Modeling artificial category learning from pixels: Revisiting Shepard, Hovland, and Jenkins (1961) with deep neural networks

Alexa R. Tartaglini (art481@nyu.edu)
Wai Keen Vong (waikeen.vong@nyu.edu)
Brenden M. Lake (brenden@nyu.edu)
Center for Data Science, 60 5th Ave,
New York University, NY, 10011

Abstract

Recent work has paired classic category learning models with convolutional neural networks (CNNs), allowing researchers to study categorization behavior from raw image inputs. However, this research typically uses naturalistic images, which assess participant responses to existing categories; yet, much of traditional category learning research has focused on using novel, artificial stimuli to examine the learning process behind how people acquire categories. In this work, we pair a CNN with ALCOVE (Kruschke, 1992), a well-known exemplar model of categorization, and attempt to examine whether this model can reproduce the classic type ordering effect from Shepard, Hovland, and Jenkins (1961) on raw images rather than abstract features. We examine this question with a variety of CNN architectures and image datasets and compare ALCOVE-CNN to two other models that lacked certain key features of ALCOVE. We found that our ALCOVE-CNN model could reproduce the type ordering effect more often than the other models we tested, but in limited situations. Our results showed that success varied greatly across the various configurations we tested, suggesting that the feature representations from CNNs provide strong constraints in properly capturing this effect.

Keywords: category learning; convolutional neural networks; exemplar models; attention

Introduction

Category learning is one of the oldest and most central areas of cognitive science, with a wide range of computational models developed to explain its many facets (Murphy, 2004). One challenge for these computational models is how to specify the underlying feature representation that is provided as input. It is common to either directly use the hard-coded, abstract features as inputs, or alternatively, one could use representations derived using multi-dimensional scaling from similarity judgments, which project a limited set of stimuli into a low-dimensional representational space (Shepard, 1980). Either way, traditional models work with a representation several steps removed from the raw stimuli.

Recent progress in computer vision opens the door to cognitive models that operate on stimuli in their raw form, just as a human participant might see them on the screen in an experiment. Rather than specifying features in advance, convolutional neural networks (CNNs) learn to extract useful high-level features from raw naturalistic images (Krizhevsky, Sutskever, & Hinton, 2012), providing a potential route to interfacing cognitive models with more raw forms of category learning stimuli.

Recent studies suggest that CNNs may see the right sorts of structure in raw stimuli for modeling human perception and categorization. For instance, CNNs pre-trained for image classification have shown success in predicting category typicality ratings (Lake, Zaremba, Fergus, & Gureckis, 2015) and similarity ratings from natural images (Peterson, Abbott, & Griffiths, 2018). More recently, researchers have begun to combine CNNs with classic prototype and exemplar models of categorization (Battleday, Peterson, & Griffiths, 2020; Guest & Love, 2019; Singh, Peterson, Battleday, & Griffiths, 2020; Nosofsky, Meagher, & Kumar, 2020), usually with the aim of predicting human categorization decisions for images of common categories such as animals and vehicles.

These successes in predicting human judgments on natural images do not, however, imply that such an approach will be a successful psychological model of categorization as studied in its most classic setting: artificial category learning tasks conducted in the lab. This is the question we take up in our work here. One of the most influential findings in this vein is the classic experiment by Shepard et al. (1961) which has been replicated many times (Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994; Crump, McDonnell, & Gureckis, 2013). Shepard et al. (1961) showed that with stimuli composed of three binary features, six distinct category types consisting of two categories with four members each, referred to as Types I-VI, can be constructed (see Table 1 for the full set of category structures, and Figure 1 for examples of the stimuli). They found a strong relationship between the types of categories and the difficulty with which they are learned by human participants: the Type I category structure is the easiest to learn. This is followed by Type II, which is consistently easier to learn than Types III, IV, and V, which are about equally difficult. Finally, Type VI was found to be the most consistently difficult to learn, as the learner needs to attend to all three dimensions and memorize all possible configurations of the stimuli.

