented groups by filtering for candidates in these groups, i.e., annotations over age 30, annotators who are not female, and annotators who are not white. The boosting process provided a noticeable benefit to the representation of the dataset. However, annotations by underrepresented groups typically took longer, and participants were less actively picking up our annotation task. This would explain some of the more dramatic demographic skews in age and gender before applying the boosting process.

controlling for any specific demographic backgrounds. In the second phase, we gather data from an additional 500 annotators by restricting certain demographic backgrounds. The process for restricting annotations from certain demographics is the same as quality related properties such as annotator approval rates. Our task is only visible to particular populations that match a pre-specified group. That is, we ask for annotators who are older than 30, who are not female, and who are not white. We keep quality related filters such as approval rates, language skills and others the same. We perform this restricted

group annotation process independently for each group. We report the demographic proportions from this process in Table 6.2. We see that the dataset is heavily skewed before this restriction process, and that the representation of smaller groups is boosted afterwards. We also report the approximate overall Prolific distribution of annotators. However, we found that underrepresented groups typically participated far less actively in the annotation process and thus amplified some dataset skews seen in age and gender.

For our study, participants who did not provide demographic information were removed from the list of pool of potential annotators. We further limited our scope of research to participants within the US to limit potential geographic effects. Including multiple geographies would also exacerbate the long-tailed distributions of demographic properties due to additional variables. However, since a large portion of the active participants on Prolific appear to live in the US, we still had a sizeable population for recruitment. Additionally, as a vast majority of the active users on Mechanical Turk are also from the US [247], we anticipated significant demographic overlap with the original labels. After applying these filters, there were approximately 50k participants who were active in the last 90 days prior to data collection.

In general we found that when boosting for specific demographics, it took longer before our annotations were completed. We also found that by only removing one demographic from consideration, such as only permitting annotators who were not female, that the other demographic properties (gender and ethnicity) were still heavily skewed. Although using more specific filters may help, such as filtering for non-female and non-white, we chose to use more general filters as some population demographics were very small that we did not want to introduce other population effects. We did not notice any differences in annotation quality for different demographics.

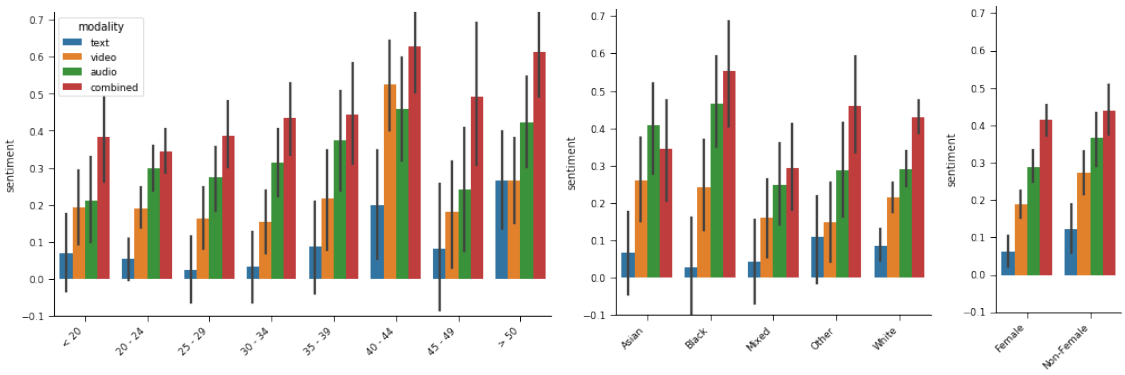
6.4 Experiments and Results

We conduct experiments and present results to first provide a summarized view of the gathered data for context to interpret the results Section 6.4.1. We present the significance of demographic effects in Section 6.4.2. We then perform a series of experiments to explore the impact of these effects. We first examine the effects of this via Monte Carlo simulation experiments in Section 6.4.3. We then analyze how inter-annotator reliability is impacted in Section 6.4.4. Lastly, we analyze the impact that this has on state-of-the-art learning algorithms for this task in Section 6.4.5.

6.4.1 Data overview

A total of 1034 annotators participated in our annotation task. We removed participants who did not complete the task by answering all 30 questions, completed the task too quickly, or experienced connectivity issues. A final 886 annotators completed the task with the distribution of population shown in Table 6.2.

We find that without controlling for demographics, our distribution skew is similar to what is found in existing literature: young, female, and educated [248, 247, 249, 250]. Without controlling for demographics, the proportion of females in the dataset was 79%, those under 30 made up 85% of the overall dataset and consisted of approximately 73% people who claimed white ethnicity. After attempting to boost the under-represented populations, we were able to obtain considerable reductions in proportion of females (73%, proportion of white ethnic background was reduced to 67%, and proportion of those under 30 was reduced to 67%. This increases the average dataset age to 29.5 years old from 25.5 years old. For comparison, the US population is approximately 62% white, 50% female, and 40% under 30. As a large portion of analysis is centered around over-represented versus under-represented groups, we will will frequently refer to under-



(a) Age Category (b) Simplified Ethnicity (c) Gender

Figure 6.2: Mean sentiment ratings broken down by demographic categories. We provide the age, ethnicity and gender charts. Component modalities (text, video, audio, and combined) are shown as different colors. Grey line illustrates 95% confidence intervals. Note that there are obvious significant differences in ratings for modality. Trends can also be observed visually.

represented groups as *other* or *non-majority* when presenting results. For example a comparison of female versus non-female or other. We do this to avoid having very small groups due to finer categorization.

Other demographic aspects of the dataset for people who gave answers were: 46% of people identify as monocultural and 41% of people identify as multi-culture for culture identity; 95% were born in the U.S., however only 89% learned English as their first language. Approximately 41% of participants were currently students where 2% of annotators had a doctorate degree, 12% had a graduate degree, 40% had an undergraduate degree, 30% had a high school degree, 14% have a technical degree, and the remainder have other or no formal qualifications. All demographic information and annotations are publicly accessible via Github for further analysis.

6.4.2 Significant effects in sentiment rating due to demographics

We wish to understand whether demographic background can cause significant differences in ratings. For example, differences between older versus younger annotators. To understand the demographic differences while accounting for the various grouping effects from samples, subjects and modalities, we conduct a linear mixed effects analysis following [251]. We construct a linear model of sentiment as a function of all gathered demographics. We modeled age, gender, ethnicity, cultural background, and education. This model was significant ($p < 0.001$). Figure 6.2 presents example annotation distributions for age, ethnicity, and gender. We find via visual inspection that despite age having a non-linear effect, the slope is still significant ($p < 0.01$).

We report mixed effects analysis results as an effect size with a standard error bound. We found that age affects sentiment (($\chi^2(1)$=12.17, $p < 0.001$)), and increasing sentiment by approximately 0.0045 ± 0.0013 (standard errors) per year. Gender was also found to have a significant effect (($\chi^2(1)$=4.33, p=0.037)), increasing rating by 0.066 ± 0.032 standard errors. Significant interaction effects were found between gender and ethnicity (($\chi^2(4)$= 11.10, p=.026)). Borderline significant interaction effects was found for age and gender (($\chi^2(1)$=3.73, p=.053)). We further test for any interaction effects between demographics and modality. As expected, testing modality is a highly significant effect (($\chi^2(1)$=11.10, $p < .001$)). Interaction tests between modality with age, ethnicity, culture and education showed no major effects, the greatest significance was between gender and modality at (($\chi^2(2)$=2.95, p=.086)).

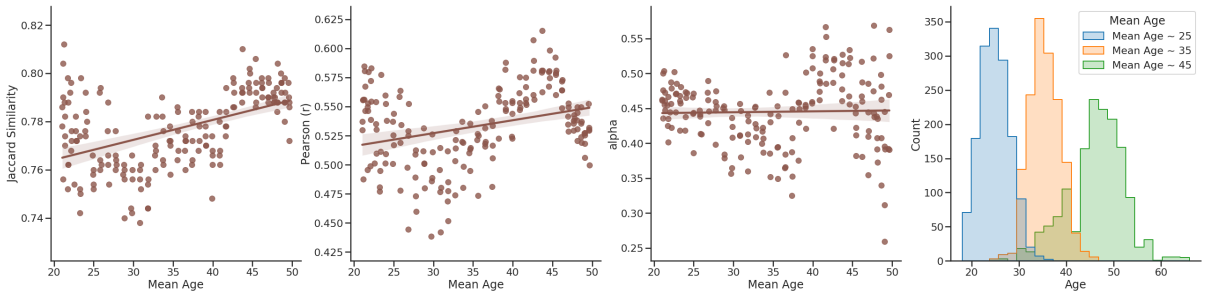
These results suggest that ratings for sentiment when annotating in text, visual, audio and all combined modalities will produce different ratings depending on which demographic is annotating. They also suggest that the demographic effect is consistent across modalities. Furthermore, the significant differences in annotations within each

modality suggest that ground truth annotations for each modality is subtly different. That is, human perceptions of text, video, audio, or combined modalities are subtly different. Therefore, when building datasets, we need to be careful whether participants are accounting for the information in the modalities we are interested in holistically. Furthermore, when developing a model for prediction, we should be wary that the predicted label matches the expected label of user interaction. That is, we do not want a model to infer the multimodal sentiment label when a user is only communicating via text, as this might result in lower perceived model effectiveness.

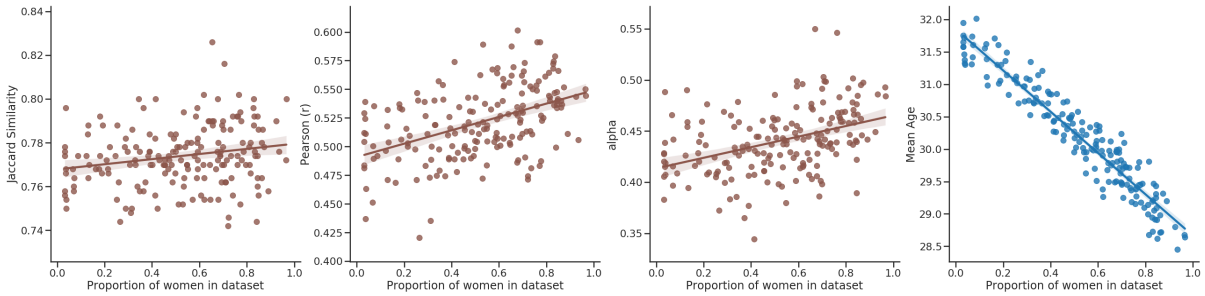
6.4.3 Monte Carlo simulation of demographic effects

We perform a Monte Carlo experiment to visualize how shifting demographics can alter the truth of labels for a dataset. The diverse demographic information in the dataset enables us to perform a Monte Carlo sampling of sub-populations to understand the effect of varying the population with differences in the dataset. We seek to empirically demonstrate how sampling various sub-populations based on demographic ratios impacts the dataset metrics. We also seek to show how using our labels can be used as an evaluation to test for the spread of performance in current algorithms due to demographic differences. We conduct the results using the combined modality for bench-marking and analysis purposes.

