Generative Artificial Intelligence for Behavioral Intent Prediction

Willa M. Mannering, Noah Ford, Justin Harsono, and John Winder
Johns Hopkins University Applied Physics Laboratory, Laurel MD
[willa.mannering][noah.ford][justin.harsono][john.winder]@jhuapl.edu

Abstract

Theory of mind is an essential ability for complex social interaction and collaboration. Researchers in cognitive science and psychology have previously sought to integrate theory of mind capabilities into artificial intelligence (AI) agents to improve collaborative abilities (Cuzzolin, Morelli, Cirstea, & Sahakian, 2020). We introduce the Recurrent Conditional Variational Autoencoder (RCVAE), a novel model which leverages the ability of generative models to learn rich abstracted representations of contextual behaviors to predict behavioral intent from human behavioral trajectories. Advancing on current concept learning models, this model allows for the discovery of latent intent in human behavior trajectories, while maintaining the scalability and performance of generative AI models. We show that in the Overcooked-AI environment, the RCVAE outperforms baseline Long Short-Term Memory (LSTM) models in predicting intent, achieving higher prediction accuracy and greater predictive stability. The implications of these results are significant; the RCVAE's proficiency in learning the relationship between basic actions and resulting contextual behaviors represents a significant advancement in concept learning for behavioral intent prediction.

Keywords: generative AI, theory of mind, imitation learning, intent prediction, variational autoencoder

Introduction

Theory of mind, the ability to predict and explain another person's actions in terms of internal mental states such as beliefs and desires, is a fundamental human ability responsible for complex social interaction and cooperative abilities (Baker, Saxe, & Tenenbaum, 2011; Perner, 1991). Theory of mind has been extensively researched in the cognitive science and psychology fields (Apperly, 2010; Wellman, Cross, & Watson, 2001) and equipping AI agents with theory of mind capabilities is becoming an increasingly popular approach in the field of machine learning (Fuchs, Walton, Chadwick, & Lange, 2021; Rabinowitz et al., 2018).

Recently, generative AI models have experienced a surge in popularity, with their impressive ability to create realistic and novel content. This increase in popularity can be at least partially attributed to the release of models such as OpenAI's ChatGPT (Brown et al., 2020) and DALL·E (Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022), which have showcased the potential of generative algorithms in diverse domains, from natural language processing to image synthesis. Beyond their creative capabilities, these models have the ability to learn rich latent representations, similar to earlier deep learning models (Engelcke, Kosiorek, Parker Jones, & Posner, 2020; Ye & Bors, 2021). Cognitive scientists have long leveraged learned latent representations to gain insights into human cognitive processes (Hills, Jones, & Todd, 2012; Jones, 2016; Kumar, Steyvers, & Balota, 2022). Thus, being at the forefront of AI research, generative AI models have the potential to provide novel insights and advanced capabilities for computational modeling in the field of cognitive science.

In this paper, we leverage the ability of generative models to learn rich abstracted representations of contextual behaviors, and introduce a novel model, the Recurrent Conditional Variational Autoencoder (RCVAE). Advancing on current concept learning models, this model allows for the discovery of latent intent in human behavior trajectories, while maintaining the scalability and performance of generative AI models. We evaluate the RCVAE's ability to predict behavioral intent when compared to established baseline models. We find that the RCVAE can predict intent with higher accuracy and consistency, paving the way for real-time intent prediction in cooperative multi-agent environments. This is a critical step towards imparting AI with theory of mind capabilities, essential for understanding the intentions of collaborative partners.

Related Work

Machine Theory of Mind

A popular method for integrating theory of mind in AI is through concept learning, which enables AI agents to comprehend and utilize human-understandable concepts (Oguntola, Campbell, Stepputtis, & Sycara, 2023; Grupen, Jaques, Kim, & Omidshafiei, 2022). In this context, a "concept" is an abstract behavior that is meaningful to humans but is not necessarily understandable to AI. Concept learning techniques allow AI models to utilize meaningful ideas which enable them to interpret and predict the beliefs and behavior of human partners.

