UNIVERSITY OF CALIFORNIA
RIVERSIDE

Learning Robust Visual-Semantic Retrieval Models with Limited Supervision

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

Doctor of Philosophy

in

Electrical Engineering

by

Niluthpol Chowdhury Mithun

June 2019

Dissertation Committee:

Dr. Amit K. Roy-Chowdhury, Chairperson
Dr. Ertem Tuncel
Dr. Evangelos Papalexakis

The Dissertation of Niluthpol Chowdhury Mithun is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

The work presented in this thesis would not have been possible without the inspiration and support of a number of wonderful individuals. I express my sincere gratitude to all of them for being part of this amazing journey and making this Ph.D. thesis possible.

First, I am deeply indebted to my advisor Dr. Amit K. Roy-Chowdhury for his fundamental role in my doctoral work, grooming me as a researcher, and supporting me through all the stumbles in between. During our course of interaction in the last five years, I have learned extensively from him, including how to look at problems holistically, how to regard an existing question from a new perspective, how to rectify things that could create challenges in formulating and solving research problems, and how to approach a problem by systematic thinking. When I felt interested to venture into new research problems, he gave me the necessary freedom, at the same time continuing to contribute valuable feedback and encouragement. I feel extremely lucky to have been a part of his group.

I would also like to express my heartfelt gratitude to my dissertation committee members, Dr. Ertem Tuncel, and Dr. Evangelos Papalexakis for giving me valuable feedback and constructive comments in improving the quality of this dissertation. Dr. Tuncel is an excellent teacher and I learned a lot from his courses. Working with Dr. Papalexakis have been invaluable experiences for me, and I cannot thank him enough for that. His guidance and motivation helped me develop a major part of my thesis as provided in Chapter 4.

Special thanks are reserved for my masters' advisor Dr. S M Mahbubur Rahman from Bangladesh University of Engineering and Technology for nurturing me as a researcher as an undergraduate, and instilling in me the curiosity to pursue a Ph.D. I also owe a lot

to all my internship collaborators Dr. Sirajum Munir, Juncheng Li from Bosch, and Dr. Han-Pang Chiu, Dr. Karan Sikka from SRI, for their support and encouragement. I would like to thank them for helping me broaden my horizons.

Completing this work would have been all the more difficult were it not for the support and friendship provided by the other members of the Video Computing Group at UC Riverside. Rameswar Panda has simultaneously been a friend, mentor, and co-author, and I am grateful for him helping me choosing research directions and ideas to pursue. I would like to thank Sujoy Paul for sharing valuable insights, and feedback over the years. I also convey my special thanks to Mahmudul Hasan, Jawadul Hasan, Tahmida Mahmud, Akash Gupta, Sudipta Paul, Ghazal Mazaheri, Cody Simons, Miraj Ahmed, Abhishek Aich, and Dripta Raychaudhury for the long intellectual discussions we had in the lab.

I would like to thank the NSF, IARPA, ONR and Volkswagen for their grants to Dr. Roy-Chowdhury, which partially supported my research. I thank Victor Hill for setting up the computing infrastructure used in most of the works presented in this thesis.

I am grateful to family, friends, and acquaintances who remembered me in their prayers for the ultimate success. I owe my deepest gratitude towards my father Nishith Chowdhury and my mother Sudipta Dhar for constantly supporting me in every possible way so that I only pay attention to the studies and achieving my objective without any obstacle on the way. I am also grateful to my loving younger brother Niladree Chowdhury for love and encouragement.

I would also like to thank my father-in-law Prodyut Bhattacharjee, mother-in-law late Dipa Bhattacharjee, sister-in-law Porna Bhattacharjee and brother-in-law Subrata

Bhowmik for their love and constant support.

And finally to my better half Suborna Bhattacharjee, for everything. Her eternal support, love, patience, and understanding of my goals and aspirations has always been my greatest strength.

Acknowledgment of previously published materials: The text of this dissertation, in part or in full, is a reprint of the material as appeared in four previously published papers that I first authored. The co-author Dr. Amit K. Roy-Chowdhury, listed in all four publications, directed and supervised the research which forms the basis for this dissertation. The papers are as follows.

1. Niluthpol Chowdhury Mithun, Sujoy Paul, Amit K. Roy-Chowdhury, "Weakly Supervised Video Moment Retrieval from Text Queries", IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019.

2. Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, Amit K Roy-Chowdhury, "Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval", ACM International Conference on Multimedia Retrieval (ICMR), 2018.

3. Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos Papalexakis, Amit K Roy-Chowdhury, "Webly Supervised Joint Embedding for Cross-Modal Image-Text Retrieval", ACM International Conference on Multimedia (ACM MM) 2018.

4. Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, Amit K Roy-Chowdhury, "Joint Embedding with Multimodal Cues for Video-Text Retrieval", International Journal of Multimedia Information Retrieval (IJMIR), 2019.

To my parents, and my better-half for their unbounded love and support.

ABSTRACT OF THE DISSERTATION

Learning Robust Visual-Semantic Retrieval Models with Limited Supervision

by

Niluthpol Chowdhury Mithun

Doctor of Philosophy, Graduate Program in Electrical Engineering
University of California, Riverside, June 2019
Dr. Amit K. Roy-Chowdhury, Chairperson

In recent years, tremendous success has been achieved in many computer vision tasks using deep learning models trained on large hand-labeled image datasets. In many applications, this may be impractical or infeasible, either because of the non-availability of large datasets or the amount of time and resource needed for labeling. In this respect, an increasingly important problem in the field of computer vision, multimedia and machine learning is how to learn useful models for tasks where labeled data is sparse. In this thesis, we focus on learning comprehensive joint representations for different cross-modal visual-textual retrieval tasks leveraging weak supervision, that is noisier and/or less precise but cheaper and/or more efficient to collect.

Cross-modal visual-textual retrieval has gained considerable momentum in recent years due to the promise of deep neural network models in learning robust aligned representations across modalities. However, the difficulty in collecting aligned pairs of visual data and natural language description and limited availability such pairs in existing datasets makes it extremely difficult to train effective models, which would generalize well to uncon-

trolled scenarios as they are heavily reliant on large volumes of training data that closely mimic what is expected in the test cases. In this regard, we first present our work on developing a multi-faceted joint embedding framework-based video to text retrieval system that utilizes multi-modal cues (e.g., objects, action, place, sound) from videos to reduce the effect of limited data. Then, we describe our approach on training text to video moment retrieval systems leveraging only video-level text descriptions without any temporal boundary annotations. Next, we present our work on learning powerful joint representations of images and text from small fully annotated datasets with supervision from weakly-annotated web images. Extensive experimentation on different benchmark datasets demonstrates that our approaches show substantially better performance compared to baselines and state-of-the-art alternative approaches.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Challenges

Cross-modal retrieval of visual data using natural language description has attracted intense attention in recent years [154, 57, 148, 149, 96, 25, 129, 100], but remains a very challenging problem [154, 28, 90] due to the gap and ambiguity between modalities. The majority of the success in different visual-semantic retrieval tasks (e.g., image to text retrieval, video to text retrieval, text to video moment retrieval) has been achieved by the joint embedding models trained in a supervised way using vision-language pairs from hand-labeled datasets. Although, these datasets cover a significant number of labeled pairs, creating a large-scale dataset by collecting such pairs is extremely difficult and labor-intensive [68]. Moreover, it is generally feasible to have only a limited number of users to annotate training data, which may lead to a biased model [134, 49, 156].

Availability of limited labeled vision-language pairs in datasets makes it extremely difficult to develop comprehensive systems by training deep neural network models for most

cross-modal visual-semantic retrieval tasks. Hence, although trained models on existing vision-language datasets show good performance on benchmark datasets, applying such models in an open-world setting is unlikely to show satisfactory cross-dataset generalization (training on a dataset, testing on a different dataset) performance. The process of developing robust algorithms with a limited degree of supervision is non-trivial and has been hardly explored for the problem of cross-modal retrieval between textual and visual queries. In this regard, we study three challenging cross-modal vision-language retrieval tasks and describe our works focusing on developing efficient solutions with limited supervision leveraging incidental signals or weak labels that is less precise but less costly to collect.

1.2 Contributions

Joint embeddings have been widely used in multimedia data mining as they enable us to integrate the understanding of different modalities together. These embedding models are usually learned by mapping inputs from two or more distinct domains (e.g., images and text) into a common latent space, where the transformed vectors of semantically associated inputs should be close. Learning an appropriate embedding is crucial for achieving high-performance in many multimedia applications involving multiple modalities. The second chapter focuses on learning effective joint embedding models for the video-text retrieval task. Most existing approaches for video-text retrieval are very similar to the image-text retrieval methods by design and we observe that simple adaptation of a state-of-the-art image-text embedding methods [28] shows better result than most existing video-text retrieval approaches [25, 99]. However, such methods ignore lots of contextual information in

video sequences such as temporal activities or audio. Hence, they often fail to retrieve the most relevant information to understand important questions for efficient matching. While developing a system without considering most available cues in the video content is unlikely to be comprehensive, an inappropriate fusion of complementary cues could adversely increase ambiguity and degrade performance. Moreover, existing hand labeled video-text datasets are very small which makes it extremely difficult to train deep neural network models to understand videos in general to develop a successful video-text retrieval system. To lessen the effect of such cases, we analyze how to judiciously utilize different available cues from videos effectively for efficient retrieval.

The text to video moment retrieval task is more challenging than the task of localizing activities in videos, which is a comparatively well-studied field [83, 143, 157, 147, 104, 123]. Recent activity localization approaches show success, but these methods are limited to a pre-defined set of activity classes. In this regard, there has been a recent interest in localizing moments in a video from natural language description [44, 34, 148, 18]. Supervision in terms of labeled text description related to parts of the video is used to train these models. However, these supervised approaches are plagued by the issue of collecting human-annotated text descriptions of the videos along with the temporal extensions of the moments corresponding to each of the descriptions of a video. Moreover, it is often difficult to mark the start and end locations of a certain moment, which introduces ambiguity in the training data. On the other hand, it is often much easier to describe the moments appearing in a video in natural language than providing exact temporal boundaries associated with each of the descriptions. Moreover, such descriptions can often be obtained easily from cap-

tions through some sources in the web. Motivated by this, we pose a question in the third chapter: *Is it possible to develop a weakly-supervised framework for video moment localization from the text, leveraging only video-level textual annotation?* Temporal localization using weak description is a much more challenging task than the supervised approaches. However, this is extremely relevant to address due to the difficulty and non-scalability of acquiring a precise frame-wise information with text descriptions which requires enormous amount of manual labor.

We also study how to utilize web images in training comprehensive joint embedding models from small clean datasets for image-text retrieval. Although existing datasets contain limited labeled image-text pairs, streams of images with noisy tags are readily available in datasets, such as Flickr-1M [54], as well as in nearly infinite numbers on the web. Developing a practical system for image-text retrieval considering a large number of web images is more likely to be robust. However, inefficient utilization of weakly-annotated images may increase ambiguity and degrade performance. Motivated by this observation, we pose an important question in the fourth chapter: *Can a large number of web images with noisy annotations be leveraged upon with a fully annotated dataset of images with textual descriptions to learn better joint embedding models?*. This is an extremely relevant problem to address due to the difficulty and non-scalability of obtaining a large amount of human-annotated training set of image-text pairs.

Main Contributions. We address three novel and practical cross-modal visual-semantic retrieval problems in this thesis as follows.

- First, how to develop a robust video-text retrieval system by utilizing multiple

salient cues from videos (different visual features and audio inputs) to deal with the issue of limited video-text pairs in existing datasets.

• Second, how to temporally localize video moments from text queries without requiring human-crafted training data consisting of videos with text-based localization of moments; rather, we achieve the same with video-level descriptions only.

• Third, how to exploit large scale web data and associated tags for learning more effective multi-modal joint embedding models without requiring a large amount of human-crafted training data.

Towards solving these problems, we develop novel frameworks that show clear performance improvement over state-of-the-art methods and baselines in the tasks.

1.3 Organization of the Thesis

We organize the rest of the thesis as follows. In Chapter 2, we present our work on developing a multi-faceted joint embedding framework for effective video-text retrieval that utilizes multiple salient cues from videos to deal with the issue of limited number of pairs in existing video-text datasets. In Chapter 3, we propose a weakly-supervised framework for text to video moment retrieval trained utilizing only video-level text descriptions without any temporal boundary annotations of the moments. In Chapter 4, we study how to leverage supervision from web images and associated tags in training robust joint embedding models for image-text retrieval from small fully annotated datasets. We conclude the thesis in Chapter 5 with concluding remarks and some future research directions.

Chapter 2

Joint Visual-Semantic Embedding for Video-Text Retrieval

2.1 Introduction

The goal of this work is to retrieve the correlated text description given a random video, and vice versa, to retrieve the matching videos provided with text descriptions (See Fig. 2.1). While several computer vision tasks (e.g., image classification [43, 94, 51], object detection [113, 112, 92]) are now reaching maturity, cross-modal retrieval between visual data and natural language description remains a very challenging problem [154, 90] due to the gap and ambiguity between different modalities and availability of limited training data. Some recent works [93, 66, 145, 57, 32] attempt to utilize cross-modal joint embeddings to address the gap. By projecting data from multiple modalities into the same joint space, the similarity of the resulting points would reflect the semantic closeness between their

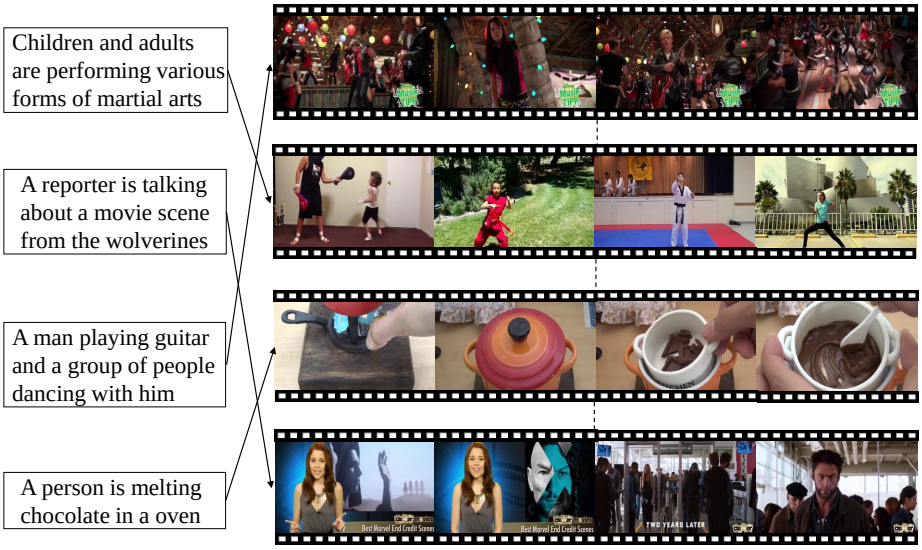

Figure 2.1: Illustration of Video-Text retrieval task: given a text query, retrieve and rank videos from the database based on how well they depict the text, and vice versa.

corresponding original inputs. In this work, we focus on learning joint video-text embedding models and combining video cues for different purposes effectively for developing robust video-text retrieval system.

The video-text retrieval task is one step further than the image-text retrieval task, which is a comparatively well-studied field. Most existing approaches for video-text retrieval are very similar to the image-text retrieval methods by design and focus mainly on the modification of loss functions [25, 150, 129, 99, 100]. We observe that simple adaptation of a state-of-the-art image-text embedding method [28] by mean-pooling features from video frames generates a better result than existing video-text retrieval approaches [25, 99]. However, such methods ignore lots of contextual information in video sequences such as temporal activities or specific scene entities, and thus they often can only retrieve some generic responses related to the appearance of static frame. They may fail to retrieve the most relevant information in many cases to understand important questions for efficient

Figure 2.2: Sample frame from two videos and associated caption to illustrate the significance of utilizing supplementary cues from videos to improve the chance of correct retrieval.

retrieval such as ‘What happened in the video’, or ‘Where did the video take place’. This greatly undermines the robustness of the systems; for instance, it is very difficult to distinguish a video with the caption “a dog is barking” apart from another “a dog is playing” based only on visual appearance (See Fig. 2.2). Associating video motion content and the environmental scene can give supplementary cues in this scenario and improve the chance of correct prediction. Similarly, to understand a video described by “gunshot broke out at the concert” may require analysis of different visual (e.g., appearance, motion, environment) and audio cues simultaneously. On the other hand, a lot of videos may contain redundant or identical contents, and hence, an efficient video-text retrieval should utilize the most distinct cues in the content to resolve ambiguities in retrieval.

While developing a system without considering most available cues in the video content is unlikely to be comprehensive, an inappropriate fusion of complementary features could adversely increase ambiguity and degrade performance. Additionally, existing hand labeled video-text datasets are very small and very restrictive considering the amount of rich descriptions that a human can compose and the enormous amount of diversity in the visual world. This makes it extremely difficult to train deep models to understand videos

in general to develop a successful video-text retrieval system. To ameliorate such cases, we analyze how to judiciously utilize different cues from videos. We propose a mixture of experts system, which is tailored towards achieving high performance in the task of cross-modal video-text retrieval. We believe focusing on three major facets (i.e., concepts for Who, What, and Where) from videos is crucial for efficient retrieval performance. In this regard, our framework utilizes three salient features (i.e., object, action, place) from videos (extracted using pre-trained deep neural networks) for learning joint video-text embeddings and uses an ensemble approach to fuse them. Furthermore, we propose a modified pairwise ranking loss for the task that emphasizes on hard negatives and relative ranking of positive labels. Our approach shows significant performance improvement compared to previous approaches and baselines.

Contributions: The main contributions of this work can be summarized as follows.

• The success of video-text retrieval depends on more robust video understanding. In this chapter, we study how to achieve the goal by utilizing multimodal features from a video (different visual features and audio inputs). Our proposed framework uses action, object, place, text and audio features by a fusion strategy for efficient retrieval.

• We present a modified pairwise loss function to better learn the joint embedding which emphasizes on hard negatives and applies a weight-based penalty on the loss based on the relative ranking of the correct match in the retrieval.

• We conduct extensive experiments and demonstrate a clear improvement over the state-of-the-art methods in the video to text retrieval tasks on the MSR-VTT dataset [149] and MSVD dataset [17].

2.2 Related Work

Image-Text Retrieval. Recently, there has been significant interest in learning robust visual-semantic embeddings for image-text retrieval [93, 58, 45, 141]. Based on a triplet of object, action and, scene, a method for projecting text and image to a joint space was proposed in early work [29]. Canonical Correlation Analysis (CCA) and several extensions of it have been used in many previous works for learning joint embeddings for the cross-modal retrieval task [124, 47, 38, 151, 109, 41] which focuses on maximizing the correlation between the projections of the modalities. In [38], authors extended classic two-view CCA approach with a third view coming from high-level semantics and proposed an unsupervised way to derive the third view from clustering the tags. In [109], authors proposed a method named MACC (Multimedia Aggregated Correlated Components) aiming to reduce the gap between cross-modal data in the joint space by embedding visual and textual features into a local context that reflects the data distribution in the joint space. Extension of CCA with deep neural networks named deep CCA (DCCA) has also been utilized to learn joint embeddings [151, 2], which focus on learning two deep neural networks simultaneously to project two views that are maximally correlated. While CCA-based methods are popular, these methods have been reported to be unstable and incur a high memory cost due to the covariance matrix calculation with large-amount of data [144, 84]. Recently, there are also several works leveraging adversarial learning to train joint image-text embeddings for cross-modal retrieval [141, 21].

Most recent works relating to text and image modality are trained with ranking loss [64, 32, 144, 28, 96, 136]. In [32], authors proposed a method for projecting words and

visual content to a joint space utilizing ranking loss that applies a penalty when a non-matching word is ranked higher than the matching one. A cross-modal image-text retrieval method has been presented in [64] that utilizes triplet ranking loss to project image feature and RNN based sentence description to a common latent space. Several image-text retrieval methods have adopted a similar approach with slight modifications in input feature representations [96], similarity score calculation [144], or loss function [28]. VSEPP model [28] modified the pair-wise ranking loss based on violations caused by the hard-negatives (i.e., non-matching query closest to each training query) and has been shown to be effective in the retrieval task. For image-sentence matching, a LSTM based network is presented in [52] that recurrently selects pairwise instances from image and sentence descriptions, and aggregates local similarity. In [96], authors proposed a multimodal attention mechanism to attend to sentence fragments and image regions selectively for similarity calculation. Our method complements these works that learn joint image-text embedding using a ranking loss (e.g., [64, 136, 28]). The proposed retrieval framework can be applied to most of these approaches for improved video-text retrieval performance.

Video Hyperlinking. Video hyperlinking is also closely relevant to our work. Given an anchor video segment, the task is to focus on retrieving and ranking a list of target videos based on the likelihood of being relevant to the content of the anchor [3, 10]. Multimodal representations have been utilized widely in video hyperlinking approaches in recent years [13, 140, 3]. Most of these approaches rely heavily on multimodal autoencoders for jointly embedding multimodal data [139, 30, 15]. Bidirectional deep neural network (BiDNN) based representations have also been shown to be very effective in video hyper-

linking benchmarks [140, 138]. BiDNN is also a variation of multimodal autoencoder, which performs multimodal fusion using a cross-modal translation with two interlocked deep neural networks [139, 138]. Considering the input data, video-text retrieval is dealing with the same multimodal input as video hyperlinking in many cases. However, video-text retrieval task is more challenging than hyperlinking since it requires to distinctively retrieve matching data from a different modality, which requires effective utilization of the correlations in between cross-modal cues.

Video-Text Retrieval. Most relevant to our work are the methods that relate video and language modalities. Two major tasks in computer vision related to connecting these two modalities are video-text retrieval and video captioning. In this work, we only focus on the retrieval task. Similar to image-text retrieval approaches, most video-text retrieval methods employ a shared subspace. In [150], authors vectorize each subject-verb-object triplet extracted from a given sentence by word2vec model [88] and then aggregate the Subject, Verb, Object (SVO) vector into a sentence level vector using RNN. The video feature vector is obtained by mean pooling over frame-level features. Then a joint embedding is trained using a least squares loss to project the sentence representation and the video representation into a joint space. Web image search results of input text have been exploited by [99], which focused on word disambiguation. In [137], a stacked GRU is utilized to associate sequence of video frames to a sequence of words. In [100], authors propose an LSTM with visual-semantic embedding method that jointly minimizes a contextual loss to estimate relationships among the words in the sentence and a relevance loss to reflect the distance between video and sentence vectors in the shared space. A method named

Word2VisualVec is proposed in [25] for the video to sentence matching task that projects vectorized sentence into visual feature space using mean squared loss. A shared space across image, text and sound modality is proposed in [5] utilizing ranking loss, which can also be applied to video-text retrieval task.

Utilizing multiple characteristics of video (e.g., activities, audio, locations, time) is evidently crucial for efficient retrieval [152]. In the closely related task of video captioning, dynamic information from video along with static appearance features has been shown to be very effective [155, 111]. However, most of the existing video-text retrieval approaches depend on one visual cue for retrieval. In contrast to the existing works, our approach focuses on effectively utilizing different visual cues and audio (if available) concurrently.

Ensemble Approaches. Our retrieval system is based on an ensemble framework [107, 31]. A strong psychological context of the ensemble approach can be found from its intrinsic connection in decision making in many daily life situations [107]. Seeking the opinions of several experts, weighing them and combining to make an important decision is an innate behavior of human. The ensemble methods hinge on the same idea and utilize multiple models for making an optimized decision, as in our case diverse cues are available from videos and we would like to utilize multiple expert models which focus on different cues independently to obtain a stronger prediction model. Moreover, ensemble-based systems have been reported to be very useful when dealing with a lack of adequate training data [107]. As diversity of the models is crucial for the success of ensemble frameworks [108], it is important for our case to choose a diverse set of video-text embeddings that are significantly different from one another.

