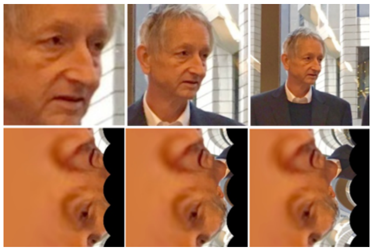
Visual expertise in an anatomically-inspired model of the visual system

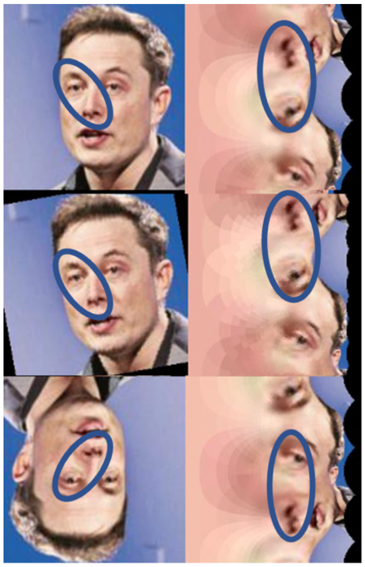
We report on preliminary results of an anatomically-inspired deep learning model of the visual system and its role in explaining the face inversion effect. Contrary to the generally accepted wisdom, our hypothesis is that the face inversion effect can be accounted for by the representation in V1 combined with the reliance on the configuration of features due to face expertise. We take two features of the primate visual system into account: 1) The foveated retina; and 2) The log-polar mapping from retina to V1. Our entire model (not used here) includes two more features, a salience or attention map for sampling the image, and dual pathways from V1, central and peripheral. Here we have just one pathway, and we sample from the image randomly.

The log-polar mapping, when used as input to a convolutional neural network (CNN), provides two kinds of invariances. Scale is just a left-right shift in this representation (see image of Geoff). Rotation in the image plane is an up-down shift (see image of Elon). Because CNNs are (somewhat) translation invariant, then the network as a whole is scale and rotation invariant. However, translation invariance is lost. We make up for this by sampling from the image at multiple points.

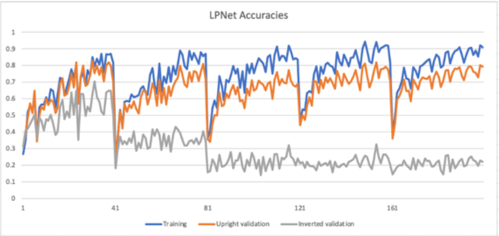


We compare the results of training our model on faces, mono-oriented objects, and non-mono-oriented objects, and test the effects of inversion on the three categories. We simulate acquisition of faces, etc., by gradually increasing the number of identities the network learns. We find that the more faces the network knows, the more the network shows the face inversion effect, while the effect on cars and non-mono-oriented objects is smaller.

It may be puzzling why a network that is rotation invariant shows any inversion effect at all. This is because there is a topological difference between the two invariances. While the shift left and right corresponds to scale, the up-down shift corresponds to rotation, and V1 is not a torus. Instead, it only ranges from 90 degrees to 270 degrees. Hence, the image "falls off" the top and the configuration of features is disrupted. As shown on the right, in the first two images, the nose is next to the right eye, but when the image is inverted, the nose is next to the left eye. Hence, as the network learns more and more faces, and configural information comes to play a role in its performance, the network is more disrupted by inversion.



The graph on the right shows the accuracy of the network on the training set (in blue) and the holdout set (orange) in the situation where we are adding 8 faces every 40 epochs to the training set. Note that inversion performance (gray) decreases as more expertise with faces is acquired. In contrast, a standard convolutional network's inversion performance drops to nearly 0 in the same situation.


LPNet Accuracies
Training
Upright validation
Inverted validation