The relative rate of learning across these six category types has provided the field with a strong empirical constraint, whereby computational models are required to capture this effect to count as a serious theoretical account of human category learning (Kruschke, 1992; Nosofsky et al., 1994; Goodman, Tenenbaum, Feldman, & Griffiths, 2008). However, to our knowledge, no existing research has attempted to examine whether categorization models interfacing with raw forms of the stimuli can capture this effect. Relative to previous studies that have focused on naturalistic images, attempting to capture this effect presents two challenges. First, the stimuli used in these tasks are more abstract compared to naturalis-

	Category Types					
Stimulus	**I**	**II**	**III**	**IV**	**V**	**VI**
[0, 0, 0]	A	A	B	B	B	B
[0, 0, 1]	A	A	B	B	B	A
[0, 1, 0]	A	B	B	B	B	A
[0, 1, 1]	A	B	A	A	A	B
[1, 0, 0]	B	B	A	B	A	A
[1, 0, 1]	B	B	B	A	A	B
[1, 1, 0]	B	A	A	A	A	B
[1, 1, 1]	B	A	A	A	B	A

Table 1: **The category structure for the six types in Shepard et al. (1961)**. Each category type (I-VI) assigns four stimuli each to either Category A or B according to different grouping patterns based on the stimulus encoding. For each type, there are six possible permutations for how the abstract encodings map onto the corresponding features.

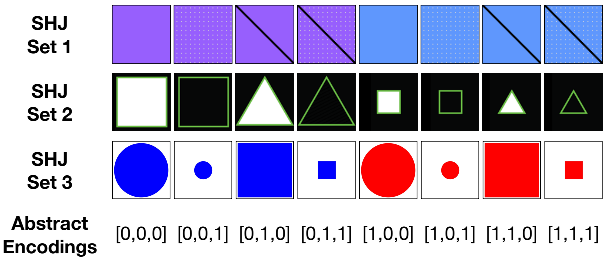

Figure 1: **The three image datasets in our simulations**. Each of the three rows depict a different set of eight stimuli. The fourth row contains the corresponding abstract representations of the stimuli. Each number encodes a value for one of the three binary dimensions in each image.

tic images. These images may not be well represented in the kinds of datasets used to train CNNs, and such models may not be able to extract the relevant features for categorization. Second, the effect is based on the relative ordering across category types. Whereas recent work using CNNs has examined categorization behavior for known categories, here, we are explicitly interested in the pattern of learning for novel categories and whether the same ordering learning curves can be reproduced.

In this paper, we test this question by creating a hybrid model that extracts features from images of psychological stimuli using a CNN, and passes this to ALCOVE (Kruschke, 1992), an exemplar category learning model with learned attentional weights, which we call CNN-ALCOVE-Attn. As a control, we contrast our model against a variant with no attentional learning (CNN-ALCOVE-No-Attn) and one where the CNN features are passed into a multi-layer perceptron (CNN-MLP). In the original ALCOVE paper, both of these variants were unable to produce the correct ordering, and serve as useful baselines here. We test these models against various CNN architectures and image datasets, and across an extensive range of hyperparameters. Our results show that while our model can reproduce this effect, the successes were only observed in a limited number of configurations tested.

Methods

Datasets We used three datasets for our simulations, each containing 8 images of stimuli spanning all possible configurations of the three binary features, as shown in Figure 1. The image-based realization of each binary feature is different for each dataset. In SHJ Set 1 (from Love (2002)), the three binary features are color (purple or blue), the presence of dots (dots or no dots), and the presence of a line (line or no line). In SHJ Set 2 (from Crump et al. (2013)), the images are varying geometric shapes on a black background with a green border, with the three binary features being size (large or small), shape (square or triangle), and color (black or white). Finally, in SHJ Set 3 (from Guest and Love (2019)), the three binary features are the same as SHJ Set 2 but with different values and visual appearance, with size (large or small), shape (circle or square), and color (red or blue). Additionally, for each dataset, there are six different mappings (permutations) between the abstract and image-based features. To computational models that take in the abstract feature representation as input, these different image-based realizations are identical. However, each of the image sets are quite perceptually different from one another, and therefore the different permutations for each of the six types may differ. Furthermore, using multiple image sets provided us with a good robustness check for whether the type ordering effect would be consistent across different image sets that are typically used in category learning experiments.