Following the mixed effects analysis, we saw a significance in the annotator gender and age categories for all the modalities. We randomly sample 3 raters per video segment with replacement from the dataset for different mean ages and for varying proportions of women. We examine metrics with respect to the original dataset [105]. We report Jaccard similarity and Pearson correlation against this dataset, similar to how previous models used these [87]. We additionally report the Krippendorf Agreement score for each



(a) Monte Carlo performance metrics by age. Sampling is performed to obtain a subset of our data to obain differing mean ages. Similarity with original annotations from MOSEI [105] is reported in the left two figures using Jaccard similarity and Pearson's r. Agreement within the Monte Carlo sampling is reported as Krippendorf alpha. We report the distribution of demographics in the right-most figure. Observe that when compared to the original dataset (right two figures) that there is a high degree of relative change. In particular, annotators aged approximately 30 had the least amount of similarity with original annotations. Stronger agreement within the dataset (third from right) appears to correspond with more similarity to original annotations.



(b) Monte Carlo performance by gender distribution. From the right two figures, we see that as the proportion of women increases, there is more similarity with the original labels. Since a majority of annotators are women, this shows that the labels are more biased towards the opinions of women. Women also tended to agree more with the annotations of other women. The right most figure demonstrates that there are more younger women than older women and helps to illustrate the co-variance between age and gender.

Figure 6.3: Visualization of Monte Carlo experiments on age and gender. Left two figures in each category show similarity with original ratings from [105]. Alpha is the agreement score of the dataset sampled via Monte Carlo. The trends observed above support the significance of effects found in Section 6.4.2

.

monte carlo sample.

As can be observed in Figure 6.3, we observe large variations in age metrics when adjusting for the mean age of the dataset. There appears to be a correlation of within dataset agreement for age with regards to agreement with previous labels. When controlling for proportion of females, we see a small improvement in metrics as the proportion

of female annotators increase. This supports works in literature on the influences of gender in emotion interpretation [252, 253, 200, 254, 255, 256]. Additionally, all these results agree with our observations for significance previously. Furthermore, they support the finding that testing for demographics can be beneficial for measuring ground truth quality in subjectively annotated datasets. In summary, these results suggests that the original ground truth labeled via Mechanical Turk likely follow the overall crowd-working demographic biases, and these effects showcase a demonstrable effect.

6.4.4 Differences in inter-annotator reliability due to demographics

In this section we provide experiments for annotation quality and reliability. We explore two questions: 1) Are annotators in certain demographics more in agreement than others? and 2) Are annotators in certain demographics more in agreement with the original dataset? To evaluate agreement within a demographic group, we use Krippendorf's alpha. Krippendorf's alpha is a metric used to measure annotation consistency among annotators and to give an indication to the quality and amount of variability present in a dataset. It can also normalize for missing data and is applicable on a variable number of coders. To measure agreement with the original dataset, we compute the Pearson correlation (r) of each demographic with the original labels from MOSEI. Correlation is reported for the combined modality only as the original dataset does not provide per-modality annotations. Results are reported in Table 6.3

The overall krippendorf agreement of our data is .48 which is good for a publicly annotated dataset, especially given the diverse population which provided annotation. Additionally, this is comparable to the .51 reported in [245]. While works such as [257] report higher agreement scores, these annotation efforts typically require a post-annotation

discussion phase to find score consensus. It is challenging for crowd-sourced annotated data to do this and thus explains much of this difference. Some works such as Schaekermann et al. [258] have studied how to effectively incorporate a deliberation process into the crowd-sourced annotation process. However, when creating larger datasets, incorporating deliberations have thus far not been used extensively. Potential for future work such as using advanced semi-supervised models on small strongly annotated datasets that incorporate deliberations exist.

We measure the agreement score among sub populations to look for any large demographic effects. No large differences in agreement were noticed in age, with the exception of the text modality and all modalities being slightly lower for annotators over 30. This demonstrates that within each age group, participants had similar opinions regarding the sentiment of a sample. However when examining the correlation, participants over 40 provided annotations that were far less correlated with the original labels. This trend is observed for non-white annotators, as well as non-female annotators. In addition to using the seven-class annotations to compute agreement, we also simplified the labels to be binary and computed the agreement. That is, all labels less than zero are considered to be negative, and all labels greater than zero are considered to be positive. We found that the agreement scores follows a similar trend of higher agreement in over-represented groups and lower agreement in under-represented groups.

These results demonstrate that certain demographic groups might agree on labels more than others. To improve this, some demographic groups may benefit from additional training due to task familiarity. Additional demographic factors such as differences in emotion interpretation due to age, gender, and culture might also be influencing the results. Furthermore, we see that the same groups that have lower agreement (older than 40, non-white, and non-female) also had lower correlation scores with the original annotations. These demographic groups are also less represented among crowdworkers.

This is further evidence that the demographics has a strong influence on ground truth labels.

6.4.5 Best/Worst case analysis by demographic

We further quantify the effect of shifting demographics on trained model performance. We experimented with the best possible (and worst possible) demographic distributions with respect to the MOSEI dataset. In other words, what is the population distribution that gives us video labels as close (or as far) as possible to MOSEI? This is relevant because MOSEI and many other sentiment prediction datasets [192] are often taken as ground truth in various works, even though annotator demographics are typically unaccounted for. We wanted to understand the possible swing in scores that could occur with an arbitrarily good (or bad) population distribution.

Sampling procedure

To perform this experiment, we first divide the population of annotators into a series of age bins from 18-20, 20-25, ..., 45-50, $>$ 50. We further breakdown the annotators into female or non-female bins. This gives us 16 bins to optimize. Given each bin, we then adjust the weights of the female and non-female annotators to match the desired target gender distribution within that age demographic. The "ground truth" labels for each video sample can then be computed using a weighted averaging of the demographic category based on the mean ratings within each demographic category:

$$v_i = \frac{\Sigma_j w_j l_i}{\Sigma_j w_j}, \tag{6.1}$$

where v is the determined ground truth label using the simulated distribution, w_j is the weight for the j-th group of annotators for the video, $l_i j$ is their label. Weights

are optimized via gradient descent using the MAE of our predictions v_i against the original MOSEI labels as loss. By maximizing the MAE with respect to MOSEI, we can obtain the worst case demographic population. We found that the best or worst case population demographics did not change between multiple optimization runs. We restrict the minimum demographic to being 1% of the overall population. For the US population data, we base the demographic weights on census data. We then compare the ratings of this hypothetical population of annotators.

Models

Three recent state of the art techniques for multimodal sentiment classification are used for evaluation:

MulT [87] is an extension of the transformer architecture to enable multimodal inputs. It incorporates elements of early feature fusion by mixed-attention of modalities and then using late fusion to combine predictions across modalities. We used the unaligned model for evaluation.

MAG-Bert [221] Enables the fusion of modalities and the use of pretrained embeddings by exploiting modality gating mechanism inspired by [81] and incorporating into a transformer architecture. State-of-the-art benchmarks were reported on multiple datasets using BERT embeddings. [259]

MAG-XLNet Utilizes the same fusion technique however uses an XLNet [222] backbone which is improvement on Bert that exploits autoregressive training, relative positioning, and segment recurrence from Transfomer-XL [260] for improved modeling.

Each algorithm is trained on the original dataset using publicly available code. The algorithm results are then measured against the ratings scores for a particular population distribution derived from the optimization procedure. We report the binary accuracy, F1 score, mean average error and pearson correlation. For binary accuracy, this value is

reported as the accuracy of positive or negative sentiment only. F1 is a harmonic mean of precision recall for positive and negative sentiment. MAE and correlation is a measure obtained from the means for a specific sample and the predicted mean rating. Note also that since our annotations are for the test set, we run for only a single trial as opposed to cross-validation. This methodology is the same as existing standards.

Classification results

We present the results in Table 6.4. As can be seen, there is a large spread in performance among different population distributions. A drop in performance is expected as the existing works do not optimize for our test condition. The most similar sampling to original labels did not improve results significantly, as expected, potentially due to crowd-working demographics being similar. Examining the US population shows that there is a drop, indicating that the demographic differences are having an effect of about 1%, when compared against a "crowdworker" demographic.

However, what is surprising is the potential effect from a demographic that is the least similar to the original annotations. While we see an approximately 3.3% drop in binary accuracy between least and most similar for when compared to the original dataset, algorithmic performance decreased much more. Algorithm performance decreased up to 4.9%. This different suggests that the algorithm is over-fitting to properties in the original dataset, and that these properties can be observed when adjusting for demographics. From an HCI perspective, the effect of this would suggest that AI systems for sentiment prediction works well for some people (e.g., younger, white, and female) and not others (e.g., older, non-white, and non-female).

Furthermore, when measured against the original annotations (Table 6.4(d)), the new models MAG-Bert and MAG-XLNet outperform the older technique (MulT). However, when the population is changed, we see that newer techniques perform much worse than

older techniques. This suggests that the newer models are matching patterns and properties in the original dataset that can be quantified via annotator demographics.

In summary, we see that by sampling for different demographics we can place a bound on the expected behavior due to variations in annotator demographics. From our experiments, this effect is almost 5% for *binary* accuracy and affects models differently. This is quite significant as we are simply evaluating based on positive and negative expressed sentiment and not at a fine grained level. For more recent models (MAG-XLNET and MAG-Bert) the performance drops *more* than the older technique (MulT). As models have become more capable of capturing dataset nuances, these effects appear to become more amplified based on this experiment. This points to the importance for more rigorous evaluation measures that include annotator demographic information. This is particularly important for annotations that have high degrees of subjectivity.

6.5 Discussion

We discuss the impact of annotator demographics on dataset biases and model efficacy, current limitations of our work, and recommendations for the research field.

6.5.1 Impact

In this work we produced a set of annotations large enough for machine learning evaluations that contains detailed demographic information. We find that there can be a nearly 5% difference (77.8%-72.9% in the case of MulT) in binary classification accuracy alone when adjusting demographics for evaluating model behavior. This difference is likely exacerbated when when examining fine-grained sentiment classification or for more controversial annotation tasks. For example, in applications such as language toxicity classification, it has been observed that real world user experience and

reported algorithmic performance is vastly different [244]. Furthermore, as models are becoming increasingly expressive and optimized with respect to original datasets (which clearly have their own demographic biases), demographic differences may make the seeming improvements much less pronounced and impressive. This can be observed in Table 6.4 where MulT, an older algorithm, outperforms the newer algorithms MAG-Bert and MAG-XLNet when labels are given by annotators which follow a US population distribution. We further contextualize this expectation via hypothetical user groups who use sentiment prediction algorithms and frameworks in Table 6.5. With no exception, we see that demographics that correspond to the majority class correlate better with previous annotations. These results mean that any user who belongs to a minority demographic group (with respect to the overall annotator demographics distribution) will perceive the sentiment rating system to perform worse than those in the majority group. Given that there are quite different biases in common machine learning corpora, namely biases towards white male populations, and common crowdsourcing populations, namely biases towards young white female populations – both uniquely problematic —, this situation is bound to occur fairly often.

