Generative Environment-Representation Instance-Based Learning: A Cognitive Model

Tyler Malloy[1], **Yinuo Du**[1,2], **Fei Fang**[2], **Cleotilde Gonzalez**[1]

[1]Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh PA
[2]Software and Societal Systems Department, Carnegie Mellon University, Pittsburgh PA
tylerjmalloy@cmu.edu, yinuodu@cmu.edu, feifang@cmu.edu, coty@cmu.edu

Abstract

Instance-Based Learning Theory (IBLT) suggests that humans learn to engage in dynamic decision making tasks through the accumulation of experiences, represented by the decision task features, the actions performed, and the utility of decision outcomes. This theory has been applied to the design of Instance-Based Learning (IBL) models of human behavior in a variety of contexts. One key feature of all IBL model applications is the method of accumulating instance-based memory and performing recognition-based retrieval. In simple tasks with few features, this knowledge representation and retrieval could hypothetically be done using all relevant information. However, these methods do not scale well to complex tasks when exhaustive enumeration of features is unfeasible. This requires cognitive modelers to design task-specific representations of state features, as well as similarity metrics, which can be time consuming and fail to generalize to related tasks. To address this issue, we leverage recent advancements in Artificial Neural Networks, specifically generative models (GMs), to learn representations of complex dynamic decision making tasks without relying on domain knowledge. We evaluate a range of GMs in their usefulness in forming representations that can be used by IBL models to predict human behavior in a complex decision making task. This work connects generative and cognitive models by using GMs to form representations and determine similarity.

Introduction

Instance Based Learning Theory (IBLT) represents the cognitive processes for human decision making based on cognitive memory mechanisms (i.e, recognition, recall, decay, noise) relevant to dynamic decision making tasks (Gonzalez, Lerch, and Lebiere 2003). IBLT brings together the following characteristics: accumulation of examples in memory through training and task repetition, development of pattern recognition and selective alternative search, similarity-based memory retrieval, gradual withdrawal of attention while increasing memory retrieval, and transition from rule-based to exemplar-based performance.

Although IBLT models have been applied to dynamic tasks involving complex information, this has previously relied on the use of hand-crafted features of the environment being represented in an IBL model and, therefore, the features are unique to each environment. Another issue of applications of IBL modeling is the requirement of a static definition of similarity in the space of environment states throughout modeling.

In contrast, Generative Models (GM) are trained to learn from a data set the underlying distribution that is causally responsible for generating those data (Salakhutdinov 2015). In other words, in GMs, the attributes are not hand-crafted, but are learned from the data. GMs have been integrated with other learning models to demonstrate impressive success in improving learning speed (Higgins et al. 2017).

One useful application of such GMs is in unsupervised and semi-supervised learning, where there data is not categorized, or only a small fraction has relevant categories (Kingma et al. 2014). The learning of representations useful for behavioral goals is an important area of research in modelling human utility-based learning (Radulescu, Shin, and Niv 2021). However, to date, the integration of GMs with cognitive models is lacking.

In this work, we propose the integration of GMs and IBLT, into a new proposed algorithm called Generative Environment-Representation Instance-Based Learning (GERIBL) (pronounced as “jur-bl”). This new algorithm seeks to enable IBLT models to leverage pre-trained models that form representations of environments for dynamic decision making. This is done by integrating IBLT with Generative Models (GMs) that are trained to learn from a data set the underlying distribution that is causally responsible for generating such data (Salakhutdinov 2015).

GMs have previously been integrated with Reinforcement Learning (RL) to predict human learning of the utility of visual stimuli (Malloy, Klinger, and Sims 2022) and fast generalization to novel tasks (Malloy et al. 2022). This integration of GMs and RL has demonstrated the usefulness of pre-trained GMs in forming representations of environments that can be used in cognitive models of learning. We expect that, a similar approach can be taken by integrating GMs into IBLT, and take advantage of the strong cognitive foundations of IBLT into cognitive architectures (i.e., ACT-R (Thomson et al. 2015)).

GERIBL is used as a test bed for the potential integration of GMs with cognitive models by comparing different GM approaches. The learning task of generative models is closely related to the human experience of making decisions based on visual information. Humans can leverage their experience observing visual information outside of the context of decision making to improve their speed of learning and high generalization (i.e., transfer of learning). Part of the reason for this is that humans observe visual information in an unsupervised context and form representations of that information that is useful for a variety of tasks. This is similar to the unsupervised training of Deep GMs which enables them to form useful representations of information that are generalizable. GERIBL leverages these useful features of GMs to integrate with the cognitive mechanisms of IBLT.

