Emulating Magnocellular and Parvocellular Parallel Stream Processing in Deep Neural Networks

Nelson Hidalgo

Department of Brain and Cognitive Science and Computer Science, MIT

Aniekan Umoren

Department of Brain and Cognitive Science and Computer Science, MIT

Aidan Cook

Department of Brain and Cognitive Science, MIT

1 Introduction

1.1 Background

The processing of the vastly complex visual world that we navigate is an ongoing topic of investigation in both computer vision and human vision studies. Building computer vision algorithms that are robust to occlusions and blurry images in the way that humans are remains an unsolved challenge as adversarial attacks that perform even minor alterations to images can result in failed classification. Unfortunately, this pitfall of computer vision is not adequately addressed by the current paradigm which overly relies on model accuracy on benchmarks like ImageNet-1K. Such datasets are not representative of natural examples, and so the performance of trained models sharply declines when applied in real-world settings (Hendrycks et al. [2021]). This decline can be partially attributed to the combinatorial complexity of natural images (Yuille and Liu [2021]). To create more robust computer vision systems, the literature suggests that novel architectures and less biased datasets are the way forward. Novel architectures will allow machine vision systems to leverage the latent structure of images. Datasets that are more representative of real-world examples will reduce errors

Preprint. Under review.

due to out-of-distribution samples. However it is challenging to optimize over the space of possible architectures and curating large, representative datasets is often costly.

1.2 Hypothesis and Related Work

Our project aims to address these challenges by taking inspiration from a known property of the human visual system and applying it to a convolutional neural network (CNN) algorithm. To do so we plan to use the parallel processing of visual stimuli through the magnocellular and parvocellular streams in the visual cortex as our guiding inspiration (Bear et al. [2007]). The magnocellular stream is color-blind and carries coarse-grained information about slight brightness changes, while the parvocellular stream carries more fine-grained details that are informative about local color differences (Bear et al. [2007]). These parallel processes interact through complex feed-forward, lateral, and top-down connectivity in the brain (Medathati et al. [2016]), but we will simplify our approach and consider the feed-forward aspects of this neural circuit. We will simulate the magnocellular and parvocellular streams by using two parallel CNNs, the outputs of which are merged through an averaging layer and passed on to a fully-connected network that produces a classification of images. Architectures similar to our implementation have been used to identify gestures by Khurana et al. [2019]. We will assess the performance of people on a subset of the testing set that our algorithm is tested on as a benchmark for optimal results. We hypothesize that our Dual-stream CNN will be more robust to occlusions and blurriness that are partly out-of-distribution compared to traditional CNNs. We also expect our algorithm to achieve closer to human performance.

2 Methods

2.1 Computer Vision Experimental Design

Our Dual-stream CNN, which is implemented at https://github.com/nelsonalbertohj/Magno-Parvo-CNN, consists of two parallel CNN networks, which are inspired on AlexNet's architecture. The Parvocellular CNN (Parvo-CNN), was designed with smaller filters 11 x 11, 5 x 5, and 3 x 3 filters, while the Magnocellular CNN (Magno-CNN) was designed with with larger filters 13 x 13, 7 x 7, and 3 x 3. As illustrated in Figure 1, the Parvo-CNN receives as input the normalized RGB image with no blur applied, while the Magno-CNN receives as input the normalized grayscale image with

a Gaussian blur applied to it. The Magnocellular portion of the network has 3 grayscale channels that are identical, which was needed in order to pass it through the CNN architecture that expects 3 channels. The outputs from each CNN are averaged in an intermediate layer that is then passed onto a fully connected network that outputs final predictions.

The Dual-stream CNN was trained on 10 image classes from the ImageNet dataset. While using a smaller portion of the dataset has limited our ability to show how the model generalizes to a larger, more complex dataset, it allows us to draw more precise conclusions about the relationships between our model, the control model, and human performance. In addition, ImageNet was chosen because it provides higher resolution images compared to other datasets, which allows for building a model that more closely approximates the high-resolution parvocellular stream versus the lower resolution magnocellular stream.

Our control model consists of a Parallel CNN network that does not apply the Magnocellular versus Parvocellular differences. The control model was trained on the same group of RGB images. Both models were trained with approximately 1000 images from each class. The training procedure followed an early stopping criteria based on validation accuracy with patience of 20, training for 60 epochs, and Adam optimizer with step size of 0.001. Both our Dual-stream CNN and the Parallel CNN were trained from scratch and were tested on unseen full-resolution images, blurred images, and uniformly occluded images.