Computational Model In this section, we describe our variant of an attention-weighted deep exemplar model based on the original ALCOVE model by Kruschke (1992), which we refer to as **CNN-ALCOVE**. ALCOVE was chosen because the original model naturally captures the type ordering effect through the use of selective attention, and is an end-to-end differentiable model that is straightforward to attach to a CNN front-end (although we leave an exploration of end-to-end training of the combined model for future work). As a baseline, we compared the CNN-ALCOVE model to a standard neural network architecture, replacing the exemplar categorization module with a set of feedforward layers, which we refer to as the **CNN-MLP** model. The set-up of our models is shown in Figure 2, and the following section provides additional details for each of these models.

CNN Architectures First, as a pre-processing step for both models, the raw image of a stimulus, denoted x, was passed through an ImageNet (Deng et al., 2009) pre-trained convolutional neural network up to the penultimate layer to extract a set of visual features,

$$f(x) = CNN(x) = [f_1, f_2, \ldots, f_D]. \quad (1)$$

This results in a high-dimensional, abstract representation of the input stimulus, where the number of features D is equal to the size of the penultimate layer of the CNN, and generally far larger than the commonly used 3-dimensional abstract binary representation. We chose to extract features from the penultimate layer because it allowed for a simpler comparison of results across different CNN architectures and may be better suited to representing abstract features like shape from the stimuli in our simulations. In this work, we tested three standard CNN architectures: VGG-16 (4096 features; Simonyan & Zisserman, 2014), ResNet-18 (512 features) and ResNet-50 (2048 features; He, Zhang, Ren, & Sun, 2016). Since each dataset only contained 8 images, the CNNs were fixed and not fine-tuned.

CNN-ALCOVE Model For the CNN-ALCOVE model, this high-dimensional representation of the stimulus is then passed to our adaptation of the original ALCOVE model. In ALCOVE, each of its input nodes are associated with a learnable attention weight vector $\alpha = [\alpha_1, \ldots, \alpha_D]$, the strength of which indicates the relevance of that particular stimulus dimension to the task. These attention weights are initialized uniformly and are individually strengthened or weakened over the course of learning. Our version of the model also included learnable attention weights, but rather than the attention being spread over three abstract features, it was spread over the much larger set of D features from the CNN.

The attention-weighted stimulus is compared to the set of eight possible exemplars, which consist of the high-dimensional representations extracted from each of the stimuli by the CNN. For an input $f(x)$, the activation a_j of the jth exemplar node h_j is calculated by the similarity between the exemplar and the input as follows:

$$a_j = \exp(-c \sum_i \alpha_i |h_{ij} - f_i(x)|), \tag{2}$$

where c is a hyperparameter referred to as specificity. The larger the specificity, c, the faster the similarity falls off as the distance between the input stimulus and the exemplar increases. Additionally, the psychological distance metric r and similarity gradient q were set to 1 as in Kruschke (1992), and not displayed above.

Finally, the CNN-ALCOVE outputs a predicted category label $\hat{y}$ by passing the exemplar node activations $a = [a_1, \ldots, a_8]$ through a single linear layer W_o and applying a sigmoid operation with slope hyperparameter ϕ as follows:

$$\hat{y} = \sigma(\phi W_o(a)). \tag{3}$$

CNN-MLP Model The CNN-MLP model accepts the same high-dimensional representation of the input stimulus that the CNN-ALCOVE model sees. However, unlike ALCOVE, the CNN-MLP does not contain any stored exemplars, nor does it learn an attention weight vector. Instead, it passes the CNN feature representation through a two-layer multilayer perceptron (with corresponding weights W_h and W_o), containing eight hidden nodes to match the number of exemplars, followed by a tanh non-linearity and a single sigmoid output node. The predicted category label $\hat{y}$ is then:

$$\hat{y} = \sigma(\phi W_o(\tanh(W_h(f(x))))), \tag{4}$$

with bias terms omitted for clarity.

Because the CNN-MLP does not contain any exemplars or attention-learning mechanisms, it provides a useful baseline to determine whether or not the extra components in the CNN-ALCOVE model are required.