A variety of models for concept learning, including concept whitening (Chen, Bei, & Rudin, 2020), concept bottleneck (Koh et al., 2020), and concept embedding models (Zarlenga et al., 2022), have been used previously to integrate theory of mind reasoning into AI. However, these models have primarily been used for predicting concepts present in static images. We aim to adapt the idea of concept learning models to understand and predict behavioral concepts

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

from sequential trajectory data. This advancement will aid AI agents in better understanding human behavior as it changes throughout a behavioral trajectory.

Imitation Learning

Imitation learning has recently attracted interest with an increase in real-world applications, with researchers using the technique for playing video games, driving autonomous vehicles, and training assistive robots (Hussein, Gaber, Elyan, & Jayne, 2017). In the imitation learning paradigm, agents observe expert trajectories in some task (such as driving a car) and attempt to develop a policy that replicates expert behavior. Imitation learning contrasts with reinforcement learning, where the objective is to learn a policy that maximizes a predefined reward function. An advantage of imitation learning is that it is not required to hand craft reward functions as learning relies on expert behavior data. This property makes it easier to scale up to real-world tasks, especially in scenarios where gathering expert behavior data is possible. The method described in this paper falls into the category of behavioral cloning (Torabi, Warnell, & Stone, 2018), with additional steps applied to extract contextual behavior predictions from the learned latent space of our model.

Variational Autoencoders

The Variational Autoencoder (VAE) (Kingma & Welling, 2013) is a class of generative model that extends the core concept of the autoencoder architecture to data generation. An autoencoder is typically comprised of two subnetworks, an encoder q_ϕ and a decoder p_θ (where ϕ and θ denote the parameterizations of the distributions q and p). The encoder transforms a sample of input data, x, into a latent representation, z. The decoder transforms the latent variables into an output, most often a reconstruction of the input. VAEs estimate mappings between distributions by incorporating auxiliary noise, denoted as ε, into the latent variables. The integration of noise encourages the model to map points nearby in the latent space, modeled by the distribution of ε, to similar reconstructions. This property enhances the model's ability to generalize from the training data and allows for smoother interpolations between different inputs (Kingma & Welling, 2013).

VAEs originally gained popularity due to their ability to produce diverse, high-quality data samples and to learn via unsupervised methods (Doersch, 2016). Their structured latent space allows for insightful data representation and manipulation, making them valuable in fields such as image generation, anomaly detection, and data analysis where understanding underlying patterns in the data is important. Considering their widespread use in AI applications, cognitive scientists should explore the benefits of utilizing these models for addressing computational modeling challenges within the field.

Modeling Intent Prediction

Recurrent Conditional Variational Autoencoder

The model introduced by this paper is a Recurrent Conditional Variational Autoencoder (RCVAE) which is a novel architecture designed to predict behavioral intent from sequential trajectory data. The RCVAE is a modification of the standard VAE that is designed to handle trajectory data rather than merely perform input reconstruction.

The Conditional Variational Autoencoder (CVAE) is a variation of the VAE architecture where the decoder predicts the output y conditioned on input x, instead of predicting decoding back to the input space, x. For this paper, x is the agent's environmental observations and y is the agent's action. Additionally, both the VAE and CVAE produce an intermediary latent variable, z. This paper will view z a representation of the higher-level task that an agent is performing.

The loss function for the CVAE is a slight variation of the VAE loss function displayed below:

$$\mathcal{L}_{CVAE}(x,y;\theta,\phi) = -\frac{1}{L}\sum_{l=1}^{L}\log p_\theta(y|x,z^{(l)}) + D_{KL}(q_\phi(z|x,y)||p_\theta(z|x)), \quad (1)$$

where $z^{(l)} \sim g_\phi(x,y,\varepsilon^{(l)})$, g is the noise-conditioned parameterization of the encoder distribution q, $\varepsilon^{(l)} \sim \mathcal{N}(0,I)$, D_{KL} is the Kullback-Leibler Divergence, and L is the batch size (Sohn, Lee, & Yan, 2015).

