# The Visual Expertise Mystery Visualized

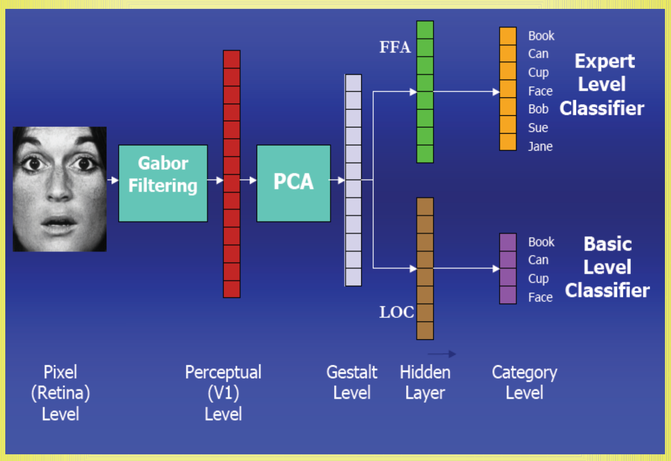
Jose Garcia Moreno-Torres. Advisor: Gary Cottrell


## Abstract

In previous works, we showed that networks trained to perform expert-level classification tasks as opposed to general classification were better suited to learn a new expert task, even if the new task did not have anything to do with the already learned one.

In this paper, we apply a reverse correlation technique that will let us visualize the features encoded in the hidden units of our neural networks; in an attempt to find an explanation for this phenomenom.


## Our Expertise Model




## Methodology

To investigate this issue, neural networks were trained on Greeble identification following various pretraining regimens. The stimulus set consisted of 300 64x64 8-bit grayscale images of human faces, books, cans, cups, and Greebles (60 images per class, 5 images of 12 individuals). The five images of each individual within each category were created by randomly moving the item 1 pixel in the vertical/horizontal plane, and rotating up to +/-3 degrees in the image plane.

Images were preprocessed by applying Gabor wavelet filters as a simple model of complex cell responses in visual cortex, extracting the magnitudes (which makes them nonlinear), normalizing via z-scoring, and reducing dimensionality to 40 via principal component analysis (PCA). Greeble images were not used to generate the principal components in order to model subjects' lack of experience with this category.

A standard feed-forward neural network architecture (40 input units, 60 standard logistic-sigmoid hidden units, variable numbers of linear output units) was used. Networks were trained using a learning rate of 0.01 and momentum of .5.

During pretraining, both the basic and face expert network learned to perform basic level categorization on all 4 non-Greeble categories. The face expert was additionaly taught to perform subordinate categorization of human faces. In phase two, the pretrained networks learned subordinate level Greeble categorization along with their original task. Both networks were trained on 30 images (3 images of 10 individuals) per class during pre-training and 30 more images of Greebles in phase 2.

Once we had the networks trained, we analyzed them. To create an image that represented accurately the features encoded in a given hidden unit (ie, drives maximally said unit), we used the input weights of said unit as a starting point. After multiplying them by an escalating factor, we inverted the PCA, and applied the algorithm developed by Shams & von der Malsburg (2002) to invert the obtained Gabor magnitudes. The basic idea behind it is to start with a random image and, using the difference between its gabor magnitudes and the target magnitudes; calculate an update to it using gradient descent.

Due to time constraints, we could not obtain the images corresponding to every hidden unit for every network; so we settled for the most interesting ones. We chose:

- The two units with the highest intra-class* variance (the class we chose for this experiment was faces) both in the face expert network and in the basic one, and both pre- and post- greeble training.
- The two units with the highest intra-class* variance (for greebles) both in the face expert network and in the basic one, and both pre- and post- greeble training.
- The unit with the highest FLD**, for every possible class pair (ie, face-greeble, face-cup, face-can, face-book, greeble-cup, ...).

(*) To measure the intra-class variance, we used the standard deviation among the outputs for all images of the class of interest. This units are particularly interesting because they show what do our nets look at when trying to distinguish between faces or greebles.