GERIBL describes the general framework for integrating environment representations learned by a generative model into an IBL model. We evaluated two approaches for GMs, AutoEncoders (AEs) which form representations of stimuli that are useful for reconstruction; and Generative Adversarial Networks (GANs), which attempt to learn to discriminate between environment stimuli not in the original data set while simultaneously learning to generate environment stimuli that are similar to the underlying data. The results show the advantages of the integration of GMs and IBLT.

Preliminaries: Instance-Based Learning Theory

In IBLT, the memory of agents consists of instances (s, a, x) defined by the state s, their action a and the outcome x (Gonzalez, Lerch, and Lebiere 2003). All instances are stored in memory as outcomes x and options $k = (s, a)$. This means that an IBL model requires the storage of all instances in memory in the form of these triplets.

At time t there may be $n_{k,t}$ generated instances $(k, x_{i,k,t})$. Calculating the expected utility of an action requires an aggregation of all similar instances to determine their memory activation and probability of retrieval.

Among a set of actions considered at each time step, agents take the action with the expected maximum utility. Expected utility is calculated through a "blending" function according to:

$$V_{k,t} = \sum_{i=1}^{n_{k,t}} p_{i,k,t} x_{i,k,t} \tag{1}$$

Where $n_{k,t}$ are the instances in memory, $x_{i,k,t}$ are the outcomes, and the probability of retrieval is $p_{i,k,t}$ is calculated as:

$$p_{i,k,t} = \frac{\exp\left(\Lambda_{i,k,t}/\tau\right)}{\sum_{j=1}^{n_{k,t}} \exp\left(\Lambda_{j,k,t}/\tau\right)} \tag{2}$$

where τ is a temperature parameter and the activation value $\Lambda_{i,k,t}$, which represents the ease of recall of a specific instance in memory, calculated according to:

$$\Lambda_{i,k,t} = \ln\left(\sum_{t' \in T_{i,k,t}} (t - t')^{-d}\right) + \alpha \sum_{j} \text{Sim}_j(f_j^k, f_j^{k_i}) + \sigma \ln \frac{1 - \xi_{i,k,t}}{\xi_{i,k,t}} \tag{3}$$

where d and σ are decay and noise parameters, and $T_{i,k,t} \subset \{0, ..., t{-}1\}$ is the previous observations of instance i. The similarity function $\text{Sim}(f, f')$ calculates the similarity of instances in memory with the current instance (Nguyen, Phan, and Gonzalez 2022). Because of the relationship between noise σ and temperature τ in IBLT, the temperature parameter τ is typically set to $\sigma\sqrt{2}$.

Pattern Recognition

One potential challenge with the use of IBL models in practice, for real-world problems, is that states can be significantly complex. This motivates the formation of hand-crafted representations of the state by cognitive modelers. A cognitive modeler often represents the features in the state of an instance by using the observable attributes in the environment that are relevant to perform a task. This has the benefit of more accurately representing cognitive realities, compared to the alternative of storing complex visual information in memory or using hand-crafted features. The model proposed in this work seeks to determine whether storing representations of complex information learned from a GM can still be useful for modeling cognition, or if the task-relevant information is lost.

Although the hand-crafted features that cognitive modelers define might be practical, a disadvantage of this approach is that they cannot be formed automatically. The representations depend on the cognitive modelers' judgment of what is important for the task. There are no general principles or guidelines to decide on the features that are relevant for the state in a task. Although cognitive modelers rely on what is "observable" in the task, the selection of features may be arbitrary, highly determined by the experience of the cognitive modeler on the task. The model proposed in this work seeks to address this requirement on cognitive modelers.

Similarity-Based Memory Retrieval

A key feature of IBLT is that the activation function depends on the similarity $\text{Sim}(f^k, f^{k_i})$ between the characteristics of the environment and the attributes of the stored instances. This means that recognition, judgment, and choice depend on the method of determining similarity (Gonzalez, Lerch, and Lebiere 2003). IBLT also proposes that decision makers learn to focus their attention on task-relevant features and, in turn, select the limited information they attend to based on this similarity (Gonzalez, Lerch, and Lebiere 2003). However, until now, there has been no principled method to achieve this goal.

Although measuring similarity is highly relevant in models designed in IBLT, relatively little work has been done in IBLT to compare different approaches to measuring similarity. The similarity function used is often linear similarity, but some times it is opted for some non-linear similarity function in a trial-and-error modeling process. In the next section, the proposed model will attempt to address the challenge of automatically producing instance attributes, and consistently and meaningfully measuring similarity, through the integration of an IBL model with a generative model.

Preliminaries: Generative Models

Generative Models (GM) are a class of machine learning methods that attempt to learn from a data set by assuming that a probability distribution generated the data and attempting to learn the underlying distribution (Harshvardhan et al. 2020). In this research, we propose a set of methods to integrate IBL models with three major classes of generative models, Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Visual Transformers (ViT) to address the current limitations of IBL models described above.