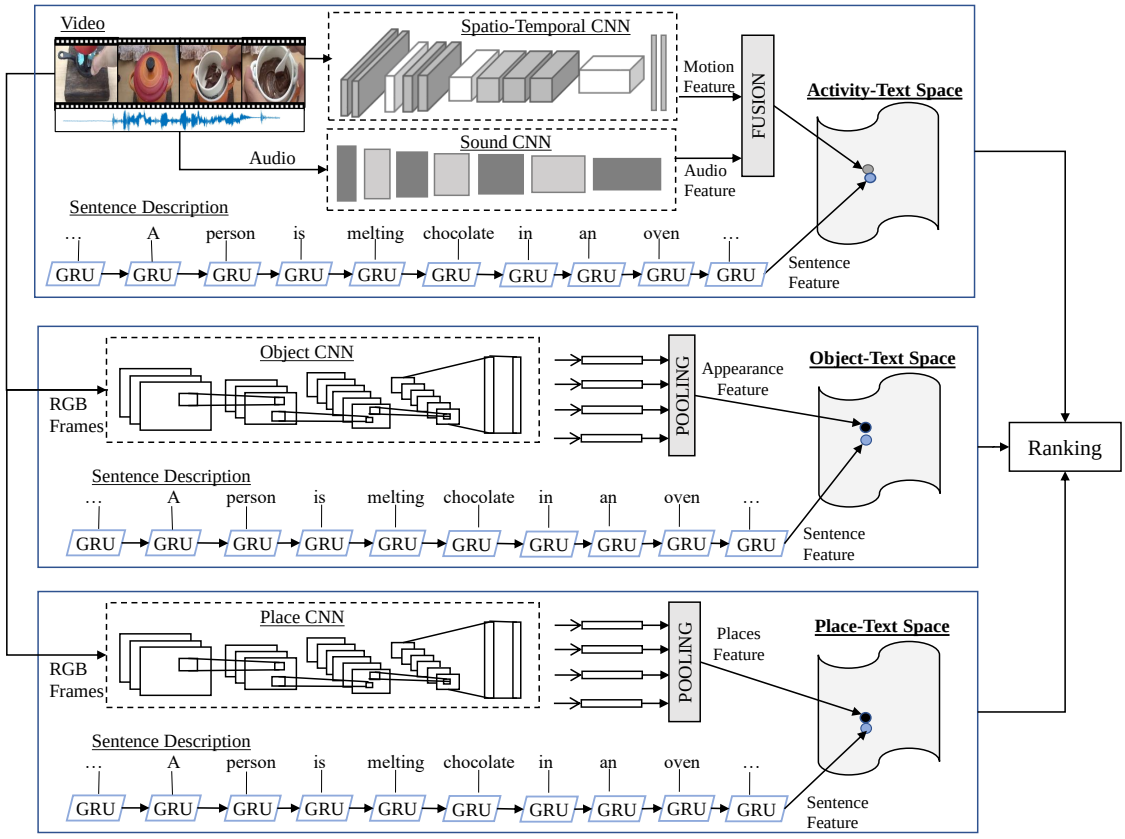

Figure 2.3: An overview of the proposed retrieval process. We propose to learn three joint video-text embedding networks as shown in the figure. Given a query sentence, we calculate the sentence's similarity scores with each one of the videos in the entire dataset in all of the three embedding spaces and use a fusion of scores for the final retrieval result.

2.3 Approach

In this section, we first provide an overview of our proposed framework (Section 2.3.1). Then, we describe the input feature representation for video and text (Section 2.3.2). Next, we describe the basic framework for learning visual-semantic embedding using pair-wise ranking loss (Section 2.3.3). After that, we present our modification on the loss function which improves the basic framework to achieve better recall (Section 2.3.4). Finally, we present the proposed fusion step for video-text matching (Section 2.3.5).

2.3.1 Overview of the Proposed Approach

In a typical cross-modal video-text retrieval system, an embedding network is learned to project video features and text features into the same joint space, and then retrieval is performed by searching the nearest neighbor in the latent space. Since in this work we are looking at videos in general, detecting most relevant information such as object, activities, and places could be very conducive for higher performance. Therefore, along with developing algorithms to train better joint visual-semantic embedding models, it is also very important to develop strategies to effectively utilize different available cues from videos for a more comprehensive retrieval system.

In this work, we propose to leverage the capability of neural networks to learn a deep representation first and fuse the video features in the latent spaces so that we can develop expert networks focusing on specific subtasks (e.g. detecting activities, detecting objects). For analyzing videos, we use a model trained to detect objects, a second model trained to detect activities, and a third model focusing on understanding the place. These heterogeneous features may not be used together directly by simple concatenation to train a successful video-text model as intra-modal characteristics are likely to be suppressed in such an approach. However, an ensemble of video-text models can be used, where a video-text embedding is trained on each of the video features independently. The final retrieval is performed by combining the individual decisions of several experts [107]. An overview of our proposed retrieval framework is shown in Fig. 2.3. In Fig. 2.3, Object-Text space is the expert in solving ambiguity related to who is in the video, whereas Activity-Text space is the expert in retrieving what activity is happening and place-Text space is the expert

in solving ambiguity regarding locations in the video. We believe that such an ensemble approach will significantly reduce the chance of poor/wrong prediction.

We follow network architecture proposed in [64] that learns the embedding model using a two-branch network using image-text pairs. One of the branches in this network takes text feature as input and the other branch takes in a video feature. We propose a modified bi-directional pairwise ranking loss to train the embedding. Inspired by the success of ranking loss proposed in [28] in image-text retrieval task, we emphasize on hard negatives. We also apply a weight-based penalty on the loss according to the relative ranking of the correct match in the retrieved result.

2.3.2 Input Feature Representation

Text Feature. For encoding sentences, we use Gated Recurrent Units (GRU) [23]. We set the dimensionality of the joint embedding space, D, to 1024. The dimension of the word embeddings that are input to the GRU is 300. Note that the word embedding model and the GRU are trained end-to-end in this work.

Object Feature. For encoding image appearance, we adopt deep pre-trained convolutional neural network (CNN) model trained on ImageNet as the encoder. Specifically, we utilize state-of-the-art 152 layer ResNet model ResNet152 [43]. We extract image features directly from the penultimate fully connected layer. We first rescale the image to 224x224 and feed into CNN as inputs. The dimension of the image embedding is 2048.

Activity Feature. The ResNet CNN can efficiently capture visual concepts in static frames. However, an effective approach to learning temporal dynamics in videos was proposed by inflating a 2-D CNN to a deep 3-D CNN named I3D in [14]. We use I3D model

to encode activities in videos. In this work, we utilize the pre-trained RGB-I3D model and extract 1024 dimensional feature utilizing continuous 16 frames of video as the input.

Place Feature. For encoding video feature focusing on scene/place, we utilize deep pre-trained CNN model trained on Places-365 dataset as the encoder [159]. Specifically, we utilize 50 layer model ResNet50 [43]. We extract image features directly from the penultimate fully connected layer. We re-scale the image to 224x224 and feed into CNN as inputs. The dimension of the image embedding is 2048.

Audio Feature. We believe that by associating audio, we can get important cues to the real-life events, which would help us remove ambiguity in many cases. We extract audio feature using state-of-the-art SoundNet CNN [4], which provides 1024 dimensional feature from input raw audio waveform. Note that, we only utilize the audio which is readily available with the videos.

2.3.3 Learning Joint Embedding

In this section, we describe the basic framework for learning joint embedding based on bi-directional ranking loss.

Given a video feature representation (i.e., appearance feature, or activity feature, or scene feature) $\overline{v}$ ($\overline{v} \in \mathbb{R}^V$), the projection for a video feature on the joint space can be derived as $v = W^{(v)}\overline{v}$ ($v \in \mathbb{R}^D$). In the same way, the projection of input text embedding $\overline{t}(\overline{t} \in \mathbb{R}^T)$ to joint space is $t = W^{(t)}\overline{t}(t \in \mathbb{R}^D)$. Here, $W^{(v)} \in \mathbb{R}^{D \times V}$ is the transformation matrix that projects the video content into the joint embedding space, and D denotes the dimension of the joint space. Similarly, $W^{(t)} \in \mathbb{R}^{D \times T}$ maps input sentence/caption embedding to the joint space. Given feature representation for words in a sentence, the

sentence embedding $\bar{t}$ is found from the hidden state of the GRU. Here, given the feature representation of both videos and corresponding text, the goal is to learn a joint embedding characterized by θ (i.e., $W^{(v)}$, $W^{(t)}$ and GRU weights) such that the video content and semantic content are projected into the joint embedding space. We keep image encoder (e.g., pre-trained CNN) fixed in this work, as the video-text datasets are small in size.

In the embedding space, it is expected that the similarity between a video and text pair to be more reflective of semantic closeness between videos and their corresponding texts. Many prior approaches have utilized pairwise ranking loss for learning joint embedding between visual input and textual input. They minimize a hinge based triplet ranking loss combining bi-directional ranking terms, in order to maximize the similarity between a video embedding and the corresponding text embedding, and while at the same time, minimize the similarity to all other non-matching ones. The optimization problem can be written as,

$$\min_{\theta} \sum_{v} \sum_{t^-} [\alpha - S(v,t) + S(v,t^-)]_+ \; + \; \sum_{t} \sum_{v^-} [\alpha - S(t,v) + S(t,v^-)]_+ \qquad (2.1)$$

where, $[f]_+ = max(0, f)$. t^- is a non-matching text embedding, and t is the matching text embedding for video embedding v. This is similar for text embedding t. α is the margin value for the pairwise ranking loss. The scoring function $S(v,t)$ is defined as the similarity function to measure the similarity between the videos and text in the joint embedded space. We use cosine similarity in this work, as it is easy to compute and shown to be very effective in learning joint embeddings. [64, 28].

In Eq. (2.1), in the first term, for each pair (v,t), the sum is taken over all non-matching text embedding t^-. It attempts to ensure that for each visual feature, matching text features should be closer than non-matching ones in the joint space. Similarly, the sec-

ond term attempts to ensure that text embedding that corresponds to the video embedding should be closer in the joint space to each other than non-matching video embeddings.

2.3.4 Proposed Ranking Loss

Recently, focusing on hard-negatives has been shown to be effective in many embedding tasks [28, 118, 85]. Inspired by this, we focus on hard negatives (i.e., the negative video and text sample closest to a positive/matching (v, t) pair) instead of summing over all negatives in our formulation. For a positive/matching pair (v, t), the hardest negative sample can be identified using $\hat{v} = \arg\max_{v^-} S(t, v^-)$ and $\hat{t} = \arg\max_{t^-} S(v, t^-)$. The optimization problem can be rewritten as following to focus on hard-negatives,

$$\min_{\theta} \sum_{v} [\alpha - S(v,t) + S(v,\hat{t})]_+ \; + \; \sum_{t} [\alpha - S(t,v) + S(t,\hat{v})]_+ \tag{2.2}$$

The loss in Eq. 2.2 is similar to the loss in Eq. 2.1 but it is specified in terms of the hardest negatives [28]. We start with the loss function in Eq. 2.2 and further modify the loss function following the idea of weighted ranking [133] to weigh the loss based on the relative ranking of positive labels.

$$\min_{\theta} \sum_{v} L(r_v)[\alpha - S(v,t) + S(v,\hat{t})]_+ \; + \; \sum_{t} L(r_t)[\alpha - S(t,v) + S(t,\hat{v})]_+ \tag{2.3}$$

where $L(.)$ is a weighting function for different ranks. For a video embedding v, r_v is the rank of matching sentence t among all compared sentences. Similarly, for a text embedding t, r_t is the rank of matching video embedding v among all compared videos in the batch. We define the weighting function as $L(r) = (1 + \beta/(N - r + 1))$, where N is the number of compared videos and β is the weighting factor. Fig. 2.4 shows an example showing the

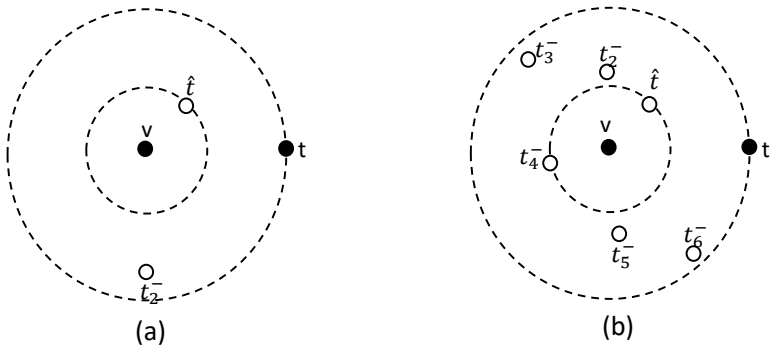

Figure 2.4: An example showing the significance of the proposed ranking loss. The idea is that if a large number of non-matching instances are ranked higher than the matching one given the current state of the model, then the model must be updated by a larger amount (Case:(b)). However, the model needs to be updated by a smaller amount if the matching instance is already ranked higher than most non-matching ones (Case:(a)) Here, the idea is illustrated with a positive/matching video-text pair $(v,\ t)$ (The cross-modal pair is shown with filled circles) and margin $\theta = 0$. For the positive pair $(v,\ t)$, the non-matching/negative examples which contributes to the loss (i.e., empty circles in the figure) are shown with t^-. $\hat{t}$ is the highest violating negative sample.

significance of the proposed ranking loss.

It is very common, in practice, to only compare samples within a mini-batch at each iteration rather than comparing the entire training set for computational efficiency [85, 118, 57]. This is known as semi-hard negative mining [85, 118]. Moreover, selecting the hardest negatives in practice may often lead to a collapsed model and semi-hard negative mining helps to mitigate this issue [85, 118]. We utilize a batch-size of 128 in our experiment.

It is evident from Eq. 2.3 that the loss applies a weight-based penalty based on the relative ranking of the correct match in retrieved result. If a positive match is ranked top in the list, then $L(.)$ will assign a small weight to the loss and will not cost the loss too much. However, if a positive match is not ranked top, $L(.)$ will assign a much larger weight to the loss, which will ultimately try to push the positive matching pair to the top of rank.

2.3.5 Matching and Ranking

The video-text retrieval task focuses on returning for each query video, a ranked list of the most likely text description from a dataset and vice versa. We believe, we need to understand three main aspects of each video: (1) Who: the salient objects of the video, (2) What: the action and events in the video and (3) Where: the place aspect of the video. To achieve this, we learn three expert joint video-text embedding spaces as shown in Fig. 2.3.

The Object-Text embedding space is the common space where both appearance features and text feature are mapped to. Hence, this space can link video and sentences focusing on the objects. On the other hand, the Activity-Text embedding space focuses on linking video and language description which emphasizes more on the events in the video. Action features and audio features both provide important cues for understanding different events in a video. We fuse action and audio features (if available) by concatenation and map the concatenated feature and text feature into a common space, namely, the Activity-Text space. If the audio feature is absent from videos, we only use the action feature as the video representation for learning the Activity-Text space. The Place-Text embedding space is the common space where visual features focusing on scene/place aspect and text feature are mapped to. Hence, this space can link video and sentences focusing on the entire scene. We utilize the same loss functions described in Sec. 2.3.4 for training these embedding models.

At the time of retrieval, given a query sentence, we compute the similarity score of the query sentence with each one of the videos in the dataset in three video-text embedding spaces and use a fusion of similarity scores for the final ranking. Conversely, given a query video, we calculate its similarity scores with all the sentences in the dataset in three

embedding spaces and use a fusion of similarity scores for the final ranking.

$$S_{v-t}(v,t) = w_1 S_{o-t} + w_2 S_{a-t} + w_3 S_{p-t} \tag{2.4}$$

It may be desired to use a weighted sum when it is necessary in a task to put more emphasis on one of the facets of the video (objects or captions or scene). In this work, we empirically found putting comparatively higher importance to S_{o-t} (Object-Text) and S_{a-t} (Activity-Text), and slightly lower importance to S_{p-t} (Place-Text) works better in evaluated datasets than putting equal importance to all. We empirically choose $w_1 = 1$, $w_2 = 1$ and $w_3 = 0.5$ in our experiments based on our evaluation on the validation set.

2.4 Experiments

In this section, we first describe the datasets and evaluation metric (Section 4.1). Then, we describe the training details. Next, we provide quantitative results on MSR-VTT dataset (Section 4.3) and MSVD dataset (Section 4.4) to show the effectiveness of our proposed framework. Finally. we present some qualitative examples analyzing our success and failure cases (Section 4.5).

2.4.1 Datasets and Evaluation Metric

We present experiments on two standard benchmark datasets: Microsoft Research Video to Text (MSR-VTT) Dataset [149] and Microsoft Video Description dataset (MSVD) [17] to evaluate the performance of our proposed framework. We adopt rank-based metric for quantitative performance evaluation.

Table 2.1: Video-to-Text and Text-to-Video Retrieval Results on MSR-VTT Dataset.

#	Method	Video-to-Text Retrieval—					Text-to-Video Retrieval				
		R@1	R@5	R@10	MedR	MeanR	R@1	R@5	R@10	MedR	MeanR
1.1	VSE (Object-Text)	7.7	20.3	31.2	28.0	185.8	5.0	16.4	24.6	47.0	215.1
	VSEPP (Object-Text)	10.2	25.4	35.1	25	228.1	5.7	17.1	24.8	65	300.8
1.2	Ours (Object-Text)	10.5	26.7	35.9	25	266.6	5.8	17.6	25.2	61	296.6
	Ours (Audio-Text)	0.4	1.1	1.9	1051	2634.9	0.2	0.9	1.5	1292	1300
	Ours (Activity-Text)	8.4	22.2	32.3	30.3	229.9	4.6	15.3	22.7	71	303.7
	Ours (Place-Text)	7.1	19.8	28.7	38	275.1	4.3	14	21.1	77	309.6
1.3	CON(Object, Activity)-Text	9.1	24.6	36	23	181.4	5.5	17.6	25.9	51	243.4
	CON(Object, Activity, Audio)-Text	9.3	27.8	38	22	162.3	5.7	18.4	26.8	48	242.5
1.4	Joint Image-Text-Audio Embedding	8.7	22.4	32.1	31	225.8	4.8	15.3	22.9	73	313.6
1.5	Fusion [Object-Text, Activity (I3D)-Text]	12.3	31.3	42.9	16	145.4	6.8	20.7	29.5	39	224.7
	Fusion [Object-Text, Activity(I3d-Audio)-Text]	12.5	32.1	42.4	16	134	7	20.9	29.7	38	213.8
	Fusion [Object-Text, Place-Text]	11.8	30.1	40.8	18	172.1	6.5	19.9	28.5	43	234.1
	Fusion [Activity-Text, Place-Text]	11	28.4	39.3	20	152.1	5.9	18.6	27.4	44	224.7
1.6	**Fusion [Object-Text, Activity-Text, Place-Text]**	**13.8**	**33.5**	**44.3**	**14**	**119.2**	**7.3**	**21.7**	**30.9**	**34**	**196.1**
1.7	Rank Fusion [Object-Text, Activity-Text, Place-Text]	12.2	31.6	42.7	16	127.6	6.8	20.5	29.4	38	204.3

MSR-VTT. The MSR-VTT is a large-scale video description dataset. This dataset contains 10,000 video clips. The dataset is split into 6513 videos for training, 2990 videos for testing and 497 videos for the validation set. Each video has 20 sentence descriptions. This is one of the largest video captioning dataset in terms of the quantity of sentences and the size of the vocabulary.

MSVD. The MSVD dataset contains 1970 Youtube clips, and each video is annotated with about 40 sentences. We use only the English descriptions. For a fair comparison, we used the same splits utilized in prior works [137], with 1200 videos for training, 100 videos for validation, and 670 videos for testing. The MSVD dataset is also used in [99] for video-text retrieval task, where they randomly chose 5 ground-truth sentences per video. We use the same setting when we compare with that approach.

Evaluation Metric. We use the standard evaluation criteria used in most prior works on image-text retrieval task [99, 64, 25]. We measure rank-based performance by $R@K$, Median Rank($MedR$) and Mean Rank ($MeanR$). $R@K$ (Recall at K) calculates the percentage of test samples for which the correct result is found in the top-K retrieved points to the query sample. We report results for $R@1$, $R@5$ and $R@10$. Median Rank calculates the median of the ground-truth results in the ranking. Similarly, Mean Rank calculates the mean rank of all correct results.

2.4.2 Training Details

We used two Titan Xp GPUs for this work. We implemented the network using PyTorch following [28]. We start training with a learning rate of 0.002 and keep the learning

rate fixed for 15 epochs. Then the learning rate is lowered by a factor of 10 and the training continued for another 15 epochs. We use a batch-size of 128 in all the experiments. The embedding networks are trained using ADAM optimizer [63]. When the L2 norm of the gradients for the entire layer exceeds 2, gradients are clipped. We tried different values for margin α in training and found $0.1 \leq \alpha \leq 0.2$ works reasonably well. We empirically choose α as 0.2. The embedding model was evaluated on the validation set after every epoch. The model with the best sum of recalls on the validation set is chosen as the final model.

2.4.3 Results on MSR-VTT Dataset

We report the result on MSR-VTT dataset [149] in Table 2.1. We implement several baselines to analyze different components of the proposed approach. To understand the effect of different loss functions, features, effect of feature concatenation and proposed fusion method, we divide the table into 7 rows (1.1-1.7). In row-1.1, we report the results on applying two different variants of pair-wise ranking loss. VSE[64] is based on the basic triplet ranking loss similar to Eq. 2.1 and VSEPP[28] is based on the loss function that emphasizes on hard-negatives as shown in Eq. 2.2. Note that, all other reported results in Table 2.1 are based on the modified pairwise ranking loss proposed in Eq. 2.3. In row-1.2, we provide the performance of different features in learning the embedding using the proposed loss. In row-1.3, we present results for the learned embedding utilizing a feature vector that is a direct concatenation of different video features. In row-1.4, we provide the result when a shared representation between image, text and audio modality is learned using proposed loss following the idea in [5] and used for video-text retrieval task. In row-1.5, we provide the result based on the proposed approach that employs two video-text joint embeddings

for retrieval. In row-1.6, we provide the result based on the proposed ensemble approach that employs all three video-text joint embeddings for retrieval. Additionally, in row-1.7, we also provide the result for the case where the rank fusion has been considered in place of the proposed score fusion.

Loss Function. For evaluating the performance of different ranking loss functions in the task, we can compare results reported in row-1.1 and row-1.2. We can choose only results based on Object-Text spaces from these two rows for a fair comparison. We see that VSEPP loss function and proposed loss function performs significantly better than the traditional VSE loss function in $R@1$, $R@5$, $R@10$. However, VSE loss function has better performance in terms of the mean rank. This phenomenon is expected based on the characteristics of the loss functions. As higher $R@1$, $R@5$ and $R@10$ are more desirable for a efficient video-text retrieval system than the mean rank, we see that our proposed loss function performs better than other loss functions in this task. We observe similar performance improvement using our loss function in other video-text spaces too.

Video Features. We can compare the performance of different video features in learning the embedding using the proposed loss from row-1.2. We observe that object feature and activity feature from video performs reasonably well in learning a joint video-text space. The performance is very low when only audio feature is used for learning the embedding. It can be expected that the natural sound associated in a video alone does not contain as much information as videos in most cases. However, utilizing audio along with i3d feature as activity features provides a slight boost in performance as shown in row-1.3 and row-1.4 of Table 2.1.

Feature Concatenation for Representing Video. Rather than training multiple video-semantic spaces, one can argue that we can simply concatenate all the available video features and learn a single video-text space using this concatenated video feature [25, 149]. However, we observe from row-1.3 that integrating complementary features by static concatenation based fusion strategy fails to utilize the full potential of different video features for the task. Comparing row-1.2 and row-1.3, we observe that a concatenation of object feature, activity feature and Audio feature performs even worse than utilizing only object feature in R@1. Although we see some improvement in other evaluation metrics, overall the improvement is very limited. We believe that both appearance and action feature gets suppressed in such concatenation as they focus on representing different entities of a video.