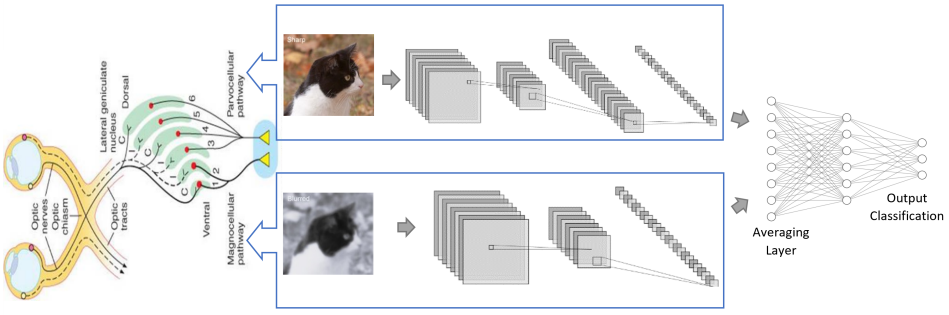

Figure 1: Diagram of Dual-stream CNN which shows the Parvo-CNN at the top, the Magno-CNN at the bottom, and the fully-connected layer that is trained for final classification of the image.

2.2 Human Experiment Design

In order to have another basis of comparison for our Dual-stream CNN, we tested human performance on 5 image classes ('Asparagus', 'Banana', 'Bee', 'Bracelet', 'Elephant') from this dataset as a way to establish a human-performance benchmark for both models. To do so, we gathered 15 images from the ImageNet database for each of our five image classes. For each class, five control images were left unedited, five were blurred, and five had a grid-based occlusion superimposed on top of them. Volunteers (N=35) were presented with a series of 20 unedited, 20 uniformly occluded, and 20 blurred images in random order and asked to classify them as either 'Asparagus', 'Banana', 'Bee', 'Bracelet', 'Elephant', or 'Other'.

3 Results

3.1 Computer Vision Experimental Results

In order to assess whether our Dual stream model's training had allowed it to become significantly more robust to out-of-distribution images, we compared its image classification abilities to those of a vanilla parallel CNN. As described in the methods, we tested each model on both unedited and blurred images from the ImageNet database in order to test our hypothesis. To this end, we analyzed the confusion matrices in Figures 3, 4 and 5 and performed a a binomial test to evaluate the statistical significance of the Confusion Difference Matrix. As shown in Figure 3, none of the elements in the confusion matrix where significantly different (alpha = (5%)/100 for Bonferroni correction) between the the Dual-stream CNN and the control CNN for the classification of full-resolution images. Similarly in Figure 4, None of the elements in the confusion matrix where significantly different

Figure 2: Examples of transformations Applied to Images for Human Trials.

(alpha = (5%)/100 as per the Bonferroni correction) between the two models for the classification of blurred images. The confusion matrices in Figure 5 indicates that both networks confused the grid-like occlusion pattern for bees and bracelets performing close to random chance, which is 0.1 in this case.

To evaluate model performance, an extension of the Matthews Correlation Coefficient (Baldi et al. [2000]) to multi-class classifiers was used. This metric suggests that the Dual-stream model is a

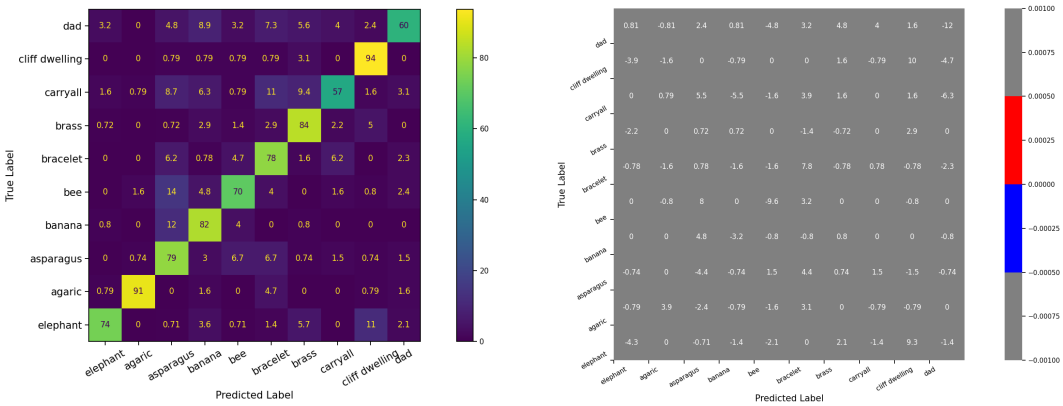

Figure 3: (Left) Vanilla Dataset Confusion Matrix for Dual-stream Model. Each element represents the empirical probability of a prediction given the true value. (Right) Confusion Matrix Difference between Dual Stream and control CNN. The control CNN confusion matrix was subtracted from the Dual Stream. Colors represent the p-value (gray signifies failure to reject null)