The RCVAE modifies the CVAE to handle trajectory data, allowing it to capture behavior and intent over time. We enhanced the architecture with recurrent layers, which allow the model to maintain information about previous states, and added a new behavioral term (see Equation 2) to the loss function, which acts as a latent space regularizer that conditions a portion of the latent space on contextual behaviors.

$$\mathcal{L}_{behavior}(x;\theta) = \frac{1}{L}\frac{1}{E}\sum_{l=1}^{L}\sum_{i=1}^{E}(z_i^{(l)} - e_i^{(l)})^2, \quad (2)$$

where $z^{(l)} \sim p_\theta(z|x)$, E is the number of behaviors, and $\{e^{(l)}\}_{i=1}^{E}$ is the one-hot encoding of the behaviors.

To learn a relationship between basic actions and contextual behaviors, we have introduced a divided training approach for the latent space. We enforce a separation of the latent space whereby half of the latent dimensions are trained using the behavioral loss and the remaining dimensions are trained exclusively on the conventional CVAE loss (Equation 1). The number of dimensions that receive the additional behavioral loss training will vary depending on the training environment.

To achieve balanced training, we employ a latent space segmentation scheme to help structure the latent space. In particular, when computing the loss, we only apply the behavioral loss to half of the latent space: for a latent z with h hidden dimension (latent features), the behavioral loss only computes mean squared error for $[z_1, z_{h/2}]$ elements. As a result, the first half of the latent space is fine-tuned for specific contextual behavior encoding, providing clear semantic interpretation, while the latter half ($[z_{(h/2)+1}, z_h]$) is trained to capture more generalized features of the basic action input data. Thus, the full loss for the RCVAE is shown in Equation 3:

$$\mathcal{L}_{RCVAE}(x, y; \theta, \phi) = \gamma \mathcal{L}_{behavior}(x; \theta) + \mathcal{L}_{CVAE}(x, y; \theta, \phi), \quad (3)$$

where γ is a hyperparameter for balancing the behavioral loss.

Extracting Meaningful Relationships

After training the RCVAE, we extract meaningful relationships between behaviors from the learned latent space of the model. To do this we utilize several unsupervised algorithms: Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, Saul, & Großberger, 2018), Density Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester, Kriegel, Sander, Xu, et al., 1996), and K-Nearest Neighbors (KNN). While the dimensionality of input data is reduced through the latent layer of the RCVAE during training, when representations of trajectories are extracted from the latent layer they are still relatively high dimensional. To further reduce the dimensionality of the latent variables we employ the UMAP algorithm. UMAP is a dimensionality reduction algorithm that generally preserves the global structure of the original data while also keeping similar data points together. Unlike simpler dimensionality reduction algorithms, UMAP has higher performance on non-linear data and manifolds. Once UMAP reduces the dimensionality of the latent variables to an interpretable size (2 or 3 dimensions), DBSCAN is used to identify potential clusters. To facilitate real-time behavioral intent prediction with our trained model, the KNN algorithm is used for online clustering where new observations are passed through the RCVAE. Then, the latents are extracted and undergo dimensionality reduction via the saved UMAP model. Finally, KNN associates the new data points with the distinct clusters identified by DBSCAN. The full pipeline for our proposed method of intent prediction is shown in Figure 1.

Experiments

Training Environment

We use the Overcooked-AI environment[1] developed by Carroll et al. (2019), a dynamic and interactive platform inspired by the cooperative cooking game Overcooked. The environment is characterized by several kitchen layouts where AI agents are tasked with preparing and serving a variety of dishes under time constraints. The action space for agents in the environment is discretely defined, comprising six possible basic actions: (move) up, down, right, left, wait (to remain stationary), and interact (to carry or drop objects). The observation space is structured as an 11x5 grid with 26 different channels. The contextual behavior space consists of eight possible events: tomato dropoff, tomato pickup, onion dropoff, onion pickup, dish dropoff, dish pickup, soup dropoff, and soup pickup.

