A Dichotomy of Visual Relations

Junkyung Kim (junkyung_kim@brown.edu)
Brown University

Matthew Ricci (matthew_ricci_1@brown.edu)
Brown University

Dan Shiebler (dan@gotruemotion.com)
True Motion

Thomas Serre (thomas_serre@brown.edu)
Brown University

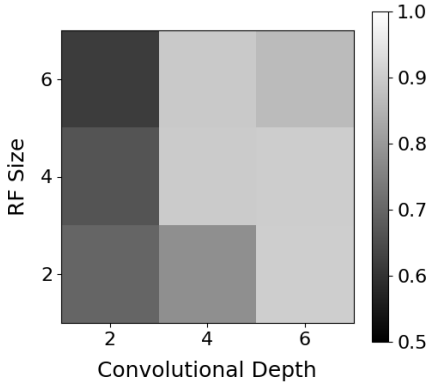
Abstract

Convolutional neural networks (CNNs) have achieved state-of-the-art performance in image classification (He et al., 2015). However, a growing body of work indicates that CNNs still struggle on visual rule-learning tasks (Fleuret et al., 2011; Gülçehre & Bengio, 2013; Ellis et al., 2015). Currently, our understanding of precisely which rule-based problems are hard or easy for CNNs is limited. Here, we conducted a systematic analysis of CNN performance on the 23 problems of the Synthetic Visual Reasoning Test (SVRT), while varying network hyperparameters. We find that one group of SVRT problems is easily solved by most networks, whereas another group is not solved at all. We propose that the soluble problems of this dichotomy rely only on *spatial relations*. Intractable problems, on other hand, require *same-different* judgments, in which image regions must be compared. We conclude by sketching a novel cognitive architecture designed to solve visual reasoning problems.

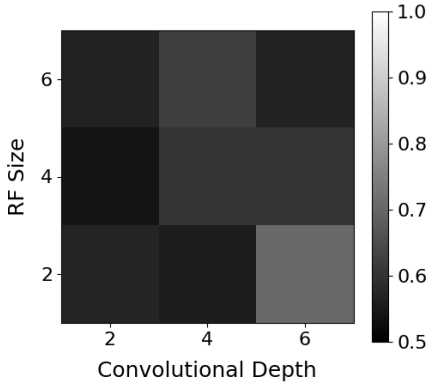
Keywords: visual relations; convolutional networks; attention; memory; reinforcement learning; mental imagery

Background

Feedforward neural networks have recently surpassed human performance in image classification (He et al., 2015). Despite this achievement, machine learning algorithms lag far behind humans on visual reasoning tasks. For example, Fleuret et al. (2011) found that black-box Adaboost and support vector machine classifiers failed on a variety of binary visual relation tasks despite massive amounts of training. Human subjects solved most of these problems effortlessly and with very few examples. Later, Ellis et al. (2015) extended this finding to CNNs, showing that these models could easily solve some problems but found others entirely intractable. Gülçehre & Bengio (2013) showed that feedforward networks struggled even to determine whether two simple shapes were the same and could only achieve above-chance performance with explicit hints. Humans, on the other hand, solve same-different problems with ease (Shepard & Metzler, 1971). Currently, there is no accepted theory of *why* some visual relation problems are much harder than others. A taxonomy of relational tasks would facilitate the construction of visual reasoning models. Furthermore, as CNNs represent the standard model for the visual cortex, understanding their strengths and weaknesses is likely to provide insight into primate vision.



(a) Average AUCs for SP problems.



(b) Average AUCs for SD problems.

Figure 1: The separability of the two classes in each SP problems was measured by calculating the area under the ROC curve on a validation set. Classes are easily separated across all parameter settings. *b*) The case is radically different for SD problems, where performance is low overall.

Method and Results

To investigate which reasoning problems are particularly difficult for CNNs, we trained a family of networks parametrized by depth and per-layer receptive field (RF) size on 1M images from each of the 23 visual reasoning problems from Fleuret et al. (2011). A problem consists of images of simple closed curves that must be classified according to whether they obey a binary visual rule (e.g. one shape inside another). Networks had 2, 4 or 6 convolutional layers followed by 3 fully connected layers. RF sizes at each convolutional layer were 2x2, 4x4 or 6x6 with a stride of 2 and 2x2 pooling kernels. Following training, ROC curves were constructed for each problem on a validation set of 1M images and the area under these curves (AUC) was calculated to measure the separability of image classes.

This analysis revealed that reasoning problems fell naturally into two groups. Models had good performance on *spatial relation* (SP) tasks (e.g. inside vs outside) with nearly perfect class separability for deeper networks with larger RF sizes (Fig. 1a). However, on *same-different* (SD) tasks (e.g. identical up to rotation), performance was poor and at chance for most parameter settings (Fig. 1b). This result indicates that a simple, coarse parameter search can find CNNs with sufficient capacity to separate image classes according to a spatial relation. Separation of same-different classes, on the other hand, likely requires massive networks or perhaps a categorically different model not relying exclusively on template matching.