Learning a Shared Space across Image, Text and Audio. Learning a shared space across image, text and sound modality is proposed for cross-modal retrieval task in [5]. Following the idea, we trained a shared space across video-text-sound modality using the pairwise ranking loss by utilizing video-text and video-sound pairs. The result is reported in row-1.4. We observe that performance in video-text retrieval task degrades after training such an joint representation across 3 modalities. Training such a representation gives the flexibility to transfer across multiple modalities. Nevertheless, we believe it is not tailored towards achieving high performance in a specific task. Moreover, aligning across 3 modalities is a more computationally difficult task and requires many more examples to train.

Proposed Fusion. The best result in Table. 2.1 is achieved by our proposed fusion approach as shown in row-1.6. We see that the proposed method achieves 31.43% improvement in R@1 for text retrieval and 25.86% improvement for video retrieval in R@1

compared to best performing Ours(Object-text) as shown in row-1.2, which is the best among the other methods which use a single embedding space for the retrieval task. In row-1.5, Fusion[Object-text & Activity(I3D-Audio)-text] differs from Fusion[Object-text & Activity(I3D)-text] in the feature used in learning the activity-text space. We see that utilizing audio in learning the embedding improves the result slightly. However, as the retrieval performance of individual audio feature is very low (shown in row-1.2), we did not utilize audio-text space separately in fusion as we found it degraded the performance significantly.

Comparing row-1.6, row-1.5 and row-1.2, we find that the ensemble approach with score fusion results in significant improvement in performance, although there is no guarantee that the combination of multiple models will perform better than the individual models in the ensemble in every single case. However, the ensemble average consistently improves performance significantly.

Rank vs Similarity Score in Fusion. We provide the retrieval result based on weighted rank aggregation of three video-text spaces in row-1.7. Comparing the effect of rank fusion in replacement of the score fusion from row-1.6 and row-1.7 in Table. 2.1, it is also evident that the proposed score fusion approach shows consistent performance improvement over rank fusion approach. It is possible that exploiting similarity score to combine multiple evidences may be less effective than using rank values in some cases, as score fusion approach independently weights scores and does not consider overall performance in weighting [70]. However, we empirically find that utilizing score fusion is more advantageous than rank fusion in our system in terms of retrieval effectiveness.

Table 2.2: Video-to-Text Retrieval Results on MSVD Dataset. We highlight the proposed method. The methods which has 'Ours' keyword are trained with the proposed loss.

Method	R@1	R@5	R@10	MedR	MeanR
Results Using Partition used by JMET and JMDV					
CCA					245.3
JMET					208.5
JMDV					224.1
W2VV-ResNet152	16.3		44.8	14	110.2
VSE (Object-Text)	15.8	30.2	41.4	12	84.8
VSEPP(Object-Text)	21.2	43.4	52.2	9	79.2
Ours(Object-Text)	23.4	45.4	53	8	75.9
Ours(Activity-Text)	21.3	43.7	53.3	9	72.2
Ours(Place-Text)	11.2	25.1	34.3	27	147.7
Ours-Fusion(O-T, P-T)	25.7	45.4	54	7	65.4
Ours-Fusion(A-T, P-T)	26	46.1	55.8	7	53.5
Ours-Fusion(O-T, A-T)	31.5	51	61.5	5	41.7
Ours-Fusion(O-T, A-T, P-T)	**33.3**	**52.5**	**62.5**	**5**	**40.2**
Rank-Fusion(O-T, A-T, P-T)	30	51.3	61.8	5	42.3
Results Using Partition used by LJRV					
ST	2.99	10.9	17.5	77	241
LJRV	9.85	27.1	38.4	19	75.2
W2VV(Object-Text)	17.9	-	49.4	11	57.6
Ours(Object-Text)	20.9	43.7	54.9	7	56.1
Ours(Activity-Text)	17.5	39.6	51.3	10	54.8
Ours(Place-Text)	8.5	23.3	32.7	26	99.3
Ours-Fusion(O-T, A-T)	25.5	51.3	61.9	5	32.5
Ours-Fusion(O-T, A-T, P-T)	**26.4**	**51.9**	**64.5**	**5**	**31.1**
Rank-Fusion(O-T, A-T, P-T)	24.3	49.3	62.4	6	34.6

2.4.4 Results on MSVD Dataset

We report the results of video to text retrieval task on MSVD dataset [17] in Table 2.2 and the results for text to video retrieval in Table 2.3.

We compare our approach with existing video-text retrieval approaches, CCA[124], ST [65], JMDV [150], LJRV [99], JMET [100], and W2VV [25]. For these approaches, we directly cite scores from respective papers when available. We report score for JMET from [25]. The score of CCA is reported from [150] and the score of ST from [99]. If scores for multiple models are reported, we report the score of the best performing method.

Table 2.3: Text-to-Video Retrieval Results on MSVD Dataset. We highlight the proposed method. The methods which has 'Ours' keyword are trained with the proposed loss.

Method	R@1	R@5	R@10	MedR	MeanR
Results Using Partition used by JMET and JMDV					
CCA					251.3
JMDV					236.3
VSE(Object-Text)	12.3	30.1	42.3	14	57.7
VSEPP(Object-Text)	15.4	39.6	53	9	43.8
Ours(Object-Text)	16.1	41.1	53.5	9	42.7
Ours(Activity-Text)	15.4	39.2	51.4	10	43.2
Ours(Place-Text)	7.9	24.5	36	21	64.6
Ours-Fusion(O-T, P-T)	17	42.2	56	8	36.5
Ours-Fusion(A-T, P-T)	17.2	42.6	55.6	8	34.1
Ours-Fusion(O-T, A-T)	20.3	47.8	61.1	6	28.3
Ours-Fusion(O-T, A-T, P-T)	**21.3**	**48.5**	**61.6**	**6**	**26.3**
Rank-Fusion(O-T, A-T, P-T)	19.4	45.8	59.4	7	29.2
Results Using Partition used by LJRV					
ST	2.6	11.6	19.3	51	106
LJRV	7.7	23.4	35	21	49.1
Ours(Object-Text)	15	40.2	51.9	9	45.3
Ours(Activity-Text)	14.6	38.9	51	10	45.1
Ours(Place-Text)	7.9	24.5	36	21	64.6
Ours-Fusion(O-T, A-T)	20.2	47.5	60.7	6	29
Ours-Fusion(O-T, A-T, P-T)	**20.7**	**47.8**	**61.9**	**6**	**26.8**
Rank-Fusion(O-T, A-T, P-T)	18.5	44.9	58.8	7	30.2

We also implement and compare results with state-of-the-art image-embedding approach VSE[64] and VSEPP[28] in the Object-Text(O-T) embedding space. Additionally, to show the impact of only using the proposed loss in retrieval, we also report results based on the Activity-Text(A-T) space and Place-Text(P-T) space in the tables. Our proposed fusion is named as Ours-Fusion(O-T,A-T,P-T) in the Table. 2.2 and Table. 2.3. The proposed fusion system utilizes the proposed loss and employs three video-text embedding spaces for calculating the similarity between video and text. As the audio is muted in this dataset, we train the Activity-Text space utilizing only I3D feature from videos. We also report results for our fusion approach using any two of the three video-text spaces in the tables. Additionally, we report results of Rank-Fusion(O-T, A-T, P-T), which uses rank in place of similarity score in combining retrieval results of three video-text spaces in fusion.

From Table 2.2 and Table 2.3, it is evident that our proposed approach performs significantly better than existing ones. The result is improved significantly by utilizing the fusion proposed in this chapter that utilizes multiple video-text spaces in calculating the final ranking. Moreover, utilizing the proposed loss improves the result over previous state-of-the-art methods. It can also be identified that our loss function is not only useful for learning embedding independently, but also it is useful for the proposed fusion. We observe that utilizing the proposed loss function improves the result over previous state-of-the-art methods consistently, with a minimum improvement of 10.38% from best existing method VSEPP(Object-Text) in Video-to-Text Retrieval and 4.55% in Text-to-Video Retrieval. The result is improved further by utilizing the proposed fusion framework in this chapter that utilizes multiple video-text spaces in an ensemble fusion approach in calculating the final

ranking, with an improvement of 57.07% from the best existing method in the video to text retrieval and 38.31% in the text to video retrieval. Among the video-text spaces, object-text and activity-text space show better performance in retrieval, compared to place-text space which indicates that the annotators focused more on object and activity aspects in annotating the videos. Similar to the results of MSR-VTT dataset, we observe that the proposed score fusion approach consistently shows superior performance than rank fusion approach in both video to text and text to video retrieval.

2.4.5 Qualitative Results

We report the qualitative results on MSVD dataset in Fig. 2.5 and the results on MSR-VTT dataset in Fig. 2.6.

MSVD Dataset. In Fig. 2.5, we show examples of a few test videos from MSVD dataset and the top 1 retrieved captions for the proposed approach. We also show the retrieval result when only one of the embeddings is used for retrieval. Additionally, we report the rank of the highest ranked ground-truth caption in the figure. We can observe from the figure that in most of the cases, utilizing cue from multiple video-text spaces helps to match the correct caption. We see from Fig. 2.5 that, among 9 videos, the retrieval performance is improved or higher recall is retained for 7 videos. Video-6 and video-9 show two failure cases, where utilizing multiple video-text spaces degrades the performance slightly than using object-text in video-6 and activity-text in video-9. These failure cases provide a future direction of this work focusing on developing more sophisticated algorithms to combine similarity scores from multiple joint spaces.

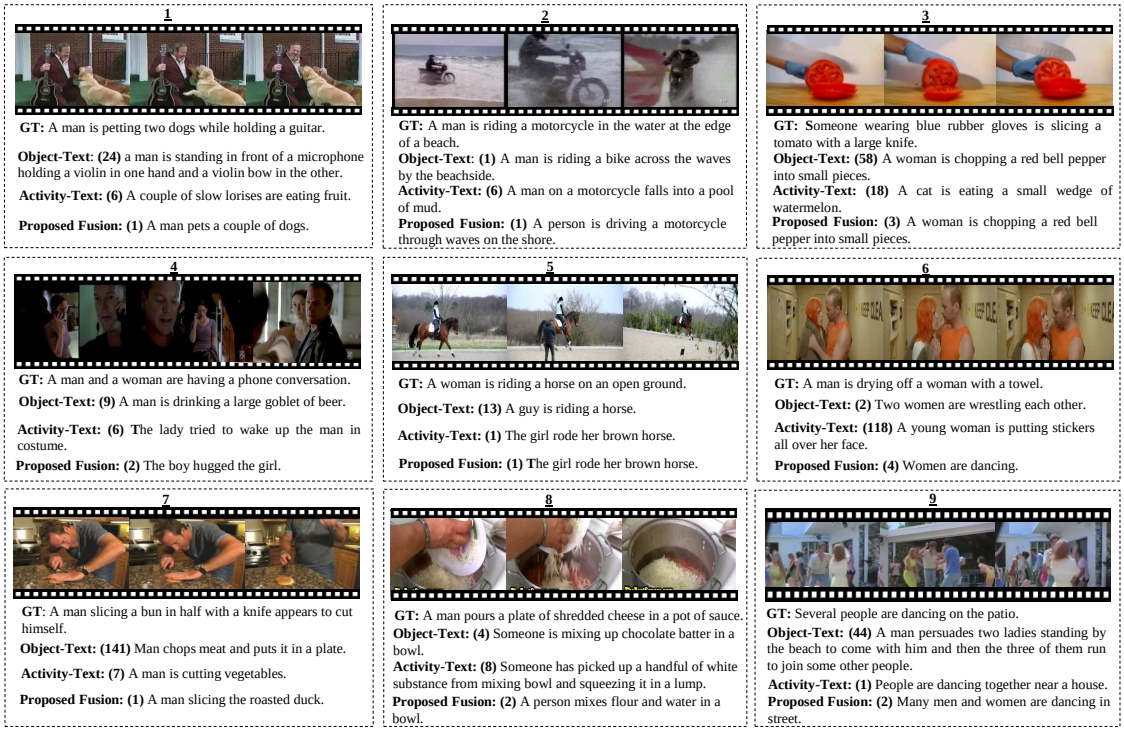

Figure 2.5: Examples of 9 test videos from MSVD dataset and the top 1 retrieved captions by using a single video-text space and the fusion approach with our loss function. The value in brackets is the rank of the highest ranked ground-truth caption. Ground Truth (GT) is a sample from the ground-truth captions. Among all the approaches, object-text (ResNet152 as video feature) and activity-text (I3D as video feature) are systems where single video-text space is used for retrieval. We also report result for the fusion system where three video-text spaces (object-text, activity-text and place-text) are used.

MSR-VTT Dataset. Similar to Fig. 2.5, we also show qualitative results for a few test videos from MSR-VTT dataset in Fig. 2.6. Video 1-6 in Fig. 2.6 shows a few examples where utilizing cue from multiple video-text spaces helps to match the correct caption compared to using only one of the video-text space. Moreover, we also see the result was improved after utilizing audio in learning the second video-text space (Activity-text space). We observe this improvement for most of the videos, as we also observe from Table. 2.1. Video 7-9 shows some failure cases for our fusion approach in Fig. 2.6. Video 7 shows a case, where utilizing multiple video-text spaces for retrieval

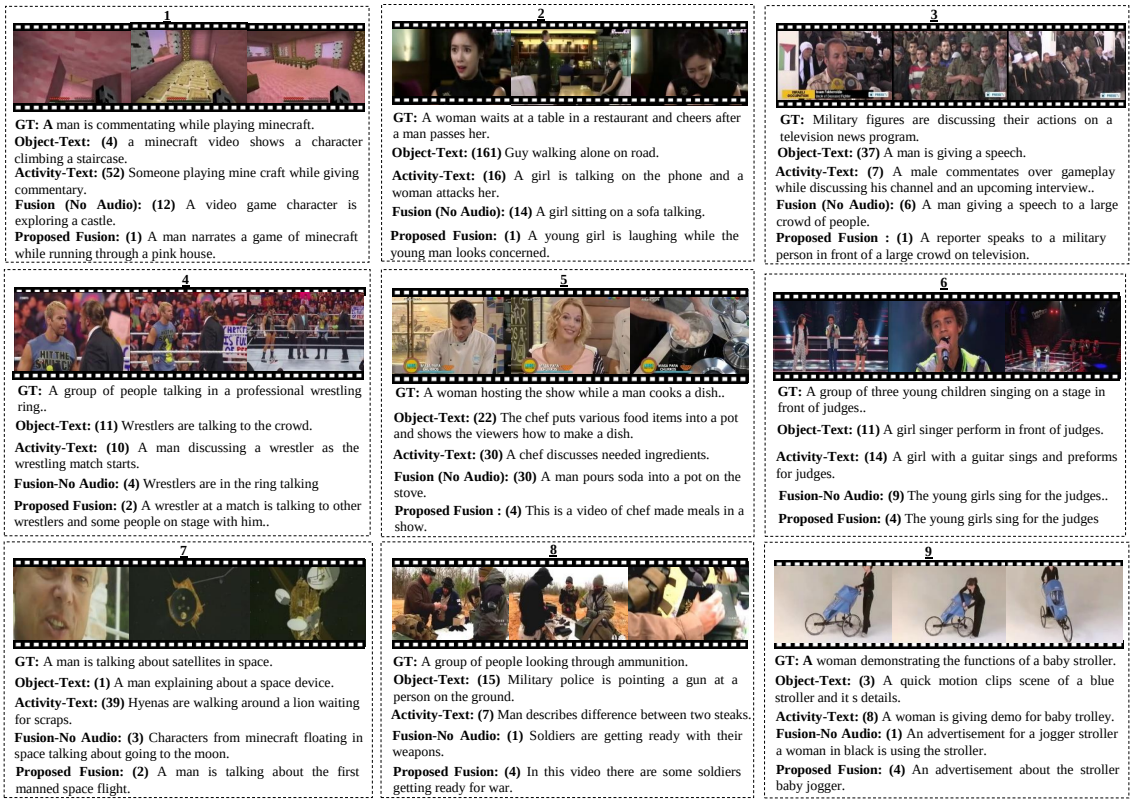

Figure 2.6: A snapshot of 9 test videos from MSR-VTT dataset with success and failure cases, the top 1 retrieved captions for four approaches based on the proposed loss function and the rank of the highest ranked ground-truth caption inside the bracket. We also report results for fusion approaches where three video-text spaces are used for retrieval. The fusion approaches use an object-text space trained with ResNet feature and place-text space trained with ResNet50(Place) feature, while in the proposed fusion, the activity-text space is trained using concatenated I3D and Audio feature.

degrades the performance slightly compared to utilizing only one of the video-text space. For Video-8 and video-9 in Fig. 2.6, we observe that the performance improves after fusion overall, but the performance is better when the audio is not used in learning video-text space. On the other hand, video 1-6 shows example of cases where utilizing audio along with visual cues helped to improve the result.

2.4.6 Discussion

The experimental results are aligned with our rationale that utilizing multiple characteristics of a video is crucial for developing an efficient video-text retrieval system. Experiments also demonstrate that our proposed ranking loss function is effective in learning video-text embeddings better than existing ones. However, we observe that major improvement in performance comes from our mixture of experts system which utilizes evidence from three complementary video-text spaces for retrieval. Our mixture of expert video-text model may not outperform the performance of a single video-text model in the ensemble in every single case, but it is evident from experiments that our system significantly reduces the overall risk of making a particularly poor decision.

From qualitative results, we observe it cannot be claimed in general that one video feature is consistently better than others for the task of video-text retrieval. It can be easily identified from the top-1 retrieved captions in Fig. 2.5 and Fig. 2.6 that the video-text embedding (Object-Text) learned utilizing object appearance feature (ResNet) as video feature is significantly different from the joint embedding (Activity-Text) learned using Activity feature (I3D) as video feature. The variation between the rank of the highest matching caption further strengthens this observation. Object-text space performs better than the activity-text space in retrieval for some videos. For other videos, the activity-text space achieves higher performance. However, it can be claimed that combining knowledge from multiple video-text embedding spaces consistently shows better performance than utilizing only one of them.

We observe from Fig. 2.6 that using audio is crucial in many cases where there is deep semantic relation between visual content and audio (e.g., the audio is from the third person narration of the video, the audio is music or song) and it gives important cues in reducing description ambiguity (e.g., video-2, video-5 and video-6 in Fig. 2.6). We observe that the performance degrades in some cases when audio is utilized in the system (e.g., video-8 in Fig. 2.6). We see an overall improvement in the quantitative result (Table 2.1) which also supports our idea of using audio. Since we did not exploit the structure of the audio and analyze the structural alignment between audio and video, it is difficult to determine whether audio is always helpful. For instance, audio can come from different things (persons, animals or objects) in a video, and it might shift our attention away from the main subject. Moreover, the captions in the datasets are provided mostly based on visual aspects, which makes information related to audio very sparse. Hence, the overall improvement using audio was limited.

2.5 Conclusion

For multimedia applications, constructing a joint representation that could carry information for multiple modalities could be very conducive for downstream use cases. In this chapter, we study how to leverage diverse video features effectively for developing a robust cross-modal video-text retrieval system. In this chapter, we study how to effectively utilize available multimodal cues from videos in learning joint representations for the cross-modal video-text retrieval task. Existing hand labeled video-text datasets are often very limited by their size considering the enormous amount of diversity the visual world contains.

This makes it extremely difficult to develop a robust video-text retrieval system based on deep neural network models. In this regard, we propose a framework that simultaneously utilizes multi-modal visual cues by a "mixture of experts" approach for retrieval. Our proposed framework learns three expert video-text embedding models focusing on three salient video cues (i.e., object, activity, place) and uses a combination of these models for high-quality prediction. A modified pair-wise ranking loss function is also proposed for better learning the joint embeddings, which focuses on hard negatives and applies a weight-based penalty based on the relative ranking of the correct match. Extensive evaluations on MSVD and MSR-VTT datasets demonstrate that our framework performs significantly better than baselines and state-of-the-art systems.

Chapter 3

Video Moment Retrieval from Text Queries with Weak Supervision

3.1 Introduction

Cross-modal retrieval of visual data using natural language description has attracted intense attention in recent years [45, 154, 64, 57, 148, 149, 96], but remains a very challenging problem [154, 28, 90] due to the differences and ambiguity between different modalities. The identification of the video moment (or segment) is important since it allows the user to focus on the portion of the video that is most relevant to the textual query, and is beneficial when the video has a lot of non-relevant portions. (See Fig. 3.1). The aforementioned approaches operate in a fully supervised setting, i.e., they have access to text descriptions along with the exact temporal location of the visual data corresponding to the descriptions. However, obtaining such annotations is tedious and noisy, requiring

Figure 3.1: Illustration of text to video moment retrieval task: given a text query, retrieve and rank videos segments based on how well they depict the text description.

multiple annotators. The process of developing algorithms which demand a weaker degree of supervision is non-trivial and is yet to be explored by researchers for the problem of video moment retrieval using text queries. In this work, we focus, particularly on this problem.

The text to video moment retrieval task is more challenging than the task of localizing categorical activities in videos, which is a comparatively well-studied field [83, 143, 157, 147, 104, 123]. Although these methods show success on activity localization, unlike text to moment retrieval, they are limited to a pre-defined set of activity classes. In this regard, there has been a recent interest in localizing moments in a video from natural language description [44, 34, 148, 18]. Supervision in terms of text description with their temporal boundaries in a video is used to train these models. However, acquiring such dense annotations of text-temporal boundary tuples are often tedious and costly, as it is difficult to mark the start and end locations of a certain moment, which may also introduce ambiguity in the training data.

On the contrary, it is often much easier to just describe the moments appearing in a video with a set of natural language sentences, than providing exact temporal boundaries associated with each of the sentences. Moreover, such descriptions can often be obtained easily from captions through some sources on the web. Motivated by this, we pose a question: *is it possible to develop a weakly-supervised framework for video moment localization from the text, leveraging only video-level textual annotation, without their temporal boundaries?* Temporal localization of moments using weak description is a much more challenging task than its supervised counterpart. It is extremely relevant to address this question, due to the difficulty and non-scalability of acquiring precise frame-wise information with text descriptions in the fully supervised setting.

Overview of the Proposed Framework. An illustration of our proposed weakly-supervised framework presented in Fig. 3.2. Given a video, we first extract frame-wise visual features from pre-trained Convolutional Neural Network architectures. We also extract features for text descriptions using Recurrent Neural Network based models. Similar to the video-text embedding model described in chapter 2, we train a joint embedding network to project video features and text features into the same space. However, as we have text descriptions for the videos as a whole and not moment-wise descriptions like in a fully supervised setting, the learning procedure for text to video moment retrieval is non-trivial.