(**) FLD stands for Fisher's Linear Discriminant, which is defined as follows:

$$(m_2 - m_1)^2 / (s_1^2 + s_2^2)$$

where $m_i$ is the mean of the outputs for class i and $s_i^2$ represents the within-class covariance. Thus value basically tells us what unit is distinguishing between the chosen classes. Usually, the highest FLD was almost twice as much as the second highest. This suggests that a single unit is taking care of the disccrimination between two given classes.

## Results

Legend: Hidden unit number - amount in which this image drives said unit

| Network | Top unit in terms of... face variability | | greeble variability | | face-cup FLD | face-can FLD | face-book FLD |
|---|---|---|---|---|---|---|---|
| Face expert, pre-train | 3 - 0.8254 | 4 - 0.9415 | 35 - 0.8324 | 4 - 0.9415 | 15 - 0.8044 | 27 - 0.8974 | 27 - 0.8974 |
| Face expert, post-train | 24 - 1.0000 | 3 - 0.9969 | 48 - 0.9632 | 35 - 0.9987 | 36 - 0.9696 | 21 - 0.8468 | 44 - 0.9936 |
| Basic, pre-train | 42 - 0.9868 | 5 - 0.9506 | 42 - 0.9868 | 60 - 0.8544 | 1 - 0.7872 | 1 - 0.7872 | 50 - 0.8110 |
| Basic, post-train | 46 - 0.9998 | 10 - 1.0000 | 39 - 0.9999 | 51 - 0.9769 | 15 - 0.9307 | 7 - 0.9343 | 38 - 0.8462 |

| Network | Top unit in terms of... cup-can FLD | cup-book FLD | can-book FLD | face-greeble FLD | cup-greeble FLD | can-greeble FLD | book-greeble FLD |
|---|---|---|---|---|---|---|---|
| Face expert, pre-train | 17 - 0.8014 | 20 - 0.8410 | 20 - 0.8410 | 5 - 0.8998 | 24 - 0.9632 | 50 - 0.8760 | 5 - 0.8998 |
| Face expert, post-train | 38 - 0.7916 | 44 - 0.9936 | 10 - 0.8955 | 21 - 0.8468 | 52 - 0.8620 | 17 - 0.9718 | 18 - 0.9986 |
| Basic, pre-train | 54 - 0.8076 | 41 - 0.6043 | 41 - 0.6043 | 50 - 0.8110 | 39 - 0.8790 | 54 - 0.8076 | 40 - 0.8024 |
| Basic, post-train | 54 - 0.7217 | 1 - 0.9809 | 9 - 0.9600 | 38 - 0.8462 | 24 - 0.8078 | 10 - 1.0000 | 52 - 0.7625 |

## Discussion

The results obtained can be qualified as encouraging. While it is true that some of these images are difficult to interpret; some others do show characteristics from which conclusions can be drawn.

As an example, it is interesting to consider the top units in terms of face variability. They show evidence of being sensitive to the eyes (apparently, our nets found useful to look at that part of the face to distinguish individuals); but they also show some cup structure. This means they are dual use units!

There is still a lot of work to be done in analyzing these images, but we can affirm the experiment was pretty successful. It provided us with a fairly accurate representation of the encoded features of the network, and presents us the new challenge of interpreting said representations.


## Acknowledgments

- CaliT2 Undergraduate Research Program, who provided the funding for this research.
- Members of Gary's Unbelievable Research Unit (GURU) for their invaluable help.


## References


- Shams, L. & von der Malsburg, C. (2002). The role of complex cells in object recognition. Vision Research. Vol. 42 (22), pp. 2547-2554.
- Carrie Joyce and Garrison W. Cottrell (2004) Solving the visual expertise mystery. In Connectionist Models of Cognition and Perception II: Proceedings of the Eighth Neural Computation and Psychology Workshop
- Tong, M H., Joyce, CA. and Cottrell, GW (2005) Are Greebles special? Or, why the Fusiform Fish Area would be recruited for sword expertise (if we had one). In Proceedings of the 27th Annual Cognitive Science Conference, La Stresa, Italy
- Tong, M.H., et al., Why is the fusiform face area recruited for novel categories of expertise? A neurocomputational investigation, Brain Res. (2007), doi:10.1016/j.brainres.2007.06.079