Many techniques have tackled the issue of mismatch in real world versus experimental metrics from different angles. We find that annotator biases quantified by demographics might be one important source of the issue. This would suggest that existing datasets, while valuable and necessary for the development of learning models, do not work well for a large portion of the population in practice. However, as previous works have pointed out, biases can be removed by increasing the number of annotators in diverse groups [213]. This suggest that one solution might be to extend current datasets with additional annotations from less represented demographics.

6.5.2 Limitations

While we see the data and analysis as highly beneficial for the domain, there are limitations to the answers that our work can provide. The data gathered was limited to a single country (US), and more work is needed to understand the effect of a wider demographic on machine labeled data. It may be for this reason that we did not see a significant effect for ethnicity, and culture across regions could change (likely amplify) the significance of certain effects. Additional effort will need to be made to examine the differences across cultures and the effects. Another demographic issue is that while we obtained data for additional gender categories, we could not obtain sufficient amounts of data to model gender as a non-binary demographic. For this reason, we resorted to analysis using a proportion of women in the overall dataset. This allowed us to show that differences do exist when accounting for gender and the degree of this effect. In our study, we compromised on these demographic choices as including them would have drastically exacerbated the long-tailed distribution of crowdworker demographics. The availability of annotators from certain demographic groups was frequently very low.

Considerations regarding dataset type should also be made. The annotations are for a specific kind of data – opinions and monologue videos. While it is an important problem, there is a lot more to sentiment recognition research beyond just talking heads and opinions. The MOSEI dataset is mostly comprised of video samples which is less controversial than topics such as hate speech [236, 237, 214, 238]. While previous works have demonstrated that demographic imparts differences in ratings for videos of differing content type, the degree of this effect does not appear to be quantified. Furthermore, labeling content that is less subjective might also demonstrate different effects. For example, in the case of determining dogs versus cats, demographic background likely play a smaller role. We provide our rich annotations for future researchers to understand

and correctly these differences.

6.5.3 Recommendations

We echo existing calls for caution when using ML systems. We recommend that ML practitioners should be cautious when implementing technologies that use sentiment prediction models and that users should take great care in interpreting model predictions. This is particularly important in high risk scenarios such as in clinical settings. We hope that current users and practitioners can use our results to interpret ML predictions in a new light. In our experiments, we found that the attributes of annotators imparted a significant difference on both the ground truth as well as the model predictions. One potential approach which follows from this is to match the demographic properties of the annotator with the users of the predictions of the models. Or, more generally, we can try to match annotator distribution with user distribution to maintain performance of any system. Because naturally, demographics do not explain all of the variability or distribution differences in annotator properties. However, by restricting the user distribution, HCI researchers are greatly restricting themselves in the experiments they can conduct and the designs they can create. As models become more capable, data variability, such as those arising due to demographics, is more readily captured by the model. Yet at the same time, without understanding what this variability is, it can be difficult to both improve the model or interpret the results correctly. For example, suppose we did not know that most dataset annotators are female, then we would be confused as to why a sentiment prediction software works better for female than for non-female users. This highlights an important need for increased collaboration between ML and HCI researchers to develop better models and to build more representative datasets.

We recommend the collection and release of properly anonymized annotator demo-

graphic information for subjective tasks such as sentiment or emotion labeling. This recommendation was also voiced by previous work (e.g. [212]), showing significance on similar tasks. Our analysis provides strong evidence that machine learning researchers in particular need to be mindful of the demographic composition of human annotators. As added evidence for the importance of this, we show that the effect on algorithms is larger than the expected effect when comparing different human annotations. Release of our data will facilitate the development of improved algorithms for predicting distributions of multimodal sentiment classification for different demographic groups. This would lead to the improved experience of users which are in different demographic groups than the majority of those who provided annotations.

We recommend the development of a richly annotated subset of data to help quantify variability or annotation noise. In this work, we examined the effect of demographics and expect to explore additional biases that can occur on different demographic dimensions. Annotating a subset of data to quantify the degree of variability due to biases is sufficient for analysis and is also considerably more cost-effective than duplication of entire datasets with additional demographic information. The annotation effort for this work for 500 samples for all modalities cost approximately 3000 USD, or approximately 750 USD per modality. We find this cost to be reasonable for any large scale annotation effort. Understanding dataset biases in this manner can substantially benefit future users from diverse groups, and the insights can likely be transferred to larger datasets.

Lastly, we recommend the balancing of demographic backgrounds of annotators during dataset creation. In our experiments, we found that certain demographic effects were amplified potentially due to the task being more appealing to certain populations. And one has to be mindful of the danger of oversampling individual annotators in highly underrepresented demographic groups. We took a less aggressive approach with regard to enforcing parity of underrepresented demographic groups and saw a benefit to the overall

representation. Such a method is easy to implement in practice and can potentially be combined with methods such as [213]. While not perfect in that it would not result in the ultimately desired (e.g. uniform) distribution, the improved representation might increase the benefit to more groups of future users, and statistical methods, paired with the insights from our work, can be used to approximate the desired distributions.

6.6 Conclusion

The goal of this work was to gain an understanding for the variability that subjectively annotated datasets might contain. Towards this goal, we present a large scale dataset that captures annotator demographics variability and contains annotations for multimodal data and its component modalities. We demonstrate the importance for understanding annotator demographics. We show that that demographic differences impute a significant effect on the ratings they provide and that these effects occur in all modalities. We verify these properties and show large algorithmic performance variability when measured against different demographic groups.

As models become more complex and capable of modeling the details of human expression, more thorough evaluations which can account for the biases in data should be conducted. As is the case here, models and evaluations weigh the opinions of those who perform crowdwork versus those who do not differently. This leaves the potential to bias evaluations and model selection to people who are not part of the annotation process. We hope our data and results can be beneficial not only for future researchers who wish to build more representative datasets, but for evaluation of algorithms and understanding annotator behavior.

Age	**Text**	**Video**	**Audio**	**All**	r
$\leq$ **20**	0.45	0.31	0.43	0.47	0.67
21-30	0.47	0.33	0.43	0.48	0.71
31-40	0.44	0.30	0.41	0.41	0.68
$>$ **40**	0.39	0.30	0.43	0.43	0.59
Overall	0.45	0.32	0.42	0.46	0.75

(a) Agreement scores broken by age and modality. Annotators over 40 had slightly lower agreement within themselves. Their predictions also correlated less with previous annotations.

Ethnicity	**Text**	**Video**	**Audio**	**All**	r
White	0.46	0.33	0.45	0.48	0.74
Other	0.44	0.30	0.36	0.42	0.67
Overall	0.45	0.32	0.42	0.46	0.75

(b) Agreement scores by ethnicity and modality. Non-white annotators had lower within-group agreement and lower correlation with previous labels. Lower agreement is observed in all modalities.

Gender	**Text**	**Video**	**Audio**	**All**	r
Female	0.48	0.32	0.44	0.47	0.76
Other	0.38	0.27	0.35	0.42	0.60
Overall	0.45	0.32	0.42	0.46	0.75

(c) Agreement scores by gender and modality. Non-female annotators had lower within-group agreement and lower correlation with previous labels. Reduction in agreement is observed in all modalities.

Table 6.3: Agreement scores by modality for age, ethnicity, and gender. For text, video, audio, and all modalities, we report Krippendorff's alpha computed using a variable number of annotators and accounts for missing data. We also report the correlation of the labels (all modalities only) provided by each demographic with the original MOSEI labels (r). Older (> 40), non-white, and non-female populations all demonstrated lower agreement with original labels. This further showcases the bias when not controlling for annotator demographics during annotation. The original annotations were obtained via Mechanical Turk which had higher proportions of younger, white and female annotators.

	Most Similar				Least Similar			
Model	**Acc_2**	**F1**	**MAE**	r	**Acc_2**	**F1**	**MAE**	r
MulT	**0.778**	0.779	0.724	0.617	**0.729**	0.734	0.905	0.506
MAG-Bert	**0.756**	0.752	0.698	0.679	**0.715**	0.714	0.879	0.551
MAG-XLNet	**0.766**	0.760	0.733	0.683	**0.729**	0.726	0.909	0.556
Human	**0.818**	0.711	0.661	0.748	**0.785**	0.685	0.845	0.618

(a) Most similar sampling versus least similar sampling performance metrics.

	US Population				Crowdworker Distribution			
Model	**Acc_2**	**F1**	**MAE**	r	**Acc_2**	**F1**	**MAE**	r
MulT	0.770	0.771	0.708	0.623	0.791	0.795	0.599	0.625
MAG-Bert	0.756	0.752	0.678	0.690	0.811	0.811	0.584	0.695
MAG-XLNet	0.754	0.748	0.727	0.689	0.861	0.860	0.551	0.746
Human	0.806	0.703	0.644	0.752	1.000	1.000	0.000	1.000

(b) US population distribution sampling compared to uncontrolled crowdworker distribution.

Table 6.4: Performance metrics when measured against different demographic distributions of our annotations. We see some recent models have reduced performance metrics when measured against different sampling techniques. Binary accuracy measured according to Tsai et al. [87]. The human model is the original human (MOSEI) annotations compared against our new annotations. The difference w.r.t. original dataset is minimal due to the difference in population and training effect differences. Demographics has a much larger effect on learned models. The difference in performance provides demographic bounds on algorithmic performance. Notice the large difference in accuracy for the same model for most similar and least similar sampling (**bold**). This indicates that demographic differences of annotators can account for more than 4.5 percent difference in performance.

	Younger Annotators (<30)				Older Annotators (≥30)			
Model	**Acc_2**	**F1**	**MAE**	r	**Acc_2**	**F1**	**MAE**	r
MulT	0.756	0.759	0.746	0.604	0.772	0.771	0.785	0.591
MAG-Bert	0.746	0.742	0.714	0.666	0.748	0.741	0.810	0.643
MAG-XLNet	0.754	0.752	0.755	0.664	0.756	0.747	0.763	0.651
Human	0.808	0.712	0.673	0.733	0.792	0.707	0.749	0.691

(a) Comparison of younger annotators (<30) versus older annotators ≥ 30.

	Female Annotators				Other Annotators			
Model	**Acc_2**	**F1**	**MAE**	r	**Acc_2**	**F1**	**MAE**	r
MulT	0.768	0.771	0.737	0.616	0.732	0.736	0.882	0.533
MAG-Bert	0.746	0.742	0.705	0.681	0.742	0.740	0.833	0.605
MAG-XLNet	0.754	0.751	0.753	0.674	0.726	0.721	0.868	0.609
Human	**0.816**	0.711	0.652	0.755	**0.760**	0.693	0.833	0.621

(b) Comparison of female versus other annotators.

	White				Other			
Model	**Acc_2**	**F1**	**MAE**	r	**Acc_2**	**F1**	**MAE**	r
MulT	0.766	0.766	0.750	0.606	0.762	0.764	0.809	0.567
MAG-Bert	0.752	0.747	0.704	0.678	0.756	0.753	0.789	0.616
MAG-XLNet	0.752	0.749	0.759	0.670	0.748	0.746	0.821	0.628
Human	0.798	0.704	0.678	0.738	0.774	0.698	0.761	0.680

(c) Comparison of ethnically white vs annotators who are not ethnically white.

