HMAX Strikes Back: Self-supervised Learning of Human-Like Scale Invariant Representations

Nishka Pant* **(nishka_pant@brown.edu)**

Ivan Felipe Rodriguez* **(ivan_felipe_rodriguez@brown.edu)**

Arjun Beniwal (ax2119@nyu.edu)

Scott Warren (ax2119@nyu.edu)

Thomas Serre (thomas_serre@brown.edu)

Abstract

Early hierarchical models of the visual cortex, such as HMAX (Serre et al., 2007; Riesenhuber & Poggio, 2002), have now been superseded by modern deep neural networks. Modern deep neural networks optimized for image categorization have been shown to outperform HMAX (and related models) significantly on image categorization tasks and to fit better neural data from the visual cortex, even though they were not explicitly constrained by neuroscience data. However, earlier hierarchical models were also trained with simpler local learning rules in the pre-deep learning era. So far, these models have yet to be updated with modern gradient-based training methods. Here, we describe a novel contrastive learning algorithm to train HMAX (CHMAX) to learn scale-invariant object representations. Unlike standard deep neural networks trained with data augmentation methods, we show that CHMAX learns visual representations that generalize to novel objects at levels of generalizations comparable to human observers. We hope our results will help spur some renewed interest in other classic biologically-inspired vision models.

Keywords: biologically-inspired vision, Hmax, scale invariance, contrastive learning

Introduction

Modern deep neural networks (DNNs) currently hold the state of the art on image classification tasks (Simonyan & Zisserman, 2014). This has been possible due to the development of training techniques that allow these models to learn from large scale datasets directly via gradient descent (Deng et al., 2009; Schuhmann et al., 2022). However, despite their advanced capabilities, DNNs have not substantially deepened our understanding of cortical processes and their alignment with primate vision remains imperfect (Bowers et al., 2022; Fel* et al., 2022; Linsley et al., 2023). Before the rise of deep learning, models such as HMAX were designed with a focus on anatomical accuracy, attempting to mimic the hierarchical organization of the visual cortex. These models, governed by simpler, local learning rules, were adept at forming stable representations with a limited scope of variation (Serre et al., 2007; Riesenhuber & Poggio, 2002; Mutch & Lowe, 2006). They were, however, not trained with methods that could scale to the complexities handled by modern DNNs.

DNNs often employ extensive data augmentation to simulate scale invariance, that generally fail to achieve genuine invariance, as their learning is restricted to the variations present within the training data (Biscione & Bowers, 2022). This is a significant shortfall, as the biological visual system reliably recognizes objects across a broad spectrum of scales, a capability attributed to true scale invariance (Biederman & Cooper, 1992; Han et al., 2020; Logothetis et al., 1995).

In this context, we revisit the HMAX model and propose the CHMAX. This new model maintain the anatomical hierachy of the original but is augmented with trainable filters as well as trained with a loss inspired in the latest development in self supervised learning Chen et al. (2020). Unlike standard DNNs, CHMAX learns visual representations that generalize to novel objects with a level of generalization comparable to human observers, as demonstrated in our experiments.

An HMAX in the deep learning era

In the bypass route of the HMAX, the simple (S) layers were composed of Gabor filters at 16 different sizes (also referred as scale bands, $S1^7$, $S1^9$...$S1^k$) with 4 orientations each. They are followed by a set of complex units (C) that apply spatial pooling and max pooling across adjacent scale bands (e.g. max(spatial pool(($S1^k$,$S1^{k+1}$)). We start by replacing these Gabor filters with trainable filters. We then adopt a pooling layer that uses adaptive strides to return an invariant feature map regardless of the scale of the input.

In Gaussian scale theory, scaling an image is analogous to scaling the filters (Jansson & Lindeberg, 2022; Mutch & Lowe, 2006). In order to optimize the memory footprint of the model, we used an image pyramid with rescaled versions of the image. The C1 layer computes a max spatial pooling that ensure that the feature map is coherent at different sizes and then a max pool over neighboring scales. A new set of convolutional kernels (S2b) further processes the information, followed by another pooling layer (C2b) where a max operation is used again between two adjacent scale bands to select the scale band that will be seen by the classification layer.

In order to mantain the invariance of the model when presented with multiple scales, we use contrastive learning

[1] These authors contributed equally to this work.