Importantly, we distinguish between the six low-level, atomic "actions" (up, down, left, right, interact, wait) defined by the environment, and the eight high-level "contextual behaviors" that we defined to capture semantic events. Behaviors emerge as a combination of the basic actions and the surrounding game context. For example, at the beginning of the game, the players are empty-handed. If the player first directs their avatar to move up then interact near a tomato, the behavior "tomato pickup" will occur. However, if the avatar is holding a tomato already and the player chooses the interact action, the "tomato dropoff" behavior will occur. Therefore, by executing the same interact action the player can trigger different behaviors depending on the current context of the game environment (where the avatar is located, whether they are holding an ingredient already or not, etc.)

Model Evaluation

To tailor the RCVAE to the Overcooked environment, we configured the model with a 16-dimensional latent space ($h = 16$), where the first eight dimensions of the latent space are trained to represent the eight possible behaviors in the Overcooked-AI environment and the other eight dimensions remain flexible to learn additional latent encodings that may be important for task performance. We enforce this separation of the latent space by using a mean squared error loss to fit the first eight dimensions of the latent space to predict the one-hot encoding of the eight behaviors.

To evaluate the behavioral intent prediction ability of the RCVAE, we compare our model to a baseline Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) predictor model which is trained specifically to predict behaviors. An LSTM is a type of recurrent neural network designed to learn sequential information that may vary over distance scales, such as time series or text data. Our LSTM model encodes the environmental observations into a latent space, passes these vectors to an LSTM, and decodes the output into an 8-dimensional vector. The model is trained using a mean squared error loss to predict the most likely behavior out of eight. Meanwhile, the RCVAE is trained to predict the basic actions and the latent space is additionally conditioned on the contextual behaviors. Our goal is to determine whether learning a relationship between basic actions and contextual behaviors imparts an increased ability to predict behavioral intention more accurately and more consistently.

Data and Automated Labeling

In the original dataset, contextual behaviors and the behavioral intentions are not explicitly labeled, so we apply automatic heuristic methods to label them for training. Our goal is to recover human-understandable subtasks undertaken by players. These subtasks typically end with a qualitative change in the environment, which we have termed a "behav-

[1] The GitHub repository for the Overcooked-AI environment can be found here: `https://github.com/HumanCompatibleAI/overcooked_ai`

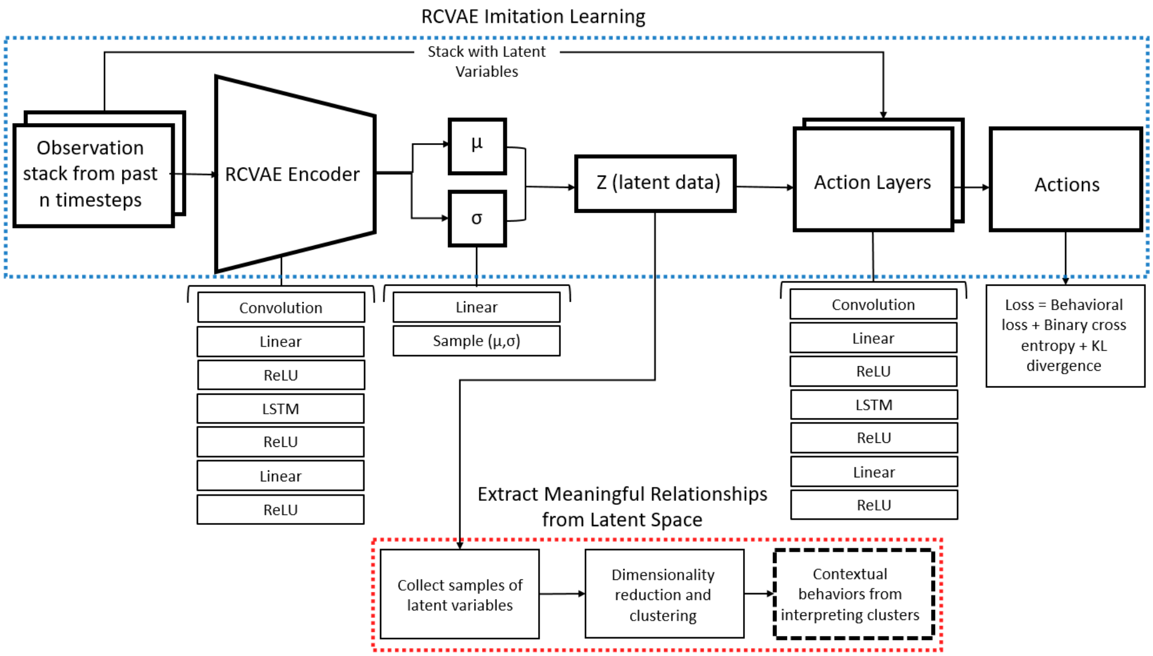
RCVAE Imitation Learning
Stack with Latent Variables
Observation stack from past n timesteps
RCVAE Encoder
μ
σ
Z (latent data)
Action Layers
Actions
Convolution
Linear
ReLU
LSTM
ReLU
Linear
ReLU
Linear
Sample (μ,σ)
Convolution
Linear
ReLU
LSTM
ReLU
Linear
ReLU
Loss = Behavioral loss + Binary cross entropy + KL divergence
Extract Meaningful Relationships from Latent Space
Collect samples of latent variables
Dimensionality reduction and clustering
Contextual behaviors from interpreting clusters