Figure 1 illustrates the proposed Generative Environment-Representation Instance-Based Learning (**GERIBL**) cognitive model. In this proposed model, the environment representation can be generated from the GM, producing an environment state that the IBL model can use to make decisions from experience. Furthermore, the figure illustrates how the execution of actions from an IBL model can influence the environment presented in the GM.

While other types of generative models exist, these two were chosen because of their general applicability to various input modalities (image, text, audio, etc.) and their usefulness in applications of the learning setting described later. The remainder of this section provides background information on these two types of generative models, as well as insight into the usefulness of representations learned by these approaches in IBL models.

AutoEncoders

Autoencoder (AE) models function by assuming that there is a set of generative factors ζ that causally explain the data in a set $x \in X$. The goal of training these models is to learn an encoding function $p(z|x)$ and a decoding function $p(x|z)$ that reflect these generative factors. The result is a model that can approximate the true environmental distribution $p^*(x)$.

When used with image data, these models typically use the general structure of Convolutional Neural Networks (CNNs) to learn low-dimensional representations of visual information that can be used to form reconstructions of unobserved visual stimuli, such as human faces (Zhang 2018).

Variational Autoencoder: (VAE) models use a deep neural network to learn an encoder function $q_\phi(z|x)$ that outputs constrained representations z of visual stimuli x (Kingma and Welling 2014). These representations define a vector of means μ_z and variances σ_z that form a Normal distribution $\mathcal{N}(\mu_z, \sigma_z)$. This distribution is sampled to form a vector that is translated through to the encoder layers $p_\theta(x|z)$ to produce a reconstruction. These VAE models are trained to minimize the difference between input and reconstruction by maximizing the objective function (Pu et al. 2016):

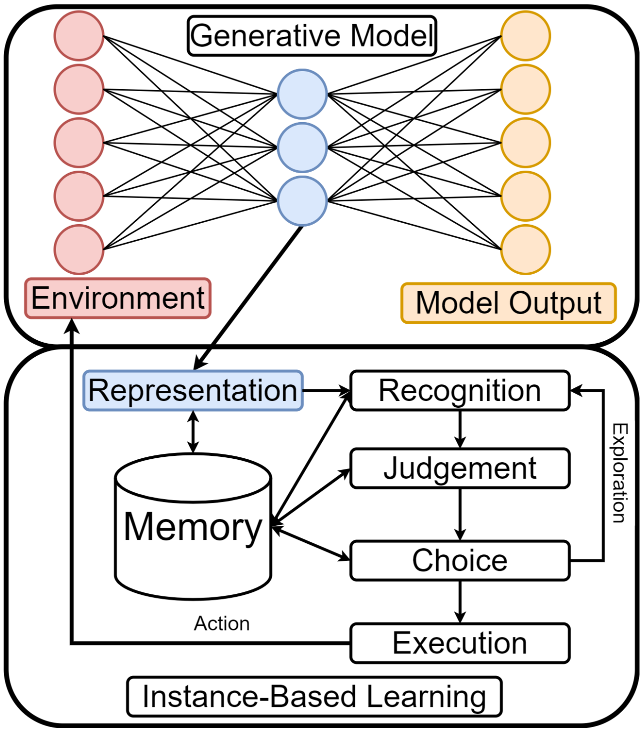

Figure 1: GERIBL: Generative Environment Representation Instance Based Learning Model consisting of a generative model producing environment stimuli representations that are used by an instance-based learning model to make decisions from experience.

$$\mathcal{L}(\theta, \phi; x, z) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \tag{4}$$

This learning objective is guaranteed to learn a generative model that will approximate the true environmental distribution $p^*(x)$. However, there is no guarantee of any meaningful connection between the learned latent representation z and the true generative factors ζ (Chen et al. 2016). This lack of connection could be problematic for decision models based on these internal representations, potentially motivating the use of alternative training (Aridor, da Silveira, and Woodford 2022).

β-Variational Autoencoder: models seek to connect generative factors ζ and latent representations z by adjusting the training of traditional VAEs by introducing a β parameter that further controls the information bottleneck (Burgess et al. 2018). This is done by penalizing a metric of informational complexity of the representations using the KL-divergence between the decoder and latent distribution, using the training function (Higgins et al. 2016):

$$\begin{aligned}\mathcal{L}(\theta, \phi; x, z, \beta) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \\ -\beta D_{KL}\big(q_\phi(z|x)||p(z)\big)\end{aligned} \tag{5}$$

The β parameter allows for additional control over the information bottleneck of the model by adding a weight to the informational complexity of the latent representations defining the multivariate Gaussian distribution. The result is that the entire model is trained to balance the accuracy of reconstruction and the complexity of latent representation in an adjustable fashion.

