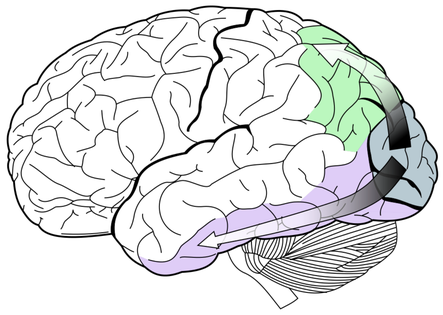
Incorporating action information into computational models of the human visual system

Justin Zhou[1], Dr. Marieke Mur[1,2]
1 Department of Psychology, Western University, London Ontario
2 Brain and Mind Institute, Western University, London, Ontario

Background

Dorsal stream: visual stream responsible for discerning the location of objects and visually-guided behaviour.[1]



"Ventral-dorsal streams.svg" by Selket is licensed under CC BY-SA 3.0.

Figure 1. The dorsal stream runs from the occipital lobe to the parietal lobe (green path) whereas the ventral stream runs along the temporal lobe (purple path).

Neural network: a computing system inspired by neuroscience. Neurons are simplified into *units*, each of which take inputs and produce outputs. Units can be linked together to model increasingly complex input-output relationships.[2]

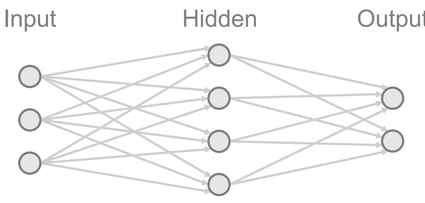


Figure 2. Illustration of a simple neural network with a single hidden layer.

Convolutional neural network: a type of neural network often used to analyze images. It uses a process called *convolution* to decrease sensitivity to changes in object position (shift invariance). Convolution involves repeatedly sampling patches of an image using filters.[2]

Deep neural network: a neural network with greater than one hidden layer.[2]

Introduction

The ventral visual stream can be modeled using Deep Convolutional Neural Networks (DCNNs).

- DCNNs approach human-level accuracy on image categorization tasks.[3,4,5]
- DCNNs predict neural representations of images.[6,7,8]

However, computational models of the dorsal visual stream remain relatively unexplored.[9] This is problematic, as:

- The ventral and dorsal streams are not entirely independent[10,11]
- The streams likely influence each other during development

Additionally, DCNN models of vision also suffer from flaws such as over-reliance on texture information.[12]

Research questions:

- Does incorporating action information improve computational models of the ventral visual system?
- How do the ventral and dorsal streams influence each other during development?

Methods

The study will involve creating three models:

1. **Two-task network:** a neural network trained with two cost functions, one approximating the function of the ventral visual stream and the other approximating the function of the dorsal stream.

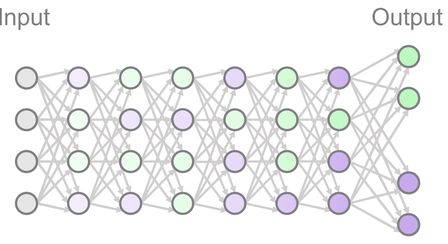


Figure 3. Simplified illustration of the two-task network. Function specificity towards the dorsal stream task is represented by green, and for the ventral stream task, by purple. Greater colour saturation represents greater function specificity.

2. **Single-task network:** a neural network trained with only a single cost function, approximating the function of the ventral stream.

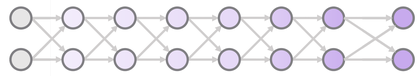


Figure 4. Simplified illustration of the single-task network.

3. **Lesioned network:** a copy of the trained two-task network, with the units that contributed the most to the dorsal stream task deactivated.

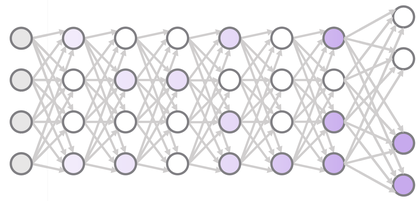


Figure 5. Simplified illustration of the lesioned network. Deactivated units are shown in white.