Figure 1: The behavioral intent prediction pipeline. This figure includes the RCVAE architecture (in the blue box) and the process for extraction and clustering of the learned latent space into human understandable clusters (in the red box).

ior" or behavioral intent. These behaviors, once defined, can be automatically detected and labeled in trajectories. In Overcooked, discrete subtasks like "take onion to stove" conclude with a qualitative environmental change, such as the "onion drop off" behavior.

We make the assumption that all basic actions performed by a player directly leading up to a contextual behavior were part of the task culminating in that behavior. This heuristic rule, while not always accurate, is reasonably justified in environments with short task horizons. In Overcooked, this assumption works well to segment the trajectories into identifiable subtasks. This method of automatic heuristic labeling enables our models to correlate sequences of basic actions with the resulting contextual behaviors to which they contribute.

To train both the LSTM and RCVAE we use the 2020 dataset collected and made available by Carroll et al. (2019). The dataset consists of two human players playing the "soup coordination" layout (see image of the game environment in Figure 2). The dataset contains a total of 15,410 steps distributed across 39 distinct trajectories (representing 39 pairs of humans playing the game).

The models were trained via the imitation learning paradigm on one human player from each trajectory. The training dataset contained 13,410 trajectory steps while the remaining 2,000 steps were set aside as a testing set to evaluate model performance. The models were trained on 15,000 batches, each containing 20 steps.

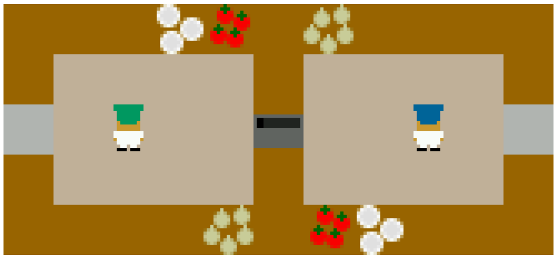

Figure 2: Overcooked "soup coordination" layout.

Results

In this section we compare the performance of the RCVAE and LSTM models in two areas: clustering ability and behavioral intent prediction. The clustering ability analysis will determine how accurately and distinctly the latent spaces of each model represent the contextual behaviors. Ideally, the models would produce eight individual clusters representing each behavior type. The behavioral intent prediction analysis will test both the accuracy and consistency of each model's clustering ability through time. This analysis will determine how far in advance each model can accurately predict the behavioral intent of a player to trigger a contextual behavior given the basic actions performed at each timestep.

