THE SELECTIVITY AND COMPETITION OF THE MIND'S EYE IN VISUAL PERCEPTION

Edward Kim[1] *Maryam Daniali*[2] *Jocelyn Rego*[1] *Garrett T. Kenyon*[3]

[1] Department of Computer Science, Drexel University, PA
[2] Department of Biomedical and Health Informatics (DBHi), Children's Hospital of Philadelphia, PA
[3] Los Alamos National Laboratory, NM

ABSTRACT

Research has shown that neurons within the brain are selective to certain stimuli. For example, the fusiform face area (FFA) region is known by neuroscientists to selectively activate when people see faces over non-face objects. While the exact mechanisms by which the primary visual system directs information to the correct higher levels of the brain are currently unknown, there are high-level neural mechanisms of perception that we can incorporate in a novel computational model - ones that utilizes lateral and top down feedback in the form of hierarchical competition. We demonstrate that these neural mechanisms provide the foundation of a novel classification framework that rivals traditional supervised learning in computer vision. Additionally, we show that the innate priors built into our architecture support out of distribution generalization on the application of face detection.

***Index Terms*—** Multiscale Sparse Coding, Dictionary Learning, Competition Pathways, Robust Classification

1. INTRODUCTION

In 2017, a 26-year old patient at Asahikawa Medical University was being treated for intractable epilepsy. This patient had subdural electrodes placed on a specific part of the brain known as the fusiform face area (FFA) [1]. When researchers artificially stimulated neurons in the FFA, the patient hallucinated faces or face parts in non-face, everyday objects. The FFA region is known by neuroscientists to selectively activate when people see faces, specifically upright faces, compared to the activations elicited by non-face objects [2]. Additional evidence exists revealing that the visual processing of faces and objects occur in different pathways. Exploring deeper into the cerebral cortex, we can see that the FFA is one of many specialized, high level areas within the brain. The FFA exists in the Inferior Temporal (IT) Cortex, the part of the brain critical for visual object memory and recognition, colloquially referred to as the *Mind's Eye*. Other specialized areas within the temporal cortex include selectivity for visual scenes or buildings (parahippocampal place area, PPA), for body parts (extrastriate body area, EBA), and for reading words (visual word form area, VWFA).

How do the low-level, primary visual areas of the brain know where to send visual input information? This would imply that the low level areas have already done some sort of recognition of the input stimulus to route the information correctly to the higher levels in IT. Some have hypothesized that there exists some low level gating mechanism that performs a rough detection and then forwards the information to specialized expert models, e.g. gating with a mixture of experts model [3]. Others used a category template model that could roughly match an input stimulus [4].

In our work, we develop a novel classification framework that can outperform traditional supervised learning by mimicking high-level neural mechanisms of perception. Here, we demonstrate that our novel algorithm and framework that incorporates lateral and top down feedback in the form of hierarchical competition can perform category level image classification with improved performance over supervised neural networks. We demonstrate our results on the problem of face detection bias that has been uncovered in deep learning.

2. BACKGROUND

Selectivity in the visual cortex can be achieved through competition. Research has shown that neurons within the brain are selective to certain stimuli. Early work by Hubel and Wiesel demonstrated that cat V1 neurons were sensitive to the placement, orientation, and direction of movement of oriented edges [5]. We see similar patterns of stimulus selectivity at higher levels of the brain i.e. in the inferior temporal gyrus or IT, where regions are selective to specific objects, faces, body parts, places, and words. Also neurons selective for different objects mutually inhibit each other in the presence of their preferred stimulus [6], evidenced by a measured reduced blood oxygen level during competitive interactions among stimuli. As a result, we observe that given a specific stimulus, only a highly selective, small subset of neurons will activate [7].

Faces are processed in a unique pathway. The area responsible for face processing was first discovered by Sergent et al. [8]. This area was later named the fusiform face area and shown to activate more when people see faces rather than general objects [2]. It is important to note that while faces

activate specialized areas with the brain, these areas are not completely silent when non-face objects are viewed. Instead, the cortical response for the preferred category is about *twice that for the non-preferred category* as consistently observed in most normal individuals [9].