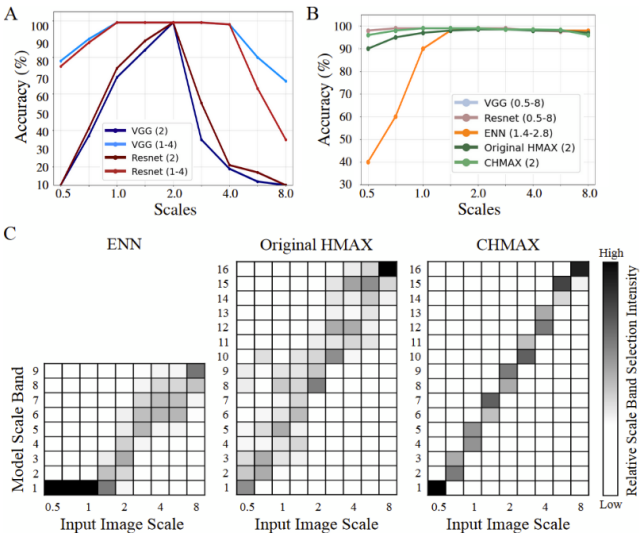

Figure 1: Top Left: Performance of ResNet50 and VGG16 models when trained on a limited subset (shown in parentheses) of the MNIST scale dataset and evaluated across all scales at inference time. Top Right: Performance of CNNs trained with data augmentation compared to HMAX models trained without augmentation. Bottom: Demonstration of the scale bands selected by the model vs. the scale of the input.

(Chopra et al., 2005) to force identical outputs from the final pooling stage for both the reference scale and the image pyramid presented to the model. We calculate the Mean Absolute Error between the C2b outputs of the two input scale to use as a penalty to regularize the classification loss.

Scale-Selection

In our comparative study on the MNIST dataset, we compare the accuracy of the proposed CHMAX architectures against VGG-16 (Simonyan & Zisserman, 2015) and ResNet50 (He et al., 2016). We train three distinct versions of both the VGG-16 and ResNet50 models: one trained solely on MNIST images of Scale 2 (center scale), one trained on scales 1-4 (limited scale range), and one trained on scale 0.5-8 (full scale range). Through these models, we aim to determine the degree to which data augmentation alone can facilitate the achievement of scale invariance in deep learning models. We also consider the Eccentricity-dependent Neural Networks (ENN) proposed by (Zhang et al., 2019) as a model of the primate visual cortex trained on scales 1.414 to 2.838.

As reflected in Figure 1B, the Original HMAX and CHMAX models are able to exhibit generalization to unseen scales during training. Furthermore, the results also demonstrate how the implementation of constrained augmentation can enhance the pre-existing scale invariance facilitated by these innate mechanisms. The ENN shows a modest degree of scale invariance but with a pronounced bias towards larger scales and diminished performance for smaller scales. Specifically, the HMAX architectures exhibit a more equitable response across the scale spectrum as seen in Figure 1C, underlining the efficient scale selection during global max pooling across scales.

The ResNet50 and VGG-16 in 1B, having been trained across the complete scale spectrum, perform commendably on the scale invariance task. However, as Figure 1A reveals, they cannot generalize to unseen scales. This suggests that models relying solely on data augmentation for achieving scale invariance are learning a separate representation for each object at a different scale. This limitation is starkly evident in the performance of the ResNet50 and VGG-16 models trained on scales 1 to 4, which perform well within this range but suffer a significant performance drop for unseen scales.

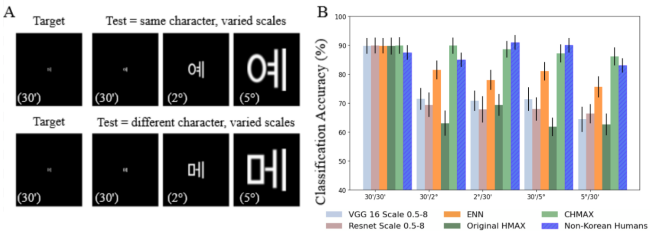

Figure 2: Top Left: Experiment conditions demonstrating the target/distractor pair. Top Right: Performance of models on one-shot learning task, including Non-Korean human performance from Han et al. (2020) for comparison.

One-Shot Learning Performance on Hangul Dataset To demonstrate the generalizability of the HMAX to out of distribution examples, we utilize a zero-shot Korean letters discrimination task presented in Han et al. (Han et al., 2020).