Given a certain text description, we obtain its similarity with the video features, which gives an indication of temporal locations which may correspond to the textual description. We call this Text-Guided Attention as it helps to highlight the relevant temporal locations, given a text description. Thereafter, we use this attention to pool the video

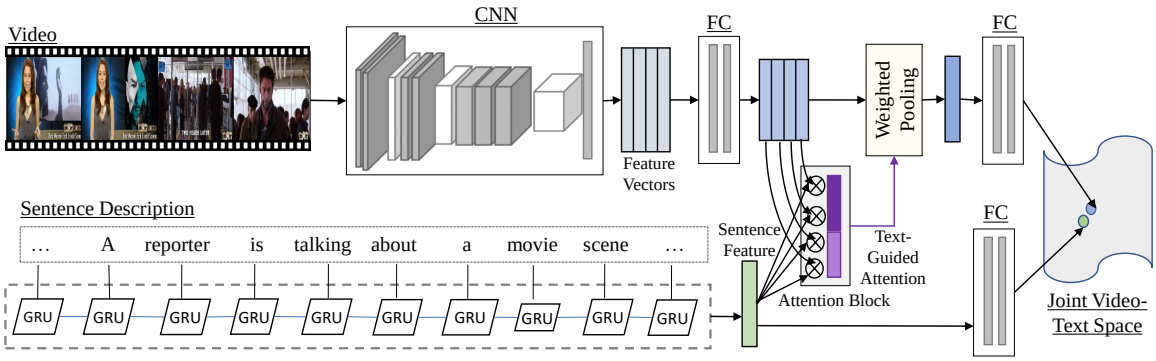

Figure 3.2: A brief illustration of our proposed weakly supervised framework for learning joint embedding model with Text-Guided Attention for text to video moment retrieval. Our framework learns a latent alignment between video frames and text corresponding to the video. This alignment is utilized for attending video features based on relevance and the pooled video feature is used for learning the joint video-text embedding. In the figure, CNN refers to a convolutional neural network, and FC refers to a fully-connected neural network.

features along the temporal direction to obtain a single text-dependent feature vector for a video. We then train the network to minimize a loss which reduces the distance between the text-dependent video feature vector and the text vector itself. We hypothesize that along with learning a shared video-text embedding, hidden units will emerge internally to learn the notion of relevance between moments of video and corresponding text description. During the testing phase, we use TGA for localizing the moments, given a text query, as it highlights the portion of the video corresponding to the query.

Contributions: The main contributions of the proposed approach are as follows.

• We address a novel and practical problem of temporally localizing video moments from text queries without requiring temporal boundary annotations of the text descriptions while training but using only the video-level text descriptions.

• We propose a joint visual-semantic embedding framework, that learns the notion of relevant moments from video using only video-level description. Our joint embedding

network utilizes latent alignment between video frames and sentence description as Text-Guided Attention for the videos to learn the embedding.

• Experiments on two benchmark datasets: DiDeMo [44] and Charades-STA [34] show that our weakly-supervised approach performs reasonably well compared to supervised baselines in the task of text to video moment retrieval.

3.2 Related Works

Image/Video Retrieval using Text Queries. Cross-modal language-vision retrieval methods focus on retrieving relevant figs/videos from a database given text descriptions. Most of the recent methods for image-text retrieval task focus on learning joint visual-semantic embedding models [58, 64, 32, 144, 28, 96, 136, 93]. Inspired by the success of these approaches, most video-text retrieval methods also employ a joint subspace model [150, 25, 137, 100, 90, 91]. In this joint space, the similarity of different points reflects the semantic closeness between their corresponding original inputs. These text-based video retrieval approaches focus on retrieving an entire video from dataset given text description. However, we focus on temporally localizing a specific moment relevant to a text query, within a given video. Similar to the video/image to text retrieval approaches, our proposed framework is also based on learning joint video-text embedding models. However, instead of focusing only on aligning video and text in the joint space as in video-text retrieval, our aim is to learn a latent alignment between video frames and text descriptions, which is used for obtaining the relevant moments corresponding to a given text query.

Activity Localization. The moment retrieval aspect of our work is related to the problem of temporal activity localization in untrimmed videos. From the perspective of our interest, the works in literature pertaining to activity localization can be categorized as either fully supervised or weakly supervised. Works in fully supervised setting include SSN [157], R-C3D [147], TAL-Net [16] among others. Most of these works structure their framework by using temporal action proposals with activity location predictors. However, in the weakly supervised setting, the exact location of each activity is unknown, and only the video-level labels are accessible during training. In order to deal with that, researchers take a Multiple Instance Learning approach [143] with constraints applied for better localization [104, 98]. Our task of video moment retrieval from text description is more challenging than the activity localization task, as our method is not limited to a pre-defined set of categories, but rather sentences in natural language.

Text to Video Moment Retrieval. Most relevant to our work are the methods that focus on identifying relevant portions from text description using fully-supervised annotations: MCN [44], CTRL [34], EFRC [148], ROLE [79], TGN [18]. These methods are severely plagued by the issue of collecting training videos with temporal natural language annotation. Temporal sliding window over videos frames [44], or hard-coded segments containing a fixed number of frames [34] has been used for generating moment candidate corresponding to a text description. Moreover, unlike in images, generating temporal proposals for videos in an unsupervised manner is itself a challenging task. In [148, 147], the authors proposed an end-to-end framework where the activity proposals are generated as one of the initial steps, but for the much easier task of activity localization. Attention

mechanism has been used in [79, 148] for the text to video moment retrieval task. Although we also use attention, our usage is significantly different from them. ROLE [79] uses attention over the words using video moment context, which they obtain from the temporal labels. EFRC [148] uses attention in training a temporal proposal network as it has access to temporal boundary annotations of the sentences. We use attention over the temporal dimension of the videos as we do not have access to the temporal boundaries. More importantly, our method is weakly-supervised, which requires only video-level text annotation during training. Hence, the data collection cost for our approach is substantially less, and it is possible to acquire and train using larger video-text captioning datasets.

A weakly supervised setting is considered in [11] for the video-text alignment task, which is to assign temporal boundaries to a set of temporally ordered sentences, whereas our task is to retrieve a portion of the video given a sentence. Moreover, [11] assumes temporal ordering between the sentences as additional supervision. Also, their method would require dense sentence annotations describing all portions of the video including tokens representing background moments (if any). The task considered in this work is a generalization of the task in [11]. We consider that there can be multiple sentences describing different temporal portions of a single video and do not consider any temporal ordering information of the sentences. The Text-Guided Attention mechanism used in our framework allows us to deal with multiple sentence descriptions during training and provide the relevant portions for each of them during testing, even with weak supervision.

3.3 Approach

In this section, we first describe the network architecture and input feature representation for representing video and text (Sec. 3.3.1). Then, we present our proposed Text-Guided Attention module (Sec. 3.3.2). Finally, we describe the framework for learning joint video-text embedding (Sec.3.3.3).

Problem Definition. In this chapter, we consider that the training set consists of videos paired with text descriptions composed of multiple sentences. Each sentence describes different temporal regions of the video. However, we do not have access to the temporal boundaries of the moments referred to by the sentences. At test time, we use a sentence to retrieve relevant portions of the video.

3.3.1 Network Structure and Features

Network Structure. The joint embedding model is trained using a two-branch deep neural network model, as shown in Fig. 3.2. The two branches consist of different expert neural networks to extract modality-specific representations from the given input. The expert networks are followed by fully connected embedding layers which focus on transforming the modality-specific representations to joint representations. In this work, we keep the pre-trained image encoder fixed as we have limited training data. The fully-connected embedding layers, the word embedding, the GRU are trained end-to-end. We set the dimensionality (D) of the joint embedding space to 1024.

Text Representation. We use Gated Recurrent Units (GRU) [23] for encoding the sentences. GRU has been very popular for generating a representation for sentences in

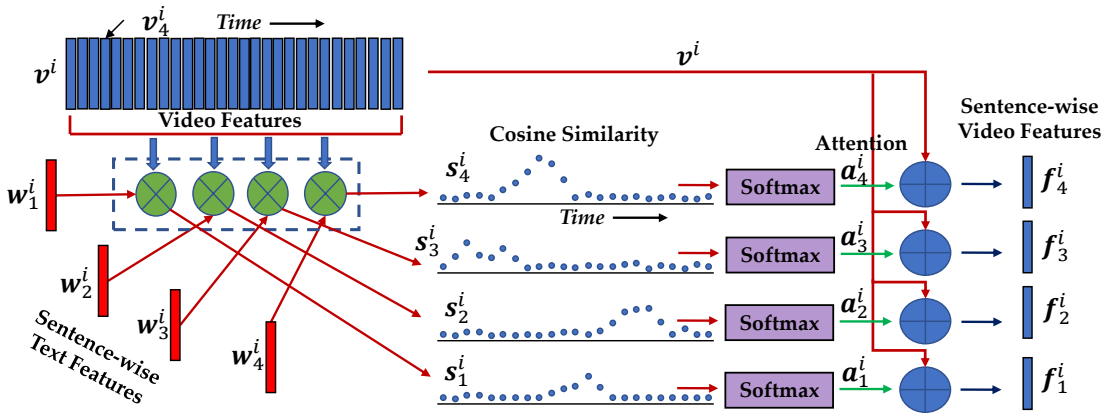

Figure 3.3: This figure presents the procedure of computing the Text-Guided Attention and using it to generate sentence-wise video features. We first obtain the cosine similarity between the features at every time instant of the video $\boldsymbol{v}_i$, and its corresponding sentences $\boldsymbol{w}_j^i$, followed by a softmax layer along the temporal dimension to obtain the sentence-wise temporal attention. Thereafter, we use these attentions to compute a weighted average of the video features to finally obtain the sentence-wise video features.

recent works [28, 64]. The word embeddings are input to the GRU. The dimensionality of the word embeddings is 300.

Video Representation. We utilize pre-trained convolutional neural network models as the expert network for encoding videos. Specifically, following [34] we utilize C3D model [131] for feature extraction from every 16 frames of video for the Charades-STA dataset. A 16 layer VGG model [122] is used for frame-level feature extraction in experiments on DiDeMo dataset following [44]. We extract features from the penultimate fully connected layer. For both the C3D and VGG16 model, the dimension of the representation from the penultimate fully connected layer is 4096.

3.3.2 Text-Guided Attention

After the feature extraction process, we have a training set $\mathcal{D} = \{\{\boldsymbol{w}_j^i\}_{j=1}^{nw_i}, \{\boldsymbol{v}_k^i\}_{k=1}^{nv_i}\}_{i=1}^{n_d}$, where n_d is the number of training pairs, $\boldsymbol{w}_j^i$ represents the j^{th} sentence feature of i^{th} video, $\boldsymbol{v}_k^i$ represent the video feature at the k^{th} time instant of the i^{th} video, nw_i and nv_i are the number of sentences in the text description and video time instants for the i^{th} video in the dataset. Please note that we do not consider any ordering in the text descriptions.

Each of the sentences provides us information about a certain part of the given video. In a fully supervised setting, where we have access to the temporal boundaries associated with each sentence, we can apply a pooling technique to first pool the relevant portion of the video features and then use a similarity measure to learn a joint video segment-text embedding. However, in our case of weakly supervised moment retrieval, we do not have access to the temporal boundaries associated with the sentences. Thus, we need to first obtain the portions of the video which are relevant to a given sentence query.

If some portion of the video frames corresponds to a particular sentence, we would expect them to have similar features. Thus, the cosine similarity between text and video features should be higher in the temporally relevant portions and low in the irrelevant ones. Moreover, as the sentence described a part of the video rather than individual temporal segments, the video feature obtained after pooling the relevant portions should be very similar to the sentence description feature. We employ this idea to learn the joint video-text embedding via an attention mechanism based on the sentence descriptions, which we name Text-Guided Attention (TGA). Note that during the test phase, we use TGA to obtain the localization.

We first apply a Fully Connected (FC) layer with ReLU [69] and Dropout [126] on the video features at each time instance to transform them into the same dimensional space as the text features. We denote these features as $\bar{\boldsymbol{v}}_k^i$. In order to obtain the sentence specific attention over the temporal dimension, we first obtain the cosine similarity between each temporal feature and sentence descriptions. The similarity between the j^{th} sentence and the k^{th} temporal feature of the i^{th} training video can be represented as follows,

$$s_{kj}^i = \frac{\boldsymbol{w}_j^{i\,T} \boldsymbol{v}_k^i}{||\boldsymbol{w}_j^i||_2 ||\boldsymbol{v}_k^i||_2} \tag{3.1}$$

Once we obtain the similarity scores for the temporal locations, we apply a softmax operation along the temporal dimension to obtain an attention vector for the i^{th} video as follows,

$$a_{kj}^i = \frac{\exp(s_{kj}^i)}{\sum_{k=1}^{nv_i} \exp(s_{kj}^i)} \tag{3.2}$$

These should have high values at temporal locations which are relevant to the given sentence vector $\boldsymbol{w}_j^i$. We consider this as local similarity because the individual temporal features may correspond to different aspects of a sentence and thus each of the temporal features might be a bit scattered away from the sentence feature. However, the feature obtained after pooling the video temporal features corresponding to the relevant locations should be quite similar to the entire sentence feature. We consider this global similarity. We use the attention in Eqn. 3.2 to obtain the pooled video feature for the sentence description $\boldsymbol{w}_j^i$ as follows,

$$\boldsymbol{f}_j^i = \sum_{k=1}^{nv_i} a_{kj}^i \boldsymbol{v}_k^i \tag{3.3}$$

Note that, this feature vector corresponds to the particular sentence description $\boldsymbol{w}_j^i$ only. In a similar procedure, we can extract the text-specific video feature vector

corresponding to the other sentences in the text descriptions of the same video and other videos as well. Fig. 3.3 presents an overview of the sentence-wise video feature extraction procedure using the video temporal features and a set of sentence descriptions for the video. We use these feature vectors to derive the loss function to be optimized to learn the parameters of the network. This is described next.

3.3.3 Training Joint Embedding

We now describe the loss function we optimize to learn the joint video-text embedding. Many prior approaches have utilized pairwise ranking loss as the objective for learning joint embedding between visual and textual input [64, 158, 145, 58]. Specifically, these approaches minimize a hinge-based triplet ranking loss in order to maximize the similarity between an image embedding and corresponding text embedding and minimize similarity to all other non-matching ones. Note that, the loss function has also been presented in Eq. 2.1 in Chapter 2 as VSE loss.

For the sake of notational simplicity, we drop the index i, j, k denoting the video number, sentence index and time instant. Given a text-specific video feature vector based on TGA, $\boldsymbol{f}$ $(\in \mathbb{R}^V)$ and paired text feature vector $\boldsymbol{w}$ $(\in \mathbb{R}^T)$, the projection for the video feature on the joint space can be derived as $\boldsymbol{v}_p = W^{(v)}\boldsymbol{f}$ $(\boldsymbol{v}_p \in \mathbb{R}^D)$. Similarly, the projection of paired text vector in the embedding space can be expressed as $\boldsymbol{t}_p = W^{(t)}\boldsymbol{w}(\boldsymbol{t}_p \in \mathbb{R}^D)$. Here, $W^{(v)} \in \mathbb{R}^{D\times V}$ is the transformation matrix that projects the video content into the joint embedding and D is the dimensionality of the joint space. Similarly, $W^{(t)} \in \mathbb{R}^{D\times T}$ maps input sentence/caption embedding to the joint space.

Using these pairs of feature representation of both videos and corresponding sentences, the goal is to learn a joint embedding such that the positive pairs are closer than the negative pairs in the feature space. Now, the video-text loss function $\mathcal{L}_{VT}$ can be expressed as follows,

$$\begin{aligned}\mathcal{L}_{VT} = \sum_{(\boldsymbol{v}_p, \boldsymbol{t}_p)} \Big\{ \sum_{\boldsymbol{t}_p^-} max\big[0, \Delta - S(\boldsymbol{v}_p, \boldsymbol{t}_p) + S(\boldsymbol{v}_p, \boldsymbol{t}_p^-)\big] \\ + \sum_{\boldsymbol{v}_p^-} max\big[0, \Delta - S(\boldsymbol{t}_p, \boldsymbol{v}_p) + S(\boldsymbol{t}_p, \boldsymbol{v}_p^-)\big] \Big\}\end{aligned} \tag{3.4}$$

where $\boldsymbol{t}_p^-$ is a non-matching text embedding for video embedding $\boldsymbol{v}_p$, and $\boldsymbol{t}_p$ is the matching text embedding. This is similar for video embedding $\boldsymbol{v}_p$ and non-matching image embedding $\boldsymbol{v}_p^-$. Δ is the margin value for the ranking loss. The scoring function $S(\boldsymbol{v}_p, \boldsymbol{t}_p)$ measures the similarity between the image embedding and text embedding in the joint space. We utilize cosine similarity in the representation space to compute similarity. Cosine similarity is widely used in learning joint embedding models in prior works on image-text retrieval [158, 64, 28, 93]. Our approach does not depend on any specific choice of similarity function.

In Eq. (3.4), the first term attempts to ensure that for each visual input, the matching text inputs should be closer than non-matching text inputs in learning the joint space. However, the second term in Eq. (3.4) attempts to ensure that for each text input, the matching image input should be closer in the joint space than the non-matching images.

3.3.4 Batch-wise Training

We train our network using Stochastic Gradient Descent (SGD) by dividing the dataset into batches. For a video with multiple sentences, we create multiple video-sentence pairs, with the same video, but different sentences in the corresponding video's text descrip-

tion. During training, our method learns to automatically identify the relevant portions for each sentence using the Text-Guided Attention. The negative instances $\boldsymbol{v}_p^-$ and $\boldsymbol{t}_p^-$ correspond to all the instances which are not positive in the current batch of data.

3.4 Experiments

We perform experiments on two benchmark datasets with the goal of comparing the performance of our weakly-supervised approach against different supervised baselines. As we introduce the problem in this work, to the best of our knowledge, ours is the first to show results on this task. Ideally, any weakly supervised methods would attempt at attaining the performance of the supervised methods, with similar features and setting.

We first describe the details on the datasets and evaluation metric in Sec. 3.4.1, followed by the training details in Sec. 3.4.2. Then, we report the results of different methods on DiDeMo and Charades-STA dataset in Sec. 3.4.3.

3.4.1 Datasets and Evaluation Metric

We present experiments on two benchmark datasets for sentence description based video moment localization, namely Charades-STA [34] and DiDeMo [44] to evaluate the performance of our proposed framework.

Charades-STA. The Charades-STA dataset for text to video moment retrieval was introduced in [34]. The dataset contains 16,128 sentence-moment pairs with 12,408 in the training set and 3,720 in the testing set. The Charades dataset was originally introduced in [121] which contains temporal activity annotation and video-level paragraph description

for the videos. The authors of [34] enhanced the dataset [121] for evaluating temporal localization of moments in videos given text queries. The video-level descriptions from the original dataset were decomposed into short sentences. Then, these sentences are assigned to segments in videos based on matching keywords for activity categories. The annotations are manually verified at last.

DiDeMo. The Distinct Describable Moments (DiDeMo) dataset [44] is one of the largest and most diverse datasets for the temporal localization of events in videos given natural language descriptions. The videos are collected from Flickr and each video is trimmed to a maximum of 30 seconds. The videos in the dataset are divided into 5-second segments to reduce the complexity of annotation. The dataset is split into training, validation and test sets containing 8,395, 1,065 and 1,004 videos respectively. The dataset contains a total of 26,892 moments and one moment could be associated with descriptions from multiple annotators. The descriptions in DiDeMo dataset are detailed and contain camera movement, temporal transition indicators, and activities. Moreover, the descriptions in DiDeMo are verified so that each description refers to a single moment.

Evaluation Metric. We use the evaluation criteria following prior works in literature [44, 34]. Specifically, we follow [44] for evaluating DiDeMo dataset and [34] for evaluating Charades-STA. We measure rank-based performance $R@K$ (Recall at K) which calculates the percentage of test samples for which the correct result is found in the top-K retrievals to the query sample. We report results for R@1, R@5, and R@10. We also calculate temporal intersection over union (tIoU) for Charades-STA dataset and mean intersection over union (mIoU) for DiDeMo dataset.

3.4.2 Implementation Details

We used two Telsa K80 GPUs and implemented the network using PyTorch [103]. We start training with a learning rate of 0.001 and keep the learning rate fixed for 15 epochs. The learning rate is lowered by a factor of 10 every 15 epochs. We tried different values for margin α in training and found $0.1 \leq \Delta \leq 0.2$ works reasonably well. We empirically choose Δ as 0.1 for Charades-STA and 0.2 for DiDeMo in the experiments. We use a batch-size of 128 in all the experiments. ADAM optimizer was used in training the joint embedding networks [63]. The model was evaluated on the validation set on the video-text retrieval task after every epoch. To deal with the over-fitting issue, we choose the best model based on the highest sum of recalls.

3.4.3 Quantitative Results

We report the experimental results on Charades-STA dataset [34] in Table 3.1 and DiDeMo dataset [44] in Table 3.3.

Results on Charades-STA Dataset

The quantitative results on Charades-STA dataset [34] are reported in Table 3.1. The evaluation setup in Charades-STA dataset [34] considers a set of IoU (Intersection over Union) thresholds. We report for IoU $0.3, 0.5$ and 0.7 in Table 3.1. For these IoU thresholds, we report the recalls - R@1, R@5, and R@10 in Table 3.1. Following [34], we use sliding windows of 128 and 256 to obtain the possible temporal segments. The segments are ranked based on the corresponding Text-Guided Attention score.

Table 3.1: This table presents the results on the Charades-STA dataset, using the evaluation protocol used in previous works. We also use C3D feature for a fair comparison. The proposed weakly-supervised approach performs significantly better that visual-semantic embedding based baselines: VSA-RNN and VSA-STV. Our approach also performs reasonably compared to state-of-the-art approaches CTRL[34] and EFRC [148].

Method	IoU=0.3			IoU=0.5			IoU=0.7		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Random	-	-	-	8.51	37.12	-	3.03	14.06	-
VSA-RNN	-	-	-	10.50	48.43	-	4.32	20.21	-
VSA-STV	-	-	-	16.91	53.89	-	5.81	23.58	-
CTRL	-	-	-	23.63	58.92	-	8.89	29.52	-
EFRC	53.00	94.60	98.50	33.80	77.30	91.60	15.00	43.90	60.90
Proposed	**32.14**	**86.58**	**99.33**	**19.94**	**65.52**	**89.36**	**8.84**	**33.51**	**53.45**

Compared Methods. We compare our approach with state-of-the-art text to video moment retrieval approaches, CTRL[34], EFRC[148], and baseline approaches, VSA-RNN[57] and VSA-STV[65]. For these methods, we directly cite performances from respective papers when available [34, 148]. We report score for VSA-RNN and VSA-STV from [34]. If the score for multiple models is reported, we select the score of the best performing method in R@1. Here, VSA-RNN (Visual-Semantic Embedding with LSTM) and VSA-STV (Visual-Semantic Embedding with Skip-thought vector) are text-based image/video retrieval baselines. We also report results for "Random" which selects a candidate moment randomly. Similar to these approaches, we also utilize the C3D model for obtaining feature representation of videos for fair comparison. We follow the evaluation criteria utilized in prior works [34, 148].

Analysis of Results. We observe that the proposed approach consistently perform comparably to the fully-supervised approaches in all evaluation metrics. Our weakly-

Table 3.2: Ablation Study of the Model on Charades-STA Dataset

Input Encoding		Margin	IoU=0.3		IoU=0.5		IoU=0.7	
Video Feature	Text Feature	(Δ)	R@1	R@5	R@1	R@5	R@1	R@5
C3D	GRU	0.05	30.6	86.4	17.7	64.9	8.1	33.4
C3D	GRU	0.15	31.5	87.3	19.4	65.9	8.2	32.9
C3D	GRU	0.20	31.7	87.7	18.9	65.5	8.4	33.8
C3D	GRU	0.10	32.1	86.6	19.9	65.5	8.9	33.5
C3D	Bi-GRU	0.10	32.5	87.9	19.9	65.6	9.2	33.5
I3D	GRU	0.10	33.1	87.5	19.7	65.4	9.3	33.2
ResNet-152	GRU	0.10	28.9	87.4	18.8	66.0	9.0	33.6
DenseNet-121	GRU	0.10	31.2	87.1	19.0	66.2	8.9	34.1

supervised TGA based approach performs significantly better than supervised visual-semantic embedding based approaches VSA-RNN and VSA-STV. We observe that the proposed method achieves a minimum absolute improvement of 13.3% in R@5 and 4.5% in R@1 from VSA-RNN. The relative performance improvement over VSA-STV is a minimum of 17.9% in R@1 and 21.5% in R@5. We also observe that the proposed approach achieves better performance than state-of-the-art method CTRL [34] on R@5 evaluation metrics with a maximum relative improvement of about 13.5% in R@5 with IoU=0.7. Our approach also shows reasonable performance compared to EFRC [148].