Image Transformers

Pre-trained transformer models have the advantage of wide applicability on a variety of different tasks and domains, particularly in the context of Natural Language Processing (NLP) (Wolf et al. 2019). However, concerns have been raised over the use and usability of massive pre-trained transformer models, suggesting that their output may be the results of spurious correlations and stochasticity (Bender et al. 2021). Part of the testing of the Transformer based GMs with GIRBL will be to compare models pre-trained using the exact same stimuli with ones trained using similar stimuli.

Image-based transformers apply transformer-based self-attention mechanisms to machine learning domains with visual data (Parmar et al. 2018; Dosovitskiy et al. 2020). The two models used to test the GERIBL model use transformers, and differ in their training methods and the size and form of their representations of visual information.

Vision Transformer VAE: Variational Autoencoders trained using transformer models are able to learn constrained representations of images of variable size that are still useful for reconstruction (He et al. 2022). These models can be integrated into the GERIBL cognitive model using the encodings learned by a Visual Transformer Variational Autoencoder (ViT-VAE) model.

The ViT-VAE model uses 4 attention heads, and 2 NN layers of 64 nodes for the multi-layer perceptron layers. The loss function is based on the difference between the input and reconstruction. The VAE encoding representation is used by the GERIBL model as an environment state representation, and takes the form of a vector of real numbers of size 100.

Attention: The second transformer based GM that is compared using the GERIBL model uses learned values from the self-attention heads of the transformer network when processing visual information, this model is referred to as the Attention model.

The Attention model has the same general structure as the ViT-VAE model with the main difference being that it is not trained to reconstruct lossy versions of input stimuli. The second difference is the form and size of the representation that is used by the GERIBL model. In the case of the Attention model, the values of the 4 self-attention heads are used as the representations for the GERIBL model.

Generative Adversarial Networks

Generative Adversarial Network (GAN) models are trained using generator and discriminator networks (Salakhutdinov 2015). The goal of the generator is to produce images that appear similar to those in the training data set so that the discriminator network cannot tell the difference. The goal of the discriminator is to determine if a given image was produced by the generator or is a genuine original data set member. These models are trained in tandem in an adversarial structure. Two GAN based models are used for comparison with integration with the proposed GERIBL model. Both models have the same general structure and training, differing only in the size of their internal representation space and other network features.

GAN Model: is the first GAN model uses representations of size 100 to complete the learning objectives of the generator and discriminator networks. This is considered to be an 'unconstrained' version of a GAN, analogous to the VAE model which has a larger representation size and information complexity compared to the β-VAE model. The calculation of similarity of the GAN model is determined by the

Constrained GAN: The second GAN-based model is motivated by a similar motivation to the β-VAE model, in using an information bottleneck to produce constrained representations that are less informationally complex, allowing for faster generalization, while still being useful for the IBL module. This is done by reducing the size of the internal representation from 100 to 3, the same size as the latent representation of the β -VAE model. Additionally, the generator and discriminator network feature map is reduced from 64 to 8, additionally imposing a stricter information bottleneck. All other model structures and hyper-parameters are kept the same.

GERIBL: Proposed Model

The proposed Generative Environment-Representation Instance-Based Learning (GERIBL) model is the integration of IBLT (the Python implementation of IBLT called PyIBL) and generative models. We compare a variety of GMs, including VAEs and GANs, in their ability to form representations of visual information that can be used in a cognitive architecture model of dynamic decision making. This change is made primarily by replacing environment state s with the corresponding GM internal representations $p(z)$. The result is a cognitive architecture that predicts human recognition, judgement, choice, and execution based on constrained representations of visual information.

Furthermore, the GERIBL model alters the IBLT activation function (Eq. 3) by replacing the feature-based similarity function $\text{Sim}(f, f')$, where similarity is based on the internal representation of the GM z and the similarity metric of the GM Sim_{GM} as follows:

$$\begin{aligned}\Lambda_{i,k,t} =& \ln\left(\sum_{t'\in T_{i,k,t}} (t-t')^{-d}\right)\\ &+ \alpha \sum_j (\text{Sim}_{\text{GM}}(p(z_k|k), p(z_{k_j}|k_j)))\\ &+ \sigma \ln \frac{1-\xi_{i,k,t}}{\xi_{i,k,t}}\end{aligned} \tag{6}$$

where $p(z_s|s)$ is the GM internal representation of observed state s and $p(z_{s_j}|s_j)$ are the GM internal representations of each instance in memory s_j. Importantly, this altered activation function avoids the necessity of storing the full original