In the Hangul dataset used in these experiments, there are 27 target/distractor pairs of visually similar characters. Accuracy is determined by correctly deciding if the test character is the same as the target. To test scale invariance, the letters were presented in 5 conditions: target character presented at 30' of visual angle, and test character presented at 2° of visual angle; 2°(target)/30'(test); 30'/5°; 5°/30'; 30'/30'. The non-Korean human results are included in 2B.

For evaluating the models, we defined a hyperparameter of 26 pixels$= 1^\circ$ of visual angle in order to analogize the model performance to human performance. We compared the features of two Korean letters from the same pair, and classified them as the same or different based on the Pearson correlation of the features extracted from the penultimate layer of the model. Two Korean letters are considered to have the same identity if their associated features have a Pearson correlation higher than a threshold. We evaluated the accuracy on the target/target and target/distractor pairs to evaluate the selectivity of the models as well as the scale invariance.

As shown in figure 2B, the CHMAX model demonstrates superior performance on the one-shot learning task. Although the ResNet50 and VGG-16 models were trained to be scale invariant utilizing extensive data augmentation, their perfor-

mance still falls short. This observation underscores the limitation of such architectures that solely depend on data augmentation for achieving scale invariance.

Conclusions

In this work we have demonstrated the ability of the CHMAX model to generalize to unseen scales in the multi-scale MNIST dataset. We also showed that this scale invariance can be exhibited in completely novel data such as the korean task, mirroring human performance. In contrasts models trained with data augmentation are able to exhibit scale invariance over the multi-scale MNIST but no in the human task.

Acknowledgments

This work was funded by ONR grant N00014-24-1-2026. We acknowledge the Cloud TPU hardware resources that Google made available via the TensorFlow Research Cloud (TFRC) program and computing hardware supported by NIH Office of the Director grant S10OD025181 through the Center for Computation and Visualization at Brown University.

References

Biederman, I., & Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(1), 121–133.

Biscione, V., & Bowers, J. S. (2022). Learning online visual invariances for novel objects via supervised and self-supervised training. *Neural Networks*, *150*, 222-236. doi: https://doi.org/10.1016/j.neunet.2022.02.017

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... Blything, R. (2022, December). Deep problems with neural network models of human vision. *Behav. Brain Sci.*, 1–74.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. *CoRR*, *abs/2002.05709*. Retrieved from https://arxiv.org/abs/2002.05709

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 ieee computer society conference on computer vision and pattern recognition (cvpr'05)* (Vol. 1, pp. 539–546).

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* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

Fel*, T., Rodriguez*, I. F., Linsley*, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. *Adv. Neural Inf. Process. Syst.*.

Han, Y., Roig, G., Geiger, G., & Poggio, T. (2020, Jan). Scale and translation-invariance for novel objects in human vision. *Sci Rep*, *10*(1), 1411. doi: 10.1038/s41598-019-57261-6

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).

Jansson, Y., & Lindeberg, T. (2022, June). Scale-Invariant Scale-Channel Networks: Deep Networks That Generalise to Previously Unseen Scales. *Journal of Mathematical Imaging and Vision*, *64*(5), 506–536. Retrieved from https://doi.org/10.1007/s10851-022-01082-2 doi: 10.1007/s10851-022-01082-2

Linsley, D., Rodriguez, I. F., Fel, T., Arcaro, M., Sharma, S., Livingstone, M., & Serre, T. (2023). *Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex.*

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552-563. doi: https://doi.org/10.1016/S0960-9822(95)00108-4

Mutch, J., & Lowe, D. G. (2006). Multiclass object recognition with sparse, localized features. In *2006 ieee computer society conference on computer vision and pattern recognition (cvpr'06)* (Vol. 1, pp. 11–18).

Riesenhuber, M., & Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology*, *12*(2), 162–168. doi: 10.1016/s0959-4388(02)00304-5

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., ... Jitsev, J. (2022). *Laion-5b: An open large-scale dataset for training next generation image-text models.*

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, *104*(15), 6424–6429. doi: 10.1073/pnas.0700622104

Simonyan, K., & Zisserman, A. (2014, September). Very deep convolutional networks for Large-Scale image recognition.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings.* Retrieved from http://arxiv.org/abs/1409.1556

Zhang, J., Han, Y., Poggio, T., & Roig, G. (2019, 09). Eccentricity dependent neural network with recurrent attention for scale, translation and clutter invariance. *Journal of Vision*, *19*, 209. doi: 10.1167/19.10.209