Ablation Study. We present a ablation study on Charades dataset in Table 3.2.The Table 3.2 shows that our method performs reasonably well over a range of parameters and feature choice. However, Δ=0.1 performs better overall compared to other margin values. Also, C3D, I3D works slightly better than ResNet, DenseNet, and Bi-GRU performs slightly better than GRU.

Results on DiDeMo Datset

Table 3.3 summarizes the results on the DiDeMo dataset [44]. DiDeMo only has a coarse annotation of moments. As the videos are trimmed at 30 seconds and the videos are divided into 5-second segments, each video has 21 possible moments. We follow the evaluation setup in [44], which is designed for evaluating 21 possible moments from sentence descriptions. Average of Text-Guided Attention scores of corresponding segments is used as the confidence score for the moments and used for ranking. Following previous works [44, 148], the performance in the dataset is evaluated based on R@1, R@5, and mean intersection over union (mIoU) criteria.

Compared Methods. In Table 3.3, we report results for several baselines to analyze the performance of our proposed approach. We divide the table into 3 rows (2.1-2.3). In row-2.1, we report the results of trivial baselines (i.e., Random and Upper-Bound) following evaluation protocol reported in [44]. In row-2.2, we group the results of LSTM-RGB-Local [44], EFRC [148], and our proposed approach for a fair comparison, as these methods are trained with only the VGG-16 RGB feature. We report the performance of the proposed approach in both validation and test set as LSTM-RGB-local model has been evaluated on validation set [44]. In row-2.3, we report results for state-of-the-art approaches MCN [44] and TGN [18]. We also report results of CCA [66] and natural language object retrieval based baseline Txt-Obj-Retrieval [48] in row-2.3. These methods additionally use optical flow feature along with VGG16 RGB feature. We report the performance of MCN [44], TGN [18] and EFRC [148] from the respective papers. The results of LSTM-RGB-Local, Txt-Obj-Retrieval, Random, and Upper-Bound are reported from [44].

Table 3.3: This table reports results on DiDeMo following the evaluation protocol in [44]. Our approach performs on par with several competitive fully-supervised approaches

#	Method	R@1	R@5	mIoU
3.3.1	Upper Bound	74.75	100	96.05
	Random	3.75	22.5	22.64
3.3.2	LSTM-RGB-Local [44]	13.10	44.82	25.13
	EFRC [148]	13.23	46.98	27.57
	Proposed (Val. Set)	**11.18**	**35.62**	**24.47**
	Proposed (Test Set)	**12.19**	**39.74**	**24.92**
3.3.3	CCA	18.11	52.11	37.82
	Txt-Obj-Retrieval [48]	16.20	43.94	27.18
	MCN [44]	27.57	79.69	41.70
	TGN [18]	28.23	79.26	42.97

Analysis of Results. Similar to the results on Charades-STA, it is evident from Table 3.3 that our proposed weakly supervised approach consistently shows comparable performance to several fully-supervised approaches. From row-2.2, we observe that our proposed approach achieves similar performance as LSTM-RGB-Local [44] and EFRC [148]. We observe that R@5 accuracy is slightly lower for our approach compared to supervised approaches. However, R@1 accuracy and mIoU is almost similar. Comparing row-2.3, we observe that the performance is comparable to CCA and Txt-Obj-Retrieval baselines. The performance is low compared to MCN [44] and TGN [18]. Both of the approaches use additional optical flow features in their framework. MCN additionally use a moment-context feature. Hence, a performance drop is not unexpected. However, we have already observed from the row-2.2 that the performance of our weakly supervised approach is comparable to the MCN baseline model of LSTM-RGB-Local which uses the same RGB feature in training as our proposed method.

3.4.4 Qualitative Results

We provide six qualitative examples of moments predicted by the proposed approach from Charades-STA dataset [34] in Fig. 3.4. In Fig. 3.4, case 1, 2, and 4 show some examples where our approach was successful in retrieving the ground truth moment with high IoU. Cases 1 and 2 are examples where the same video has been used to retrieve different moments based on two different text descriptions. We see our text-aware attention module was successful in finding the correct segment of the video in both the cases.

While our method retrieves the correct moment from sentence description many cases, it fails to retrieve the correct moment in some cases (e.g., case 3, 5, and 6). Among these three cases, case 3 presents an ambiguous query where the person stands on the doorway but does not enter into the room. The GT moment covers a smaller segment, while our system predicts a longer one. We observe the performance of our system suffers when important visual contents occupy only small portions in frames, e.g., case 5 and 6. In case 6, a sandwich is mentioned in the query which occupies a small portion of frames initially and our framework shifted the start time of the moment to a much later time instant than in the ground truth. Similarly, in case 5, our system was only successful in identifying the person laughing into a blanket after the scene is zoomed in. We believe these are difficult to capture without additional spatial attention modeling or generating region proposals. Moreover, utilizing more cues from videos (e.g., audio, and context) may be helpful in reducing ambiguity in these cases.

3.5 Conclusion

There have been a few recent methods proposed in text to video moment retrieval using natural language queries, but requiring full supervision during training. However, acquiring a large number of training videos with temporal boundary annotations for each text description is extremely time-consuming and often not scalable. In order to cope with this issue, in this work, we introduce the novel problem of learning from weak labels for the task of text to video moment retrieval. The weak nature of the supervision is because, during training, we only have access to the video-text pairs rather than the temporal extent of the video to which different text descriptions relate. We propose a joint visual-semantic embedding based framework that learns the notion of relevant segments from video using only video-level sentence descriptions. Specifically, our main idea is to utilize latent alignment between video frames and sentence descriptions using Text-Guided Attention (TGA). TGA is then used during the test phase to retrieve relevant moments. Our formulation of the task makes it more realistic compared to existing methods in the literature which require supervision as temporal boundaries or temporal ordering of the sentences. Moreover, the weak nature of the task allows it to learn from easily available web data, which requires minimal effort to acquire compared to manual annotations. Experiments on two benchmark datasets demonstrate that our method in spite of being weakly supervised performs comparably to several fully supervised approaches.

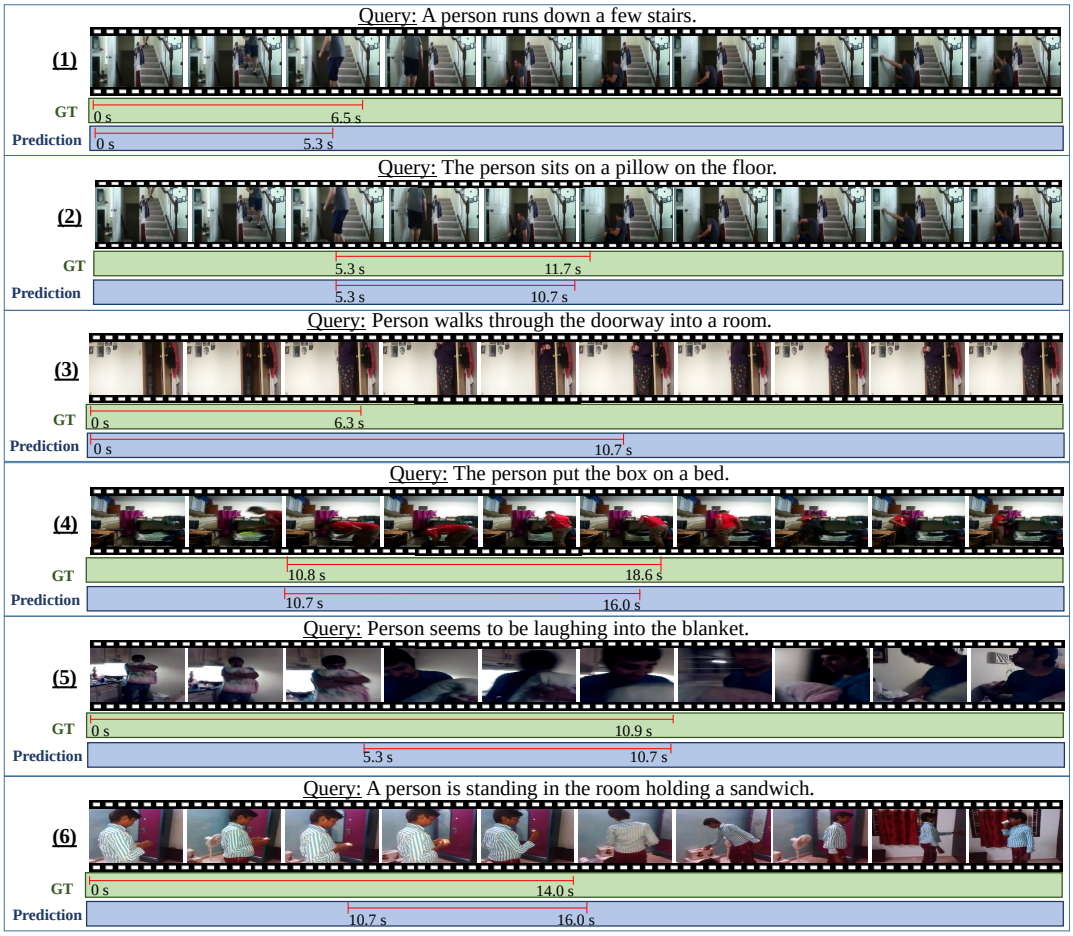

Figure 3.4: A snapshot of six queries and test videos from Charades-STA dataset with success and failure cases. GT is a ground-truth annotation and Prediction is the moment predicted by the proposed approach. Queries 1, 2, and 4 show cases where our approach was successful in retrieving the GT moment with very high temporal intersection over union (IoU). However, queries 3, 5, and 6 show cases where our approach was not successful in retrieving the GT moment with high IoU.

Chapter 4

Web-Supervised Joint Embedding for Cross-Modal Image-Text Retrieval

4.1 Introduction

Joint embeddings have been widely used in multimedia data mining as they enable us to integrate the understanding of different modalities together. These embeddings are usually learned by mapping inputs from two or more distinct domains (e.g., images and text) into a common latent space, where the transformed vectors of semantically associated inputs should be close. Learning an appropriate embedding is crucial for achieving high-performance in many multimedia applications involving multiple modalities. In this work, we focus on the task of cross-modal retrieval between images and language (See Fig. 4.1),

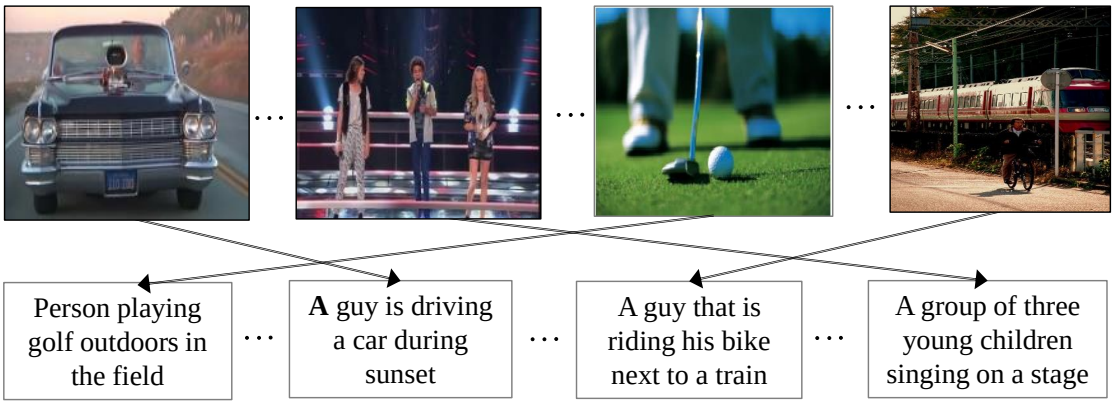

Figure 4.1: Illustration of Image-Text retrieval task: Given a text query, retrieve and rank images from the database based on how well they depict the text or vice versa.

i.e., the retrieval of images given sentence query, and retrieval of text from a query image.

The majority of the success in image-text retrieval task has been achieved by the joint embedding models trained in a supervised way using image-text pairs from hand-labeled image datasets (e.g., MSCOCO [20], Flickr30k [106]). Although, these datasets cover a significant number of images (e.g., about 80k in MSCOCO and 30K in Flickr30K), creating a larger dataset with image-sentence pairs is extremely difficult and labor-intensive [68]. Moreover, it is generally feasible to have only a limited number of users to annotate training images, which may lead to a biased model [134, 49, 156]. Hence, while these datasets provide a convenient modeling assumption, they are very restrictive considering the enormous amount of rich descriptions that a human can compose [57]. Accordingly, although trained models show good performance on benchmark datasets for image-text retrieval task, applying such models in the open-world setting is unlikely to show satisfactory cross-dataset generalization (training on a dataset, testing on a different dataset) performance.

On the other hand, streams of images with noisy tags are readily available in datasets, such as Flickr-1M [54], as well as in nearly infinite numbers on the web. Developing a practical system for image-text retrieval considering a large number of web images is more likely to be robust. However, inefficient utilization of weakly-annotated images may increase ambiguity and degrade performance. Motivated by this observation, we pose an important question: *Can a large number of web images with noisy annotations be leveraged upon with a fully annotated dataset of images with textual descriptions to learn better joint embeddings?* Fig. 4.2 shows an illustration of this scenario. This is an extremely relevant problem to address due to the difficulty and non-scalability of obtaining a large amount of human-annotated training set of image-text pairs. In this work, we study how to judiciously utilize web images to develop a successful image-text retrieval system. We propose a novel framework that can augment any ranking loss based supervised formulation with weakly-supervised web data for learning robust joint embeddings.

The raw tags associated with web images are often incomplete and error-prone. Hence, directly utilizing such data without any refinement in the objective of webly supervised learning may lead to an increased ambiguity and degraded performance. Moreover, the learning approach should be able to deal with huge amount missing information as encountered frequently in our setting (i.e., most social media images may not contain many relevant tags). These challenges make the problem of learning robust joint embedding models using web images extremely difficult when the amount of noisy tags associated with web images is unexpectedly high compared to clean relevant tags. In this regard, we also explore the research question - *Based on a limited fully annotated set of images with textual*

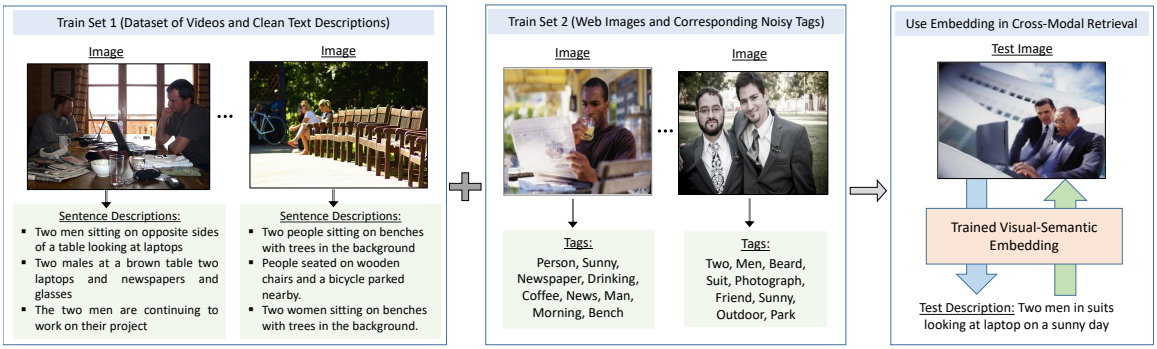

Figure 4.2: The problem setting of our work. Our goal is to utilize web images associated with noisy tags to learn a robust visual-semantic embedding from a dataset of clean images with ground truth sentences. We test the learned latent space by projecting images and text descriptions from the test set in the embedding and perform cross-modal retrieval.

descriptions, is it possible to refine the tags of web image and utilize them in boosting the performance of joint image-text embedding models? For example, can we build a reasonable joint image-text embedding model when we have access to only 5% of labeled data from image-text datasets (e.g., MSCOCO) and the remaining 95% data are weakly annotated? Although, existing largest image-text datasets cover a limited number of images (e.g., about 80k in MSCOCO and 30K in Flickr30K), it is critical to consider availability of a significantly smaller number of cross-modal pairs (e.g., 2K pairs) focusing on specific practical applications, such as cross-modal retrieval focusing on a sudden emergency scenario. In such a case, it is extremely crucial to complement scarcer clean set of pairs with freely available web images to improve the performance of image-text embedding models. However, availability of a small clean set makes it extremely difficult to train a reliable model, considering significantly high amount of noisy and missing entries typical in web image tagging.

4.1.1 Overview of the Proposed Webly Supervised Embedding Approach

In the cross-modal image-text retrieval task, an embedding network is learned to project image features and text features into the same joint space, and then the retrieval is performed by searching the nearest neighbor in the latent space. In this work, we attempt to utilize web images annotated with noisy tags for improving joint embeddings trained using a dataset of images and ground-truth sentence descriptions. However, combining web image-tag pairs with image-text pairs in training the embedding is non-trivial. The greatest obstacle arises from noisy tags and the intrinsic difference between the representation of sentence description and tags. A typical representation of text is similar to, and yet very different from the representation of tags. Sentences are usually represented using RNN-based encoder with word-to-vec (Word2Vec) model, providing sequential input vectors to the encoder. In contrast, tags do not have sequential information and a useful representation of tags can be tf-idf weighted BOW vectors or the average of all Word2Vec vectors corresponding to the tags.

To bridge this gap, we propose a two-stage approach that learns the joint image-text representation. Firstly, we use a supervised formulation that leverages the available clean image-text pairs from a dataset to learn an aligned representation that can be shared across three modalities (e.g., image, tag, text). As tags are not available directly in the datasets, we consider nouns and verbs from a sentence as dummy tags (Fig. 4.3). We leverage ranking loss based formulation with image-text and image-tags pairs to learn a shared representation across modalities. Secondly, we utilize weakly-annotated image-tags pairs from the web (e.g., Flickr) to update the previously learned shared representation,

which allows us to transfer knowledge from thousands of freely available weakly annotated images to develop a better cross-modal retrieval system. Our proposed approach is also motivated by learning using privileged information (LUPI) paradigm [135, 119] and multi-task learning strategies in deep neural networks [114, 9] that share representations between closely related tasks for enhanced learning performance.

4.1.2 Overview of the Proposed Image-Tag Refinement Approach

The idea is to first refine the tags of weakly annotated web image collection utilizing their latent relationships with the small clean set of images. The two set of image collections can be inter-related easily based on associated tags, however, we can only have partial observations of the relationships due to the noisy nature of web image tags. We propose to utilize the observed incomplete relationships in a tensor completion framework to predict the missing tags and remove the noisy ones. The proposed image tag refinement approach is motivated by the success of tensor completion approaches in multi-way data analysis [110, 128, 97]. In this work, we formulate the web image-tag refinement as a CP decomposition based tensor completion approach that leverages ternary interactions among dataset images, tags and web images in refining web image tags. To efficiently recover missing dynamics, we also incorporate intra-modal similarity as auxiliary information to regularize the tensor completion problem. Refined web images are then used with webly supervised learning frameworks for training joint image-text embeddings.

4.1.3 Contributions

We address a novel and practical problem in this chapter—how to exploit large scale web data for learning an effective joint visual-semantic embedding models without requiring large amount of human-crafted training data. Towards solving this problem, we make the following main contributions.

• We propose a webly supervised approach utilizing web image collection with associated noisy tags, and a clean dataset containing images and ground truth sentence descriptions for learning robust joint representations.

• We develop a novel framework with ranking loss for augmenting a typical supervised method with weakly-supervised web data to learn a more robust joint embedding.

• We also present an extension of our webly supervised image-text embedding framework in the presence of very limited clean labeled data and web images containing significant noise. In the framework, the web images associated with noisy tags are first refined using proposed tensor completion approach and then used with a small clean dataset in webly supervised learning frameworks for training joint image-text embedding models.

• We propose to refine tags of web images by modeling the inter-relation between web image collection and clean dataset images (based on associated tags) as a tensor and utilizing intra-modal similarity as side information in a CP decomposition based tensor completion framework.

• We demonstrate clear performance improvement in image-text retrieval using proposed web-supervised approach on standard benchmark image-text retrieval datasets, e.g., Flickr30K [106] and MSCOCO [77].

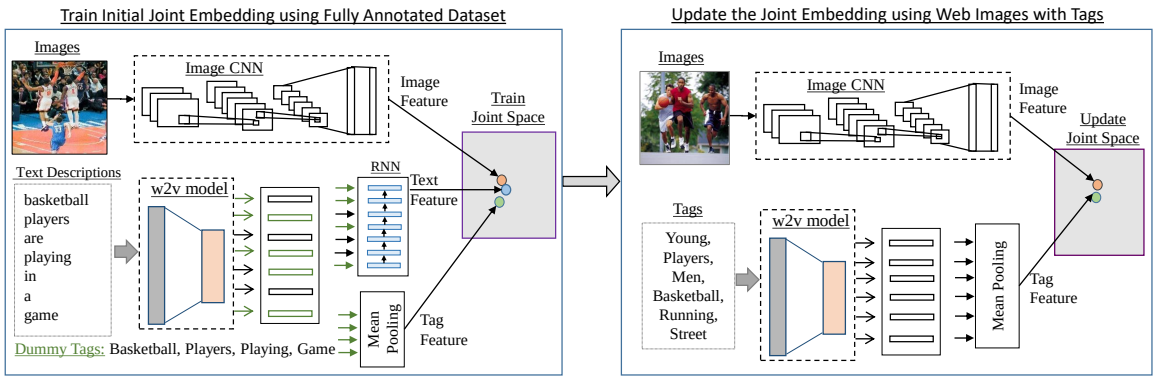

Figure 4.3: A brief illustration of our proposed framework for learning visual-semantic embedding model utilizing image-text pairs from a dataset and image-tag pairs from the web. First, a dataset of images and their sentence descriptions are used to learn an aligned image-text representation. Then, we update the joint representation using web images and corresponding tags. The trained embedding is used in image-text retrieval task.

4.2 Related Work

Visual-Semantic Embedding: Joint visual-semantic models have shown excellent performance on several multimedia tasks, e.g., cross-modal retrieval [145, 66, 50, 90], image captioning [86, 57], image classification [53, 32, 38] video summarization [22, 105]. Cross-modal retrieval methods require computing semantic similarity between two different modalities, i.e., vision and language. Learning joint visual-semantic representation naturally fits to our task of image-text retrieval since it is possible to directly compare visual data and sentence descriptions in such a joint space [28, 96].

Image-Text Retrieval: Recently, there has been significant interest in developing powerful image-text retrieval methods in multimedia, computer vision and machine learning communities [58, 45]. In [29], a method for mapping visual and textual data to a common space based on extracting a triplet of object, action, and scene is presented. A number of image-text embedding approaches has been developed based on Canonical

Correlation Analysis (CCA) [151, 124, 47, 38]. Ranking loss has been used for training the embedding in most recent works relating image and language modality for image-text retrieval [64, 32, 144, 28, 96]. In [32], words and images are projected to a common space utilizing a ranking loss that applies a penalty when an incorrect label is ranked higher than the correct one. A bi-directional ranking loss based formulation is used to project image features and sentence features to a joint space for cross-modal image-text retrieval in [64].

Several image-text retrieval methods extended this work [64] with slight modifications in the loss function [28], similarity calculation [136, 144] or input features [96]. In [28], authors modified the ranking loss based on violations incurred by relatively hard negatives. An embedding network is proposed in [144] that uses the bi-directional ranking loss along with neighbourhood constraints. Multi-modal attention mechanism is proposed in [96] to selectively attend to specific image regions and sentence fragments and calculate similarity. A multi-modal LSTM network is proposed in [52] that recurrently select salient pairwise instances from image and text, and aggregate local similarity measurement for image-sentence matching. Our method complements the works that projects words and images to a common space utilizing a bi-directional ranking loss. The proposed formulation could be extended and applied to most of these approaches with little modifications.