Clustering Ability

Five distinct clusters were extracted from the RCVAE latent space and six clusters were extracted from the LSTM latent space. The additional LSTM cluster consisted only of behav-

Table 1: RCVAE Cluster Information.

Cluster ID	Cluster Size	Unique Behaviors	Cluster Purity	Intra-Cluster Distance
0	54	5	0.5	7.2
1	26	1	1	0.5
2	10	1	1	1.7
3	11	1	1	0.6
4	1	1	1	0

Table 2: LSTM Cluster Information.

Cluster ID	Cluster Size	Unique Behaviors	Cluster Purity	Intra-Cluster Distance
0	36	3	0.6	5.4
1	30	3	0.9	2.4
2	14	2	0.9	2.1
3	11	1	1	0.7
4	11	2	0.9	1.6

ioral intents without corresponding behaviors. This observation suggests that the LSTM was unable to learn sufficient similarity between the intents and their respective behaviors, and failed to group them into the same cluster.

The following analysis of clustering ability includes only contextual behaviors and does not include timesteps leading up to the behaviors (labeled as behavioral intents). Five distinct clusters containing only contextual behaviors were identified from the latent space of each model. To compare the clustering capabilities of each model, we computed several metrics: cluster purity, the degree to which clusters are comprised of a single behavior; intra-cluster distance, the mean distance between points within the same cluster; and inter-cluster distance, the average distance between points across different clusters. The purity and intra-cluster distances for each cluster for both the RCVAE and LSTM models are shown in Table 1 and Table 2, respectively. The RCVAE model had an average cluster purity of 0.9, an average intra-cluster distance of 1.9, and an average inter-cluster distance of 22.8. The LSTM model had an average cluster purity of 0.86, an average intra-cluster distance of 2.4, and an average inter-cluster distance of 19.7.

The analysis of cluster metrics supports the finding that the RCVAE achieves a greater clustering ability. Higher clustering purity and lower intra-cluster distance indicate that the RCVAE is able to more accurately group similar behaviors together while the higher inter-cluster distance implies more separation between unrelated behaviors. Overall, the combination of these results suggests that the RCVAE learns more discernable differences between behavior types, which may lead to better behavioral intent prediction.

Behavioral Intent Prediction

To evaluate the behavioral intent prediction performance of the RCVAE and LSTM models, we have employed two metrics: weighted clustering accuracy and clustering uncertainty. Weighted clustering accuracy quantifies the precision with which each model groups intents, calculated by averaging the proportion of actual behaviors in a cluster relative to the cluster's total behavior count. This metric favors models that can accurately segregate behaviors into homogeneous clusters and penalizes those that produce heterogeneous clusters. This property is important, as a model which produces only a single, heterogeneous cluster for all behavior types will be able to cluster behavioral intents with 100% accuracy. For each trajectory in the test data, each intent timestep is assigned a weighted accuracy score depending on the assigned cluster. The weighted clustering accuracy by timestep leading up to a behavior is displayed in Figure 3.

The RCVAE shows improvement over the baseline LSTM model in accurately predicting clusters for intents. We evaluated the overall difference in accuracy between models by collapsing across timesteps. The RCVAE and LSTM distributions were tested for normality with the Kolmogorov-Smirnov test for goodness of fit and were found to be non-normal. We employed the non-parametric Mood's median test to compare the overall weighted clustering accuracy of both models and found that the RCVAE median weighted clustering accuracy score (Mdn = 0.49) was significantly different than the LSTM score (Mdn = 0.34) where $\chi^2 = 42.05$ and $p < 0.001$.

Clustering uncertainty measures the frequency that a model revises its decisions regarding the clustering of intents and is an indication of a model's overall predictive stability. A model with high clustering uncertainty is less decisive, and often reclassifies intents before a contextual behavior occurs. Therefore, a lower clustering uncertainty is desirable as it indicates the ability to assign intents to the correct cluster well before the behavior occurs, without wavering. Figure 4 shows the clustering uncertainty for each model across timesteps leading up to a contextual behavior.