Faces are processed in a holistic, coarse-to-fine manner. In addition to the unique pathway for face processing, the holistic manner by which the face is recognized has also been discovered. Maurer et al. [10] notes that face stimuli are processed as a gestalt, and holistic processing occurs with the internal structure of the face and with the external contour. Even simple circles containing three curved lines, if shaped like a smiling face, triggered a holistic face response. The holistic response was thought to contribute to an early stage process so that one could distinguish faces from other competing objects [11].

The holistic approach of face recognition is also consistent with a related theory where research shows that the visual system integrates visual input in a coarse-to-fine manner (CtF) [12], i.e. low frequency information is processed quickly first, which then projects to high level visual areas. Critically, the high level areas *generate a feedback signal* that guides the processing of the high frequency (details) of the image. These dynamics suggest a feedback or competitive process resulting in a winner-take-all situation [3].

In computer vision, there are only a few works that have addressed lateral and top-down feedback explicitly in a model. In one case, top-down feedback was implemented using a parallel neural network to provide feedback to a standard CNN [13]. Elsayed et al. [14] showed that adversarial examples can fool time-limited humans, but not no-limit humans, stating no-limit humans are fundamentally more robust to adversarial examples and achieve this robustness via top-down or lateral connections. Given both inhibitory and excitatory top-down feedback in a generative model, immunity to adversarial examples was demonstrated [15].

3. METHODOLOGY

3.1. Sparse Coding for Selectivity and Competition

The main algorithm that underlies our framework is sparse coding. Sparse coding was first introduced by Olshausen and Field [16], in order to explain how the primary visual cortex efficiently encodes natural images. Sparse coding seeks a minimal set of generators that most accurately reconstruct each input image. Each generator adds its associated feature vector to the reconstructed image with an amplitude equal to its activation.

Mathematically, sparse coding can be defined as follows. Assume we have some input variable $x^{(n)}$ from which we are attempting to find a latent representation $a^{(n)}$ (we refer to as "activations") such that $a^{(n)}$ is sparse, e.g. contains many zeros, and can reconstruct the input, $x^{(n)}$, with high fidelity.

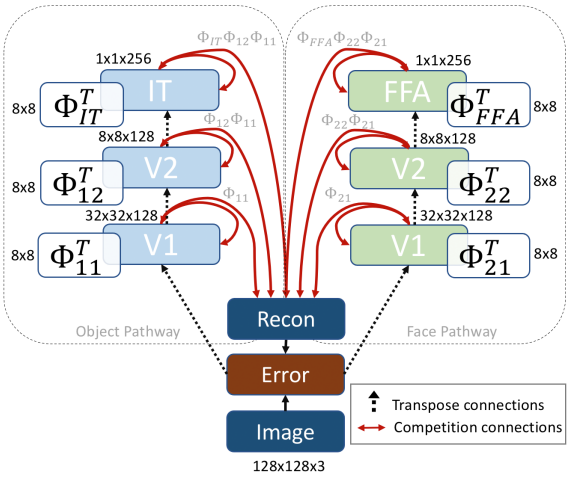


Fig. 1. Our multipath deconvolutional competitive algorithm (MDCA) model consists of two distinct pathways, one for faces and one for general objects. Not only do the neurons in each of the layers compete to represent an input stimuli, every hierarchical layer in both pathways compete in the reconstruction of the input image.

The sparse coding algorithm is defined as,

$$\min_{\Phi} \sum_{n=1}^{\mathcal{N}} \min_{a^{(n)}} \frac{1}{2} \|x^{(n)} - \Phi a^{(n)}\|_2^2 + \lambda \|a^{(n)}\|_1 \qquad (1)$$

Where Φ is the overcomplete dictionary, and $\Phi a^{(n)} = \hat{x}^{(n)}$, the reconstruction of $x^{(n)}$. The λ term controls the sparsity penalty, balancing the reconstruction versus sparsity term. $\mathcal{N}$ is the total training set, where n is one element of training. Φ represents a dictionary composed of small kernels that share features across the input signal.