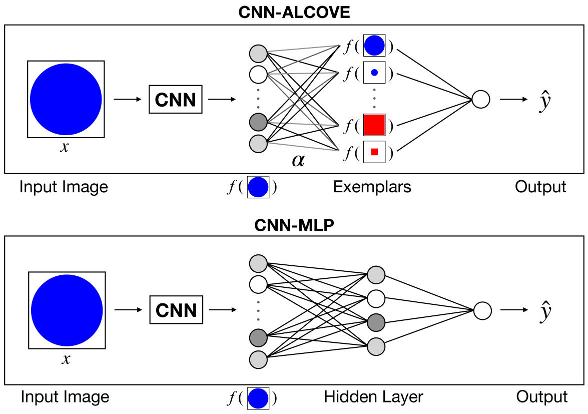

Figure 2: **Architecture for the two categorization models**. In both models, the image stimulus x is first passed through a pre-trained CNN, which extracts a high-level feature representation. For the CNN-ALCOVE model (top), this high-level representation is then reweighted based on the learnable attention parameters and its similarity computed to all of the other exemplars, then passed into a feedforward layer to predict its category label. For the CNN-MLP model (bottom), the extracted features are passed into a simple two-layer neural network which computes the category label without any attention or exemplar comparisons.

Abstract Models For an additional comparison, we also tested variants of the two models that were trained using the standard binary feature encodings of SHJ stimuli as inputs rather than raw pixel images, replacing the CNN front-end with an input layer of three units. These are equivalent to the models in Kruschke (1992), with the exception of the loss function as described below. We referred to these models as Abstract-ALCOVE and Abstract-MLP respectively.

Training Details Each model was trained for 128 epochs. We used a binary cross-entropy loss for all simulations[1], and gradient updates were performed in batches containing all eight stimuli. Optimization was performed using stochastic gradient descent, using separate learning rates for the attention and associative weights, based on the hyperparameter ranges outlined below.

We tested each model, image set and loss combination on a range of hyperparameters centered around Kruschke's (1992) original values[2]. For CNN-ALCOVE and Abstract-ALCOVE, this meant varying c, ϕ, the attention learning rate, and the association learning rate, while for the CNN-MLP and Abstract-MLP models, we varied ϕ and the association learning rate. For the CNN-ALCOVE model, we varied the attention learning rate over the values $[0, 0.0025, 0.005]$, where models trained with a positive attention learning rate were further grouped as **CNN-ALCOVE-Attn** models, while

[1] We tested a number of other loss types, including the humble teacher loss originally described in Kruschke (1992), but found that the binary cross-entropy loss yielded the best results.

[2] The original hyperparameter values reported were attention learning rate = 0.0033, association learning rate = 0.03, c = 6.5, and ϕ = 2.0.