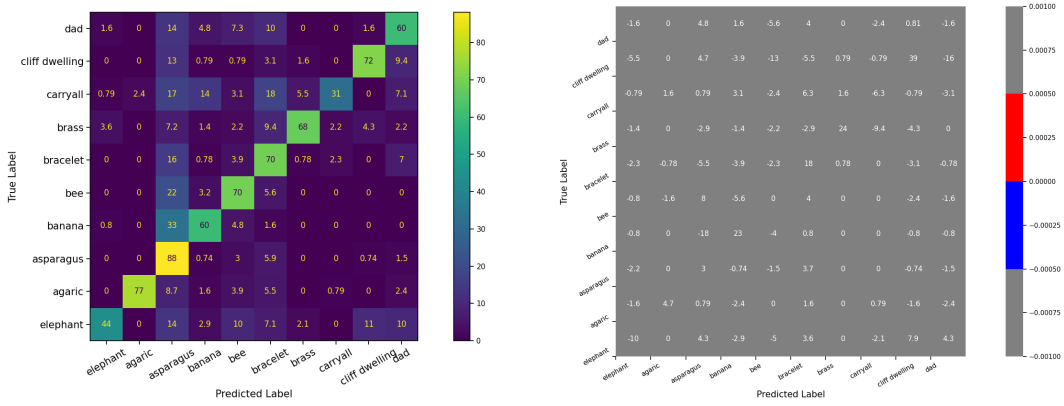

Figure 4: (Left) Blur Dataset Confusion Matrix for Dualstream Model. Each element represents the empirical probability of a prediction given the true value. (Right) Confusion matrix difference between the Dual Stream and control CNN. The control CNN confusion matrix was subtracted from the Dual Stream. Colors represent the p-value (gray signifies failure to reject null).

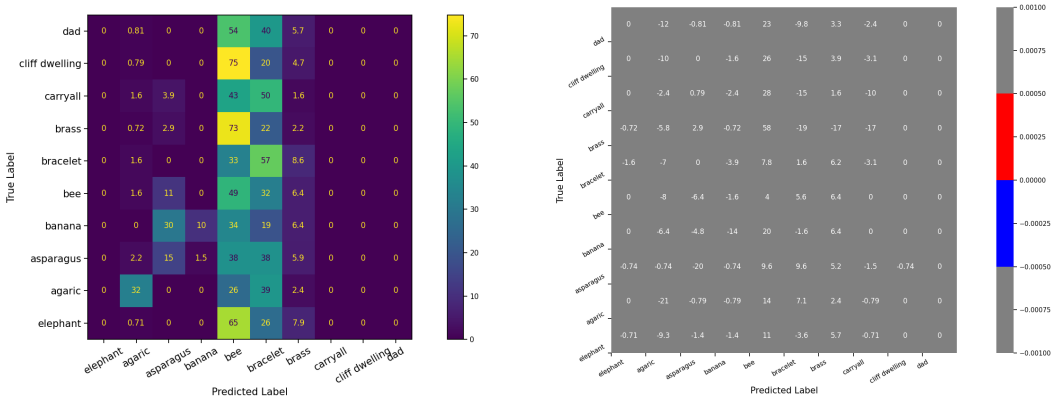

Figure 5: (Left) Occlusion Dataset Confusion Matrix for Dualstream Model. Each element represents the empirical probability of a prediction given the true value. (Right) Confusion matrix difference between the Dual Stream and control CNN. The control CNN confusion matrix was subtracted from the Dual Stream. Colors represent the p-value (gray signifies failure to reject null).

better classifier than the parallel CNN model (MCCDual = 0.608, MCCParallel = 0.501) on the blurred dataset as shown in Figure 6. The models perform similarly on classifying full resolution images (MCCDual = 0.746, MCCParallel = 0.749) and both perform close to chance when classifying occluded images (MCCDual = 0.16, MCCParallel = 0.24).

3.2 Human Vision Experiment Results

After testing our Dual Stream model against a parallel CNN model, we decided to further contextualize the performance of the Dual Stream model by testing humans on their ability to classify some of the same images. In order to assess how well our volunteers (N=35) were able to classify images that had gone through various transformations, we split our classification results into non-edited, occluded, and blurred image groups. Participants were recruited from MIT dorms and asked to complete our image classification test. We created confusion matrices and recorded the average response time it took a participant to classify an image for each group (Table 1). In addition to

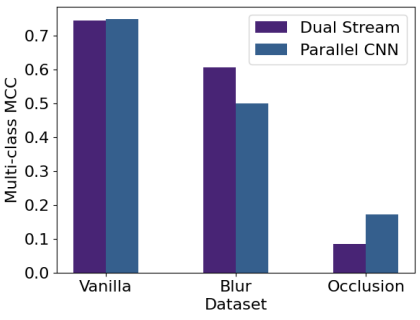

Figure 6: Bar graph of multi-class Matthews Correlation Coefficient on full resolution, blurred, an occluded test sets.