There are a number of solvers for Equation 1, but we selected the solver that is biologically informed. This solver is the Locally Competitive Algorithm (LCA) [17] that evolves the dynamical variables (neuron's membrane potential) when presented with some input stimulus. Of particular importance is that the activations of neurons in this model compete and laterally inhibit units within the layer to prevent them from firing. The input potential e.g. excitatory drive to the neuron state is proportional to how well the image matches the neuron's dictionary element, while the inhibitory strength is proportional to the similarity of the current neuron and competing neuron's convolutional patches, forcing the neurons to be decorrelated. The LCA model is an energy based model similar to a Hopfield network where the neural dynamics can be represented by a nonlinear ordinary differential equation.

3.2. Construction of Face and Object Pathways

In the context of face recognition, there is strong suggestive evidence that some component of face processing is innate to the human visual system. Thus, for our model, we choose

to pre-train each pathway to reflect the propensity of neural pathways we see in the visual cortex. One pathway is tuned to reconstruct faces, and the other pathway tuned to reconstruct general objects. The pathway consists of a 3-layer hierarchical, multiscale, convolutional sparse coding network as shown delineated by the dotted lines in Figure 1. The training procedure involved showing 10,000 images from the ImageNet [18] dataset and 10,000 images from the Celeb-A [19] dataset to the respective pathways. At the pre-training stage, the pathways are independent from each other and do not compete.

The images shown to the network have been resized to 128x128x3. The dictionary sizes, activation maps, and architecture are identical in the two pathways. There are 128, 8x8xC, (C being the number of input channels), dictionary elements in each of the respective V1 layers, Φ_{11} and Φ_{21} in Figure 1. We stride by 4 throughout the hierarchy, thus increasing the receptive field of neurons by a factor of 4x at each layer. We keep the same size dictionary patches for V2, but at the top layer, FFA and IT, we expand the number of neurons to 256. While this number is empirically chosen, it does have a biological connection as it has been shown that faces can be linearly reconstructed using responses of approximately 200 face cells [20]. Self-supervised learning of features can be obtained by taking the gradient of the reconstruction error with respect to Φ, resulting in a biologically plausible local Hebbian learning rule. The dictionary can be updated via Stochastic Gradient Descent (SGD). Examples of the dictionary learned can be seen in Figure 2(a)(c).

(a) Face V1, Φ_{21}

(b) Activity avg of the top (FFA) layer face pathway.

(c) Obj V1, Φ_{11}

(d) Activity avg of the top (IT) layer object pathway.

Fig. 2. Visualization of the learned dictionary elements at V1 and activity triggered averages of our top level (b) face and (d) object pathway. The face pathway was shown images from Celeb-A and the object pathway was shown ImageNet.

3.3. Multipath Deconvolutional Competitive Algorithm (MDCA)

We combine paths together in a multiscale, hierarchical competitive structure that we refer to as the Multipath Deconvoloutional Competitive Algorithm (MDCA). The learning of elements at multiple scales was explored by [21] and used for image and video restoration. Our network consists of a deep deconvolutional sparse coding network similar to the work of Zeiler [22] and Paiton [23].