Table 6.5: Performance metrics when measured against different demographic distributions of our annotations. We compare against specific demographics to observe differences for certain user groups. Assuming that users of a certain demographic prefer annotations from the same demographic, we can see a difference in performance for different user groups. Tables on left (a, c, e) present results that represents the majority of crowdworker demographics. We can see our scores from demographics which match the majority demographic population far better than those that match the minority. Reduction in model performance is also pronounced for female vs non-female annotators (c vs d). The trend of decrease between majority vs minority demographics is obvious via MAE and correlation.

Chapter 7

Conclusion

7.1 Summary

In this thesis, we make contributions in two broad directions: 1) multimodal machine learning, and 2) addressing variability arising from differences in human opinions modeling human-centric data. We further examine how multimodality and variability are inter-related. We do this to bring together the domains of multimodality and variability in hopes of addressing critical and common problems relevant to human-centric computing such as understanding human behavior and emotions.

In Part I, we examined how we can improve multimodal machine learning models. We presented work which made use of multiple modalities including EEG, eye-tracking, motion, video, audio, and text. We demonstrated several ways in which the modeling of these signals can be improved to overcome motion and computational constraints. In Part II, we examined the theme of variability and discussed how differences in human opinion can impact multimodal problems. We presented methods for how we can alleviate issues of imperfect labels in a unimodal scheme, and examine how imperfect labels are manifested in a multimodal scheme.

7.1.1 Multimodality

Since human interaction is inherently multimodal, creating better multimodal models is critical to building AIs that can better interact with humans. In this thesis, we made advancements in two key areas of multimodal machine learning. First we presented a state-of-the-art method to interpret human EEG and motion signals. In particular, we are targeting the situation where the user is in motion. Such a requirement is important as we cannot expect users to be stationary during interaction. Furthermore, EEG sensors were not designed to operate in motion and thus incorporating it as a sensor for in-motion settings remains a challenge. We further provide a demonstration of how a multimodal system can be used for authentication to an AR/VR device using EEG and other sensors.

Continuing along the lines of building a convenient and multimodal human-AI interface, we then discussed our work on sparsifying inputs during fusion of inputs. As we wish to build natural interfaces, we cannot expect to have high-powered devices available to use at all times. We show that by incorporating the sparsification process during multimodal fusion, we can maintain algorithmic performance while reducing computational cost. Traditionally multimodal signals have been used to overcome per-modality signal inaccuracies, we flip this intuition and propose that modern sensors are more detailed than necessary. By taking advantage of this, we can prune out unnecessary information to save on computational cost.

7.1.2 Variability

Another key property for human behavior is that we are all variable in the way we behave. This highlights the importance of studying techniques that can tolerate and take advantage of this property. In Part II, we first explore this question from the angle of affect classification. Affect classification encompasses many properties of human behavior

including variability of human reactions. We develop a way to model the distribution of responses that are induced by viewing affect-rich videos on the internet. Motivated by this, we explored ways to improve techniques that can address errors in human responses. This led us to our work on noisy labels in which we presented a state of the art technique for modeling imperfect datasets.

One of the key goals of this thesis is to bring together the ideas of multimodality and variability because they are both critical pieces for understanding human behavior. To do this, we developed an evaluation dataset based on an existing dataset for multimodal sentiment analysis by number of labels. We obtained a high number of annotations per sample to fully capture human behavior characteristics and multimodality effects. This led to our discovery that demographic traits can account for larger differences in performance metrics than algorithmic improvements. We use this work and thesis as the foundation for future work in the study of multimodality and variability in human-centric machine learning.

7.2 Future Work

A broad goal of mine has always been to enable AI to understand human behaviors and emotions. This is useful as it would enable much richer interaction and collaboration between the AI and human, such as for personal health support. To effectively address these goals, I believe an end-to-end holistic approach is needed. I identify three broad fundamental future research areas that will serve as the foundation for my future research pursuits. First, data quality requirements for factors such as privacy and diversity are stringent for tasks involving human content. Second, the development of algorithms is needed to analyze the complex temporal, cross-modal and causal relationships in human-centered data. Third, how can we effectively integrating these new AI capabilities into

our daily lives.

7.2.1 Data Challenges

To enable ML algorithms requires data gathered from humans for analysis. Important considerations for privacy and data diversity should be made. For example, personal information has the potential to be remembered by a deep learning model. Additionally, content diversity can potentially bias the dataset to be exclusive to groups not represented by the dataset. Lastly, the subjective qualities of datasets for tasks such as intention or emotion prediction introduces large variability. All these challenges are compounded by the fact that machine learning requires copious amounts of data.

A promising angle will be to address these issues by exploring privacy protecting ways for building datasets, as well as improving the representation qualities. As one of the foundational building blocks for data-driven machine learning models, understanding biases and the representation that datasets impart on the model is central to our pursuit of fair, equal, and accessible AI. A particular question of interest is to explore the question of how content may impart different types of biases on different annotators. For example a male annotator rating a female subject might be different than a female annotator rating the same female subject.

Another future project would be to curate a large dataset suitable for training general purpose feature extraction algorithms on human data, and gather smaller, more personalized data for optimization for a particular user for evaluation purposes. Lastly, gathering quality data from humans requires the development of quality perceptual metrics that can can quantify behavior between multiple people and for individuals over time. All of these tasks coupled with multimodal influences are rich for exploration.

7.2.2 Model Challenges

Human data is highly variable, temporal, and typically involves complex multimodal and causal relationships. Modeling this data effectively in addition to addressing concerns such as privacy are critical and important in machine learning. An natural direction is to focus on taking existing approaches for addressing the modeling of data in the unimodal realm and adapt it for the multimodal, human-centric realm.

To address these issues, future work will continue to explore the fundamental machine learning problem of learning with noisy labels. We will further explore techniques that can learn by exploiting the disagreement between multiple annotators. Additionally, we will explore techniques that can learn better representations through unsupervised or self-supervised methodologies for this task.

We also wish to explore ways to address concerns for privacy from the model perspective by enabling better machine learning in low-resource environments. This would remove the need to send personal data to the cloud for training. We will explore how federated learning techniques can be applied here. Lastly, to effectively model and support human-centered data, developing explainable models is of critical importance.

7.2.3 Interaction Challenges

I believe that there are additional ways to leverage EEG and other psychophysiological data for both understanding fundamental processes of human behavior, as well as inform fundamental development of better neural network algorithms. However, the development of more intelligent systems is useless without the continuous feedback of users. An area of investigation is in how we can take advantage of the capabilities offered by how AI's understanding of humans to build novel interfaces involving humans in the loop. A few key questions of interest arise here: 1) How should an AI exhibit proactive-

ness? 2) How should an AI respond when it senses ambiguity or sarcasm? 3) What might the AI be able to sense about user's current state to improve and potentially offload a task at hand? Critically important questions relevant to ethics for interaction are also beginning to surface: How should we handle over-reliance (accidents in self driving cars)? Who is responsible for mistakes made by an artificial agent? All of these questions must be answered before AI can be effectively integrated into society.

In addition to answering fundamental questions for how we can interact with an increasingly advanced system, I believe there is room to incorporate the feedback of potential future users of this technology. A few "testbed" applications that would also be highly beneficial to society, I believe lie in the health and education domains. In the future, I hope to partner with education and health researchers to explore the limits of current systems and opportunities for improvement. Questions such as how we can use AI to improve the education experience and to support learning for people of all ages are great next steps for research.

I believe that there is now and in the future a need to address AI research and HCI research in a wholistic manner. This fact is evident in the rising popularity of the nascent field of Human-Centered AI. It is an exciting time to be working at the intersection of AI and HCI – building models that can address the concerns necessary for better Human-AI interaction; developing novel interactive technologies that can fully take advantage of model capabilities; and making improvements to each while respecting the data privacy, diversity, and quantity concerns.

Bibliography

[1] H. McGurk and J. MacDonald, *Hearing lips and seeing voices*, *Nature* **264** (1976), no. 5588 746–748. 3

[2] R. Heale and A. Twycross, *Validity and reliability in quantitative studies*, *Evidence-based nursing* **18** (2015), no. 3 66–67. 4

[3] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, *Brain–computer interfaces for communication and control*, *Clinical neurophysiology* **113** (2002), no. 6 767–791. 8

[4] A. B. Schwartz, X. T. Cui, D. J. Weber, and D. W. Moran, *Brain-controlled interfaces: movement restoration with neural prosthetics*, *Neuron* **52** (2006), no. 1 205–220. 8

[5] G. Fischer, *User modeling in human–computer interaction*, *User modeling and user-adapted interaction* **11** (2001), no. 1-2 65–86. 9

[6] A. Cockburn, C. Gutwin, J. Scarr, and S. Malacria, *Supporting novice to expert transitions in user interfaces*, *ACM Computing Surveys (CSUR)* **47** (2015), no. 2 31. 9

[7] L. R. Krol and T. O. Zander, *Passive bci-based neuroadaptive systems.*, in *GBCIC*, 2017. 9

[8] M. Rötting, T. Zander, S. Trösterer, and J. Dzaack, *Implicit interaction in multimodal human-machine systems*, in *Industrial Engineering and Ergonomics*, pp. 523–536. Springer, 2009. 9

[9] N. Friedman, T. Fekete, Y. K. Gal, and O. Shriki, *Eeg-based prediction of cognitive load in intelligence tests*, *Frontiers in Human Neuroscience* **13** (2019) 191. 9

[10] N. Kumar and J. Kumar, *Measurement of cognitive load in hci systems using eeg power spectrum: an experimental study*, *Procedia Computer Science* **84** (2016) 70–78. 9

[11] J. Atkinson and D. Campos, *Improving bci-based emotion recognition by combining eeg feature selection and kernel classifiers*, *Expert Systems with Applications* **47** (2016) 35–41. 9

[12] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *Deap: A database for emotion analysis; using physiological signals*, *IEEE transactions on affective computing* **3** (2011), no. 1 18–31. 9, 61, 64, 72, 110

[13] N.-H. Liu, C.-Y. Chiang, and H.-C. Chu, *Recognizing the degree of human attention using eeg signals from mobile sensors*, *Sensors* **13** (2013), no. 8 10273–10286. 9

[14] C.-M. Chen, J.-Y. Wang, and C.-M. Yu, *Assessing the attention levels of students by using a novel attention aware system based on brainwave signals*, *British Journal of Educational Technology* **48** (2017), no. 2 348–369. 9

[15] F.-R. Lin and C.-M. Kao, *Mental effort detection using eeg data in e-learning contexts*, *Computers & Education* **122** (2018) 63–79. 9

[16] L.-D. Liao, C.-Y. Chen, I.-J. Wang, S.-F. Chen, S.-Y. Li, B.-W. Chen, J.-Y. Chang, and C.-T. Lin, *Gaming control using a wearable and wireless eeg-based brain-computer interface device with novel dry foam-based sensors*, *Journal of neuroengineering and rehabilitation* **9** (2012), no. 1 5. 9

