Combining Convolutional Neural Networks and Cognitive Models to Predict Novel Object Recognition in Humans

Jeffrey Annis (jeff.annis@vanderbilt.edu)
Vanderbilt University, 111 21st Avenue South
Nashville, TN 37240 USA

Thomas J. Palmeri (thomas.j.palmeri@vanderbilt.edu)
Vanderbilt University, 111 21st Avenue South
Nashville, TN 37240 USA

Abstract:

Cognitive models have become ubiquitous in cognitive science and cognitive neuroscience, playing a key role in understanding visual cognition, providing insights into how we recognize, remember, and categorize objects. Cognitive models are often relatively abstract, instantiating high-level aspects of visual cognition, such as how visual evidence is represented, how it is accumulated, and how response bias and caution combine to predict errors and response times in perceptual decisions. Many such models instantiate mechanisms flowing from an object representation to a perceptual decision but do not specify how an object representation is created from the visual image of an object. Convolutional Neural Networks (CNNs) have become successful at visual tasks like classifying objects in real-world images. We explore if CNN object representations, built up over a network hierarchy from object images, can be used as input to a cognitive model to predict human recognition performance. We specifically use CNN representations to drive a cognitive model of decision making, the Linear Ballistic Accumulator (LBA), to predict a range of performance in a visual matching task with novel objects.

Keywords: Convolutional Neural Network, cognitive model

Introduction

Our understanding of visual cognition has been enhanced by the development of cognitive models that instantiate hypothesized mechanisms involved in visual recognition, categorization, and memory, and are tested on how well they predict the detailed patterns of errors and response times observed in visual tasks. Some of the most successful have been sequential sampling models. These assume that after an object is perceptually encoded with respect to a visual task, noisy evidence accumulates towards decision thresholds defined by the task. When evidence reaches threshold, a response is made. Such models predict human performance in a wide range of visual tasks and provide important insights into how the brain performs these tasks (Forstmann, Ratcliff, & Wagenmakers, 2016).

Many such cognition models begin in the "middle" of processing. They assume, for example, that objects are represented as a vector of features in a multidimensional psychological space, with the values of those features determined by physical properties, psychophysical measures, or psychological scaling. These object representations are used to generate evidence that drives a decision regarding recognition, memory, or categorization; in some cases, the evidence itself is a freely estimated parameter of the model. Rarely are these models presented the same images of objects that are shown to human observers. They are not equipped with an explicit visual "front end" for perceptual processing.

Recently, a class of neural networks known as Convolutional Neural Networks (CNNs) have become increasingly successful at certain visual tasks such as classifying objects in real-world images (e.g., Krizhevsky, Sutskever, & Hinton, 2012), and have, in some cases, exceeded human performance (e.g., He, Zhang, Ren, & Sun, 2015). This is largely due to their ability to learn useful representations through training on large image databases (Russakovsky et al., 2015), which enable them to generalize to new images at testing. While CNNs are inspired by the deep, hierarchical, convolutional nature of the primate visual system, they are not designed to be models of primate vision. CNNs are designed with the goal of achieving the highest level of accuracy possible on a given recognition task rather than achieving accurate predictions of human response time and choice patterns. That said, they have provided insights into high-level object representations in the brain (Kriegeskorte, 2015; Yamins et al., 2014).

CNNs and cognitive models have complimentary strengths: CNNs are capable of forming representations of complex visual images that have proven useful for tasks like object classification, while cognitive models are capable of predicting human response time and choice in a wide range of tasks. We ask whether CNN representations can be used as input to cognitive models to predict human performance.

Here, we connect CNNs to cognitive models by using the representations formed by the CNNs to drive evidence accumulation to predict details of observed response times and accuracies on a matching task using novel objects. Novel objects have played a significant role in our understanding of visual cognition (e.g., Gauthier, Williams, Tarr, & Tanaka, 1998; Richler & Palmeri, 2014). One reason for this is that novel objects allow for the control of prior experience. In the present case, novel objects also allow us to better equate experience between CNNs and human observers in that both

have been exposed to millions of exemplar images of everyday objects but neither CNNs nor humans have ever seen the novel objects. Basic questions we explore are whether CNNs are capable of generalizing to novel object recognition and whether their representations can be used to drive a cognitive model to predict the range of observed human performance.