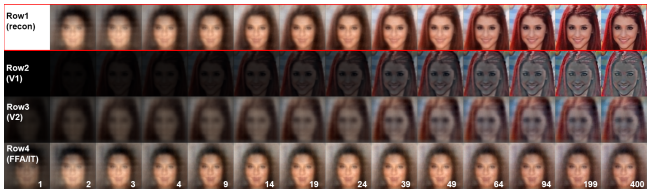
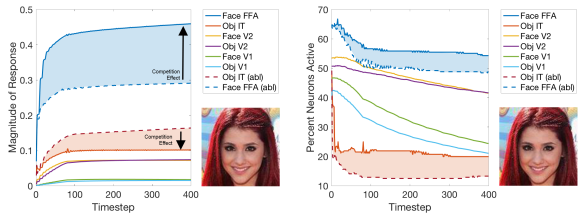


Fig. 3. Illustration of the coarse-to-fine reconstruction over 400 timesteps. **(Row 1)** shows the total reconstruction at the given timestep. **(Row 2)** is the summed contribution from V1 in both pathways. **(Row 3)** is the summed contribution V2 in both pathways, and **(Row 4)** is the summed contribution of FFA and IT. The numbers in the bottom right corner indicate the timestep of the each column.

At a conceptual level, our model is implementing the following process illustrated in Figure 1. An input stimulus is presented, and very quickly sent up the hierarchy of both pathways. One could think of this initial step as a feed forward pass in a typical deep learning model. Each neuron in every layer is "charged up" by the input stimulus, where neurons at higher levels have larger receptive fields, and neurons at the top level see the entire input stimulus. As each neuron passes threshold, they add their respective feature to the reconstruction via deconvolution. Thus, the reconstruction layer is not only influenced by fine, high-frequency features from the lower layer, but also guided by the large, low spatial-frequency activated features of the higher layer. As the stimulus is reconstructed over time, the network computes the error between the input and reconstruction at each timestep. This error is then forwarded up the hierarchy, driving the neurons to compete for the remaining residual representation. In our experiments, we evolve our recurrent network over t=400 timesteps. Mathematically, we define the reconstruction, $\hat{x}$, in our Multipath Deconvolutional Competitive Algorithm as the following,

$$\hat{x}^{(n)} = \sum_{m=1}^{M} (\sum_{k=1}^{K} (\prod_{l=1}^{k} \Phi_{m,l}) a_{m,k}^{(n)}) \tag{2}$$

where $m \in M$ is the number of paths, and $l, k \in K$ is the number of multiscale layers in the neural network. In our case shown in Figure 1, we have $M = 2$ e.g. object pathway and face pathway, and three multiscale layers, $K = 3$.



(a) Mag. Response (b) Neuron Activity

Fig. 4. We plot the (a) magnitude of response and (b) percent of active neurons in all of the MDCA layers over 400 timesteps. The magnitude of response at the V1 and V2 levels are similar; however, there is over a 4x response in the Face FFA region compared to the Object IT region in (a).

4. EXPERIMENTS AND RESULTS

4.1. Coarse-to-fine Information Flow

In our first experiment, we investigate the activity of the MDCA network as it processes input stimuli. Given an input image, the objective of the network is simply to minimize reconstruction error. We can visualize the process of reconstruction in Figure 3, and quantify the response at all levels in all pathways of our network (Figure 4). In the reconstruction, we can see that our **MDCA network is reconstructing the image in a holistic, coarse-to-fine manner**, consistent with the CtF neuroscience literature [12]. We can clearly visualize that low frequency information is processed quickly first and high level areas *generate a competitive feedback signal* that guides the processing of the high frequency (details) of the image [24].

4.2. Ablation of Pathway Competition

We perform an ablation study to confirm that the selectivity of the neurons in our network is mainly due to the competition of multiple pathways. In Figure 4(a) we see the magnitude of response of the FFA increases nearly 2x, and suppresses the IT response on its preferred stimulus. We see a similar pattern with Object IT when presented its preferred stimulus. From these results, we gain insight on how the model (as well as the visual cortex) is able to respond to a specific stimulus.

4.3. Robust Category Level Classification

We challenged our MDCA model to generalize in the task of face detection out of distribution. We compared our framework to a standard deep learning CNN model and fine-tuned, off the shelf modles (ResNet-50 [25], VGG16[26], Vision Transformer ViT [27]) trained with data from the FairFace dataset [28] and ImageNet. The FairFace dataset contains images of people from seven ethnicity groups, across a wide range of variations.