[17] T. O. Zander, L. R. Krol, N. P. Birbaumer, and K. Gramann, *Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity*, *Proceedings of the National Academy of Sciences* **113** (2016), no. 52 14898–14903. 9

[18] P. Aricò, G. Borghini, G. Di Flumeri, N. Sciaraffa, and F. Babiloni, *Passive bci beyond the lab: current trends and future directions*, *Physiological measurement* **39** (2018), no. 8 08TR02. 9

[19] M. Lohani, B. R. Payne, and D. L. Strayer, *A review of psychophysiological measures to assess cognitive states in real-world driving*, *Frontiers in human neuroscience* **13** (2019). 9

[20] S. L. Shishkin, Y. O. Nuzhdin, E. P. Svirin, A. G. Trofimov, A. A. Fedorova, B. L. Kozyrskiy, and B. M. Velichkovsky, *Eeg negativity in fixations used for gaze-based control: Toward converting intentions into actions with an eye-brain-computer interface*, *Frontiers in neuroscience* **10** (2016) 528. 9

[21] A. Appriou, A. Cichocki, and F. Lotte, *Towards robust neuroadaptive hci: exploring modern machine learning methods to estimate mental workload from eeg*

signals, in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, p. LBW615, ACM, 2018. 9

[22] G. D. Abowd and E. D. Mynatt, *Charting past, present, and future research in ubiquitous computing*, *ACM Transactions on Computer-Human Interaction (TOCHI)* **7** (2000), no. 1 29–58. 10

[23] A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, *A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals*, *Journal of Neural engineering* **4** (2007), no. 2 R32. 10

[24] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, *A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update*, *Journal of neural engineering* **15** (2018), no. 3 031005. 10, 11

[25] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, *Deep learning with convolutional neural networks for EEG decoding and visualization*, *Human brain mapping* **38** (2017), no. 11 5391–5420. 10, 12

[26] G. E. Hinton, S. Osindero, and Y.-W. Teh, *A fast learning algorithm for deep belief nets*, *Neural computation* **18** (2006), no. 7 1527–1554. 10

[27] P. Simard, D. Steinkraus, and J. Platt, *Best practices for convolutional neural networks applied to visual document analysis*, . 11

[28] H. Cecotti, A. Marathe, and A. Ries, *Optimization of single-trial detection of event-related potentials through artificial trials*, *IEEE Trans. on Biomedical Engineering* **62** (2015), no. 9 2170–6. 11

[29] H. Raza, S. Rathee, D.and Zhou, H. Cecotti, and G. Prasad, *Covariate shift estimation based adaptive ensemble learning for handling non-stationarity in motor imagery related EEG-based brain-computer interface*, *Neurocomputing* **343** (2019) 154–166. 11

[30] J. Polich and J. Criado, *Neuropsychology and neuropharmacology of p3a and p3b.*, *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* **60 2** (2006) 172–85. 11, 15

[31] T. Bullock, H. Cecotti, and B. Giesbrecht, *Multiple stages of information processing are modulated during acute bouts of exercise*, *Neuroscience* **307** (2015) 138–150. 11, 14

[32] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, *A review of classification algorithms for EEG-based brain–computer interfaces*, *Journal of neural engineering* **4** (2007), no. 2 R1. 11

[33] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, *Multiclass brain–computer interface classification by riemannian geometry*, *IEEE Transactions on Biomedical Engineering* **59** (2012), no. 4 920–928. 12, 21

[34] H. Cecotti and A. Graeser, *Convolutional neural network with embedded fourier transform for EEG classification*, in *2008 19th International Conference on Pattern Recognition*, pp. 1–4, IEEE, 2008. 12

[35] H. Cecotti and A. Graser, *Convolutional neural networks for p300 detection with application to brain-computer interfaces*, *IEEE transactions on pattern analysis and machine intelligence* **33** (2011), no. 3 433–445. 12

[36] H. Cecotti, M. P. Eckstein, and B. Giesbrecht, *Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering*, *IEEE transactions on neural networks and learning systems* **25** (2014), no. 11 2030–2042. 12

[37] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, *EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces*, *Journal of neural engineering* **15** (2018), no. 5 056013. 12, 14, 21, 22

[38] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, *Mobilenets: Efficient convolutional neural networks for mobile vision applications*, *arXiv preprint arXiv:1704.04861* (2017). 12

[39] Z. Yin and J. Zhang, *Cross-session classification of mental workload levels using EEG and an adaptive deep learning model*, *Biomedical Signal Processing and Control* **33** (2017) 30–47. 12

[40] J. Polich, *Updating p300: an integrative theory of p3a and p3b*, *Clinical neurophysiology* **118** (2007), no. 10 2128–2148. 12

[41] A. B. Said, A. Mohamed, T. Elfouly, K. Harras, and Z. J. Wang, *Multimodal deep learning approach for joint EEG-EMG data compression and classification*, in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, IEEE, 2017. 12

[42] K. Nathan and J. L. Contreras-Vidal, *Negligible motion artifacts in scalp electroencephalography (EEG) during treadmill walking*, *Frontiers in human neuroscience* **9** (2016) 708. 13

[43] R. C. Browning, E. A. Baker, J. A. Herron, and R. Kram, *Effects of obesity and sex on the energetic cost and preferred speed of walking*, *Journal of applied physiology* **100** (2006), no. 2 390–398. 13

[44] Y. Yagi, K. L. Coburn, K. M. Estes, and J. E. Arruda, *Effects of aerobic exercise and gender on visual and auditory P300, reaction time, and accuracy*, *European journal of applied physiology and occupational physiology* **80** (1999), no. 5 402–408. 13

[45] F. Grego, J.-M. Vallier, M. Collardeau, S. Bermon, P. Ferrari, M. Candito, P. Bayer, M.-N. Magnié, and J. Brisswalter, *Effects of long duration exercise on cognitive function, blood glucose, and counterregulatory hormones in male cyclists*, *Neuroscience letters* **364** (2004), no. 2 76–80. 13

[46] T. Bullock, J. C. Elliott, J. T. Serences, and B. Giesbrecht, *Acute exercise modulates feature-selective responses in human cortex*, *Journal of cognitive neuroscience* **29** (2017), no. 4 605–618. 13

[47] L. F. Ciria, P. Perakakis, A. Luque-Casado, and D. Sanabria, *Physical exercise increases overall brain oscillatory activity but does not influence inhibitory control in young adults*, *Neuroimage* **181** (2018) 203–210. 13

[48] G. Cheron, G. Petit, J. Cheron, A. Leroy, A. Cebolla, C. Cevallos, M. Petieau, T. Hoellinger, D. Zarka, A.-M. Clarinval, *et. al.*, *Brain oscillations in sport: toward EEG biomarkers of performance*, *Frontiers in psychology* **7** (2016) 246. 13

[49] J. T. Gwin, K. Gramann, S. Makeig, and D. P. Ferris, *Removal of movement artifact from high-density EEG recorded during walking and running*, *Journal of neurophysiology* **103** (2010), no. 6 3526–3534. 13

[50] J. A. Urigüen and B. Garcia-Zapirain, *EEG artifact removal—state-of-the-art and guidelines*, *Journal of neural engineering* **12** (2015), no. 3 031001. 14

[51] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. Miller, G. Mueller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, *Review of the BCI competition iv*, *Frontiers in Neuroscience* **6** (2012) 55. 14

[52] P. Margaux, M. Emmanuel, D. Sébastien, B. Olivier, and M. Jérémie, *Objective and Subjective Evaluation of Online Error Correction during P300-Based Spelling*, *Advances in Human-Computer Interaction* **2012** (dec, 2012) 1–13. 14

[53] G. Borg, *Perceived exertion as an indicator of somatic stress.*, *Scandinavian Journal of Rehabilitation Medicine* **2** (1970), no. 2 92–98. cited By 2796. 15

[54] P. Reis, F. Kluge, F. Gabsteiger, V. Tscharner, and M. Lochmann, *Methodological aspects of eeg and body dynamics measurements during motion.*, *Frontiers in human neuroscience* **8** (03, 2014) 156. 16

[55] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, *Brain–computer interfaces for communication and control*, *Clinical Neurophysiology* **113** (June, 2002) 767–791. 28

[56] S. N. Abdulkader, A. Atia, and M.-S. M. Mostafa, *Brain computer interfacing: Applications and challenges*, *Egyptian Informatics Journal* **16** (July, 2015) 213–230. 28

[57] H. Van Dis, M. Corner, R. Dapper, G. Hanewald, and H. Kok, *Individual differences in the human electroencephalogram during quiet wakefulness*, *Electroencephalography and Clinical Neurophysiology* **47** (July, 1979) 87–94. 28

[58] A. K. Jain, K. Nandakumar, and A. Nagar, *Biometric Template Security*, *EURASIP J. Adv. Signal Process* **2008** (Jan., 2008) 113:1–113:17. 28

[59] R. Moskovitch, C. Feher, A. Messerman, N. Kirschnick, T. Mustafic, A. Camtepe, B. Lohlein, U. Heister, S. Moller, L. Rokach, and Y. Elovici, *Identity theft, computers and behavioral biometrics*, in *2009 IEEE International Conference on Intelligence and Security Informatics*, pp. 155–160, June, 2009. 28

[60] C. D. Holland and O. V. Komogortsev, *Complex eye movement pattern biometrics: Analyzing fixations and saccades*, in *2013 International Conference on Biometrics (ICB)*, pp. 1–8, June, 2013. 28

[61] M. Wang, H. A. Abbass, and J. Hu, *Continuous authentication using EEG and face images for trusted autonomous systems*, in *2016 14th Annual Conference on Privacy, Security and Trust (PST)*, pp. 368–375, Dec., 2016. 28

[62] K. Revett, F. Deravi, and K. Sirlantzis, *Biosignals for User Authentication - Towards Cognitive Biometrics?*, in *2010 International Conference on Emerging Security Technologies*, pp. 71–76, Sept., 2010. 29

[63] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, *Multimodal emotion recognition using eeg and eye tracking data*, in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5040–5043, IEEE, 2014. 29

[64] S. Marcel and J. D. R. Millan, *Person Authentication Using Brainwaves (EEG) and Maximum A Posteriori Model Adaptation*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29** (Apr., 2007) 743–752. 29

[65] J. Hu, *New biometric approach based on motor imagery EEG signals*, in *2009 International Conference on Future BioMedical Information Engineering (FBIE)*, pp. 94–97, Dec., 2009. 29

[66] R. Das, E. Maiorana, and P. Campisi, *Motor Imagery for Eeg Biometrics Using Convolutional Neural Network*, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2062–2066, Apr., 2018. 29

[67] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, *BCI2000: a general-purpose brain-computer interface (BCI) system*, *IEEE Trans Biomed Eng* **51** (June, 2004) 1034–1043. 29

[68] Goldberger Ary L., Amaral Luis A. N., Glass Leon, Hausdorff Jeffrey M., Ivanov Plamen Ch., Mark Roger G., Mietus Joseph E., Moody George B., Peng Chung-Kang, and Stanley H. Eugene, *PhysioBank, PhysioToolkit, and PhysioNet*, *Circulation* **101** (June, 2000) e215–e220. 29