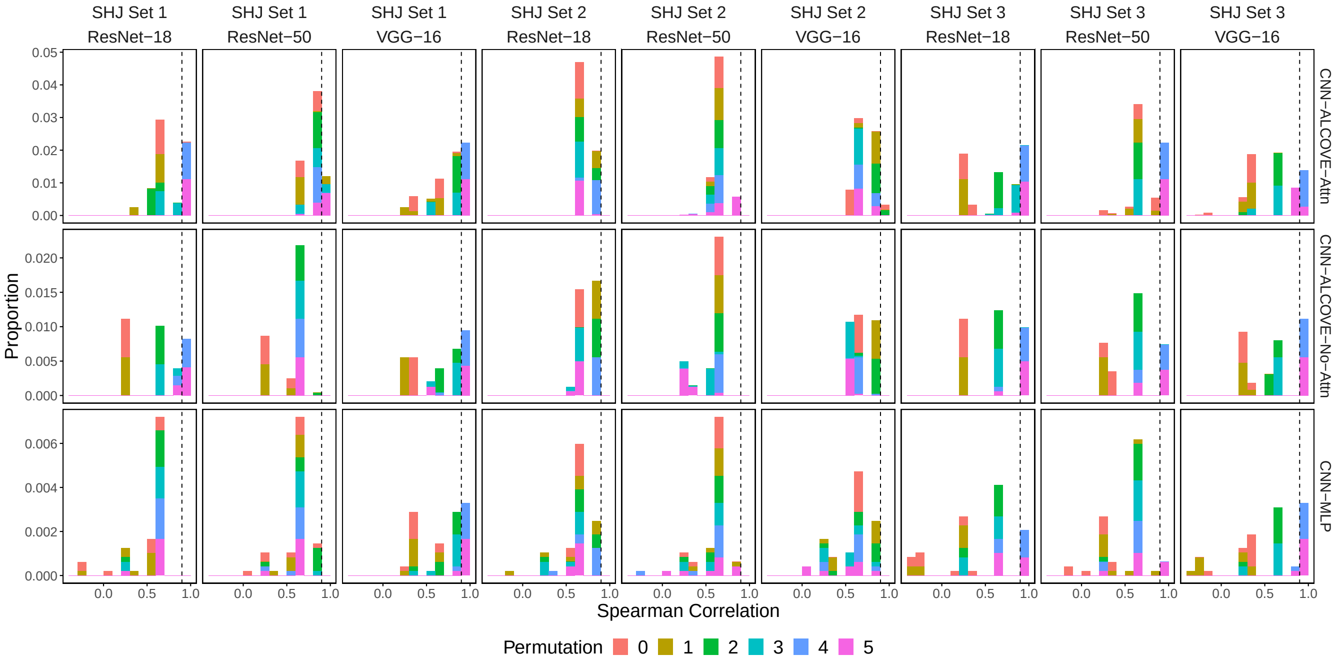

Figure 3: **Histogram of correlation scores across all models, datasets and CNN architectures**. Each panel shows the histogram of correlation scores for that simulation, with rows indicating the three models, and columns indicating each possible CNN and dataset combination. The vertical dashed line indicates the threshold for a correlation to represent a successful type ordering, thus the colored bars to the right of this line represent the set of simulations that produced correct type orderings. Our results show that all three models can produce correct type orderings with most CNNs, but only for two out of three datasets and for a subset of permutations.

the subset with a zero attention learning rate was grouped as **CNN-ALCOVE-No-Attn**. We included this set of models with no attention learning in order to assess the importance of CNN-ALCOVE's attention weights for reproducing the type ordering. The Abstract-ALCOVE simulations were similarly grouped as **Abstract-ALCOVE-Attn** and **Abstract-ALCOVE-No-Attn**. The association learning rate was varied over the values $[0.01, 0.025, 0.05]$; c over $[2.5, 5.0, 7.5]$; and ϕ over $[1.0, 2.5, 5.0]$. Each simulation consisted of training the model with six category types, across six permutations for a total of 36 separate models trained per simulation.

Scoring the Results In order to determine whether the CNN-ALCOVE or CNN-MLP models could capture the type ordering effect, we first needed an automated method to determine whether a given simulation's learning curves matched the intended type ordering. For each simulation, we calculated the learning curves for each of the six category types as the average probability of assigning the correct category label to each stimulus per batch across all 128 epochs. For each of the six types, we then calculated the integral, or the area under the learning curve, using the trapezoidal rule, whose value determined the rate at which a given category type was learned (see Figure 5 for examples of some learning curves). Generally, a category type that was learned faster than other types would have produced a larger integral, while a category type that was learned slower would have produced a smaller integral. We chose the integral as our learning metric because we were primarily interested in the type orderings of the learning curves rather than their precise shapes. A more precise comparison would be difficult due to differences in the number of training epochs. These values were calculated for two separate cases: one where the learning curves were first averaged across all six permutations within a given simulation, and another where the learning curves for each permutation were treated as distinct simulations.

After obtaining the integrals for each category type, we computed the ordering of the integrals to determine which category types were learned in the inverse order (so that the fastest learned type would have a value of 1, and the slowest learned type would have a value of 6). Then, we set the true ordering to be the vector $[1, 2, 4, 4, 4, 6]$, treating Types III, IV and V as equal reflecting the interchangability of these types, and calculated Spearman's rank correlation coefficient between our inverted ordering and this true ordering. Our results showed that all six possible type orderings that were consistent with the SHJ type ordering effect achieved a Spearman rank correlation coefficient of 0.94, while a single incorrect ranking produced a Spearman rank correlation of 0.82. Therefore, we set a threshold of 0.9, and this value served as the floor for detecting whether a given simulation resulted in correct ordering.