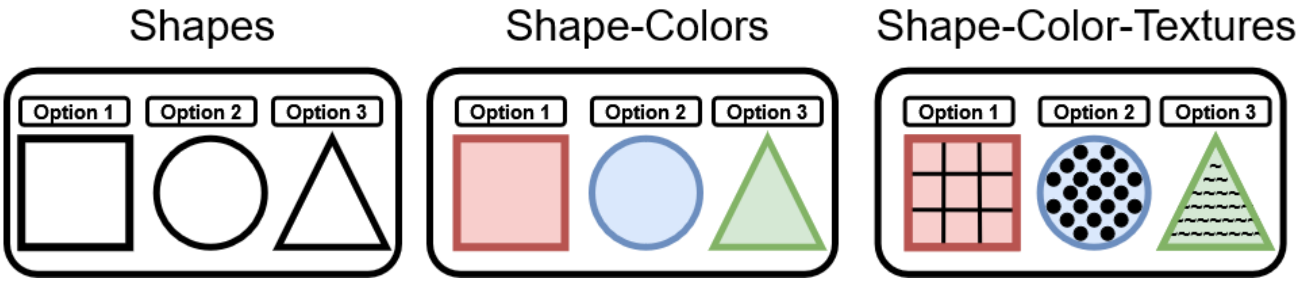

Figure 2: Contextual bandit learning task stimuli used in Experiment 1 (Right Panel) (https://nivlab.princeton.edu/data) and in Experiment 2 (Left, Middle, and Right Panel). Left panel: The first set of stimuli shown to participants in Experiment 2. Middle panel: The second set of stimuli shown to participants in Experiment 2. Right panel: the third and final set of stimuli shown to participants in Experiment 2. This is also the stimulus used in Experiment 1, to learn which of the 9 possible features (shape,color,texture) was associated with a higher reward.

environment stimuli, instead allowing for cognitive mechanisms to use low-dimensional representations of environments.

The type of GM that is used in the GERIBL model results in differences based on how the internal representations of each GM are formed and how those models determine representation similarity. For example, the β-VAE determines similarity based on the loss in Eq. 5, according to the KL divergence between the two representation distributions and their informational complexities.

Model Representations

Another benefit of using GM-acquired representations as instances of IBL models is that they can be updated as the IBL model learns the utility of choice options. This can reflect the tendency of decision makers to attend to features that are more relevant for a task at hand, which in turn changes how they represent information internally. Previous work has compared how β-VAE model representations can change as utility is learned in a bandit task involving images of human faces (Malloy et al. 2022). This is integrated into the proposed model by training the generative model with feedback from the GIRBL model blending function $V_{k,t}$ which uses the activation function 6 according to:

$$\mathcal{L}(v,k) = v\big(V_{k,t} - x_k\big)^2 \tag{7}$$

Where $V_{k,t}$ is the predicted utility of the IBL model before choice selection, and x_k is the true observed outcome. This functionality of the proposed model allows for the updating of representation of environments as the relevance to utility of different features is learned. This utility-based training of generative model representations has demonstrated more human-like decision-making, reproducing biases in utility selection (Aridor, da Silveira, and Woodford 2022), and fast generalization (Malloy et al. 2022).

Learning Tasks

Experiment 1: Visual Utility Learning

The first learning task was originally described (Niv et al. 2015) collected by the Princeton University Niv Lab and made publicly available on their lab website[1]. The experiment study was approved by the Princeton University Internal Review Board.

This task consisted of a contextual n-armed bandit in which participants were shown 3 different choice options consisting of a shape (square, circle, triangle), color (red, green, blue), and texture (hatched, dotted, wavy), as shown in Figure 2 (Right panel). On each trial, the color, shape, and texture of each option are randomized, with one instance of each feature type occurring across the stimuli options (i.e., there is always 1 green option, 1 square option, etc.).

Experiment trials were variable lengths of roughly 20-25 stimuli decision trials in which the same 1 of the 9 possible features was associated with a higher probability (75% vs. 25%) of observing a reward of 1 instead of a reward of 0. Data from 22 participants were collected in this task, each making a total of 500 choice selections.

Experiment 2: Transfer of Learning

This second experiment was originally collected and detailed in (Malloy et al. 2023) by the Dynamic Decision Making lab at Carnegie Mellon University, and made publicly available on OSF[2]. 60 participants were recruited online through Amazon Mechanical Turk. The experiment was pre-registered on OSF and approved by the Carnegie Mellon University Internal Review Board. For full methods see (Malloy et al. 2023).

This experiment sought to test human Transfer of Learning (ToL), referring to the application of previously learned skills onto a new task. The learning task in Experiment 2 involves ToL in which participants first learned the values associated with shapes alone, then shapes and colors, and finally the same shape-color-texture features described in (Niv et al. 2015). The rewards ranged from roughly 4-6 points, determined by the features of the chosen option, with random noise added to the reward.