Discussion

Our result suggests that trying to learn features for a visual relation between arbitrary objects can rapidly exceed the capacity of feedforward networks. The strain on a network is particularly acute for same-different relations, in which case it cannot rely on coarse spatial templates. This limitation of CNNs is seemingly overcome by humans via attentional and mnemonic processes occurring after the feedforward sweep of the cortex (Logan, 1994; Franconeri et al., 2012). Hence, we also propose a cognitive architecture (Fig. 2) relying on hierarchical reinforcement learning (Vezhnevets et al., 2017) that uses a shifting window of attention and working memory to detect visual relations. Initial tests indicate that this model is suited to learning challenging same-different tasks.

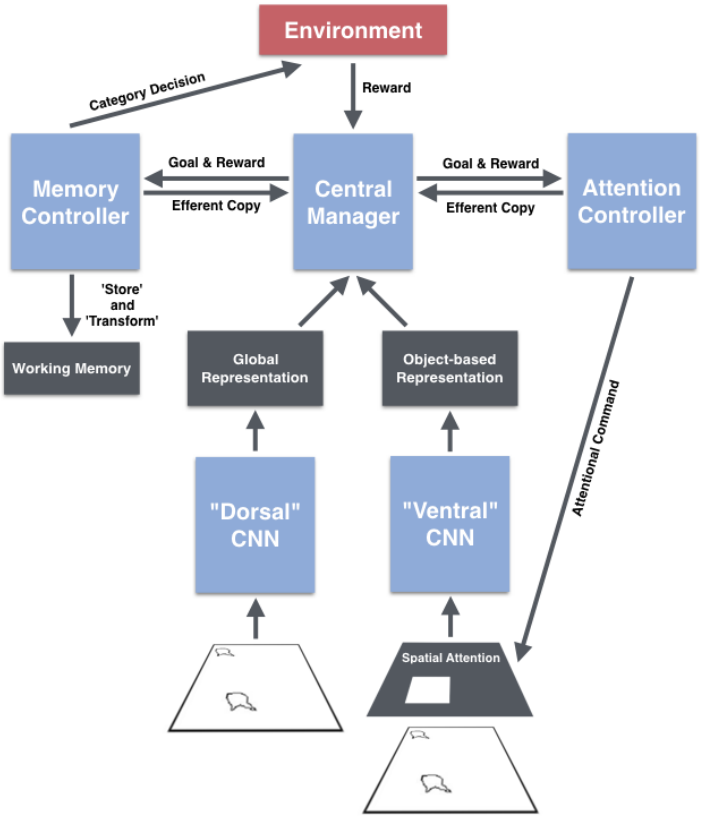


Figure 2: A hierarchical reinforcement learner for visual relations. A CNN is augmented by spatial attention and working memory, which are controlled by a recurrent worker-manager architecture.

References

Ellis, K., Solar-lezama, A., & Tenenbaum, J. B. (2015). Unsupervised Learning by Program Synthesis. *Nips*, 1–9.

Fleuret, F., Li, T., Dubout, C., Wampler, E. K., Yantis, S., & Geman, D. (2011). Comparing machines and humans on a visual categorization test. *Proc. Natl. Acad. Sci. U. S. A.*, *108*(43), 17621–5. Retrieved from http://www.pubmedcentral.nih.gov/ doi: 10.1073/pnas.1109168108

Franconeri, S. L., Scimeca, J. M., Roth, J. C., Helseth, S. A., & Kahn, L. E. (2012). Flexible visual processing of spatial relationships. *Cognition*, *122*(2), 210–227. Retrieved from http://dx.doi.org/10.1016/j.cognition.2011.11.002 doi: 10.1016/j.cognition.2011.11.002

Gülçehre, Ç., & Bengio, Y. (2013). Knowledge Matters : Importance of Prior Information for Optimization. *arXiv Prepr. arXiv1301.4083*, 1–12. Retrieved from http://arxiv.org/abs/1301.4083

He, K., Zhang, X., Ren, S., & Sun, J. (2015, feb). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. Retrieved from http://arxiv.org/abs/1502.01852

Logan, G. D. (1994). Spatial attention and the apprehension of spatial relations. *J. Exp. Psychol. Hum. Percept. Perform.*, *20*(5), 1015–36. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7964527 doi: 10.1037/0096-1523.20.5.1015

Shepard, R. N., & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science (80-.).*, *171*(3972), 701–703. doi: 10.1126/science.171.3972.701

Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., & Kavukcuoglu, K. (2017). FeUdal Networks for Hierarchical Reinforcement Learning. Retrieved from http://arxiv.org/abs/1703.01161