The RCVAE demonstrates consistently lower clustering uncertainty, indicating higher predictive stability. To evaluate the overall difference in clustering uncertainty between models, we repeated our previous analysis. Once again, the distribution of the clustering uncertainty scores were both non-normal leading us to employ the Mood's median test. The RCVAE median clustering uncertainty score (Mdn = 0.31) was significantly different than the LSTM score (Mdn = 0.97) where $\chi^2 = 76.05$ and $p < 0.001$.

Discussion

This paper presents the RCVAE (Recurrent Conditional Variational Autoencoder) model, advancing concept learning models to account for behavioral trajectory data and taking advantage of the capability of generative models to learn complex latent abstractions. Our findings indicate that the

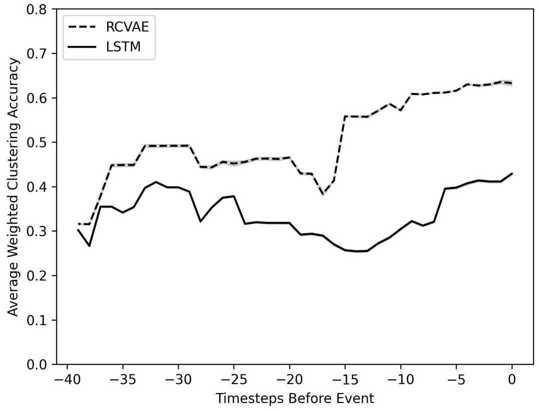

Figure 3: The weighted clustering accuracy of the RCVAE and LSTM models by timesteps leading up to a contextual behavior, averaged over 50 UMAP latent reductions to account for randomness in the latent space introduced by the UMAP algorithm. The weighted clustering accuracy quantifies the precision with which each model groups intents, calculated by averaging the proportion of actual behaviors in a cluster relative to the cluster's total behavior count.

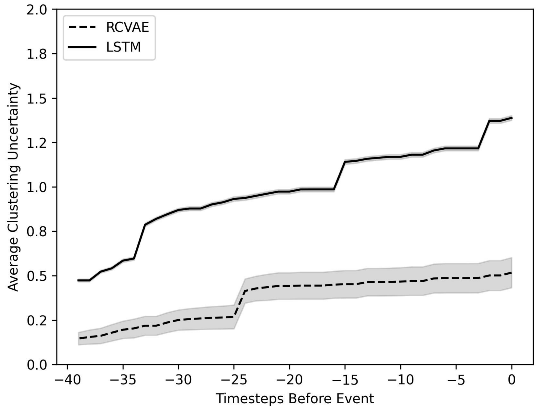

Figure 4: The clustering uncertainty of the RCVAE and LSTM models by timestep leading up to a contextual behavior, averaged over 50 UMAP latent reductions to account for randomness in the latent space introduced by the UMAP algorithm. Clustering uncertainty measures the frequency that a model revises its decisions regarding the clustering of intents and is an indication of a model's overall predictive stability.

RCVAE model outperforms traditional LSTM models in two key areas: clustering ability and behavioral intent prediction. The analysis of clustering ability suggests that the RCVAE can discern more pronounced differences among behavior types compared to the baseline model. Greater clustering purity and reduced intra-cluster distance imply that the RCVAE more effectively groups similar behaviors, while the increased inter-cluster distance indicates a clearer distinction between unrelated behaviors. These findings demonstrate the RCVAE's ability to disentangle and categorize behavior types more effectively within its latent space.

When analyzing behavioral intent prediction ability, the RCVAE model outperformed the LSTM. The RCVAE not only showed higher accuracy in clustering intents but also demonstrated greater predictive stability, as evidenced by its lower clustering uncertainty. This suggests that the RCVAE can more accurately and consistently predict intent leading up to a behavior. The results of our analysis indicate that the RCVAE model learns a meaningful relationship between basic actions and contextual behaviors within the Overcooked environment that is useful for predicting intent with greater accuracy and consistency.