Method

Matching Task

Data were from 215 participants who observed a novel-object matching task (Richler et al., under review). Five categories of novel objects from previous research were used: two types of *greebles*, two types of *ziggerins*, and *sheinbugs* (see Figure 1); each category contained 50 exemplar objects. Participants completed five matching-task blocks, each with 180 trials. Each block used a different category of novel objects. On each trial, the participant was presented with a novel object for 150 ms followed by a mask for 500 ms. Then, the participant was presented with another novel object that was either the same or different. The test object varied in viewpoint (same or different) and size (same or different). Participants were instructed to respond "same" if the two objects had the same identity, regardless of viewpoint or size, and to respond "different" if the identity of the objects was different. Each condition (viewpoint, size, same/different identity) was presented an equal number of times and the order randomized. The order of the trials was the same for all participants.

Modeling Methods

CNN. CNN models assume a deep hierarchy of processing layers. In many models, initial layers are combinations of *convolutional layers* and *pooling layers*, and later layers are *fully connected* (*fc*) layers (see Panel A of Figure 1. Convolutional layers are essentially learned filters (or feature detectors) that operate over the input from the previous layer (mirroring the mathematical operation of *convolution*) and produce a new 2-dimensional activation map. With training, a CNN essentially learns banks of filters that respond to different features in the image, such as blobs and edges in early layers or more complex object features in later layers. After each (or some) of the convolutional layers, a *pooling layer* is used to downsample the representation; pooling both helps create translation invariance and reduces the number of parameters and computational load.

After several (sometimes dozens of) convolution and pooling layers, the representation is (often) passed to *fully-connected (fc) layers*, which have connections to all units in the previous layer (like a traditional feed-forward neural network model). Typically, the units in the last fc layer are connected to an output layer with each unit corresponding to a category. The unit with the highest activation in this final layer corresponds to the best classification by the CNN.

We tested three different CNN architectures: VGG-16 (Simonyan & Zisserman, 2014), ResNet50 (He, Zhang, Ren, & Sun, 2016), and Inception v3 (Szegedy et al., 2015). These networks were all pre-trained on a standard large-scale corpus of real-world images of objects. No additional training with novel objects was done. Images of novel objects (greebles, ziggerins, and sheinbugs) were simply presented to the network(s) and we assumed the penultimate network representation (before the final classification layer) to be the CNN novel object representation.

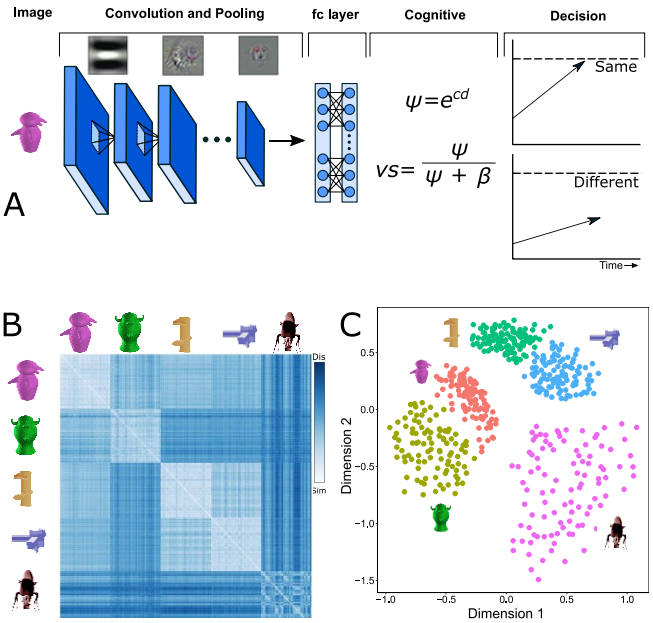

Figure 1. Panel A: Connecting a CNN to a cognitive model. Panel B: Representational Distance Matrix (RDM) for novel objects. Panel C: Multidimensional Scaling (MDS) solution for CNN novel object representations.

Decision Level. We first describe the decision model before describing how we connect it to the CNN via cognitive-level operations. The particular decision model we used is the LBA (Brown & Heathcote, 2008), illustrated on the right of Panel A of Figure 1. LBA is a type of *sequential sampling model*, which assume that decisions are made by accumulating evidence to threshold over time. Evidence is sampled from internal representations at a given rate, called the *drift rate*, and a response is made when the amount of accumulated evidence reaches a predetermined *response threshold*.