The experiment episodes consisted of 14 trials of each type in the order shown in Figure 2. During one set of trials, one of the three feature options was associated with a higher

[1] https://nivlab.princeton.edu/data

[2] https://osf.io/mt4ws/

reward (roughly 7 vs. 5). As the experiment progressed, the previously high-valued feature continued to indicate that an option had a higher value. For example, if a square is associated with a higher expected utility initially, then red squares will have a higher expected utility than red triangles for the remainder of the experiment block. The same is true for the higher utility color once the texture is introduced.

Model and Human Performance

This section compares the 6 previously mentioned GMs in their ability to be integrated with the proposed GERIBL model. These GMs are pre-trained with a subset of the stimuli shown in Figure 2, either the 3 shape stimuli, 9 shape-color stimuli, or 27 shape-color-texture stimuli. After this pre-training, the models are used to produce a representation that the IBL module of the GERIBL model takes in as an environment state. We use the two learning experiments to compare human participant performance, the 6 proposed GM instantiations of GERIBL, and a handcrafted version of the IBL model.

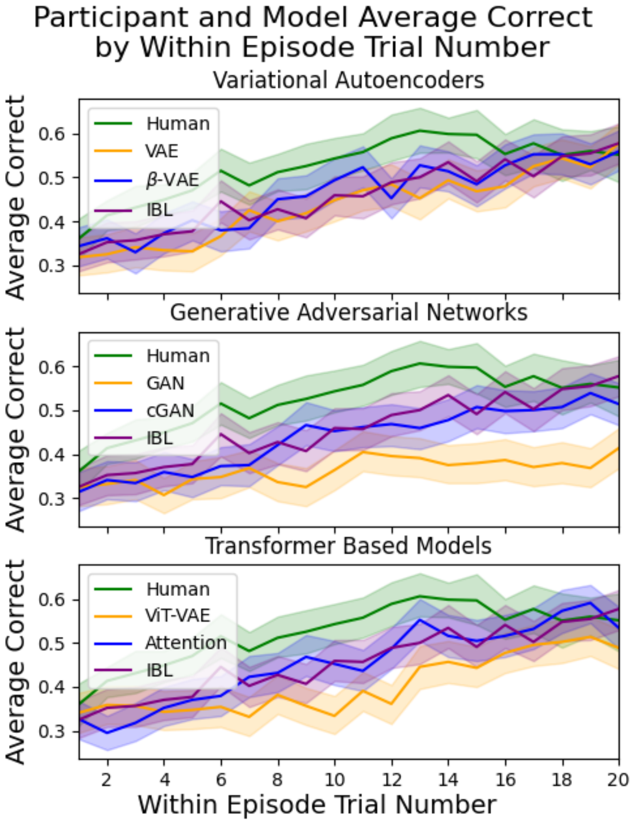

Figure 3: Model and participant average probability of selecting the correct option in the contextual bandit task by within episode, chance rate is at 1/3.

Visual Utility Learning

In the first experiment on visual utility learning, GMs are pre-trained using only the shape-color-texture stimuli set of 27 images.The results in 3 compare the three types of GMs (VAE, Transformer, and GAN) with human performance and an IBL model using hand-crafted features. These results demonstrate that all GMs roughly emulate human-like performance, with the worst performing GMs being the GAN and ViT-VAE model.

In Figure 3, the blue models correspond to the GMs with smaller representation sizes than the orange models which correspond to the GMs with larger representation sizes. As shown, the GMs with smaller representations are a better fit to human behavior compared to those with larger representations. This is likely due to the fact that smaller representations are less informationally complex and thus are easier to quickly generalize. These results indicate that one important factor of GMs when integrating them in the GERIBL model is the informational complexity of representations.

However, when using simple representations it is important to retain enough information for behavioral goals. If the GMs representations were too simple, they could remove information relevant to the task, making it difficult for the IBL module to learn. This would be a detriment to applying the GERIBL model, since the main benefit is the possibility of automatically generating environment features, as well as a metric for comparing them.

Transfer of Learning

Transfer of learning is related to the goals of applying GMs onto cognitive modeling in the potential application of pre-trained models onto novel environments. To compare the ability of GM representations to be applied onto new tasks, we limit the training data-sets in Experiment 2 by including only the shape images, only the shape-color images, and finally only the shape-color-texture images (see Figure 2). This produces 3 sets of GMs for each image type, that are used to produce representations of the visual information used to make decisions in the other two types of tasks.