[69] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen, *MNE software for processing MEG and EEG data*, *NeuroImage* **86** (Feb., 2014) 446–460. 30

[70] P. Kasprowski, O. V. Komogortsev, and A. Karpov, *First eye movement verification and identification competition at btas 2012*, in *2012 IEEE fifth international conference on biometrics: theory, applications and systems (BTAS)*, pp. 195–202, IEEE, 2012. 30

[71] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, *Integrating multimodal information in large pretrained transformers*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 2359–2369, Association for Computational Linguistics, July, 2020. 37

[72] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, *Attention bottlenecks for multimodal fusion*, *arXiv preprint arXiv:2107.00135* (2021). 37, 39, 48, 49, 109

[73] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, *CoRR* **abs/2010.11929** (2020) [arXiv:2010.1192]. 37, 39, 109

[74] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, *Training data-efficient image transformers & distillation through attention*, in *Proceedings of the 38th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, PMLR, 18–24 Jul, 2021. 37, 40

[75] S. Atito, M. Awais, and J. Kittler, *SiT: Self-supervised vision transformer*, *CoRR* **abs/2104.03602** (2021) [arXiv:2104.0360]. 37

[76] H. Bao, L. Dong, and F. Wei, *BEiT: BERT pre-training of image transformers*, 2021. 37

[77] Z. Pan, B. Zhuang, J. Liu, H. He, and J. Cai, *Scalable vision transformers with hierarchical pooling*, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 377–386, 2021. 37, 40, 50

[78] Y. Wang, R. Huang, S. Song, Z. Huang, and G. Huang, *Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition*, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 37

[79] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, *Multimodal machine learning: A survey and taxonomy*, *IEEE transactions on pattern analysis and machine intelligence* **41** (2018), no. 2 423–443. 38, 109

[80] B. Yuhas, M. Goldstein, and T. Sejnowski, *Integration of acoustic and visual speech signals using neural networks*, *IEEE Communications Magazine* **27** (1989), no. 11 65–71. 39

[81] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, *Words can shift: Dynamically adjusting word representations using nonverbal behaviors*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7216–7223, 2019. 39, 80, 109, 114, 128

[82] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, *Memory fusion network for multi-view sequential learning*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 39

[83] C. Feichtenhofer, A. Pinz, and A. Zisserman, *Convolutional two-stream network fusion for video action recognition*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941, 2016. 39

[84] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, in *Advances in neural information processing systems*, pp. 5998–6008, 2017. 39, 42, 109

[85] A. Božič, P. Palafox, J. Thies, A. Dai, and M. Nießner, *TransformerFusion: Monocular RGB scene reconstruction using transformers*, *Proc. Neural Information Processing Systems (NeurIPS)* (2021). 39

[86] N. Stier, A. Rich, P. Sen, and T. Höllerer, *VoRTX: Volumetric 3D reconstruction with transformers for voxelwise view selection and fusion*, in *3DV*, 2021. 39

[87] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, *Multimodal transformer for unaligned multimodal language sequences*, *arXiv preprint arXiv:1906.00295* (2019). 39, 49, 80, 109, 113, 114, 123, 128, 137

[88] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira, *Perceiver: General perception with iterative attention*, *arXiv preprint arXiv:2103.03206* (2021). 39

[89] M. Zhang and Y. He, *Accelerating training of transformer-based language models with progressive layer dropping*, *arXiv preprint arXiv:2010.13369* (2020). 40

[90] J. Ren, S. Rajbhandari, R. Y. Aminabadi, O. Ruwase, S. Yang, M. Zhang, D. Li, and Y. He, *Zero-offload: Democratizing billion-scale model training*, *arXiv preprint arXiv:2101.06840* (2021). 40

[91] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, *DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters*, in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506, 2020. 40

[92] Z. Wang, P. Ng, X. Ma, R. Nallapati, and B. Xiang, *Multi-passage BERT: A globally normalized BERT model for open-domain question answering*, *arXiv preprint arXiv:1908.08167* (2019). 40

[93] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, *Big Bird: Transformers for longer sequences*, in *NeurIPS*, 2020. 40

[94] N. Kitaev, Ł. Kaiser, and A. Levskaya, *Reformer: The efficient transformer*, *arXiv preprint arXiv:2001.04451* (2020). 40

[95] J. W. Rae, A. Potapenko, S. M. Jayakumar, and T. P. Lillicrap, *Compressive transformers for long-range sequence modelling*, *arXiv preprint arXiv:1911.05507* (2019). 40

[96] Z. Ye, Q. Guo, Q. Gan, X. Qiu, and Z. Zhang, *BP-transformer: Modelling long-range context via binary partitioning*, *arXiv preprint arXiv:1911.04070* (2019). 40

[97] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, *Dropout: a simple way to prevent neural networks from overfitting*, *The journal of machine learning research* **15** (2014), no. 1 1929–1958. 40

[98] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, in *ICLR*, 2019. 40

[99] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, *mixup: Beyond empirical risk minimization*, *arXiv preprint arXiv:1710.09412* (2017). 40, 44, 85, 88, 94, 99

[100] K. Nishi, Y. Ding, A. Rich, and T. Hollerer, *Augmentation strategies for learning with noisy labels*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8022–8031, 2021. 40, 111

[101] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, *Mixmatch: A holistic approach to semi-supervised learning*, in *Advances in Neural Information Processing Systems*, pp. 5049–5059, 2019. 40, 91, 94

[102] V. Verma, A. Lamb, C. Beckham, A. Najafi, A. Courville, I. Mitliagkas, and Y. Bengio, *Manifold mixup: learning better representations by interpolating hidden states*, . 40, 42, 44

[103] R. Child, S. Gray, A. Radford, and I. Sutskever, *Generating long sequences with sparse transformers*, *URL https://openai.com/blog/sparse-transformers* (2019). 42

[104] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, *VGGSound: A large-scale audio-visual dataset*, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, IEEE, 2020. 45, 46

[105] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, *Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph*, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018. 45, 46, 48, 105, 106, 107, 110, 111, 114, 116, 123, 124

[106] J. Carreira and A. Zisserman, *Quo vadis, action recognition? a new model and the kinetics dataset*, in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017. 48

[107] S. Goyal, A. R. Choudhury, S. M. Raje, V. T. Chakaravarthy, Y. Sabharwal, and A. Verma, *PoWER-BERT: Accelerating BERT inference via progressive word-vector elimination*, *arXiv preprint arXiv:2001.08950* (2020). 56

[108] R. W. Picard, *Affective computing*. MIT press, 2000. 61

[109] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, *Deep Affect Prediction in-the-wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond*, *International Journal of Computer Vision* **127** (June, 2019) 907–929, [arXiv:1804.1093]. 61, 64

[110] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, *Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph*, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne,

Australia), pp. 2236–2246, Association for Computational Linguistics, July, 2018. 61, 64, 72

[111] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, *Generating sentences from a continuous space*, *arXiv preprint arXiv:1511.06349* (2015). 63, 67, 74

[112] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, *arXiv preprint arXiv:1312.6114* (2013). 63

[113] P. Ekman and D. Keltner, *Universal facial expressions of emotion*, *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture* (1997) 27–46. 64

[114] J. Russell, *A Circumplex Model of Affect*, *Journal of Personality and Social Psychology* **39** (Dec., 1980) 1161–1178. 64, 66

[115] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford University Press, Sept., 1990. 64

[116] B. Liu, *Sentiment analysis and opinion mining*, *Synthesis lectures on human language technologies* **5** (2012), no. 1 1–167. 64

[117] P. Kuppens, F. Tuerlinckx, J. A. Russell, and L. F. Barrett, *The relation between valence and arousal in subjective experience.*, *Psychological bulletin* **139** (2013), no. 4 917. 64

[118] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, *Learning factorized multimodal representations*, *arXiv preprint arXiv:1806.06176* (2018). 64, 81

[119] M. Martin, *On the induction of mood*, *Clinical Psychology Review* **10** (Jan., 1990) 669–697. 64

[120] A. M. Bhatti, M. Majid, S. M. Anwar, and B. Khan, *Human emotion recognition and analysis in response to audio music using brain signals*, *Computers in Human Behavior* **65** (2016) 267–275. 64

[121] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, *Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages*, *IEEE Intelligent Systems* **31** (Nov., 2016) 82–88. 64

[122] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 527–536, Association for Computational Linguistics, July, 2019. 64

[123] A. Yazdani, J.-S. Lee, J.-M. Vesin, and T. Ebrahimi, *Affect recognition based on physiological changes during the watching of music videos*, *ACM Transactions on Interactive Intelligent Systems* **2** (Mar., 2012) 1–26. 64

[124] W. Liu, W.-L. Zheng, and B.-L. Lu, *Emotion Recognition Using Multimodal Deep Learning*, in *Neural Information Processing* (A. Hirose, S. Ozawa, K. Doya, K. Ikeda, M. Lee, and D. Liu, eds.), Lecture Notes in Computer Science, (Cham), pp. 521–529, Springer International Publishing, 2016. 65

[125] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, *Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset.*, in *Twenty-Ninth IAAI Conference*, Feb., 2017. 65

[126] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, *A neural probabilistic language model*, *Journal of machine learning research* **3** (2003), no. Feb 1137–1155. 65

[127] A. Mnih and G. Hinton, *Three new graphical models for statistical language modelling*, in *Proceedings of the 24th International Conference on Machine Learning - ICML '07*, (Corvalis, Oregon), pp. 641–648, ACM Press, 2007. 65

[128] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, *Recurrent neural network based language model*, in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, vol. 2, pp. 1045–1048, Jan., 2010. 65

[129] R. Kiros, R. Salakhutdinov, and R. Zemel, *Multimodal neural language models*, in *International conference on machine learning*, pp. 595–603, 2014. 65

[130] S. Ahn, H. Choi, T. Pärnamaa, and Y. Bengio, *A Neural Knowledge Language Model*, *arXiv:1608.00318 [cs]* (Mar., 2017) [arXiv:1608.0031]. 65

[131] X. Liu, P. He, W. Chen, and J. Gao, *Multi-Task Deep Neural Networks for Natural Language Understanding*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 4487–4496, Association for Computational Linguistics, July, 2019. 65

[132] S. Tang, H. Jin, C. Fang, Z. Wang, and V. R. de Sa, *Speeding up context-based sentence representation learning with non-autoregressive convolutional decoding*, *arXiv preprint arXiv:1710.10380* (2017). 65

[133] T. Zhao, K. Lee, and M. Eskenazi, *Unsupervised discrete sentence representation learning for interpretable neural dialog generation*, *arXiv preprint arXiv:1804.08069* (2018). 65

[134] A. Nie, E. D. Bennett, and N. D. Goodman, *Dissent: Sentence representation learning from explicit discourse relations*, *arXiv preprint arXiv:1710.04334* (2017). 65

[135] Y. Jernite, S. R. Bowman, and D. Sontag, *Discourse-based objectives for fast unsupervised sentence representation learning*, *arXiv preprint arXiv:1705.00557* (2017). 65