Results

Average Learning Curves We first report results on our simulations where we calculated learning curves averaged across permutations (counter-balancing the assignment of physical to abstract features). We found that for the CNN-ALCOVE-Attn model, only 19 (3.9%) of the 486 simulations

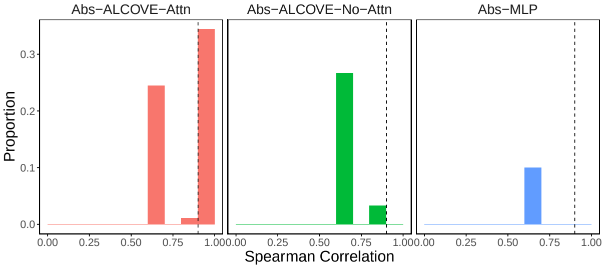

Figure 4: **Histogram of correlation scores for the abstract simulations**. Each panel shows the histogram of correlation scores for each of the abstract models. The vertical dashed line indicates the threshold for a correlation to represent a successful type ordering, and thus bars to the right of this line represent the set of simulations that produced correct type orderings.

resulted in correct orderings. Further inspection revealed that 100% of these correct orderings were produced within one CNN architecture (ResNet-18) and from one dataset (SHJ Set 1). For the CNN-ALCOVE-No-Attn and CNN-MLP models, we found that 0% of the simulations resulted in successful orderings. However, we noticed that many of the remaining simulations produced *almost* correct orderings where two adjacent ranks were swapped, which were almost always instances where Type II was learned more slowly one of the three Types III, IV or V. For CNN-ALCOVE-Attn, we found that 25.5% (166/486) simulations were almost correct in this manner.

On the other hand, when using the abstract feature representations, we found that 57% of the simulations from the Abstract-ALCOVE-Attn model resulted in correct orderings, showing robust success across a wide range of hyperparameters, and much more so than the CNN variants.[3] Furthermore, 0% of the Abstract-ALCOVE-No-Attn simulations and Abstract-MLP simulations resulted in correct type orderings, matching what Kruschke (1992) observed and highlighting the importance of attentional learning. A histogram showing the correlation scores and proportion of correct orderings is shown in Figure 4.

There were some positive signals in the current results; by modifying an existing, well-known model of category learning with a CNN front-end, we observed runs where we were able to reproduce the same type ordering as observed in Shepard et al. (1961), albeit not reliably so. Moreover, the successful simulations that displayed the effect were from our CNN-ALCOVE-Attn model, highlighting the importance of both the exemplar model representation and the attentional learning components of the model relative to the CNN-ALCOVE-No-Attn and CNN-MLP models, which lacked one or both of these components. However, there was also an unsatisfactory aspect to these results: why was the type ordering effect only observed in a very limited subset of CNN simulations? The lack of generalization across architectures and image sets suggests serious robustness issues in using pre-trained CNNs as a means of feature extraction for models of categorization.

Individual Learning Curves Rather than taking the resulting learning curves by averaging across all possible permutations per simulation, we also applied the same ranking procedure to obtain correlation scores for each possible permutation separately for the CNN simulations. One reason why examining permutations separately might be useful has to do with how the Type II rule works. The Type II rule is governed by an exclusive-or (XOR) rule, which requires attending to two out of three features for correct classifications. Traditionally, when modeling this task using the abstract encodings, there was no difference in the various possible permutations, and thus it was common to average results. However, when looking at the raw images as is relevant for the CNN models here, there are three separate possible combinations of two-out-of-three features. For example, in SHJ Set 1, the Type II rule across the permutations may involving attending to either color and dots, or color and a line, or dots and a line, and there is the possibility that the feature representations for different permutations result in large learning differences.

Figure 3 shows the histogram containing these correlation scores, with each row depicting a different model, and each column depicting a different combination of stimuli and CNN architecture used. Here, a very different pattern emerges, with three interesting findings. First, in contrast to above, all models show evidence of successful type orderings in some of the conditions, suggesting that even the CNN-MLP architecture is sufficient to produce a correct type ordering result. However, we found that the proportion of successful orderings was highest for the CNN-ALCOVE-Attn model with 19.6% of simulations producing correct type orderings. This was followed by the CNN-ALCOVE-No-Attn with success in 15.4% runs, and then the CNN-MLP with 9.3% of runs. Furthermore, the set of successful runs across the three models were generally from the same image set and CNN architecture combinations, suggesting that the initial feature representation extracted from the pre-trained CNN plays a key role in whether a type ordering is observed, but that adding the additional components of ALCOVE such as the exemplar layer and attentional learning can enhance the success rate. While the total portion of successful simulations was still quite low, overall the correlation scores were quite high and produced many close orderings.