Specifically, for the custom CNN, we built a 3-layer CNN binary classifier (face/not face) that matches the architecture, size, and parameters of a single pathway of our MDCA model. We trained with a biased and unbalanced dataset consisting of 800 White-Male faces and 10,000 ImageNet images. Our in-distribution test set contained 200 White-Male faces and 1,000 ImageNet images. We trained the CNN for 35 epochs, fine-tuned ResNet50, Vision Transformer, and VGG16 for 35 epochs, and pre-trained the MDCA pathways with the same images and for an approximately equivalent number of image impressions.

For the in-distribution test set, the custom CNN, ResNet50, VGG16, ViT, and MDCA performed very well as expected, 97.17%, 99.67%, 99.83%, 98.57%, and 98.25% respectively. However, as noted in previous literature, the deep learning models struggled to generalize their understandings in face detection of one ethnicity and gender to other categories, failing in over 36% of the cases on Black males in the custom CNN, and more than 2.8%, 5.5%, 4.6% drop in ResNet50, VGG16, ViT respectively. MDCA, on the other hand, was capable of detecting faces of every ethnicity and gender with nearly perfect accuracy in all categories, see Table 1.

Table 1. Classification accuracy on different ethnicity categories and genders, a comparison among 4 different models, custom CNN, ResNet50, VGG16, and our model, MDCA.

Ethnicity/Gender	#img	CNN	ResNet50	VGG16	ViT	MDCA
Black/F	757	73.84	97.09	95.64	92.73	**99.33**
Black/M	799	63.2	97.12	94.49	95.37	**99.87**
East-Asian/F	773	83.31	95.60	95.21	92.26	**99.61**
East-Asian/M	777	77.09	96.14	96.78	97.30	**99.87**
Indian/F	763	88.33	96.59	96.20	94.89	**100**
Indian/M	753	88.34	95.88	94.95	98.03	**100**
Latino-Hispanic/F	830	86.86	96.63	98.07	97.54	**99.63**
Latino-Hispanic/M	793	83.98	95.88	96.47	98.87	**100**
Middle-Eastern/F	396	83.83	94.19	94.95	94.65	**100**
Middle-Eastern/M	813	82.41	96.06	95.82	98.30	**99.38**
Southeast-Asian/F	680	85.00	97.21	95.74	93.74	**99.85**
Southeast-Asian/M	735	81.49	98.23	97.01	97.82	**99.59**

5. CONCLUSION AND FUTURE WORK

In conclusion, our primary goal is to abstract themes from neuroscience in order to improve artificial intelligence. To this end, we created a computational model that incorporates important thematics observed in the brain, selectivity through competition, dedicated pathways, holistic/coarse-to-fine processing, and top-down feedback. We demonstrate that these neural mechanisms provide the foundation of a novel, robust classification framework that rivals traditional supervised learning in computer vision. In the example of machine learning bias, we demonstrate that an attractor model rooted in competition can out perform a supervised deep learning model. The method itself is learning features in a self-supervised manner segregated via supervision.

6. REFERENCES

[1] Gerwin et al. Schalk, "Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain," *Proceedings of the National Academy of Sciences*, vol. 114, no. 46, pp. 12285–12290, 2017.

[2] Nancy Kanwisher, "Functional specificity in the human brain: a window into the functional architecture of the mind," *Proceedings of the National Academy of Sciences*, vol. 107, no. 25, pp. 11163–11170, 2010.

[3] Doris Y Tsao and Margaret S Livingstone, "Mechanisms of face perception," *Annu. Rev. Neurosci.*, vol. 31, pp. 411–437, 2008.

[4] Kendrick N Kay and Jason D Yeatman, "Bottom-up and top-down computations in word-and face-selective cortex," *Elife*, vol. 6, pp. e22341, 2017.