The LBA assumes that evidence accumulation begins after an item is perceptually encoding with time, τ_e. Accumulators corresponding to each response alternative, r_m, begin to accumulate evidence with drift rate, δ_m, towards a response threshold, b_m. Drift rates are sampled on each trial from a normal distribution with mean, v_m, and standard deviation, s_m. The starting point for each accumulator is sampled from a uniform distribution between 0 and A_m, where $A_m < b_m$. Accumulation terminates when the first response threshold is

reached. A response corresponding to that accumulator is made after motor execution time, t_r (where $\tau = \tau_e + \tau_r$).

Cognitive Level. Decision models like LBA are generic in that drift rates are free parameters and the accumulators can correspond to any response alternatives. Here, we assume that the responses are "same" versus "different" and we create a theory of the LBA drift rates based on CNN novel object representations. To simulate a trial of object matching, the two objects are presented to the CNN, the distance between these representations are computed, these are turned into similarities, which are then scaled relative to a criterion to create drift rates driving same vs. different accumulators.

First, Euclidean distance between CNN representations of novel objects i and j in category k, d_{ijk}, was computed. Panel B of Figure 1 illustrates the *representational distance matrix* (RDM) formed by all pairwise distances between ResNet50 representations; dark color represents distant novel objects and light color represents close novel objects. To help assess whether novel object distances produced by the CNN were sensible, we obtained a *multidimensional scaling* (MDS) solution of the RDM in two dimensions. Panel C of Figure 1 shows that the CNN sensibly clusters different types of novel objects in the MDS.

Next, distance between i and j is transformed to similarity, ψ, for participant, p, in category, k:

$$\psi_{pijk} = \exp(-c_{pk} d_{ijk}),$$

where c_{pk} is a subject-specific sensitivity parameter that governs how similarity decreases with increases in distance.

This similarity is then used in a function that rescales values between 0 and 1 to produce the mean drift rate for the "same" accumulator:

$$vs_{pijk} = \frac{\psi_{pijk}}{\psi_{pijk} + \beta_{pk}},$$

where β_{pk} is a subject-specific criterion that governs the bias to respond "different"; the mean drift rate for the "different" accumulator is set to $1 - vs$.

Bayesian Implementation and Fitting Methods

We used three different CNNs to derive novel object RDMs: VGG-16, ResNet50, and Inception v3. With the Keras package (Allaire & Chollet, 2018), RDMs were derived from each CNN by presenting each CNN with each novel object, obtaining the penultimate CNN representation, and finding all pairwise distances between pairs of representations.

We tested two additional models as benchmarks. The first benchmark had exactly the same structure as the CNN-based models, except that distances were derived from raw images, rather than CNN representations. This *pixel-based model* is one that any viable CNN-based model must surpass. The second benchmark we refer to as a *free model*. This is a purely abstract cognitive model in that drift rates were simply free parameters, not constrained by CNN representations.

All models were implemented in a Bayesian hierarchical framework. For each of the CNN and pixel-based models, participant, p, presented with stimulus pair (i,j) in category, k, choice response time pairs, $\boldsymbol{RT}$, were assumed to be distributed according to the LBA:

$$\boldsymbol{RT}_{pijk} \sim LBA(A_{pk}, b_{pk}, vs_{pijk}, \tau_{pk}, ss_{pk}, sd_{pk}),$$

where ss and sd are the standard deviations of the drift rates of the "same" and "different" accumulators, respectively. The prior on each participant-level parameters followed a truncated normal (from 0 to infinity) with its own respective mean, μ_{pk}, and variance, σ_{pk}. Group-level priors on the means and variances were as follows:

$$\mu^b_{pk}, \sigma^b_{pk} \sim TN(1.4, 1.4)$$
$$\mu^\tau_{pk}, \sigma^\tau_{pk} \sim TN(.3, .3)$$
$$\mu^A_{pk}, \sigma^A_{pk}, \mu^{ss}_{pk}, \sigma^{ss}_{pk}, \mu^{sd}_{pk}, \sigma^{sd}_{pk} \sim TN(1,1).$$

Group-level priors on means and variances of parameters involved in transformation also followed a truncated normal:

$$\mu^c_{pk}, \sigma^c_{pk}, \mu^\beta_{pk}, \sigma^\beta_{pk}, \mu^{w_l}_{pk}, \sigma^{w_l}_{pk} \sim TN(1,1).$$

For the free model, the "same" accumulator for participant p given category k, is allowed to freely vary across conditions of object identity, o (same vs. different), and viewpoint, h (same vs. different):

$$\boldsymbol{RT}_{pohk} \sim LBA(A_{pk}, b_{pk}, vs_{pohk}, \tau_{pk}, ss_{pk}, sd_{pk}).$$

All priors were the same as those just described with an additional prior placed on vs:

$$vs_{pohk} \sim TN(\mu^{vs}_{pohk}, \sigma^{vs}_{pohk}),$$

where

$$\mu^{vs}_{pohk}, \sigma^{vs}_{pohk} \sim TN(3,3).$$

Note, for simplicity, we do not explicitly model size conditions (same vs. different) in any of the models (in part because there was almost no behavioral effect of size).