Second, the successes are concentrated among certain permutations, following our earlier hypothesis. For example, for SHJ Set 1 and SHJ Set 3, almost all of the successes are due to permutations 4 and 5, where the relevant features are color with line and color with shape respectively. However, the fact that we observed failures with the remaining permutations suggests that the initial feature representation from the CNN for both image sets was still insufficient for learning a Type II rule quickly in the other permutations. Additionally, we find very few successes for SHJ Set 2 (with the exception of a few runs with the CNN-ALCOVE-Attn model with

[3] One third of the failures occurred in the simulations where $c = 2.5$, suggesting that the failure occurred due to the generalization gradient being too wide.

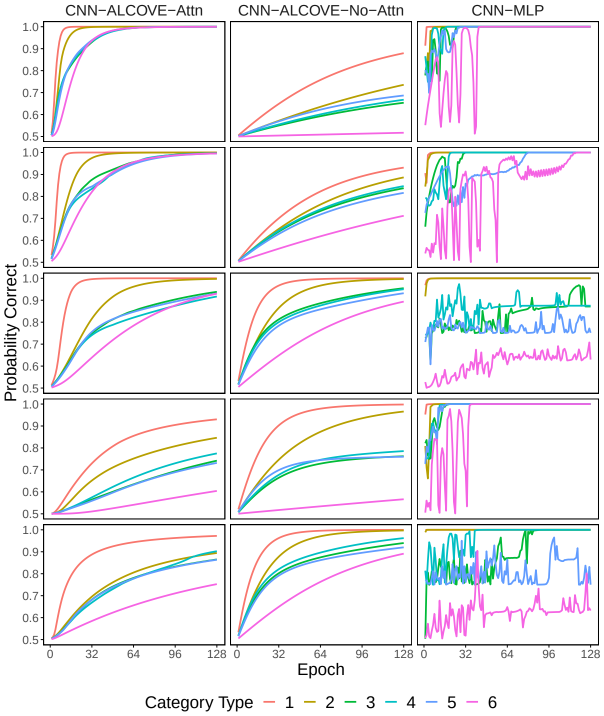

Figure 5: **Sample learning curves from each model**. In this figure, we display five randomly selected successful runs of each model where the type ordering effect was observed. While the CNN-ALCOVE models show smooth learning curves, the learning curves from the CNN-MLP were disjoint despite a correct ordering.

VGG-16), despite having the same set of abstract features as SHJ Set 3. These results suggest a failure at the feature representation level from the feature extraction process from the CNNs, affecting learning for all possible permutations.

Finally, some of the the individual learning curves for a single simulation are shown in Figure 5. We noticed that the learning curves from the CNN-ALCOVE models looked smooth and qualitatively similar to human data on this task. On the other hand, in the CNN-MLP model, despite the successful ordering based on our rank correlation analysis, there were very large oscillations the learning curves for certain category types.

Discussion

In this paper, we presented the CNN-ALCOVE model, an extension of the original connectionist category learning model ALCOVE proposed by Kruschke (1992) that adds a convolutional neural network as a front-end to facilitate learning from raw images as input. We examined whether such a model could reproduce the classic type ordering effect from Shepard et al. (1961). We compared this model (CNN-ALCOVE-Attn) to a variant of our model without attentional learning (CNN-ALCOVE-No-Attn) as well as a model without attentional learning or exemplar comparison (CNN-MLP), and variants of these using the abstract feature encodings; these models served as controls and allowed us to assess the importance of CNN-ALCOVE's attention learning and exemplar comparison. We tested each of these models on three different image datasets as well as a wide range of hyperparameters. Overall, our results presented a mixture of success and failure. When examining for successful type orderings by averaging learning curves across all permutations, we only observed successes with the CNN-ALCOVE-Attn model for one CNN architecture with one dataset, although many of the other simulations produced almost correct orderings. On the other hand, the equivalent model with the abstract feature encodings (Abs-ALCOVE-Attn) produced a high proportion of correct type orderings.