[136] B. Kratzwald, S. Ilic, M. Kraus, S. Feuerriegel, and H. Prendinger, *Deep learning for affective computing: Text-based emotion recognition in decision support*, *Decision Support Systems* **115** (Nov., 2018) 24–35, [arXiv:1803.0639]. 65

[137] X. Dong and G. de Melo, *A Helping Hand: Transfer Learning for Deep Sentiment Analysis*, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 2524–2534, Association for Computational Linguistics, July, 2018. 65

[138] S. Zhu, S. Li, and G. Zhou, *Adversarial Attention Modeling for Multi-dimensional Emotion Regression*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 471–480, Association for Computational Linguistics, July, 2019. 65

[139] Y. Zhang and Y. Zhang, *Tree communication models for sentiment analysis*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3518–3527, 2019. 65

[140] S. Ghosh, M. Chollet, E. Laksana, L.-P. Morency, and S. Scherer, *Affect-LM: A Neural Language Model for Customizable Affective Text Generation*, *arXiv:1704.06851 [cs]* (Apr., 2017) [arXiv:1704.0685]. 65, 75

[141] X. Zhou and W. Y. Wang, *MojiTalk: Generating Emotional Responses at Scale*, *arXiv:1711.04090 [cs]* (May, 2018) [arXiv:1711.0409]. 65

[142] Z. Song, X. Zheng, L. Liu, M. Xu, and X. Huang, *Generating Responses with a Specific Emotion in Dialog*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 3685–3695, Association for Computational Linguistics, July, 2019. 65

[143] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou, *Variational deep embedding: An unsupervised and generative approach to clustering*, *arXiv preprint arXiv:1611.05148* (2016). 65

[144] Q. Zhao, N. Honnorat, E. Adeli, A. Pfefferbaum, E. V. Sullivan, and K. M. Pohl, *Variational autoencoder with truncated mixture of gaussians for functional connectivity analysis*, in *International Conference on Information Processing in Medical Imaging*, pp. 867–879, Springer, 2019. 66

[145] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, *Semi-supervised learning with deep generative models*, in *Advances in neural information processing systems*, pp. 3581–3589, 2014. 66

[146] C. Doersch, *Tutorial on variational autoencoders*, *arXiv preprint arXiv:1606.05908* (2016). 67

[147] R. Hyman, *Stimulus information as a determinant of reaction time.*, *Journal of experimental psychology* **45** (1953), no. 3 188. 68

[148] J. R. Hershey and P. A. Olsen, *Approximating the kullback leibler divergence between gaussian mixture models*, in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–317, IEEE, 2007. 69

[149] J. Pennington, R. Socher, and C. D. Manning, *Glove: Global vectors for word representation*, in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. 73

[150] I. Sutskever, O. Vinyals, and Q. V. Le, *Sequence to sequence learning with neural networks*, in *Advances in neural information processing systems*, pp. 3104–3112, 2014. 73

[151] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic inquiry and word count: Liwc 2001*, in *Mahway: Lawrence Erlbaum Associates*, 2001. 76

[152] J. Berger, *Arousal increases social transmission of information*, *Psychological science* **22** (2011), no. 7 891–893. 76

[153] C. J. Hutto and E. Gilbert, *Vader: A parsimonious rule-based model for sentiment analysis of social media text*, in *Eighth international AAAI conference on weblogs and social media*, 2014. 76

[154] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, *Self-training with noisy student improves imagenet classification*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10687–10698, 2020. 84

[155] K. Sohn, Z. Zhang, C.-L. Li, H. Zhang, C.-Y. Lee, and T. Pfister, *A simple semi-supervised learning framework for object detection*, *arXiv preprint arXiv:2005.04757* (2020). 84

[156] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, *Autoaugment: Learning augmentation strategies from data*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 113–123, 2019. 84, 87, 93

[157] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, *Randaugment: Practical automated data augmentation with a reduced search space*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020. 84, 87, 93, 101

[158] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, *Augmix: A simple data processing method to improve robustness and uncertainty*, *arXiv preprint arXiv:1912.02781* (2019). 84

[159] T. DeVries and G. W. Taylor, *Improved regularization of convolutional neural networks with cutout*, *arXiv preprint arXiv:1708.04552* (2017). 84

[160] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, *Learning with noisy labels*, in *Advances in neural information processing systems*, pp. 1196–1204, 2013. 85

[161] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, *A closer look at memorization in deep networks*, in *ICML*, 2017. 85, 87, 89, 90, 104

[162] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, *Joint optimization framework for learning with noisy labels*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5552–5560, 2018. 85, 86, 87, 95, 100

[163] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, *Co-teaching: Robust training of deep neural networks with extremely noisy labels*, in *NeurIPS*, pp. 8535–8545, 2018. 85, 86, 87, 90, 94, 102

[164] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, *Unsupervised label noise modeling and loss correction*, *arXiv preprint arXiv:1904.11238* (2019). 85, 90, 97, 99, 100, 102, 103

[165] J. Li, R. Socher, and S. C. Hoi, *Dividemix: Learning with noisy labels as semi-supervised learning*, *arXiv preprint arXiv:2002.07394* (2020). 85, 91, 93, 95, 96, 97, 98, 99, 100, 101, 102

[166] E. Malach and S. Shalev-Shwartz, *Decoupling "when to update" from "how to update"*, in *NIPS*, 2017. 86, 87

[167] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, *Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels*, in *ICML*, 2018. 86, 87, 94

[168] P. Chen, B. B. Liao, G. Chen, and S. Zhang, *Understanding and utilizing deep neural networks trained with noisy labels*, in *ICML*, 2019. 86, 87

[169] S. E. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, *Training deep neural networks on noisy labels with bootstrapping*, in *ICLR*, 2015. 86, 87

[170] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, *Making deep neural networks robust to label noise: a loss correction approach*, in *CVPR*, 2017. 86, 87

[171] J. Goldberger and E. Ben-Reuven, *Training deep neural-networks using a noise adaptation layer*, in *ICLR*, 2017. 86, 87

[172] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. M. Erfani, S.-T. Xia, S. Wijewickrema, and J. Bailey, *Dimensionality-driven learning with noisy labels*, in *ICML*, 2018. 86, 87

[173] M. Sajjadi, M. Javanmardi, and T. Tasdizen, *Regularization with stochastic transformations and perturbations for deep semi-supervised learning*, in *Advances in Neural Information Processing Systems*, 2016. 88

[174] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, *Unsupervised data augmentation for consistency training*, in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6256–6268, Curran Associates, Inc., 2020. 88

[175] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, *Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring*, in *ICLR*, 2020. 88

[176] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, *Fixmatch: Simplifying semi-supervised learning with consistency and confidence*, *arXiv preprint arXiv:2001.07685* (2020). 88, 104

[177] Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang, *Smooth neighbors on teacher graphs for semi-supervised learning*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8896–8905, 2018. 88

[178] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal loss for dense object detection*, in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017. 92

[179] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, *Regularizing neural networks by penalizing confident output distributions*, *arXiv preprint arXiv:1701.06548* (2017). 94

[180] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, *Learning to learn from noisy labeled data*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5051–5059, 2019. 95, 99, 100

[181] K. He, X. Zhang, S. Ren, and J. Sun, *Identity mappings in deep residual networks*, in *European conference on computer vision*, pp. 630–645, Springer, 2016. 95

[182] X. Yu, B. Han, J. Yao, G. Niu, I. W. Tsang, and M. Sugiyama, *How does disagreement help generalization against label corruption?*, *arXiv preprint arXiv:1901.04215* (2019). 97, 99, 102, 103

[183] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich, *Training deep neural networks on noisy labels with bootstrapping*, *arXiv preprint arXiv:1412.6596* (2014). 99

[184] K. Yi and J. Wu, *Probabilistic end-to-end noise correction for learning with noisy labels*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7017–7025, 2019. 99, 100

[185] W. Zhang, Y. Wang, and Y. Qiao, *Metacleaner: Learning to hallucinate clean representations for noisy-labeled visual recognition*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7373–7382, 2019. 100

[186] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, *Early-learning regularization prevents memorization of noisy labels*, *arXiv preprint arXiv:2007.00151* (2020). 100

[187] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, *Learning from massive noisy labeled data for image classification*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2691–2699, 2015. 100

[188] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, *Iemocap: Interactive emotional dyadic motion capture database*, *Language resources and evaluation* **42** (2008), no. 4 335–359. 105, 110

[189] L.-P. Morency, R. Mihalcea, and P. Doshi, *Towards multimodal sentiment analysis: Harvesting opinions from the web*, in *Proceedings of the 13th international conference on multimodal interfaces*, pp. 169–176, 2011. 105

[190] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, *Utterance-level multimodal sentiment analysis*, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 973–982, 2013. 105, 110

[191] A. Zadeh, Y. S. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, *Cmu-moseas: A multimodal language dataset for spanish, portuguese, german and french*, in *Proceedings of the Conference on Empirical Methods in Natural*

Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2020, p. 1801, NIH Public Access, 2020. 105, 106, 110

[192] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, *Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020. 105, 106, 110, 111, 127

[193] L. Wang, D. Wang, F. Tian, Z. Peng, X. Fan, Z. Zhang, M. Yu, X. Ma, and H. Wang, *Cass: Towards building a social-support chatbot for online health community*, *Proceedings of the ACM on Human-Computer Interaction* **5** (2021), no. CSCW1 1–31. 106

[194] F. Yang, X. Peng, G. Ghosh, R. Shilon, H. Ma, E. Moore, and G. Predovic, *Exploring deep multimodal fusion of text and photo for hate speech classification*, in *Proceedings of the third workshop on abusive language online*, pp. 11–18, 2019. 106

[195] C. Giordano, M. Brennan, B. Mohamed, P. Rashidi, F. Modave, and P. Tighe, *Accessing artificial intelligence for clinical decision-making*, *Frontiers in Digital Health* **3** (2021) 65. 106

[196] C. Wolf and J. Blomberg, *Evaluating the promise of human-algorithm collaborations in everyday work practices*, *Proceedings of the ACM on Human-Computer Interaction* **3** (2019), no. CSCW 1–23. 106

[197] R. Snow, B. O'connor, D. Jurafsky, and A. Y. Ng, *Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks*, in *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263, 2008. 106

[198] W. Mason and S. Suri, *Conducting behavioral research on amazon's mechanical turk*, *Behavior research methods* **44** (2012), no. 1 1–23. 106

[199] M. K. Scheuerman, A. Jiang, K. Spiel, and J. R. Brubaker, *Revisiting gendered web forms: An evaluation of gender inputs with (non-) binary people*, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2021. 106

[200] E. A. Plant, J. S. Hyde, D. Keltner, and P. G. Devine, *The gender stereotyping of emotions*, *Psychology of Women Quarterly* **24** (2000), no. 1 81–92. 106, 110, 125

[201] E. Davis, E. Greenberger, S. Charles, C. Chen, L. Zhao, and Q. Dong, *Emotion experience and regulation in china and the united states: how do culture and gender shape emotion responding?*, *International Journal of Psychology* **47** (2012), no. 3 230–239. 106, 110