The first noticeable aspect of these results is that the majority of GMs had a higher transfer of learning compared to the IBL model with hand-crafted features. This can be observed by the asymptotic reward (measured by the average reward on the final 5 trials) of each GM trained on a subset of stimuli and tested in each of the experiment task conditions. Of these GMs, the best performing is the Transformer model using Attention values as its representation, which matches human performance regardless of the stimuli it was trained on. This indicates that this model has learned an efficient representation of the stimuli applicable to related tasks.

In addition to testing GMs in their ability to be applied onto a novel experiment task, these results strengthen the two other motivations of GIRBL, in automatically determining relevant stimuli features and a metric of similarity. If GMs required a unique training approach for each stimuli space limited to that task, then the applicability of pre-trained models would be significantly diminished. We show that GMs with small representation spaces can be applied onto producing human-like learning patterns even with novel stimuli. This supports GMs as a tool for applications beyond IBL, such as in cognitive architecture based approaches.

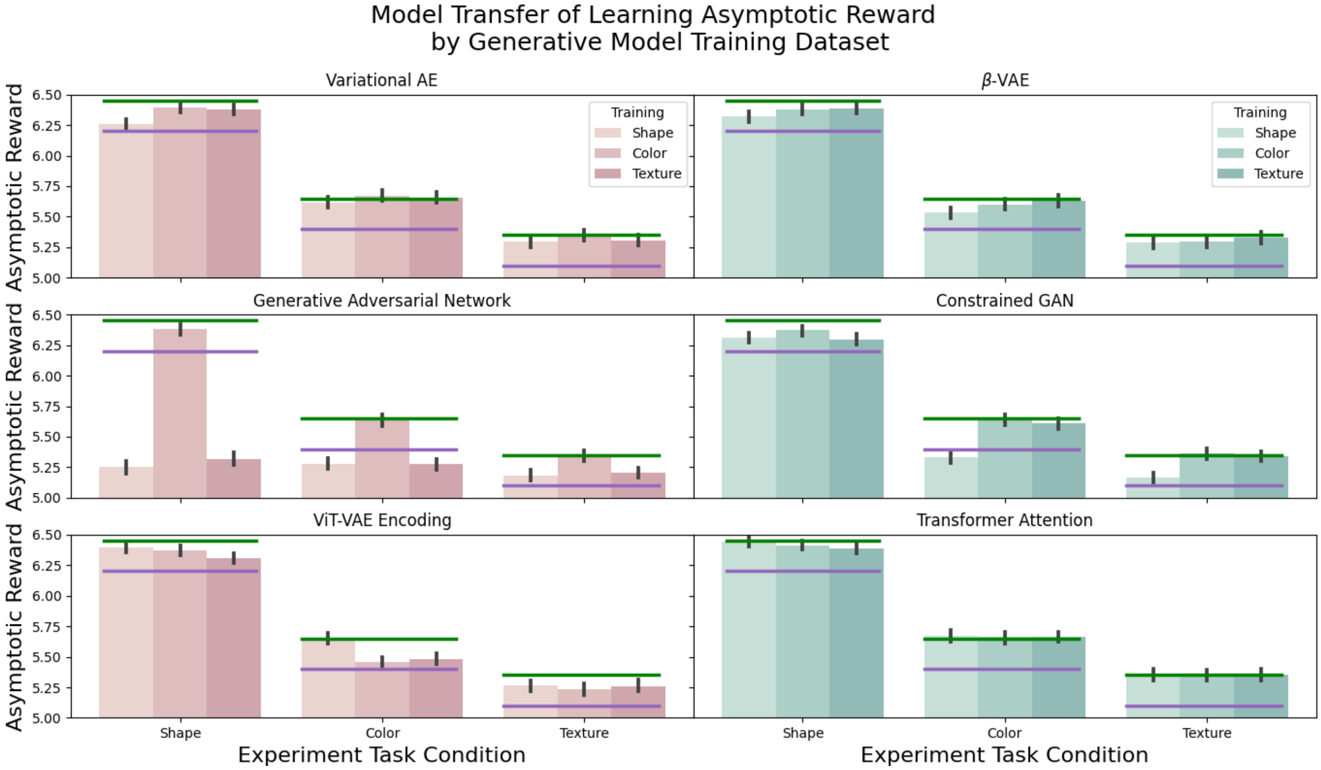

Figure 4: GERIBL model average asymptotic reward in the second experiment separated by the experiment condition. Generative Models were trained only on a subset of the stimuli space indicated by color shade. Purple lines represent IBL model performance using hand-crafted features. Green lines represent participant performance.

Conclusions

The GERIBL model incorporates GMs into an IBL model, which had three main benefits. Firstly, it uses representations of task environments that are generated automatically, without requiring cognitive modellers to develop a feature set for each new task. Secondly, it allows for an objective metric of similarity defined by the GM itself. Thirdly, it allows for improved prediction of human behavior, in a transfer of learning setting.