A different pattern emerged in our analysis when considering each permutation of the stimuli as distinct. Here, we found many successful orderings for all three models and for each of the CNN architectures on two out of the three image datasets. The highest number of correct orderings was observed in the CNN-ALCOVE-Attn model, followed by the variant without attentional learning (CNN-ALCOVE-No-Attn), and then the CNN-MLP, confirming previous findings that both exemplar comparison and attentional learning are crucial ingredients (Kruschke, 1992). However, we failed to observe successful orderings for certain permutations of the SHJ stimuli, and very few runs with the SHJ Set 2 images yielded success. This work was motivated by the question of whether ImageNet pre-trained CNNs could be used as a drop-in visual front-end to replicate patterns of human category learning for novel categories, and the answer appears to be no, at least not reliably so in the cases we examined.

There are also a few potential explanations for this. First, the feature representations extracted from the pre-trained CNNs may have not been conducive to learning in certain category structures or permutations. Recent work has shown that pre-trained CNNs are in some ways quite unlike human vision, such as a preference for texture rather than shape (Geirhos et al., 2018), or being manipulated through adversarial examples (Szegedy et al., 2013). Since these extracted feature representations are the starting representation for all of our models, one can imagine that a poor input representation would cause downstream difficulties in accounting for human performance.

Second, instead of distributing attention over three interpretable features as the original ALCOVE model does, the CNN-ALCOVE models must learn to distribute attention over a much larger set of distributed, high-dimensional features that do not correspond one-to-one with the relevant abstract dimensions of the problem to be learned. Additionally, attending to the underlying abstract feature (e.g., color, shape, size) may be easier or harder depending on the initial feature representation the model is working with. Our results suggest that perhaps a combination of these factors may be at play here, and future work should examine both more human-like CNN architectures (Kubilius et al., 2019), or to use CNN architectures that have been pre-trained on larger datasets containing both natural and abstract images.

Another possible explanation is that the type ordering effect is a more complicated phenomena than previously assumed. Despite the original finding replicating multiple times, the empirical data on human type orderings is quite mixed, with a body of work demonstrating that the original SHJ type orderings are not always reliably reproduced in human participants. Particularly relevant is the work by Kurtz, Levering, Stanton, Romero, and Morris (2013), which showed that human participants only learn Type II categories reliably faster than Type IV when they are told this is a rule-learning task. This seems to parallel the majority of our CNN-ALCOVE results, where the model learns Type II slower than Types III-V.

The success of convolutional neural networks have been instrumental in stimulating a variety of deep categorization models that capture patterns of human categorization with naturalistic stimuli (Battleday et al., 2020; Singh et al., 2020; Nosofsky et al., 2020). In this paper, we add to this line of research by proposing an attention-weighted exemplar model with a CNN front-end which we call CNN-ALCOVE. However, rather than testing our model with naturalistic images where previous deep categorization models have been more successful, we examine the conditions that can reproduce the classic type ordering effect from Shepard et al. (1961), where rule-based categories play a more important role. Overall, our results suggest that these approaches hold promise, but there are difficulties in applying CNNs out of the box to interface with raw experimental stimuli. We also hope our work provides a useful starting point for examining how attention can be deployed in neural networks for categorization (Lindsay, 2020).

Acknowledgments

Thanks to Kanishk Gandhi for helpful feedback on the draft. This work was supported by NIH grant R90DA043849.

References

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2020). Capturing human categorization of natural images by combining deep networks and cognitive models. *Nature communications*, *11*(1), 1–14.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Guest, O., & Love, B. C. (2019). Levels of representation in a deep learning model of categorization. *BioRxiv*, 626374.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Kubilius, J., Schrimpf, M., Kar, K., Hong, H., Majaj, N. J., Rajalingham, R., ... others (2019). Brain-like object recognition with high-performing shallow recurrent anns. *arXiv preprint arXiv:1909.06161*.

Kurtz, K. J., Levering, K. R., Stanton, R. D., Romero, J., & Morris, S. N. (2013). Human learning of elemental category structures: Revising the classic result of shepard, hovland, and jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 552.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In *Cogsci*.

Lindsay, G. W. (2020). Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, *14*, 29.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic bulletin & review*, *9*(4), 829–835.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961). *Memory & cognition*, *22*(3), 352–369.

Nosofsky, R. M., Meagher, B., & Kumar, P. (2020). Contrasting exemplar and prototype models in a natural-science category domain.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, *42*(8), 2648–2669.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singh, P., Peterson, J. C., Battleday, R. M., & Griffiths, T. L. (2020). End-to-end deep prototype and exemplar models for predicting human behavior. *arXiv preprint arXiv:2007.08723*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.