Although the RCVAE represents an advancement in concept learning for human behavioral trajectories, like all concept learning models, the RCVAE model requires hand-labeled data to accurately learn behavioral concepts. While our automatic heuristic labeling method lessened the burden of hand-labeling data, this method may be more difficult to apply in more complex scenarios or in scenarios where hierarchical behaviors exist. Thus, further research must be conducted to develop effective methods for automatically labeling data in scenarios where labeled data is scarce or difficult to obtain.

Future Directions

A logical next step in this line of research is to employ the trained RCVAE model for behavioral intent prediction in multi-agent settings. This application could lead to enhanced coordination and interaction among AI agents, offering significant improvements in fields like robotics, autonomous vehicles, and collaborative AI systems. Additionally, because the RCVAE is a generative model, future research could investigate the ability to generate novel behavioral trajectories and more precisely control agent behavior conditioned on learned behavioral concepts.

The potential applications of the RCVAE model extend beyond controlled or "toy" environments. One such application is counterfactual analysis, where the model could be used to predict alternative outcomes based on varying initial conditions or decisions. This capability would be valuable in strategic planning and decision-making processes across various sectors, including business, healthcare, and public policy. Another promising application is in the field of anomaly detection. The RCVAE's ability to understand and predict intent could be leveraged to identify abnormal patterns or behaviors, which is crucial for security, fraud detection, and maintaining the integrity of complex systems.

In conclusion, the RCVAE model represents a significant advancement in concept learning for behavioral intent prediction. Outperforming traditional LSTM models in both intent prediction accuracy and clustering ability, this model paves the way for novel research and practical applications across various fields.

References

Apperly, I. (2010). *Mindreaders: the cognitive basis of "theory of mind"*. Psychology Press.

Baker, C., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877–1901.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, *32*.

Chen, Z., Bei, Y., & Rudin, C. (2020). Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, *2*(12), 772–782.

Cuzzolin, F., Morelli, A., Cirstea, B., & Sahakian, B. J. (2020). Knowing me, knowing you: theory of mind in ai. *Psychological medicine*, *50*(7), 1057–1061.

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

Engelcke, M., Kosiorek, A., Parker Jones, O., & Posner, H. (2020). Genesis: generative scene inference and sampling of object-centric latent representations. *Proceedings of the ICLR 2020*.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, pp. 226–231).

Fuchs, A., Walton, M., Chadwick, T., & Lange, D. (2021). Theory of mind for deep reinforcement learning in hanabi. *arXiv preprint arXiv:2101.09328*.

Grupen, N., Jaques, N., Kim, B., & Omidshafiei, S. (2022). Concept-based understanding of emergent multi-agent behavior. In *Deep reinforcement learning workshop neurips 2022*.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hussein, A., Gaber, M. M., Elyan, E., & Jayne, C. (2017). Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, *50*(2), 1–35.

Jones, M. N. (2016). *Big data in cognitive science*. Psychology Press.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In *International conference on machine learning* (pp. 5338–5348).

Kumar, A. A., Steyvers, M., & Balota, D. A. (2022). A critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science*, *14*(1), 54–77.

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, *3*(29).

Oguntola, I., Campbell, J., Stepputtis, S., & Sycara, K. (2023). Theory of mind as intrinsic motivation for multi-agent reinforcement learning. *arXiv preprint arXiv:2307.01158*.

Perner, J. (1991). *Understanding the representational mind*. The MIT Press.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S. A., & Botvinick, M. (2018). Machine theory of mind. In *International conference on machine learning* (pp. 4218–4227).

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, *28*.

Torabi, F., Warnell, G., & Stone, P. (2018). Behavioral cloning from observation. In *Proceedings of the 27th international joint conference on artificial intelligence* (pp. 4950–4957).

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child development*, *72*(3), 655–684.

Ye, F., & Bors, A. G. (2021). Learning joint latent representations based on information maximization. *Information Sciences*, *567*, 216–236.

Zarlenga, M. E., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., ... others (2022). Concept embedding models. In *Neurips 2022-36th conference on neural information processing systems*.