acting as a general indicator of accuracy, the confusion matrices also allowed us to assess whether any particular category of image was easier or harder than the other categories to classify in a given group. Overall, humans performed at 98.2 % accuracy at recognizing full-resolution images, 82.5 % accuracy at recognizing occluded images, and 85.7 % accuracy at recognizing blurred images. Furthermore, we were also interested to see how the average response time would vary with the type of transformations performed on our test images. The results in Table 1 indicate that response time, which is analogous to task difficulty, increased from full-resolution (2.07 s) to blurred (3.62 s) to occluded (4.03 s) images.

A)

	Asparagus	Banana	Bee	Bracelet	Elephant	Other
Asparagus	100%	0%	0%	0%	0%	0%
Banana	0%	98.3%	0%	0%	0%	1.7%
Bee	0%	0%	94.4%	0%	0%	5.6%
Bracelet	0%	0%	0%	98.3%	0%	1.7%
Elephant	0%	0%	0%	0%	100%	0%

Average Normal Image response time: 2.07 seconds

B)

	Asparagus	Banana	Bee	Bracelet	Elephant	Other
Asparagus	50.8%	16.6%	1.1%	2.9%	0%	28.6%
Banana	0%	88.0%	0%	0%	1.1%	10.9%
Bee	0%	1.1%	80%	1.1%	0%	17.8%
Bracelet	0%	0%	0%	96.0%	0%	4.0%
Elephant	0%	0%	0%	0%	97.7%	2.3%

Average Occluded Image response time: 4.03 seconds

C)

	Asparagus	Banana	Bee	Bracelet	Elephant	Other
Asparagus	66.3%	4.6%	8.6%	0%	0%	20.5%
Banana	0%	78.9%	1.1%	0%	0%	20%
Bee	0%	0%	89.1%	0.6%	0%	10.3%
Bracelet	0%	0%	0%	97.1%	0%	2.9%
Elephant	0%	0%	0%	0%	97.1%	2.9%

Average Blurred Image response time: 3.62 seconds

Table 1: Confusion Matrices and Average Response Time of Human Visual Classification Trials with the actual image identity on the leftmost column of the Confusion matrix, and the human identification of the image on the top. A) Performance from viewing unedited images B) Performance from viewing occluded images C) Performance from viewing blurred images

4 Discussion

While our Dual Stream model appears representationally similar to the vanilla model, this may be because we didn't constrain the model to meaningfully utilize the two pathways beyond priming those pathways with different size filters analogous to the information processed by the parvocellular and magnocellular stream (there was no incentive to utilize impoverished input). We would recommend that future research on this topic address this potential issue by implementing additional constraints that would make each stream of the Dual-stream network more closely resemble the parvocellular and magnocellular streams in the human brain. One such constrain could be training the network such that the magno-cellular stream is trained first, then frozen, and later the the parvo-cellular stream is trained.

Human performance from our experiment indicate that humans vastly outperform both algorithms tested in this experiment. However, there were some similarities in terms of difficulty of classifying blurry images, where for humans it took longer to classify those images and performance dropped by 15.7 % on average. Our Dual-stream CNN's accuracy dropped by 13.8 % while the Parallel CNN dropped by 24.8 % from classification of full-resolution to blurry images. These results indicate that the Dual-stream CNN more closely parallels human performance, but this hypothesis needs to be further tested with different blurs and out-of-distribution testing examples. In addition, our results indicate that the asparagus class was difficult to classify for both humans and the computer-vision models as this class had the lowest performance across.

We believe that this research shows promise as our Dual-stream model outperformed the parallel CNN model on blurred images with a 0.10 MCC increase and performed closer to humans in the aforementioned metrics. Furthermore, we believe future research of this type to encourage robustness to occluded and blurred images to be important in allowing computer vision to become more robust to classifying the vast diversity of real-world images.

References

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

Alan L Yuille and Chenxi Liu. Deep nets: What have they ever done for vision? *International Journal of Computer Vision*, 129(3):781–802, 2021.

MF Bear, BW Connors, and MA Paradiso. *The central visual system*. 2007.

NV Kartheek Medathati, Heiko Neumann, Guillaume S Masson, and Pierre Kornprobst. Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. *computer vision and image understanding*, 150:1–30, 2016.

Rajat Khurana, Alok Kumar Singh Kushwaha, et al. Delving deeper with dual-stream cnn for activity recognition. In *Recent Trends in Communication, Computing, and Electronics*, pages 333–342. Springer, 2019.

Pierre Baldi, Søren Brunak, Yves Chauvin, Claus AF Andersen, and Henrik Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.