[5] David H Hubel and Torsten N Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[6] Samuel V Norman-Haignere, Gregory McCarthy, Marvin M Chun, and Nicholas B Turk-Browne, "Category-selective background connectivity in ventral visual cortex," *Cerebral Cortex*, vol. 22, no. 2, pp. 391–402, 2012.

[7] Shy Shoham, Daniel H O'Connor, and Ronen Segev, "How silent is the brain: is there a "dark matter" problem in neuroscience?," *Journal of Comparative Physiology A*, vol. 192, no. 8, pp. 777–784, 2006.

[8] Justine Sergent, Shinsuke Ohta, and Brennan Macdonald, "Functional neuroanatomy of face and object processing: a positron emission tomography study," *Brain*, vol. 115, no. 1, pp. 15–36, 1992.

[9] David C Plaut and Marlene Behrmann, "Complementary neural representations for faces and words: A computational exploration," *Cognitive neuropsychology*, vol. 28, no. 3-4, pp. 251–275, 2011.

[10] Daphne Maurer, Richard Le Grand, and Catherine J Mondloch, "The many faces of configural processing," *Cognitive sciences*, vol. 6, no. 6, pp. 255–260, 2002.

[11] Jessica Taubert, D. Apthorp, David Aagten-Murphy, and David Alais, "The role of holistic processing in face perception: Evidence from the face inversion effect," *Vision research*, vol. 51, no. 11, pp. 1273–1278, 2011.

[12] Moshe Bar, Karim S Kassam, Avniel Singh Ghuman, Jasmine Boshyan, Annette M Schmid, Anders M Dale, Matti S Hämäläinen, Ksenija Marinkovic, Daniel L Schacter, Bruce R Rosen, et al., "Top-down facilitation of visual recognition," *Proceedings of the national academy of sciences*, vol. 103, no. 2, pp. 449–454, 2006.

[13] Deepak Babu Sam and R Venkatesh Babu, "Top-down feedback for crowd counting convolutional neural network," *arXiv preprint arXiv:1807.08881*, 2018.

[14] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein, "Adversarial examples that fool both computer vision and time-limited humans," in *Advances in Neural Information Processing Systems*, 2018, pp. 3910–3920.

[15] Edward Kim, Jocelyn Rego, Yijing Watkins, and Garrett T Kenyon, "Modeling biological immunity to adversarial examples," in *CVPR*, 2020, pp. 4666–4675.

[16] B. A Olshausen and DJ Field, "Sparse coding with an overcomplete basis set: A strategy employed by v1?," *Vision research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[17] Christopher Rozell, Don Johnson, Richard Baraniuk, and Bruno Olshausen, "Locally competitive algorithms for sparse approximation," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*. IEEE, 2007, vol. 4, pp. IV–169.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[20] Le Chang and Doris Y Tsao, "The code for facial identity in the primate brain," *Cell*, vol. 169, no. 6, pp. 1013–1028, 2017.

[21] Julien Mairal, Guillermo Sapiro, and Michael Elad, "Multiscale sparse image representationwith learned dictionaries," in *2007 IEEE International Conference on Image Processing*. IEEE, 2007, vol. 3, pp. III–105.

[22] Matthew D Zeiler, Graham W Taylor, and Rob Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *ICCV*, 2011, pp. 2018–2025.

[23] Dylan Paiton, Sheng Lundquist, William Shainin, Xinhua Zhang, Peter Schultz, and Garrett Kenyon, "A deconvolutional competitive algorithm for building sparse hierarchical representations," in *Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, 2016, pp. 535–542.

[24] Kirsten Petras, Sanne Ten Oever, Christianne Jacobs, and Valerie Goffaux, "Coarse-to-fine information integration in human vision," *NeuroImage*, vol. 186, pp. 103–112, 2019.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[28] Kimmo Kärkkäinen and Jungseock Joo, "Fairface: Face attribute dataset for balanced race, gender, and age," *arXiv preprint arXiv:1908.04913*, 2019.