Results from comparisons of 6 different GM methods in two experiment paradigms demonstrated a close correspondence with GERBIL model and human behavior. However, a general trend showed some insights that are useful for cognitive modellers interested in incorporating GMs in cognitive modeling. In both the GAN and ViT models, when incorporated into the GERBIL framework, higher performance and a closer fit to human behavior was achieved by using a smaller representation size. While the VAE and β-VAE models did not replicate this general trend, there is an inherent information bottleneck in both models.

This trend reveals insight into applying GMs to predicting behavior, specifically the usefulness of reduced representation sizes. However, there is likely a balance required to ensure that representations are large enough to effectively train GMs while learning representations useful for predicting behavior. The nature of this balance points to possible future research in the functioning of GMs applied to cognitive models.

Of the GMs tested using the GERIBL model, the β-VAE based model has the closest connection to biological visual processing, which has been related to the disentanglement objective (Higgins et al. 2021). However, performing a complete analysis and comparison of different types of GMs provides support of our proposed model as a general framework integration into cognitive models and architectures.

In addition to these main benefits, the results shown here point towards future research investigating the impact of utility on the representations learned by GMs. This could be one area where GMs differ highly in their connection to human cognition, as they would likely react differently to training that incorporated utility prediction. Previous work has compared GM representations as utility is learned in simulated settings (Malloy, Klinger, and Sims 2022), but not yet compared to behavior from human participants

Acknowledgements

This research was sponsored by the Army Research Office and accomplished under Australia-US MURI Grant Number W911NF-20-S-000 and by the Army Research Laboratory under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA)

References

Aridor, G.; da Silveira, R. A.; and Woodford, M. 2022. Information-Constrained Coordination of Economic Behavior. In *Neural Information Processing Systems Workshop: Information-Theoretic Principles in Cognitive Systems*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β-VAE. *Conference on Neural Information Processing Systems*.

Chen, X.; Kingma, D. P.; Salimans, T.; Duan, Y.; Dhariwal, P.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Variational Lossy Autoencoder. In *International Conference on Learning Representations*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Gonzalez, C.; Lerch, J. F.; and Lebiere, C. 2003. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4): 591–635.

Harshvardhan, G.; Gourisaria, M. K.; Pandey, M.; and Rautaray, S. S. 2020. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38: 100285.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Higgins, I.; Chang, L.; Langston, V.; Hassabis, D.; Summerfield, C.; Tsao, D.; and Botvinick, M. 2021. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1): 6456.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

Higgins, I.; Pal, A.; Rusu, A.; Matthey, L.; Burgess, C.; Pritzel, A.; Botvinick, M.; Blundell, C.; and Lerchner, A. 2017. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, 1480–1490. PMLR.

Kingma, D. P.; Mohamed, S.; Jimenez Rezende, D.; and Welling, M. 2014. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. *stat*, 1050: 1.

Malloy, T.; Klinger, T.; and Sims, C. R. 2022. Modeling human reinforcement learning with disentangled visual representations. In *Reinforcement Learning and Decision Making*.

Malloy, T.; Sims, C. R.; Klinger, T.; Riemer, M. D.; Liu, M.; and Tesauro, G. 2022. Learning in Factored Domains with Information-Constrained Visual Representations. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*.

Malloy, T.; Yinuo, D.; Fei, F.; and Cleotilde, G. 2023. Accounting for Transfer of Learning using Human Behavior Models. In *AAAI Human Computation and Crowdsourcing*.

Nguyen, T. N.; Phan, D. N.; and Gonzalez, C. 2022. SpeedyIBL: A comprehensive, precise, and fast implementation of instance-based learning theory. *Behavior Research Methods*, 1–24.

Niv, Y.; Daniel, R.; Geana, A.; Gershman, S. J.; Leong, Y. C.; Radulescu, A.; and Wilson, R. C. 2015. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21): 8145–8157.

Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; and Tran, D. 2018. Image transformer. In *International conference on machine learning*, 4055–4064. PMLR.

Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; and Carin, L. 2016. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems*, 29.

Radulescu, A.; Shin, Y. S.; and Niv, Y. 2021. Human representation learning. *Annual Review of Neuroscience*, 44: 253–273.

Salakhutdinov, R. 2015. Learning deep generative models. *Annual Review of Statistics and Its Application*, 2: 361–385.

Thomson, R.; Lebiere, C.; Anderson, J. R.; and Staszewski, J. 2015. A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *Journal of Applied Research in Memory and Cognition*, 4(3): 180–190.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zhang, Y. 2018. A better autoencoder for image: Convolutional autoencoder. In *International Conference on Neural Information Processing ICONIP17-DCEC*.