Webly Supervised Learning: The method of manually annotating images for training does not scale well to the open-world setting as it is impracticable to collect and annotate images for all relevant concepts [72, 94]. Moreover, there exists different types of bias in the existing datasets [134, 130, 62]. In order to circumvent these issues, several recent studies focused on using web images and associated metadata as auxiliary source of

information to train their models [73, 37, 127]. Although web images are noisy, utilizing such weakly-labeled web images in training has been shown to be very effective in several multimedia tasks [39, 73, 56].

Our work is motivated by these works on learning more powerful models by realizing the potential of web data. As the largest MSCOCO dataset for image-sentence retrieval has only 80K training images, we believe it is extremely crucial and practical to complement scarcer clean image-sentence data with web images to improve the generalization ability of image-text embedding models. Most relevant to our work is [39], where authors constructed a dictionary by taking a few thousand most common words and represent text as tf-idf weighted bag of words (BoW) vectors that ignore word order and represents each caption as a vector of word frequencies. Although, such a feature representation allows them to utilize the same feature extractor for sentences and set of tags, it fails to consider the inherent sequential nature present in sentences in training joint embedding models.

Tensor completion for multi-modal data analysis. Tensor completion approaches focus on estimating the missing elements of partially observed tensors [125]. CP decomposition [46, 42] and Tucker decomposition [132, 24] are most widely used approaches for low-rank decomposition of tensors. There are several works on completing tensors to estimate missing data based on tensor decomposition [97, 80, 116]. In this work, we develop a tensor decomposition based tensor completion approach. We specifically use CP decomposition as it has been found that Tucker decomposition based approaches are computationally less flexible than CP decomposition approaches in handling large datasets in a distributed manner as it needs to deal with complex core tensor [125].

There have been a few works on exploiting tensor decomposition based approaches in tag refinement [116, 128, 117]. These works assume the availability of additional user information along with images and tags and utilize Tucker decomposition based approach for tag refinement. Although user information may provide important cues in refining tags, user information is unlikely to be available in most cases. In this work, we explore the use of a small clean dataset containing images and tags in refining web image tags so that we can limit the propagation of noisy tags in recovering missing tags. Several previous works have shown that utilizing relationships among data as auxiliary information helps to improve the quality of tensor decomposition significantly when limited entries are observed [97, 160, 35, 142]. Inspired by these works, we use intra-modal similarity matrices as side information in the proposed approach to deal with a high ratio of missing entries.

4.3 Learning Webly Supervised Image-Text Embedding

In this section, we first describe the network structure (Section 4.3.1). Then, we revisit the basic framework for learning image text mapping using pair-wise ranking loss (Section 4.3.2). Finally, we present our proposed strategy to incorporate the tags in the framework to learn an improved embedding (Section 4.3.3).

4.3.1 Network Structure and Input Feature

Network Structure: Similar to the two-branch network utilized in Chapter 2 and Chapter 3, we again learn our joint embedding model using a deep neural network framework. As shown in Fig. 4.3, our model has three different branches for utilizing

image, sentence, and tags. Each branch has different expert network for a specific modality followed by two fully connected embedding layers. The idea is that the expert networks will focus on identifying modality-specific features at first and the embedding layers will convert the modality-specific features to modality-robust features. The parameters of these expert networks can be fine-tuned together with training the embedding layers. For simplicity, we keep image encoder (e.g., pre-trained CNN) and tag encoder (e.g., pre-trained Word2Vec model) fixed in this work. The word embedding and the GRU for sentence representation are trained end-to-end.

Text Representation: For encoding sentences, we use Gated Recurrent Units (GRU) [23], which has been used for representing sentence in many recent works [28, 64]. We set the dimensionality of the joint embedding space, D, to 1024. The dimensionality of the word embeddings that are input to the GRU is 300.

Image Representation: For encoding image, we adopt a deep CNN model trained on ImageNet dataset as the encoder. Specifically, we experiment with state-of-the-art 152 layer ResNet model [43] and 19 layer VGG model [122] in this work. We extract image features directly from the penultimate fully connected layer. The dimension of the image embedding is 2048 for ResNet152 and 4096 for VGG19. We first re-scale the image to 256x256 and 224x224 center crop is feed into CNNs as inputs.

Tag Representation: We generate the feature representation of tags by summing over the Word2Vec [87] embeddings of all tags associated with an image and then normalizing it by the number of tags. Averaged word vectors has been shown to be a strong feature for text in several tasks [153, 61, 60].

4.3.2 Train Joint Embedding with Ranking Loss

We now describe the basic framework for learning joint image-sentence embedding based on bi-directional ranking loss. Many prior approaches have utilized pairwise ranking loss as the objective for learning joint embedding between visual input and textual input [64, 158, 145, 58]. Specifically, these approaches minimize a hinge-based triplet ranking loss in order to maximize the similarity between an image embedding and corresponding text embedding and minimize similarity to all other non-matching ones.

Given a image feature representation $\overline{i}$ ($\overline{i} \in \mathbb{R}^V$), the projection on the joint space can be derived as $i = W^{(i)}\overline{i}$ ($i \in \mathbb{R}^D$). Similarly, the projection of input text embedding $\overline{s}$ ($\overline{s} \in \mathbb{R}^T$) to joint space can be derived by $s = W^{(s)}\overline{s}$ ($s \in \mathbb{R}^D$). Here, $W^{(i)} \in \mathbb{R}^{D \times V}$ is the transformation matrix that maps the visual content into the joint space and D is the dimensionality of the space. In the same way, $W^{(s)} \in \mathbb{R}^{D \times T}$ maps input sentence embedding to the joint space. Given feature representation for words in a sentence, the sentence embedding $\overline{s}$ is found from the hidden state of the GRU. Here, given the feature representation of both images and corresponding text, the goal is to learn a joint embedding characterized by θ (i.e., $W^{(i)}$, $W^{(s)}$ and GRU weights) such that the image content and semantic content are projected into the joint space. Now, the image-sentence loss function $\mathcal{L}_{IS}$ can be written as following,

$$\mathcal{L}_{IS} = \sum_{(i,s)} \Big\{ \sum_{s^-} max\big[0, \Delta - f(i,s) + f(i,s^-)\big] + \sum_{i^-} max\big[0, \Delta - f(s,i) + f(s,i^-)\big] \Big\} \quad (4.1)$$

where s^- is a non-matching text embedding for image embedding i, and s is the matching text embedding. This is similar for image embedding i and non-matching image embedding

i^-. Δ is the margin value for the ranking loss. The scoring function $f(i, s)$ measure the similarity between the images and text in the joint embedded space. In this work, we use cosine similarity in the representation space to calculate similarity, which is widely used in learning image-text embedding and shown to be very effective in many prior works [158, 64, 28]. Our approach does not depend on any particular choice of similarity function.

The first term in Eq. (4.1) represent the sum over all non-matching text embedding s^- which attempts to ensure that for each visual feature, corresponding/matching text features should be closer than non-matching ones in the joint space. Similarly, the second term attempts to ensure that text embedding that corresponds to the image embedding should be closer in the joint space to each other than non-matching image embeddings.

Recently, focusing on hard-negatives has been shown to be effective in learning joint embeddings [28, 158, 118, 85]. Subsequently, the loss in Eq. 4.1 is modified to focus on hard negatives (i.e., the negative closest to each positive (i, s) pair) instead of sum over all negatives in the formulation. For a positive pair (i, s), the hardest negative sample can be identified using $\hat{i} = \arg\max_{i^-} f(s, i^-)$ and $\hat{s} = \arg\max_{s^-} f(i, s^-)$. Hence, the ranking loss function can be written as following,

$$\mathcal{L}_{IS} = \sum_{(i,s)} \Big\{ max\big[0,\ \Delta - f(i,s) + f(i,\hat{s})\big] + max\big[0,\ \Delta - f(s,i) + f(s,\hat{i})\big] \Big\} \tag{4.2}$$

We name Eq. 4.1 as VSE loss and Eq. 4.2 as VSEPP loss. We utilize both of these loss functions in evaluating our proposed approach.

4.3.3 Training Joint Embedding with Web Data

In this work, we try to utilize image-tag pairs from the web for improving joint embeddings trained using a clean dataset with images-sentence pairs. Our aim is to learn a good representation for image-text embedding that ideally ignores the data-dependent noise and generalizes well. Utilization of web data effectively increases the sample size used for training our model and can be considered as implicit data augmentation. However, it is not possible to directly update the embedding (Sec. 4.3.2) using image-tag pairs. GRU based approach is not suitable for representing tags since tags do not have any semantic context as in the sentences.

Our task can also be considered from the perspective of learning with side or privileged information strategies [135, 119], as in our case an additional tag modality is available at training time and we would like to utilize this extra information to train a stronger model. However, directly employing LUPI strategies are also not possible in our case as the training data do not provide three modality information at the same time. The training datasets (e.g., MSCOCO, Flickr30K) provide only image-sentence pairs and does not provide tags. On the other hand, web source provides images with tags, but no sentence descriptions. To bridge this gap, we propose a two-stage approach to train the joint image-text representation. In the first stage, we leverage the clean image-text pairs from a dataset to learn an aligned representation shared across three modalities (e.g., image, tag, text). In the second stage, we adapt the model trained in the first stage with web data and tags.

Stage I: Training initial Embedding. We leverage image-text pairs from an annotated dataset to learn a joint embedding for image, tag and text. As tags are not

available directly in the datasets, we consider nouns and verbs from relevant sentence as dummy tags for an image (Fig. 4.3). For learning the shared representation, we combine the image-text ranking loss objective (Sec. 4.3.2), with image-tag ranking loss objective. We believe combining image-tag ranking loss objective provides a regularization effect in training that leads to more generalized image-text embedding.

Now the goal is to learn a joint embedding characterized by θ (i.e., $W^{(i)}$, $W^{(t)}$, $W^{(s)}$ and GRU weights) such that the image, sentence and tags are projected into the joint space. Here, $W^{(t)}$ projects the representation of tags $\bar{t}$ on the joint space as, $t = W^{(t)}\bar{t}$. The resulting loss function can be written as following,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{IS} + \lambda_2 \mathcal{L}_{IT} \tag{4.3}$$

where, $\mathcal{L}_{IT}$ represent image-tag ranking loss, which is similar to image-sentence ranking loss objective $\mathcal{L}_{IS}$ in Sec. 4.3.2. Similar to VSEPP loss in Eq. 4.2, $\mathcal{L}_{IT}$ can be written as,

$$\mathcal{L}_{IT} = \sum_{(i,t)} \Big\{ max\big[0,\ \Delta - f(i,t) + f(i,\hat{t})\big] + max\big[0,\ \Delta - f(t,i) + f(t,\hat{i})\big] \Big\} \tag{4.4}$$

where for a positive image-tag pair (i, t), the hardest negative sample tag representation can be identified as $\hat{t}$. Note that all tags associated with a image is considered for generating tag representation in creating a image-tag pair rather than considering a single tag related to that image. In Eq. 4.3, λ_1 and λ_2 are predefined weights for different losses. In the first training stage, both losses are used ($\lambda_1 = 1$ and $\lambda_2 = 1$) while in the second stage, image-text loss is not used ($\lambda_1 = 0$ and $\lambda_2 = 1$).

Stage II: Model Adaptation with Web Data. After Stage I converges, we have a shared representation of image, sentence description and tags with a learned image-tag embedding model. In Stage II, we utilize weakly-annotated image-tags pairs from

Flickr to update the previously learned embedding network using $\mathcal{L}_{IT}$ loss. This enables us to transfer knowledge from thousands of freely available weakly annotated images in learning the embedding. We utilize a smaller learning rate in Stage II, as network achieves competitive performance after Stage I and tuning the embedding network with a high learning rate from weakly-annotated data may lead to catastrophic forgetting [59].

As web data is very prone to label noise, we found it is extremely hard to learn good representation for our task in many cases. Hence, in Stage II, we adopt a curriculum learning-based strategy in training. Curriculum learning allows the model to learn from easier instances first so they can be used as building blocks to learn more complex ones, which leads to a better performance in the final task. It has been shown in many previous works that appropriate curriculum strategies guide the learner towards better local minima [8]. Our idea is to gradually inject difficult information to the learner such that in the early stages of training, the network is presented with images related to frequently occurring concepts/keywords in the clean training set. Images related to rarely occurring concepts are presented at a later stage. Since the network trained in Stage I is more likely to have learned well about frequently occurring concepts in the dataset, label noise is less likely to affect the network adversely.

4.4 Refinement of Tags of Web Image Collection

Our webly supervised joint embedding learning framework may suffer when the amount of clean fully annotated images is very low. In such a case, we propose to include a web image tag refinement approach in the webly supervised joint image-text embedding

framework. The framework attempts at attaining better performance compared to the image-text embedding baselines that directly uses raw image and tags in training without any refinement. We optimize ranking loss function in learning webly supervised joint embedding models to show the benefits of the proposed tag refinement step in the overall image-text retrieval performance.

The intuition is that the multi-dimensional relation that exists between web image with noisy tags and images with clean tags can be modeled as a multi-dimensional tensor. Analyzing the multi-dimensional relation tensor can be beneficial in refining tags of web images. We consider that we have three types of entities (i.e., web images, dataset images, and selected tags) and the ternary relationship (based on tag association) among the entities is modeled as a tensor. We propose a CP decomposition based tensor completion approach to complete the observed tensor to recover the missing relationships. A brief illustration of our proposed tensor completion approach is shown in Fig. 4.4. We start by giving notations and then present the approach.

Preliminaries: Throughout this chapter, we use calligraphic bold uppercase letters to denote tensors, uppercase letters to denote matrices and lowercase letters to denote vectors. For a third order tensor $\boldsymbol{\mathcal{X}}$, its entries are denoted by $\mathcal{X}_{ijk}$. The Frobenius norm of $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{|D|\times|W|\times|T|}$ is defined as $||\boldsymbol{\mathcal{X}}||_F = \sqrt{\sum_{i=1}^{|D|}\sum_{j=1}^{|W|}\sum_{k=1}^{|T|}\mathcal{X}_{ijk}^2}$.

The CP tensor decomposition aims to approximate an order-N tensor with R latent factors as a sum of R rank-one tensors [67, 120]. For a third order tensor $\boldsymbol{\mathcal{X}}$, it can be written as :

$$\boldsymbol{\mathcal{X}} \approx \tilde{\boldsymbol{\mathcal{X}}} = [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]$$

Here, $[[Z^{(1)}, Z^{(2)}, Z^{(3)}]]$ represents a weighted sum of rank-1 tensors where the vectors that specify the rank-1s are columns of the factor matrices $Z^{(1)}$, $Z^{(2)}$, and $Z^{(3)}$.

4.4.1 Tag Refinement using CP tensor completion model.

We consider that we have access to three types of data, i.e., the images from dataset $D = \{d_i\}_{i=1}^{|D|}$ (for which we know the associated tags correctly), the images from the web $W = \{n_i\}_{j=1}^{|W|}$ (for which we know *some* associated *noisy* tags), and the selected tag set $T = \{t_i\}_{k=1}^{|T|}$. $\boldsymbol{\mathcal{X}}$ denotes the tensor with complete tri-modal dynamics. Since very few tags are found in most images, $\boldsymbol{\mathcal{X}}$ is likely to be sparse and low-rank. If the i-th image from the dataset and the j-th image from web image collection are both annotated with the kth tag from the selected tag set, $\mathcal{X}_{ijk} = 1$. Otherwise, $\mathcal{X}_{ijk} = 0$. However, as web images mostly have a few associated noisy tags, we only have a partial observation of $\boldsymbol{\mathcal{X}}$ at the start.

In the image-tag refinement, our goal is to refine tags of web image set W by predicting missing tags and removing noisy tags. We propose to model the recovery of missing tags and removing noisy tags based on a tensor completion framework. Our model is built following CP tensor completion model [36] as follows:

$$\begin{gathered}\min_{Z^{(n)},\boldsymbol{\mathcal{X}}} \ \frac{1}{2}||\boldsymbol{\mathcal{X}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 + \frac{\lambda}{2}\sum_{n=1}^{3}||Z^{(n)}||_F^2; \\ \text{s.t. } \boldsymbol{\Omega} * \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{T}}, \mathbf{Z^{(n)}} \geq \mathbf{0}\end{gathered} \tag{4.5}$$

$\boldsymbol{\mathcal{T}}$ denotes the observations we have for $\boldsymbol{\mathcal{X}}$. The latent factor matrices for the clean dataset images, web image set, and tag set are denoted respectively by $Z^{(1)} \in \mathbb{R}^{|D|\times R}$, $Z^{(2)} \in \mathbb{R}^{|W|\times R}$ and $Z^{(3)} \in \mathbb{R}^{|T|\times R}$. Here, R is the number of latent factors. $\boldsymbol{\Omega}$ is a non-negative weight tensor with the same size as $\boldsymbol{\mathcal{X}}$. If $\mathcal{X}_{ijk}$ is observed, $\boldsymbol{\Omega_{ijk}} = \mathbf{1}$. Otherwise $\boldsymbol{\Omega_{ijk}} = \mathbf{0}$.

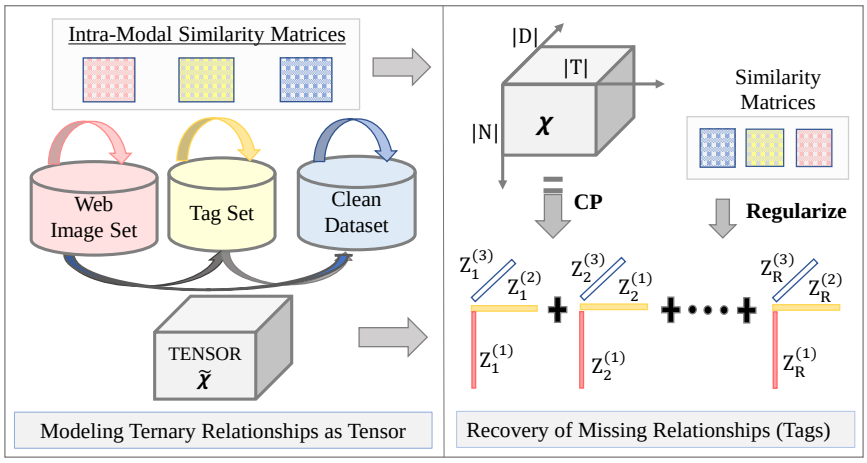

Figure 4.4: Brief Illustration of our CP decomposition based Tensor Completion approach for Image-Tag Refinement.

Our goal is to seek an estimated $\boldsymbol{\mathcal{X}}$ for recovering the missing dynamics of tags based upon the partially observed data. However, we need more information to recover $\boldsymbol{\mathcal{X}}$. In this regard, we also consider intra-relationships in the three types of data as side information as described below.

4.4.2 Regularize CP model with auxiliary information.

We believe that using intra-modal relations between entities can help as additional side information in our tensor factorization framework. We can calculate the intra-modal relationship between images based on image similarity measures. Similarly, we can model the relationship between tags by calculating the similarity between tags. In this work, we use the cosine similarity measure. Given, feature representation of images, the similarity between the images of the dataset can be calculated as follows:

$$\Theta_{Dataset}(i,j) = \frac{d_i^T d_j}{||d_i||_2 \; ||d_j||_2} \tag{4.6}$$

Θ_{Web} and Θ_{Tag} similarity matrices are also calculated in a similar fashion to Eq. 4.5 using cosine similarity measure. We utilize the similarity matrices as auxiliary information in our CP completion model. The idea is that if two images are similar, the latent representations of these two images should be similar. Therefore, we want to make the latent representations of two similar entities to be close. This can be obtained by minimizing the following:

$$\begin{aligned} L_{AUX} &= \sum_{i,j} \Theta(i,j) \; ||Z^{(n)}_{i,:} - Z^{(n)}_{j,:}||^2 \\ &= \sum_{i,j} Z^{(n)^T}_{i,:} \Theta(i,j) Z^{(n)}_{i,:} - \sum_{i,j} Z^{(n)^T}_{i,:} \Theta(i,j) Z^{(n)}_{j,:} \\ &= tr(Z^{(n)^T} \mathcal{L} \; Z^{(n)}) \end{aligned}$$

where $Z^{(n)}_{i,:}$ is the ith row of the factor matrix $Z^{(n)}$ for the nth mode of tensor $\boldsymbol{\mathcal{X}}$ ($n \in \{1,2,3\}$). D is a diagonal matrix with $D_{ij} = \sum_j \; \Theta_{ij}$ and $\mathcal{L} = D - \Theta$ is the Laplacian of similarity matrix Θ. Now, adding the auxiliary information, the Eq. 4.5 becomes:

$$\begin{aligned} \min_{Z^{(n)}, \boldsymbol{\mathcal{X}}} \; & \frac{1}{2} ||\boldsymbol{\mathcal{X}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]] \; ||^2_F + \frac{\lambda}{2} \sum_{n=1}^{3} ||Z^{(n)}||^2_F \\ & + \sum_{n=1}^{3} \alpha_n tr(Z^{(n)^T} \mathcal{L}_n \; Z^{(n)}); \\ & \text{s.t. } \boldsymbol{\Omega} * \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{T}}, \mathbf{Z^{(n)}} \geq \mathbf{0} \end{aligned} \tag{4.7}$$

α is a hyper-parameter to control the weight of auxiliary information from different factors.

4.4.3 ADMM Optimization.

In this section, we present details about the alternating direction method of multipliers (ADMM) approach [12, 76, 36] to solve our optimization problem in Eq.4.7. Specifically, the overall procedure of the ADMM algorithm consists of three main steps following [36]. First, an auxiliary variable is introduced to separate the objective function into two

different objectives. Second, an augmented Lagrangian is formed with combining both linear and quadratic terms through a scaled dual variable. Third, the augmented Lagrangian is minimized iteratively with respect to the primal variables and the dual variable until convergence. To facilitate the optimization, we consider an equivalent form of Eq. 4.7 by introducing an auxiliary variable U:

$$\begin{aligned} \min_{Z^{(n)},\boldsymbol{\mathcal{X}}} \ & \frac{1}{2}||\boldsymbol{\mathcal{X}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 + \frac{\lambda}{2}\sum_{n=1}^{3}||Z^{(n)}||_F^2 \\ & + \sum_{n=1}^{3}\alpha_n tr(U^{(n)^T}\mathcal{L}_n\ U^{(n)}); \\ & \text{s.t. } \boldsymbol{\Omega} * \boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{T}}, \mathbf{Z^{(n)}} = \mathbf{U^{(n)}} \geq \mathbf{0} \end{aligned} \tag{4.8}$$

The objective function in Eq. 4.8 is not convex together. We can form the augmented Lagrangian $L_\mu(U^{(n)}, Z^{(n)}, \Lambda^{(n)})$ with both linear and quadratic terms as follows:

$$\begin{aligned} L_\mu(U^{(n)}, Z^{(n)}, \Lambda^{(n)}, \boldsymbol{\mathcal{X}}) &= \frac{1}{2}||\boldsymbol{\mathcal{X}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 \\ + \frac{\lambda}{2}\sum_{n=1}^{3}||Z^{(n)}||_F^2 + \sum_{n=1}^{3}\alpha_n tr(U^{(n)^T}\mathcal{L}_n\ U^{(n)}) &+ \frac{1}{2}||\boldsymbol{\Omega} * \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{T}}||_{\mathbf{F}}^{\mathbf{2}} \\ + \sum_{n=1}^{3} < \Lambda^{(n)}, U^{(n)} - Z^{(n)} > &+ \sum_{n=1}^{3}\frac{\mu}{2}||U^{(n)} - Z^{(n)}||_F^2 \end{aligned} \tag{4.9}$$

where Λ is a dual variable, $\langle ., . \rangle$ denote the inner product, $\|.\|_F$ is the Frobenius norm and $\mu > 0$ is a penalty parameter.