[202] A. H. Fischer, P. M. Rodriguez Mosquera, A. E. Van Vianen, and A. S. Manstead, *Gender and culture differences in emotion.*, *Emotion* **4** (2004), no. 1 87. 106, 110

[203] L. R. Brody, *Gender and emotion: Beyond stereotypes*, *Journal of Social issues* **53** (1997), no. 2 369–393. 106, 110

[204] R. L. Mitchell, R. A. Kingston, and S. L. Barbosa Bouças, *The specificity of age-related decline in interpretation of emotion cues from prosody.*, *Psychology and aging* **26** (2011), no. 2 406. 106, 110, 111

[205] M. Riediger, M. C. Voelkle, N. C. Ebner, and U. Lindenberger, *Beyond "happy, angry, or sad?": Age-of-poser and age-of-rater effects on multi-dimensional emotion perception*, *Cognition & emotion* **25** (2011), no. 6 968–982. 106, 111

[206] M. Adachi, S. E. Trehub, and J.-I. Abe, *Perceiving emotion in children's songs across age and culture 1*, *Japanese Psychological Research* **46** (2004), no. 4 322–336. 106, 110

[207] C. Kim-Prieto and E. Diener, *Religion as a source of variation in the experience of positive and negative emotions*, *The Journal of Positive Psychology* **4** (2009), no. 6 447–460. 106

[208] P. Łowicki, M. Zajenkowski, and P. Van Cappellen, *It's the heart that matters: The relationships among cognitive mentalizing ability, emotional empathy, and religiosity*, *Personality and Individual Differences* **161** (2020) 109976. 106

[209] E. Kim, D. Bryant, D. Srikanth, and A. Howard, *Age bias in emotion detection: an analysis of facial emotion recognition performance on young, middle-aged, and older adults*, in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 638–644, 2021. 106

[210] K. Yang, K. Qinami, L. Fei-Fei, J. Deng, and O. Russakovsky, *Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy*, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 547–558, 2020. 106

[211] T. Xu, J. White, S. Kalkan, and H. Gunes, *Investigating bias and fairness in facial expression recognition*, in *European Conference on Computer Vision*, pp. 506–523, Springer, 2020. 106

[212] V. Prabhakaran, A. M. Davani, and M. Díaz, *On releasing annotator-level labels and information in datasets*, *arXiv preprint arXiv:2110.05699* (2021). 106, 112, 134

[213] C. Hube, B. Fetahu, and U. Gadiraju, *Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments*, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019. 106, 112, 131, 135

[214] H. Al Kuwatly, M. Wich, and G. Groh, *Identifying and measuring annotator bias based on annotators' demographic characteristics*, in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 184–190, 2020. 106, 112, 132

[215] C. Clark, M. Yatskar, and L. Zettlemoyer, *Don't take the easy way out: Ensemble based methods for avoiding known dataset biases*, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), pp. 4067–4080, Association for Computational Linguistics, 2019. 107

[216] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, *Improving fairness in machine learning systems: What do industry practitioners need?*, in *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–16, 2019. 107

[217] J. N. Yan, Z. Gu, H. Lin, and J. M. Rzeszotarski, *Silva: Interactively assessing machine learning fairness using causality*, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020. 107

[218] R. Benjamin, *Race after technology: Abolitionist tools for the new jim code*, *Social forces* (2019). 107

[219] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, *Contextual inter-modal attention for multi-modal sentiment analysis*, in *proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3454–3466, 2018. 109

[220] Y. Ding, A. Rich, M. Wang, N. Stier, P. Sen, M. Turk, and T. Höllerer, *Sparse fusion for multimodal transformers*, *arXiv preprint arXiv:2111.11992* (2021). 109, 114

[221] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, *Integrating multimodal information in large pretrained transformers*, in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020, p. 2359, NIH Public Access, 2020. 109, 113, 114, 128

[222] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, *Xlnet: Generalized autoregressive pretraining for language understanding*, *Advances in neural information processing systems* **32** (2019). 109, 128

[223] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. W. Schuller, *et. al.*, *Sewa db: A rich database for audio-visual emotion and sentiment research in the wild*, *IEEE transactions on pattern analysis and machine intelligence* (2019). 110, 111

[224] J. Posner, J. A. Russell, and B. S. Peterson, *The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology*, *Development and psychopathology* **17** (2005), no. 3 715–734. 110

[225] S. E. Snodgrass, *Women's intuition: The effect of subordinate role on interpersonal sensitivity.*, *Journal of Personality and Social Psychology* **49** (1985), no. 1 146. 110

[226] J. Quoidbach, E. W. Dunn, K. V. Petrides, and M. Mikolajczak, *Money giveth, money taketh away: The dual effect of wealth on happiness*, *Psychological science* **21** (2010), no. 6 759–763. 110

[227] D. Keltner, D. H. Gruenfeld, and C. Anderson, *Power, approach, and inhibition.*, *Psychological review* **110** (2003), no. 2 265. 110

[228] T. Ruffman, J. D. Henry, V. Livingstone, and L. H. Phillips, *A meta-analytic review of emotion recognition and aging: Implications for neuropsychological models of aging*, *Neuroscience & Biobehavioral Reviews* **32** (2008), no. 4 863–881. 111

[229] U. Gadiraju, A. Checco, N. Gupta, and G. Demartini, *Modus operandi of crowd workers: The invisible role of microtask work environments*, *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* **1** (2017), no. 3 1–29. 111

[230] U. Gadiraju, B. Fetahu, R. Kawase, P. Siehndel, and S. Dietze, *Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks*, *ACM Transactions on Computer-Human Interaction (TOCHI)* **24** (2017), no. 4 1–26. 111

[231] D. Karger, S. Oh, and D. Shah, *Iterative learning for reliable crowdsourcing systems*, *Advances in neural information processing systems* **24** (2011). 111

[232] F. L. Wauthier and M. Jordan, *Bayesian bias mitigation for crowdsourcing*, *Advances in neural information processing systems* **24** (2011) 1800–1808. 111

[233] A. D. Shaw, J. J. Horton, and D. L. Chen, *Designing incentives for inexpert human raters*, in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 275–284, 2011. 111

[234] J. J. Y. Chung, J. Y. Song, S. Kutty, S. Hong, J. Kim, and W. S. Lasecki, *Efficient elicitation approaches to estimate collective crowd answers*, *Proceedings of the ACM on Human-Computer Interaction* **3** (2019), no. CSCW 1–25. 112

[235] Q. Z. Chen, D. S. Weld, and A. X. Zhang, *Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty*, *Proceedings of the ACM on Human-Computer Interaction* **5** (2021), no. CSCW2 1–25. 112

[236] M. Geva, Y. Goldberg, and J. Berant, *Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets*, *arXiv preprint arXiv:1908.07898* (2019). 112, 132

[237] M. Wich, H. Al Kuwatly, and G. Groh, *Investigating annotator bias with a graph-based approach*, in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 191–199, 2020. 112, 132

[238] S. Larimore, I. Kennedy, B. Haskett, and A. Arseniev-Koehler, *Reconsidering annotator disagreement about racist language: Noise or signal?*, in *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pp. 81–90, 2021. 112, 132

[239] B. Kennedy, M. Atari, A. M. Davani, L. Yeh, A. Omrani, Y. Kim, K. Coombs, S. Havaldar, G. Portillo-Wightman, E. Gonzalez, *et. al.*, *The gab hate corpus: A collection of 27k posts annotated for hate speech*, . 112

[240] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith, *Annotators with attitudes: How annotator beliefs and identities bias toxic language detection*, *arXiv preprint arXiv:2111.07997* (2021). 112

[241] M. Díaz, I. Johnson, A. Lazar, A. M. Piper, and D. Gergle, *Addressing age-related bias in sentiment analysis*, in *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14, 2018. 112

[242] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, *GoEmotions: A dataset of fine-grained emotions*, . 112

[243] M. M. Bhuiyan, A. X. Zhang, C. M. Sehat, and T. Mitra, *Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria*, *Proceedings of the ACM on Human-Computer Interaction* **4** (2020), no. CSCW2 1–26. 112

[244] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein, *The disagreement deconvolution: Bringing machine learning performance metrics in line with reality*, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021. 112, 131

[245] P. P. Liang, R. Salakhutdinov, and L.-P. Morency, *Computational modeling of human multimodal language: The mosei dataset and interpretable dynamic fusion*, in *First Workshop and Grand Challenge on Computational Modeling of Human Multimodal Language*, 2018. 113, 116, 125

[246] W. Mason and D. J. Watts, *Financial incentives and the" performance of crowds"*, in *Proceedings of the ACM SIGKDD workshop on human computation*, pp. 77–85, 2009. 117

[247] D. Difallah, E. Filatova, and P. Ipeirotis, *Demographics and dynamics of mechanical turk workers*, in *Proceedings of the eleventh ACM international conference on web search and data mining*, pp. 135–143, 2018. 119, 120

[248] S. Uzor, J. T. Jacques, J. J. Dudley, and P. O. Kristensson, *Investigating the accessibility of crowdwork tasks on mechanical turk*, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021. 120

[249] J. Howe *et. al.*, *The rise of crowdsourcing*, *Wired magazine* **14** (2006), no. 6 1–4. 120

[250] J. Feyrer, *Demographics and productivity*, *The Review of Economics and Statistics* **89** (2007), no. 1 100–109. 120

[251] B. Winter, *Linear models and linear mixed effects models in r with linguistic applications*, *arXiv preprint arXiv:1308.5499* (2013). 122

[252] S. B. Algoe, B. N. Buswell, and J. D. DeLamater, *Gender and job status as contextual cues for the interpretation of facial expression of emotion*, *Sex roles* **42** (2000), no. 3 183–208. 125

[253] J. Condry and S. Condry, *Sex differences: A study of the eye of the beholder*, *Child development* (1976) 812–819. 125

[254] M. D. Robinson, J. T. Johnson, and S. A. Shields, *The gender heuristic and the database: Factors affecting the perception of gender-related differences in the experience and display of emotions*, *Basic and Applied Social Psychology* **20** (1998), no. 3 206–219. 125

[255] S. C. Widen and J. A. Russell, *Gender and preschoolers' perception of emotion*, *Merrill-Palmer Quarterly (1982-)* (2002) 248–262. 125

[256] E. A. Plant, K. C. Kling, and G. L. Smith, *The influence of gender and social role on the interpretation of facial expressions*, *Sex roles* **51** (2004), no. 3 187–196. 125

[257] K. Saha, A. Yousuf, L. Hickman, P. Gupta, L. Tay, and M. De Choudhury, *A social media study on demographic differences in perceived job satisfaction*, *Proceedings of the ACM on Human-Computer Interaction* **5** (2021), no. CSCW1 1–29. 125

[258] M. Schaekermann, J. Goh, K. Larson, and E. Law, *Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work*, *Proceedings of the ACM on Human-Computer Interaction* **2** (2018), no. CSCW 1–19. 126

[259] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, *arXiv preprint arXiv:1810.04805* (2018). 128

[260] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, *Transformer-xl: Attentive language models beyond a fixed-length context*, *arXiv preprint arXiv:1901.02860* (2019). 128