We used Differential Evolution MCMC (DE-MCMC; Turner, Sederberg, Brown, & Steyvers, 2013) to estimate the joint posterior distribution. We used two times the number of subject-level parameters as the number of chains and ran the sampler for a total of 3000 iterations discarding the first 1000 as burn-in. Chains were visually inspected for convergence.

Model selection was performed by computing the Bayes factor, a ratio of the evidence provided by the data in favor one model over another. We report $\log_{10}$ Bayes factors in terms of the free model. Bayes factors less than 0 indicate evidence in favor of the free model, while Bayes factors greater than 0 indicate evidence against.

Results

Panel A of Figure 2 shows the rank ordering of models in terms of the Bayes factor (compared to the free model). The data provided decisive evidence in favor of the models using the RDM derived from ResNet50 and Inception over the free model. Defining drift rates for the LBA from RDMs derived from ResNet50 and Inception resulted in an improvement over the LBA alone (free model). Models using similarities derived from VGG-16 or from raw images (pixel-based) did worse than the free model.

Panel B of Figure 2 shows predicted vs. observed accuracy (left panel) and response time quantiles (right panel) for the

best-firtting model, ResNet50. ResNet50 underestimates the accuracy in the same-object different-viewpoint condition, which requires the representation of the CNN to be viewpoint invariant (a deviance we continue to investigate).

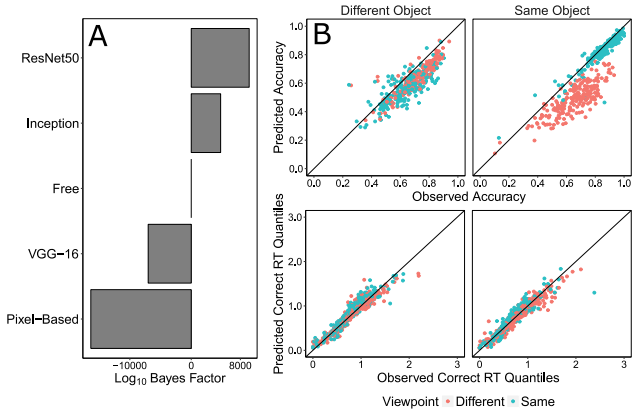

Figure 2. Panel A: Rank ordering of models in terms of the Bayes factor. Panel B: Predicted vs. observed accuracy and response times for ResNet50.

Conclusion

We used CNN representations to drive the rate of evidence accumulation, drift rate, in a sequential sampling model, known as the Linear Ballistic Accumulator (LBA; Brown & Heathcote, 2008), to predict choice response times on a matching task using novel objects. We chose to use a task with novel objects because of the importance of novel object recognition in research and to better equate experience between the CNNs and human participants. We used three different CNNs: VGG-16, Inception v3, and ResNet50. We obtained the penultimate representations for each novel object, whose pairwise distances formed a Representational Distance Matrix (RDM). We then used the distances from the RDM to derive similarities between objects. These similarities were then scaled and used as the mean of the drift rates in the LBA. We found the data provided decisive evidence in favor of combining CNNs with the LBA over the LBA alone. Specifically, the data provided the most evidence for the LBA driven by the RDM from ResNet50.

Acknowledgements: Support by NSF grant SMA-1640681, NEI core grant P30-EY008126, and NEI training grant T32-EY007135.

References

Allaire, J. J., & Chollet, F. (2018). keras: R Interface to "Keras." Retrieved from https://keras.rstudio.com

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*.

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. W. (1998). Training greebles'experts: a framework for studying expert object recognition processes. *Vision Research*, *38*(15–16), 2401–2428.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026–1034).

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).

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446.

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

Richler, J. J., & Palmeri, T. J. (2014). Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*, 75–94.

Richler, J. J., Tomarken, A. J., Sunday, M. A., Vickery, T. J., Ryan, K. F., Floyd, R. J., … Wong, A. C. N. (submitted). Individual differences in object recognition.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

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

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Computer Vision and Pattern Recognition*. Retrieved from http://arxiv.org/abs/1409.4842

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368–384.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.