To solve the problem in Eq. 4.9 at each iteration t, ADMM updates the variables in alternating fashion as:

$$U_{t+1}^{(n)} = \arg\min_{U^{(n)}} L_\mu(U_t, Z_t, \Lambda_t, \boldsymbol{\mathcal{X}}_t) \tag{4.10}$$

$$Z_{t+1}^{(n)} = \arg\min_{Z^{(n)}} L_\mu(U_{t+1}, Z_t, \Lambda_t, \boldsymbol{\mathcal{X}}_t) \tag{4.11}$$

$$\boldsymbol{\mathcal{X}}_{t+1} = \arg\min_{\boldsymbol{\mathcal{X}}} L_\mu(U_{t+1}, Z_{t+1}, \Lambda_{t+1}, \boldsymbol{\mathcal{X}}_t) \tag{4.12}$$

$$\Lambda^{(n)}_{t+1} = \arg\min_{\Lambda} L_\mu(U_{t+1}, Z_{t+1}, \Lambda_t, \boldsymbol{\mathcal{X}}_t) \tag{4.13}$$

In the following, we present the derivation of specific update rules for Eq. 4.11, Eq. 4.10, Eq. 4.13 and Eq.4.18.

Update $U^{(n)}$ when fixing others: To update $U^{(n)}$ (e.g., $U^{(1)}$ or $U^{(2)}$ or $U^{(3)}$) after ignoring the variables that are irrelevant to $U^{(n)}$, the problem (4.11) becomes:

$$\min_{U^{(n)}} \ \alpha_n tr(U^{(n)^T} \mathcal{L}\ U^{(n)}) + < \Lambda^{(n)}, U^{(n)} - Z^{(n)} > + \frac{\mu}{2} ||U^{(n)} - Z^{(n)}||_F^2$$

On combining both linear and quadratic error terms into a single term by scaling the dual variable Λ, we get the following form :

$$\min_{U^{(n)}} \ \alpha_n tr(U^{(n)^T} \mathcal{L}_n\ U^{(n)}) + \frac{\mu}{2} \|U^{(n)} - Z^{(n)} + \Lambda^{(n)}/\mu\|_F^2 \tag{4.14}$$

Note that, it is a convex quadratic problem. Solving for $U^{(n)}$ yields:

$$U^{(n)}_{t+1} = (\mu I + \alpha_n \mathcal{L}_n)^{-1} (\mu U^{(n)}_t - \Lambda^{(n)}_t) \tag{4.15}$$

Update $Z^{(n)}$ when fixing others: To update $Z^{(n)}$ ($n \in 1, 2, 3$), the method alternates among the modes, fixing every factor matrix but $Z^{(n)}$ and solving for it. The objective function can be written as follows:

$$\begin{aligned} \min_{Z^{(n)}} \ & \frac{1}{2} ||\mathcal{X}_{(n)} - Z^{(n)} A^{(n)^T}||_F^2 + \frac{\lambda}{2} ||Z^{(n)}||_F^2 \\ & + \frac{\mu}{2} \|U^{(n)} - Z^{(n)} + \Lambda^{(n)}/\mu\|_F^2 \end{aligned} \tag{4.16}$$

Here, $\mathcal{X}_{(n)}$ represents the mode-n matrix unfolding of tensor $\boldsymbol{\mathcal{X}}$. the mode-n matricization of $\tilde{\boldsymbol{\mathcal{X}}}$ can be written in terms the factor matrices as $\tilde{\mathcal{X}_{(n)}} = Z^{(n)} A^{(n)^T}$ where $A^{(n)} = (Z^{(M)} \odot$

$..Z^{(n+1)} \odot Z^{(n-1)} \odot .. \odot Z^{(1)})|_{M=3}$. Here, $\odot$ denotes Khatri-Rao product. Now, solving Eq. 4.16 for $Z^{(n)}$ yields:

$$Z_{t+1}^{(n)} = (A^{(n)}\ A^{(n)^T} + \lambda I + \mu I)^{-1}\ (\mathcal{X}_{(n)}^{t}\ A^{(n)^T} + \mu U_{t+1}^{(n)} + \Lambda_t^{(n)}) \tag{4.17}$$

Update $\mathcal{X}$: To solve for $\boldsymbol{\mathcal{X}}$, we can write the objective in Eq. 4.9 as follows:

$$\min_{\boldsymbol{\mathcal{X}}}\ \frac{1}{2}||\boldsymbol{\mathcal{X}} - [[Z^{(1)}, Z^{(2)}, Z^{(3)}]]\ ||_F^2 + \frac{1}{2}||\boldsymbol{\Omega} * \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{T}}||_{\mathbf{F}}^{\mathbf{2}} \tag{4.18}$$

Now solving for $\boldsymbol{\mathcal{X}}$ yields:

$$\boldsymbol{\mathcal{X}}_{t+1} = \boldsymbol{\mathcal{T}} + (1 - \boldsymbol{\Omega}) * [[\mathbf{Z^{(1)}}, \mathbf{Z^{(2)}}, \mathbf{Z^{(3)}}]] \tag{4.19}$$

Update $\Lambda^{(n)}$: Having (U, Z) fixed, perform a gradient ascent update with step size of μ on the Lagrange multipliers as

$$\Lambda_{t+1}^{(n)} = \Lambda_t^{(n)} + \mu(U_{t+1}^{(n)} - Z_{t+1}^{(n)}) \tag{4.20}$$

The overall ADMM procedure is shown in Algo. 1. After convergence, we have the final completed tensor $\boldsymbol{\mathcal{X}}$. From $\boldsymbol{\mathcal{X}}$, we can recover the tags for our web image collection. Summing $\boldsymbol{\mathcal{X}}$ over dataset image dimensions, we can have a matrix whose values indicate the strength of association between web images and tags.

4.5 Experiments

We perform experiments on two standard benchmark datasets with the main goal of analyzing the performance of different supervised methods by utilizing large scale web data using our curriculum guided webly supervised approach. Ideally, we would expect an

Algorithm 1 An ADMM solver for (Eq. 4.9)

Input: $\boldsymbol{\mathcal{T}}$, $\boldsymbol{\Omega}$, $\Theta^{(n)}$ and λ, $N = 3$, $n = 1 : N$, $\mu > 0$, $Th = 10^{-5}$, $nIter = 1000$

Initialization: Initialize $U^{(n)}, Z^{(n)}, \Lambda^{(n)}, iter$ to zero, $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{T}}$.

while $(\max\{||Z^{(n)} - U^{(n)}||_F; n = 1, ..., N\} < Th)$ **or** $(iter \leq nIter)$ **do**

$U_{t+1}^{(n)} \leftarrow (\mu I + \alpha_n \mathcal{L}_n)^{-1}(\mu Z_t^{(n)} - \Lambda_t^{(n)})$;

$Z_{t+1}^{(n)} \leftarrow (A^{(n)}\ A^{(n)^T} + \lambda I + \mu I)^{-1}\ (\mathcal{X}_{(n)}^t\ A^{(n)^T} + \mu U_{t+1}^{(n)} + \Lambda_t^{(n)})$;

$\boldsymbol{\mathcal{X}}_{t+1} \leftarrow \boldsymbol{\mathcal{T}} + (1 - \boldsymbol{\Omega}) * [[\mathbf{Z^{(1)}}, \mathbf{Z^{(2)}}, \mathbf{Z^{(3)}}]]$;

$\Lambda_{t+1}^{(n)} \leftarrow \Lambda_t^{(n)} + \mu(U_{t+1}^{(n)} - Z_{t+1}^{(n)})$;

$iter \leftarrow iter + 1$;

end while

Output: Tensor $\boldsymbol{\mathcal{X}}$, Factor Matrices $Z^{(1)}$, $Z^{(2)}$ and $Z^{(3)}$.

improvement in performance irrespective of the loss function and features used to learn the embedding in Sec. 4.3.

We first describe the details on the datasets, evaluation metric and training details in Sec. 4.5.1. We report the results of different methods on MSCOCO dataset in Sec. 4.5.2 and results on Flickr30K dataset in Sec. 4.5.2.

4.5.1 Datasets and Implementation Details

We present cros-modal retrieval experiments on standard benchmark datasets containing images with corresponding text descriptions: MSCOCO [20] and Flickr30K [106] to evaluate the performance of proposed framework.

MSCOCO. The MSCOCO is a large-scale image description dataset. This is the largest image captioning dataset in terms of the number of sentences and the size of the

vocabulary. This dataset contains around 123K images. Each image comes with 5 captions. Following [57], we use the training, testing and validation split. In this split, the training set contains 82, 783 images, 5000 validation images and 5000 test images. However, there are also 30, 504 images from the original validation set of MS-COCO which have been left out in this split. We refer to this set as restval(RV). Some papers use RV with training set for training to improve accuracy. We report results using RV. In most of the previous works, the results are reported by averaging over 5 folds of 1K test images [64, 144, 26].

Flickr30K. Flickr30K is another very popular image description dataset. Flickr30K has a standard 31, 783 images for training. Each image comes with 5 captions, annotated by AMT workers. We follow the dataset division provided in [57]. In this dataset split, the training set contains 29,000 images, 1000 validation images and 1000 test images.

Web Image Collection. We use photo-sharing website Flickr to retrieve web images with tags and use those images without any additional manual labeling. To collect images, we create a list of 1000 most occurring keywords in MSCOCO and Flickr30K dataset text descriptions and sort them in descending order based on frequency. We remove stop-words and group similar words together after performing lemmatization. We then use this list of keywords to query Flickr and retrieve around 200 images per query, together with their tags. In this way, we collect about 210,000 images with tags. We only collect images having at least two English tags and we don't collect more than 5 images from a single owner. We also utilize first 5 tags to remove duplicate images.

Evaluation Metric. We use the standard evaluation criteria used in most prior work on image-text retrieval task [64, 28, 25]. We measure rank-based performance

by $R@K$ and Median Rank($MedR$). $R@K$ (Recall at K) calculates the percentage of test samples for which the correct result is ranked within the top-K retrieved results to the query sample. We report results for $R@1$ and $R@10$. Median Rank calculates the median of the ground-truth results in the ranking.

Training Details. We start training with a learning rate of 0.0002 and keep the learning rate fixed for 10 epochs. We then lower the learning rate by a factor of 10 every 10 epochs and continue training for 30 epochs. During updating the learned model in Stage I with web images in Stage II, we start training with a learning rate of 0.00002. The embedding networks are trained using ADAM optimizer [63]. Gradients are clipped when the $L2$ norm of the gradients(for the entire layer) exceeds 2. We tried different values for margin Δ in training and empirically choose Δ as 0.2, which we found performed well consistently on the datasets. We evaluate the model on the validation set after every epoch. The best model is chosen based on the sum of recalls in the validation set to deal with the over-fitting issue. We use a batch-size of 128 in the experiment. We also tried with other mini-batch sizes of 32 and 64 but didn't notice significant impact on the performance. We used two Telsa K80 GPUs and implemented the network using PyTorch toolkit.

4.5.2 Comparative Evaluations on Benchmark Datasets

Results on MSCOCO Dataset

We report the result of testing on MSCOCO dataset [77] in Table 4.1. To understand the effect of the proposed webly supervised approach, we divide the table in 3 rows (1.1-1.3). We compare our results with several representative image-text retrieval ap-

Table 4.1: Image-to-Text Retrieval Results on MSCOCO Dataset.

#	Method	Image-to-Text Retrieval			Text-to-Image Retrieval		
		R@1	R@10	Med R	R@1	R@10	Med R
1.1	Embedding-Net	54.9	92.2	-	43.3	87.5	-
	2Way-Net	55.8	-	-	39.7	-	-
	Sm-LSTM	53.2	91.5	1	40.7	87.4	2
	Order-Embedding	46.7	88.9	2	37.9	85.9	2
	SAE-VGG19	46.8	87.7	2	35.8	82.9	2.4
	SAE-ResNet152	59.2	95.2	1	44.7	88.4	2
1.2	VSE-VGG19	46.8	89	1.8	34.2	83.6	2.6
	VSEPP-VGG19	51.9	90.4	1	39.5	85.6	2
	VSE-ResNet152	52.7	91.8	1	36	85.5	2.2
	VSEPP-ResNet152	58.3	93.3	1	43.6	87.8	2
1.3	Ours (VSE-VGG19)	47.2	90.9	1.6	35.1	85.3	2
	Ours (VSEPP-VGG19)	53.7	92.5	1	41.2	89.7	2
	Ours (VSE-ResNet152)	52.9	94.3	1	42.2	89.1	2
	Ours (VSEPP-ResNet152)	61.5	96.1	1	46.3	89.4	2

proaches, Embedding-Net [144], 2Way-Net [26], Sm-LSTM [52], Order-Embedding [136], SAE [39], VSE [64] and VSEPP [28]. For these approaches, we directly cite scores from respective papers when available and select the score of the best performing method if score for multiple models are reported.

In row-1.2, we report the results on applying two different variants of ranking loss based baseline VSE and VSEPP with two different feature representation from [28]. VSE [64] is based on the triplet ranking loss similar to Eq. 4.1 and VSEPP[28] is based on the loss function that emphasizes on hard-negatives as shown in Eq. 4.2. We consider VSE and VSEPP loss based formulation as the baseline for this work. Finally, in row-1.3, results using the proposed approach are reported. To enable a fair comparison, we apply our webly supervised method using the same VSE and VSEPP loss used by methods in row-1.2.

Table 4.2: Image-to-Text Retrieval Results on Flickr30K Dataset.

#	Method	Image-to-Text Retrieval			Text-to-Image Retrieval		
		R@1	R@10	Med R	R@1	R@10	Med R
2.1	Embedding-Net	43.2	79.8	-	31.7	72.4	-
	2Way-Net	49.8	-		36	-	-
	Sm-LSTM	42.5	81.5	2	30.2	72.3	3
	Order-Embedding	43.8	83	2	32.7	73.9	4
	SAE VGG19	32.8	70.3	3	25.2	63.5	5
	SAE ResNet152	43.4	80.7	2	31	71.3	3
2.2	VSE -VGG19	29.8	71.9	3	23	61	6
	VSEPP -VGG19	31.9	68	4	26.8	66.8	4
	VSE-ResNet152	38.2	80.8	2	26.6	67	4
	VSEPP-ResNet152	43.7	82.1	2	32.3	72.1	3
2.3	Ours (VSE -VGG19)	32.4	74.1	3	24.9	64.3	5
	Ours(VSEPP -VGG19)	37.8	77.1	3	27.9	68.9	4
	Ours(VSE–ResNet152)	41.4	84.5	2	29.7	71.9	4
	Ours (VSEPP-ResNet152)	47.4	85.9	2	35.2	74.8	3

Effect of Proposed Webly Supervised Training. For evaluating the impact of our approach, we compare results reported in row-1.2 and row-1.3. Our method utilizes the same loss functions and features used in row-1.2 for a fair comparison. From Table 4.1, We observe that the proposed approach improves performance consistently in all the cases. For image-to-text retrieval task, the average performance increase in text-to-image retrieval is 7.5% in R@1 and 3.2% in R@10.

We also compare proposed approach with web supervised approach SAE[39] reported in row-1.1. In this regard, we implement SAE based webly supervised approach following [39] with our data. We use the same feature and VSEPP ranking loss for a fair comparison and follow the exact same settings for experiments. We observe that our approach consistently performs better.

Figure 4.5: Examples of 4 test images from Flickr30K dataset and the top 1 retrieved captions for our web supervised VSEPP-ResNet152 and standard VSEPP-ResNet as shown in Table. 4.2. The value in brackets is the rank of the highest ranked ground-truth caption in retrieval. Ground Truth (GT) is a sample from the ground-truth captions. Image 1,2 and 4 show a few examples where utilizing our approach helps to match the correct caption, compared to using the typical approach.

Effect of Loss Function. While evaluating the performance of different ranking loss, we observe that our webly supervised approach shows performance improvement for both VSE and VSEPP based formulation, and the performance improvement rate is similar for both VSE and VSEPP (See row-1.2 and row-1.3). Similar to the previous works [28, 158], we also find that methods using VSEPP loss performs better than VSE loss. We observe that in the image-to-text retrieval task, the performance improvement using VSEPP based formulation is higher and in the text-to-image retrieval task, the performance improvement for VSE based formulation is higher.

Effect of Feature. For evaluating the impact of different image feature in our web-supervised learning, we compare VGG19 feature based results with ResNet152 feature based results. We find consistent performance improvement using both VGG19 and ResNet152

feature. However, the performance improvement is slightly more when ResNet152 feature is used. In image-to-text retrieval, the average performance improvement in R@1 using ResNet152 feature is 4%, compared to 2.3% using VGG19 feature. In text-to-image retrieval task, the average performance improvement in R@1 using ResNet152 feature is 11.18%, compared to 3.5% using VGG19 feature.

Our webly supervised learning approach is agnostic to the choice loss function used for cross-modal feature fusion and we believe more sophisticated ones will only benefit our approach. We use two different variants of pairwise ranking loss (VSE and VSEPP) in the evaluation and observe that our approach improves the performance in both cases irrespective of the feature used to represent the images.

Results on Flickr30K Dataset

Table 4.2 summarizes the results on Flickr30K dataset [106]. Similar to Table 4.1, we divide the table in 3 rows (2.1-2.3) to understand the effect of the proposed approach compared to other approaches. From Table 4.2, we have the following key observations: (1) Similar to the results on MSCOCO dataset, our proposed approach consistently improves the performance of different supervised method(row-2.2 and row-2.3) in image-to-text retrieval by a margin of about 3%-6% in R@1 and 3%-9% in R@10. The maximum improvement of 6%-9% is observed in the VSEPP-VGG19 case while the least mean improvement of 4.8% is observed in VSE-VGG19 case. (2) In text-to-image retrieval task, the average performance improvement using our webly-supervised approach are 2.25% and 3.25% in R@1 and R@10 respectively. These improvements once again show that learning by utilizing large scale web data covering a wide variety of concepts lead to a robust embedding for

cross-modal retrieval tasks. In Fig. 4.5, we show examples of few test images from Flickr30K dataset and the top 1 retrieved captions for the VSEPP-ResNet152 based formulations.

4.5.3 Comparative Evaluation with Image-Tag Refinement

In this section, we first provide details about data preparation and implementation details related to image-teag refinement experiments. Then, we provide experimental results on Flickr30K and MSCOCO dataset to evaluate the impact of the refinement step on the final performance.

Data Preparation. We are interested in estimating the influence of noisy or missing tags on the performance of our approach. However, it is very difficult to collect a large number of web images with tags and label them. Hence, we create synthetic data based on image-text pairs from datasets (e.g., Flickr30K) to evaluate the effect of our image-tag refinement approach. First, we create a synthetic clean image-tag dataset from the training sets of the datasets. For each image, we collect unique nouns and verbs as image tags from the associated 5 sentences. We retain only the top 1000 occurring words in the train set.

We then create a noisy image-tag dataset from the synthetic clean set based on the missing ratio of tags (e.g., 30%, 50%, 70%) we would like to consider in evaluating the approach. In this regard, given a missing (%) we randomly select the overall number of tags to be replaced. We remove most of the tags and replace a few tags with random English words from the dictionary. In this way, we create several noisy image-tag datasets based on different missing ratios. These noisy image-tag datasets are considered as our observed set. From the synthetic clean Image-Tag datasets, we utilize the first 1K images from the

training set as our small clean image-text set as D in tensor completion (1000=1000). The noisy image-tag dataset is created from the remaining training images and these images are considered as images from web W. The top 1000 occurring words in the training set is considered as the tags set T ($|T|$=1000).

Implementation Details. The tensor completion approach is implemented using Matlab tensor toolbox [6, 1]. In the constructed observed tensor $\mathcal{T}$, we only know the observed non-zero entries. However, we do not have any prior information about zero entries whether they are missing or not relevant. However, for a good reconstruction of the tensor, a certain amount of observed entries is often required [125]. We randomly sample zeros from remaining entries to have an equal observed ratio as non-zeros. We vary tensor rank from 10 to 20, and empirically fix the rank as 20, which we found to be consistently performing well in terms of lower relative standard error in tensor completion. We utilize ranking loss function in Eq. 4.2 in training joint embedding models.

Results Analysis

In this section, we report image-text retrieval results on Flickr30K dataset and MSCOCO dataset varying the percentage of missing tags. We also evaluate the proposed tensor completion approach based on relative error difference between the predicted tensor and the observed tensor.

Flickr30K Dataset. We report the image-to-text retrieval and text-to-image retrieval results on Flickr 30K Dataset in Table 4.3. To understand the effect of the proposed tag refinement approach in overall performance, we report performance varying the ratio

Table 4.3: This table presents the results on the Flickr30K dataset. Actual indicates the initial synthetic clean image-tag set created by extracting unique noun and verbs from captions associated with images as tags. Observed indicates the synthetic noisy web image-tag set constructed by removing tags based on a given missing ratio. Predicted indicates the refined image-tag set obtained by refining observed set applying the proposed tensor completion approach. Following [64, 28], we use VGG16 feature and VSEPP pairwise ranking loss for training joint embedding models.

Retrieval	Tag Quality	Missing Data (%) = 30				Missing Data (%) = 50				Missing Data (%) = 70			
		R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR
Image to	Actual (No Missing)	5.5	17.2	24.7	57.0	5.5	17.2	24.7	57.0	5.5	17.2	24.7	57.0
Text	Observed (Missing (%) of Actual)	4.7	12.9	18.7	97.0	1.4	5.5	9.3	280.0	1.8	5.4	7.7	365.0
Retrieval	**Predicted Set (Proposed)**	**4.3**	**14.8**	**21.2**	**79.0**	**2.8**	**8.9**	**14.7**	**186.0**	**2.5**	**9.0**	**14.7**	**163.0**
Text to	Actual Tags (No Missing)	2.8	9.4	14.7	137.0	2.8	9.4	14.7	137.0	2.8	9.4	14.7	137.0
Image	Observed (Missing (%) of Actual)	1.6	6.1	10.1	200.0	0.7	2.8	4.3	327.0	0.4	2.6	4.4	338.0
Retrieval	**Predicted Set (Proposed)**	**2.3**	**7.3**	**11.4**	**191.0**	**1.1**	**4.4**	**7.5**	**230.0**	**0.5**	**2.0**	**3.6**	**385.0**

Table 4.4: This table presents the results on the MSCOCO dataset. Similar to Table 4.2, we use VGG16 feature and VSEPP pairwise ranking loss for training joint embedding. In the Table, Actual indicates the initial synthetic clean image-tag set created by extracting unique noun and verbs from captions associated with images as tags. Observed indicates the synthetic noisy web image-tag set constructed by removing tags based on a given missing ratio. Predicted indicates the refined image-tag set obtained by refining observed set applying the proposed tensor completion approach.

Retrieval	Tag Quality	Missing Data (%) = 30				Missing Data (%) = 50				Missing Data (%) = 70			
		R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR
Image to	Actual (No Missing)	9.7	29.8	40.6	17.0	9.7	29.8	40.6	17.0	9.7	29.8	40.6	17.0
Text	Observed (Missing (%) of Actual)	8.8	27.4	37.5	20.0	8.6	23.7	33.7	27.0	3.8	13.4	19.3	136.0
Retrieval	**Predicted Set (Proposed)**	**9.7**	**27.4**	**40.0**	**19.0**	**9.2**	**24.8**	**35.4**	**25.0**	**6.8**	**19.7**	**28.9**	**34.0**
Text to	Actual Tags (No Missing)	6.0	19.7	30.2	35.0	6.0	19.7	30.2	35.0	6.0	19.7	30.2	35.0
Image	Observed (Missing (%) of Actual)	5.2	16.0	22.8	70.0	4.0	14.0	20.9	67.0	2.7	9.9	16.6	107.0
Retrieval	**Predicted Set (Proposed)**	**4.9**	**15.3**	**22.8**	**69.0**	**3.7**	**12.4**	**18.0**	**105.0**	**3.2**	**9.8**	**14.9**	**110.0**

of missing data. From Table 4.3, we have several key observations. First, we observe that the predicted set shows better performance compared to the observed set in almost all evaluation metrics. We find in case of 30% missing data, observed set performs better than predicted set in R@1 in the image to text retrieval. However, the predicted set shows performance improvement in the other metrics in both image-to-text and text-to-image retrieval. We observe similar improvement in case of other missing data ratios. Second, in image-to-text retrieval, we see that as we increase the percentage of missing data, the model learned using the predicted set performs significantly better than the model learned using the observed set. Initially, in the case of 30% missing data, the performance of predicted and observed set is comparable. As the missing data percentage increases, the observed set shows a significant drop in performance which is expected. However, the predicted set is able to limit the performance drop by recovering some related tags.