Each neural network will be evaluated using *performance metrics* and *representational metrics*. Performance metrics include:

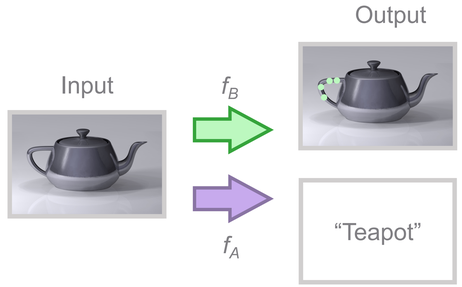
- Accuracy when evaluated with ImageNet
- Accuracy when evaluated with Stylized-ImageNet
- Transfer learning
- Robustness against distortions

Representational metrics include:

- Representation contribution analysis
- Representational similarity analysis

Training tasks:

To approximate ventral stream function, networks will be trained to perform object recognition. To approximate dorsal stream function, networks will be trained to generate realistic human grasp points.



"Utah teapot simple 2.png" by Dhatfield is licensed under CC BY-SA 3.0

Figure 6. An example of the two outputs generated by the two-task network from each input.

Hypotheses

1. The two-task network will do better on performance measures than the lesioned network and single-task network.
2. In the two-task network, more units will contribute towards the grasp point generation task than object recognition.
3. Representations in the two-task network will be more like human data than the single-task network.

Next Steps

Before creating the neural networks, a dataset must first be assembled for training and validation.

Dataset criteria:

- 12 object categories
- Minimum of 50 examples per category
- Multiple viewpoints of each object
- Must have object identity labels
- Must have human grasp point labels

Currently, efforts are being made to fulfill these criteria by collecting virtual 3D models from online and using a normative grasp model[13] to generate realistic grasp point labels.

Additionally, the dataset will be augmented using naturalistic data augmentation, such as by altering lighting or camera movement.

References

1. Pinel, J. P. J., & Barnes, S. J. (2018). The visual system. In *Biopsychology* (10th ed., pp. 132–163). Pearson.
2. 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. https://doi.org/10.1146/annurev-vision-082114-035447
3. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.
4. Simonyan, K., & Zisserman, A. (2015, April 10). Very deep convolutional networks for large-scale image recognition. *ArXiv:1409.1556 [Cs]*. http://arxiv.org/abs/1409.1556
5. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *ArXiv:1512.03385 [Cs]*. http://arxiv.org/abs/1512.03385
6. 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. https://doi.org/10.1073/pnas.1403112111
7. Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*(11), 1-29. https://doi.org/10.1371/journal.pcbi.1003915
8. Guclu, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015
9. Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C. C., & Richards, B. A. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning [Preprint]. *bioRxiv*. https://doi.org/10.1101/2021.06.18.448989
10. Mahon, B. Z., Milleville, S. C., Negri, G. A. L., Rumiati, R. I., Caramazza, A., & Martin, A. (2007). Action-related properties shape object representations in the ventral stream. *Neuron*, *55*(3), 507–520. https://doi.org/10.1016/j.neuron.2007.07.011
11. Cloutman, L. L. (2013). Interaction between dorsal and ventral processing streams: Where, when and how? *Brain and Language*, *127*(2), 251–263. https://doi.org/10.1016/j.bandl.2012.08.003
12. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv:1811.12231 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1811.12231
13. Klein, L. K., Maiello, G., Paulun, V. C., & Fleming, R. W. (2020). Predicting precision grip grasp locations on three-dimensional objects. *PLOS Computational Biology*, *16*(8). https://doi.org/10.1371/journal.pcbi.1008081

Acknowledgements

This research was supported by the Western Undergraduate Summer Research Internships (USRI) program. Special thanks to Melvyn Goodale, Jody Culham, and Jonathan Michaels for their insightful feedback.