Results on MSCOCO Dataset. Table 4.4 summarizes the image-to-text retrieval and text-to-image retrieval results on the MSCOCO dataset. Similar to Table 4.2, we compare retrieval results based on the joint embedding models trained using the actual set, the observed set, and the predicted set. It is evident from the Table that our proposed tag refinement approach helps to improve performance over directly using images with raw tags (observed set). As expected, the performance drops for both observed set and predicted set as the percentage of missing entries from the actual set increase. Similar to the observations from Table 4.3, we again see that in case of a low missing ratio in the observed set, the observed and predicted set shows comparable performance. However, the performance of the prediction method is very promising as it shows significant improvement compared to the observed set when the missing data percentage is high (70%). We see a 3% absolute

Table 4.5: Relative errors for recovering missing tags (before and after tensor completion) for different percentage of missing entries. We observe that the predicted tensor gives on average 11.4% improvement over the observed tensor

	Flickr30K			MSCOCO		
	30%	50%	70%	30%	50%	70%
Observed	0.563	0.721	0.8391	0.534	0.703	0.838
Predicted	0.514	0.649	0.7621	0.463	0.635	0.751
Improvement	**9.53%**	**11.09%**	**10.10%**	**15.33%**	**10.71%**	**11.58%**

improvement in R@1 and 102 point decrease in median rank using the proposed approach in the image to text retrieval with 70% missing data.

Relative Errors in Tensor Completion

Relative Error is one of the most commonly used evaluation metrics in evaluating the performance of tensor completion algorithms. The relative error for predicted tensor is calculated by the standard error in tensor prediction (Frobenius norm difference between ground-truth tensor and the predicted tensor), divided by the Frobenius norm of the ground-truth tensor. The relative error for observed tensor is calculated in a similar way. In Table 4.5, we compare relative errors of predicted tensor and observed tensor for different percentage of missing entries (i.e., 30%, 50%, and 70%). From the Table 4.5, we find that the predicted tensor results in consistently decreasing the relative error significantly compared to the observed tensor across datasets and missing percentage. The average improvement using the proposed prediction approach in relative error is about 11.4%. The maximum improvement of 15.33% is observed in MSCOCO dataset with 30% missing data and the minimum improvement of 9.53% is observed in Flickr30K dataset with 30% missing data.

4.6 Conclusion

In this work, our goal is to leverage web images with tags to assist training robust image-text embedding models for target task of image-text retrieval that has limited labeled data. While recent image-text retrieval methods offer great promise by learning deep representations aligned across modalities, most of these methods are plagued by the issue of training with small-scale datasets covering a limited number of images with ground-truth sentences. Moreover, it is extremely expensive to create a larger dataset by annotating millions of images with sentences and may lead to a biased model. Inspired by the recent success of web-supervised learning in deep neural networks, we attempt to capitalize readily-available web images with noisy annotations to learn robust image-text joint representation. We propose a two-stage approach for the task that can augment a typical supervised pair-wise ranking loss based formulation with weakly-annotated web images to learn a more robust visual-semantic embedding. Experiments on two standard benchmark datasets demonstrate that our method achieves a significant performance gain in image-text retrieval compared to state-of-the-art approaches.

We also address the problem that directly using web images with raw tags in training may hurt the performance of the webly supervised approaches significantly when the ratio of missing tags is high and available clean labeled data is very limited. In this regard, we propose a CP decomposition based tensor completion approach to refine tags of web images by modeling the ternary inter-relation between the web image collection and the clean dataset images (based on associated tags) as a tensor and utilizing intra-modal similarity as side information to regularize the tensor completion problem. Our image tag

refinement approach combined with supervised image-text embedding approaches provide a way for improving the learning of joint embedding models in the presence of significant noise from web data and limited clean labeled data. Experiments on two benchmark image-text datasets with different percentage of missing data demonstrate that the proposed approach can successfully recover more than 10% missing data on average and consequently helps to achieve a consistent performance gain in cross-modal image-text retrieval task.

Chapter 5

Conclusions

5.1 Thesis Summary

One increasingly important problem for most computer vision tasks in the light of data-hungry deep neural network models is how to learn useful models with limited labeled training data. Developing robust models with a limited degree of supervision could be extremely useful for cross-modal visual-semantic retrieval tasks as collecting pairs of visual data and natural language description is extremely labor-intensive and prone to significant errors. However, developing effective algorithms with limited supervision is non-trivial and has been hardly explored for the problem of cross-modal retrieval between textual and visual queries. In this thesis, we explore several cross-modal vision-language retrieval tasks (i.e., image-text retrieval, video-text retrieval and text to video moment retrieval) focusing on developing efficient solutions leveraging available incidental signals or weak labels.

In Chapter 2, we present an efficient framework for cross-modal video-text retrieval utilizing three salient video cues (i.e., object, activity, place) simultaneously by a

mixture of expert joint embedding approach. In Chapter 3, we introduce a novel problem of learning from weak labels for the task of text to video moment retrieval and propose a joint embedding based framework that learns the notion of relevant segments from video using only video-level sentence descriptions without any temporal boundary annotations. In Chapter 4, we present a novel webly supervised joint visual-semantic embedding learning approach that provides a way to augment a typical supervised learning approach with weakly-supervised web data to learn robust joint embedding models. Experimental results show that our methods achieve excellent performance gain over existing approaches and baselines in standard benchmark datasets.

5.2 Future Research Directions

5.2.1 Cross-Modal Retrieval for Visual Localization

In this thesis, we mainly focus on vision-language retrieval tasks. However, our proposed approaches and ideas can be adapted to improve several other multi-modal retrieval and analysis tasks, e.g., cross-modal geo-localization. Developing approaches for cross-modal matching between images of different modality and viewpoint (e.g., ground to aerial image matching) would be very helpful for vision-based localization across autonomous platforms and can be an interesting future research direction.

5.2.2 Moment Retrieval using Text Queries from Video Collection

We have addressed the problem of retrieving matching videos from a database based on text description in Chapter 1 and retrieving moments in a long video using textual

queries in Chapter 2. One natural extension would be to retrieve relevant moments from a video collection using natural language queries. One simple baseline solution to the problem would be to utilize the method in Chapter 2 followed by the method in Chapter 3. Effectively localizing video moments from large untrimmed video collections is an interesting future direction of our work and can very helpful in many computer vision applications.

5.2.3 Tensor Embedding for Fusing Multimodal Cues

We have shown in Chapter 2 that integrating information from different video cues yields robust, and more effective retrieval performance compared to using a single cue. While we have explored fusion approaches (e.g., feature concatenation and late fusion) for fusing cues from visual data, these approaches can not model both intra-modal and inter-modal dynamics efficiently. Our proposed approach consists of training several joint embeddings independently and performing a decision voting which prevents the retrieval model from learning inter-modality dynamics in an efficient way. Developing a tensor fusion based embedding approaches can be an interesting and more comprehensive approach to model both intra-modal and inter-modal dynamics for more effective retrieval.

5.2.4 Text Description Generation with Active Learning

Recent advancements in visual-textual retrieval and analysis tasks have been plagued significantly by the challenging and labor-intensive nature of annotating images/videos with text descriptions. Hence, existing datasets have a limited number of labeled vision-language pairs, which makes it very difficult to develop effective retrieval systems by training deep neural network models. On the other hand, active learning approaches have been shown to

be very effective in constructing high-quality image data-sets with limited human labeling. However, prior approaches have mainly focused on the issue of annotating images with a single label. Active learning to generate natural language descriptions of visual data can be a very interesting and challenging future research direction.

Bibliography

[1] Evrim Acar, Daniel M. Dunlavy, Tamara G. Kolda, and Morten Mørup. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56, March 2011.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, pages 1247–1255, 2013.

[3] George Awad, Asad Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Martial Michel, Alan F Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, et al. Trecvid 2017: Evaluating ad-hoc and instance video search, events detection, video captioning and hyperlinking. In *Proceedings of TRECVID*. National Institute of Standards and Technology, 2017.

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 892–900, 2016.

[5] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[6] Brett W. Bader, Tamara G. Kolda, et al. Matlab tensor toolbox version 2.6. Available online, February 2015.

[7] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006.

[8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48. ACM, 2009.

[9] Joachim Bingel and Anders Søgaard. Identifying beneficial task relations for multi-task learning in deep neural networks. In *European Chapter of the Association for Computational Linguistics*, volume 2, pages 164–169, 2017.

[10] Rémi Bois, Vedran Vukotić, Anca-Roxana Simon, Ronan Sicre, Christian Raymond, Pascale Sébillot, and Guillaume Gravier. Exploiting multimodality in video hyperlinking to improve target diversity. In *International Conference on Multimedia Modeling*, pages 185–197. Springer, 2017.

[11] Piotr Bojanowski, Rémi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Weakly-supervised alignment of video with text. In *International Conference on Computer Vision (ICCV)*, 2015.

[12] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[13] Mateusz Budnik, Mikail Demirdelen, and Guillaume Gravier. A study on multimodal video hyperlinking with visual aggregation. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2018.

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *arXiv preprint arXiv:1705.07750*, 2017.

[15] Miriam Cha, Youngjune Gwon, and HT Kung. Multimodal sparse representation learning and applications. *arXiv preprint arXiv:1511.06238*, 2015.

[16] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1130–1139, 2018.

[17] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 190–200. Association for Computational Linguistics, 2011.

[18] Jingyuan Chen, Xinpeng Chen, Lin Ma, Zequn Jie, and Tat-Seng Chua. Temporally grounding natural sentence in video. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 162–171, 2018.

[19] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6298–6306, 2017.

[20] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[21] Jingze Chi and Yuxin Peng. Dual adversarial networks for zero-shot cross-media retrieval. In *IJCAI*, pages 663–669, 2018.

[22] Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Textually customized video summaries. *arXiv preprint arXiv:1702.01528*, 2017.

[23] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[24] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[25] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Word2visualvec: Image and video to sentence matching by visual feature prediction. *CoRR, abs/1604.06838*, 2016.

[26] Aviv Eisenschtat and Lior Wolf. Linking image and text with 2-way nets. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4601–4611, 2017.

[27] Ehsan Elhamifar, Guillermo Sapiro, and S Shankar Sastry. Dissimilarity-based sparse subset selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2182–2197, 2016.

[28] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improved visual-semantic embeddings. *British Machine Vision Conference (BMVC)*, 2018.

[29] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision (ECCV)*, pages 15–29. Springer, 2010.

[30] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *22nd ACM international conference on Multimedia*, pages 7–16. ACM, 2014.

[31] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012.

[32] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2121–2129, 2013.

[33] Paul Furgale and Timothy D Barfoot. Visual teach and repeat for long-range rover autonomy. *Journal of Field Robotics*, 27(5):534–560, 2010.

[34] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5277–5285, 2017.

[35] Hancheng Ge, James Caverlee, and Haokai Lu. Taper: A contextual tensor-based approach for personalized expert recommendation. In *ACM Conference on Recommender Systems*, pages 261–268. ACM, 2016.

[36] Hancheng Ge, James Caverlee, Nan Zhang, and Anna Squicciarini. Uncovering the spatio-temporal dynamics of memes in the presence of incomplete information. In *ACM International on Conference on Information and Knowledge Management*, pages 1493–1502. ACM, 2016.

[37] Dihong Gong, Daisy Zhe Wang, and Yang Peng. Multimodal learning for web information extraction. In *ACM International Conference on Multimedia (ACM MM)*, pages 288–296. ACM, 2017.

[38] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision (IJCV)*, 106(2):210–233, 2014.

[39] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision (ECCV)*, pages 529–545. Springer, 2014.

[40] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2712–2719, 2013.

[41] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[42] Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an explanatory multi-modal factor analysis. 1970.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[44] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5803–5812, 2017.

[45] Christian Andreas Henning and Ralph Ewerth. Estimating the information gap between textual and visual representations. In *International Conference on Multimedia Retrieval (ICMR)*, pages 14–22. ACM, 2017.

[46] Frank L Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1-4):164–189, 1927.

[47] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[48] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4555–4564, 2016.

[49] Zeyuan Hu and Julia Strout. Exploring stereotypes and biased data with the crowd. *arXiv preprint arXiv:1801.03261*, 2018.

[50] Feiran Huang, Xiaoming Zhang, Zhoujun Li, Tao Mei, Yueying He, and Zhonghua Zhao. Learning social image embedding with deep multimodal attention networks. In *Thematic Workshops of ACM Multimedia 2017*, pages 460–468. ACM, 2017.

[51] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269. IEEE, 2017.

[52] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2310–2318. IEEE, 2017.

[53] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3571–3580, 2017.

[54] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *International Conference on Multimedia Information Retrieval (ICMR)*, pages 39–43. ACM, 2008.

[55] Mark J Huiskes, Bart Thomee, and Michael S Lew. New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In *International Conference on Multimedia Retrieval (ICMR)*, pages 527–536. ACM, 2010.

[56] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision (ECCV)*, pages 67–84. Springer, 2016.

[57] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137. IEEE, 2015.

[58] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1889–1897, 2014.

[59] Ronald Kemker, Angelina Abitino, Marc McClure, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. *arXiv preprint arXiv:1708.02072*, 2017.

[60] Tom Kenter, Alexey Borisov, and Maarten de Rijke. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*, 2016.

[61] Tom Kenter and Maarten de Rijke. Short text similarity with word embeddings. In *ACM Int. Conf. Information and Knowledge Management*, pages 1411–1420, 2015.

[62] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision (ECCV)*, pages 158–171. Springer, 2012.

[63] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[64] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[65] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3294–3302, 2015.

[66] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4437–4446. IEEE, 2015.

[67] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[68] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision (ECCV)*, pages 301–320. Springer, 2016.

[69] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012.

[70] Joon Ho Lee. Analyses of multiple evidence combination. In *ACM SIGIR Forum*, volume 31, pages 267–276. ACM, 1997.

[71] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*, 2018.

[72] Ang Li, Allan Jabri, Armand Joulin, and Laurens van der Maaten. Learning visual n-grams from web data. In *International Conference on Computer Vision (ICCV)*, 2017.

[73] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Attention transfer from web images for video recognition. In *ACM International Conference on Multimedia (ACM MM)*, pages 1–9. ACM, 2017.

[74] Qin Li, Ke Li, Xiong You, Shuhui Bu, and Zhenbao Liu. Place recognition based on deep feature and adaptive weighting of similarity matrix. *Neurocomputing*, 199:114–127, 2016.

[75] Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM transactions on graphics (TOG)*, 34(6):234, 2015.

[76] Athanasios P Liavas and Nicholas D Sidiropoulos. Parallel algorithms for constrained tensor factorization via alternating direction method of multipliers. *IEEE Transactions on Signal Processing*, 63(20):5450–5463, 2015.

[77] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[78] Jialu Liu. Image retrieval based on bag-of-words model. *arXiv preprint arXiv:1304.5168*, 2013.

[79] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *ACM International Conference on Multimedia (ACMMM)*, 2018.

[80] Yuanyuan Liu, Fanhua Shang, Licheng Jiao, James Cheng, and Hong Cheng. Trace norm regularized candecomp/parafac decomposition with missing data. *IEEE Transactions on Cybernetics*, 45(11):2437–2448, 2015.

[81] David G Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, 1999.

[82] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016.

[83] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1942–1950, 2016.

[84] Zhuang Ma, Yichao Lu, and Dean Foster. Finding linear structure in large datasets with scalable canonical correlation analysis. In *International Conference on Machine Learning (ICML)*, pages 169–178, 2015.

[85] R Manmatha, Chao-Yuan Wu, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2859–2867. IEEE, 2017.

[86] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[87] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[88] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3111–3119, 2013.

[89] Niluthpol C Mithun, Cody Simons, Robert Casey, Stefan Hilligardt, and Amit Roy-Chowdhury. Learning long-term invariant features for vision-based localization. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2038–2047. IEEE, 2018.

[90] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *International Conference on Multimedia Retrieval (ICMR)*, pages 19–27. ACM, 2018.

[91] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Joint embeddings with multimodal cues for video-text retrieval. *International Journal of Multimedia Information Retrieval*, pages 1–16, 2019.

[92] Niluthpol Chowdhury Mithun, Sirajum Munir, Karen Guo, and Charles Shelton. Odds: real-time object detection using depth sensors on embedded gpus. In *ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pages 230–241. IEEE Press, 2018.

[93] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos Papalexakis, and Amit Roy-Chowdhury. Webly supervised joint embedding for cross-modal image-text retrieval. In *ACM International Conference on Multimedia*, 2018.

[94] Niluthpol Chowdhury Mithun, Rameswar Panda, and Amit K Roy-Chowdhury. Generating diverse image datasets with limited labeling. In *ACM International Conference on Multimedia (ACM MM)*, pages 566–570. ACM, 2016.

[95] Niluthpol Chowdhury Mithun, Rameswar Panda, and Amit K Roy-Chowdhury. Construction of diverse image datasets from web collections with limited labeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[96] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 299–307, 2017.

[97] Atsuhiro Narita, Kohei Hayashi, Ryota Tomioka, and Hisashi Kashima. Tensor factorization using auxiliary information. *Data Mining and Knowledge Discovery*, 25(2):298–324, 2012.

[98] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6752–6761, 2018.

[99] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision (ECCV)*, pages 651–667. Springer, 2016.

[100] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4594–4602, 2016.

[101] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K Roy-Chowdhury. Weakly supervised summarization of web videos. In *International Conference on Computer Vision (ICCV)*, pages 3677–3686. IEEE, 2017.

[102] Rameswar Panda, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. Diversity-aware multi-video summarization. *IEEE Transactions on Image Processing*, 26(10):4712–4724, 2017.

[103] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS 2017 Workshop*, 2017.

[104] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *European Conference on Computer Vision (ECCV)*, pages 588–607. Springer, 2018.

[105] Bryan Plummer, Matthew Brown, and Svetlana Lazebnik. Enhancing video summarization via vision-language embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[106] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *International Conference on Computer Vision (ICCV)*, pages 2641–2649. IEEE, 2015.

[107] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[108] Robi Polikar. Bootstrap inspired techniques in computational intelligence: ensemble of classifiers, incremental learning, data fusion and missing features. *IEEE Signal Processing Magazine*, 24(4):59–72, 2007.

[109] Thi Quynh Nhi Tran, Hervé Le Borgne, and Michel Crucianu. Aggregating image and text quantized correlated components. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2046–2054, 2016.

[110] Dimitrios Rafailidis and Alexandros Nanopoulos. Modeling users preference dynamics and side information in recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(6):782–792, 2016.

[111] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. Multimodal video description. In *ACM International Conference on Multimedia (ACMMM)*, pages 1092–1096. ACM, 2016.

[112] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788. IEEE, 2016.

[113] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 91–99, 2015.

[114] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[115] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[116] Jitao Sang, Jing Liu, and Changsheng Xu. Exploiting user information for image tag refinement. In *ACM International Conference on Multimedia (ACMMM)*, pages 1129–1132. ACM, 2011.

[117] Jitao Sang, Changsheng Xu, and Jing Liu. User-aware image tag refinement via ternary semantic analysis. *IEEE Transactions on Multimedia*, 14(3):883–895, 2012.

[118] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823. IEEE, 2015.

[119] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H Lampert. Learning to rank using privileged information. In *International Conference on Computer Vision (ICCV)*, pages 825–832. IEEE, 2013.

[120] Nicholas D Sidiropoulos, Lieven De Lathauwer, Xiao Fu, Kejun Huang, Evangelos E Papalexakis, and Christos Faloutsos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582, 2017.

[121] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, 2016.

[122] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[123] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1961–1970, 2016.

[124] Richard Socher and Li Fei-Fei. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 966–973. IEEE, 2010.

[125] Qingquan Song, Hancheng Ge, James Caverlee, and Xia Hu. Tensor completion algorithms in big data analytics. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 13(1):6, 2019.

[126] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

[127] Chen Sun, Sanketh Shetty, Rahul Sukthankar, and Ram Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *ACM International Conference on Multimedia (ACM MM)*, pages 371–380. ACM, 2015.

[128] Jinhui Tang, Xiangbo Shu, Guo-Jun Qi, Zechao Li, Meng Wang, Shuicheng Yan, and Ramesh Jain. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1662–1674, 2017.

[129] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016.

[130] Antonio Torralba, Alexei Efros, et al. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528. IEEE, 2011.

[131] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[132] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[133] Nicolas Usunier, David Buffoni, and Patrick Gallinari. Ranking with ordered weighted pairwise classification. In *International Conference on Machine Learning (ICML)*, pages 1057–1064. ACM, 2009.

[134] Emiel van Miltenburg. Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*, 2016.

[135] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

[136] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

[137] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4534–4542, 2015.

[138] Vedran Vukotić, Christian Raymond, and Guillaume Gravier. Bidirectional joint representation learning with symmetrical deep neural networks for multimodal and crossmodal applications. In *ACM International Conference on Multimedia Retrieval (ICMR)*, pages 343–346. ACM, 2016.

[139] Vedran Vukotić, Christian Raymond, and Guillaume Gravier. Generative adversarial networks for multimodal representation learning in video hyperlinking. In *2017 ACM on International Conference on Multimedia Retrieval (ICMR)*, pages 416–419. ACM, 2017.

[140] Vedran Vukotić, Christian Raymond, and Guillaume Gravier. A crossmodal approach to multimodal fusion in video hyperlinking. *IEEE MultiMedia*, 25(2):11–23, 2018.

[141] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACM International Conference on Multimedia (ACMMM)*, pages 154–162. ACM, 2017.

[142] Hua Wang, Feiping Nie, and Heng Huang. Low-rank tensor completion with spatio-temporal consistency. In *AAAI Conference on Artificial Intelligence*, 2014.

[143] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[144] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[145] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5005–5013. IEEE, 2016.

[146] Jason Weston, Samy Bengio, and Nicolas Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, volume 11, pages 2764–2770, 2011.

[147] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: region convolutional 3d network for temporal activity detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5794–5803, 2017.

[148] Huijuan Xu, Kun He, Leonid Sigal, Stan Sclaroff, and Kate Saenko. Text-to-clip video retrieval with early fusion and re-captioning. *arXiv preprint arXiv:1804.05113v1*, 2018.

[149] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016.

[150] Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, volume 5, page 6, 2015.

[151] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3441–3450, 2015.

[152] Rong Yan, Jun Yang, and Alexander G Hauptmann. Learning query-class dependent weights in automatic video retrieval. In *ACM International Conference on Multimedia (ACMMM)*, pages 548–555. ACM, 2004.

[153] Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*, 2014.

[154] Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian. Multi-networks joint learning for large-scale cross-modal retrieval. In *ACM International Conference on Multimedia (ACMMM)*, pages 907–915. ACM, 2017.

[155] Xishan Zhang, Ke Gao, Yongdong Zhang, Dongming Zhang, Jintao Li, and Qi Tian. Task-driven dynamic fusion: Reducing ambiguity in video description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3713–3721, 2017.

[156] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[157] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2914–2923, 2017.

[158] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. Dual-path convolutional image-text embedding. *arXiv preprint arXiv:1711.05535*, 2017.

[159] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[160] Huibin Zhou, Dafang Zhang, Kun Xie, and Yuxiang Chen. Spatio-temporal tensor completion for imputing missing internet traffic data. In *International Conference on Performance Computing and Communications Conference*, pages 1–7. IEEE, 2015.