Optimizing For Possible Feature Combinations in Discriminative Vision Models

Chris Hamblin (chrishamblin@fas.harvard.edu)

George Alvarez

Talia Konkle

Harvard Psychology – William James Hall, 33 Kirkland St, Cambridge, MA 02138

Abstract

In this work we leverage feature visualization to probe the bounds of feature combinations in the InceptionV1 object recognition model (Szegedy et al., 2015). While our technique also yields conventional/viewable feature visualizations, we demonstrate how such optimizations can reveal contingencies between feature pairs that are difficult to infer from their responses to natural images alone. We propose a data visualization motif that is ideal for quickly assessing the relations between arbitrary feature pairs.

Keywords: Object recognition, visual features, interpretability, feature visualization

Figure 1: What's this?

Introduction

What do you see in Figure 1? According to Dalle-3 (Betker et al., 2023), it's a depiction of a 'broccoli elephant'. Despite the unnatural feature combination, the model represents it 'correctly', in that we perceive 'broccoli' and 'elephant' simultaneously in the generated image. While it's easy to probe generative models for unusual feature combinations like this (just give them zany prompts), it's not obvious how to do so in discriminative models. That said, we know discriminative models have the potential to be very expressive in their feature combinations; after all, our own visual system has no problem representing the broccoli elephant as such.

This raises an important question when assessing features in discriminative models; namely, what features combination are *possible* for the model, and how do we disentangle possibility from the feature covariances introduced by the input data generating process? It may be that some feature combinations are possible but others aren't, given the way that each feature is computed. For example, can a percept be simultaneously pointy and rounded? What if 'pointy' and 'rounded' are computed like the functions $pointy(x) = x$ and $rounded(x) = -x$? There could be no such image x that was pointed and rounded in such a case.

Feature Combinations with Optimization

In this work, we will test if feature combinations are possible by optimizing the model's inputs. Given a model, let's denote a function that computes a set of features $\boldsymbol{f}(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ is the domain of the model. Here we will consider features that correspond to latent neurons in a neural network, but the technique we introduce is extendable to features defined in other ways, so long as $\boldsymbol{f}$ is differentiable. Additionally, we'll constrain ourselves to the simple case of feature pairs; i.e. activations $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^2$. Given our features are continuous, we can formalize feature combinations as all the directions in which the feature vector $\boldsymbol{y}$ can point. In $\mathbb{R}^2$, direction can be parameterized by a single value, the angle θ between $\boldsymbol{y}$ and $[1,0]$. We can optimize for any θ (feature combination) by maximizing the cosine similarity between $\boldsymbol{y}$ and the unit vector $[sin(\theta), cos(\theta)]$. However, this objective may not be sufficient for generating feature combinations, as cosine similarity is invariant to changes in the *magnitude* of $\boldsymbol{y}$. We wouldn't want our probe for feature combinations to result in 'solutions' that yield small activation for both features. Luckily, the $cosdot$ objective has previously been proposed/utilized for the purposes of feature visualization (Carter, Armstrong, Schubert, Johnson, & Olah, 2019; Mordvintsev, Pezzotti, Schubert, & Olah, 2018; Olah, Mordvintsev, & Schubert, 2017), and is well suited to our needs. The $cosdot$ objective multiplies the dot product of two vectors by their cosine similarity; thus one can optimize one vector, $\boldsymbol{y}$, with respect to a target vector $\boldsymbol{h}$, such that the optimized vector is encouraged to both decrease its angle with the target (cosine similarity) and increase its overall magnitude (dot product). Given some target feature combination θ, the $cosdot$ ($\mathcal{C}$) objective yields;

$$\mathcal{C}(\boldsymbol{y}, \theta; p) := \frac{(\boldsymbol{y} \cdot \boldsymbol{h})^{p+1}}{(||\boldsymbol{y}|| \cdot ||\boldsymbol{h}||)^p} \quad with \quad \boldsymbol{h} := [cos(\theta), sin(\theta)]$$

p is a hyperparameter that controls how much weight is to be placed on the cosine similarity (direction). We set $p = 4$ for all experiments conveyed in this paper. Given this objective, we can attempt to optimize for images that yield arbitrary feature combinations;

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathcal{X}}{\arg\max}\, \mathcal{C}(\boldsymbol{f}(\boldsymbol{x}); \theta, p)$$

We can optimize for $\boldsymbol{x}^*$ with gradient ascent, augmented with additional feature visualization tricks (Fel et al., 2023).

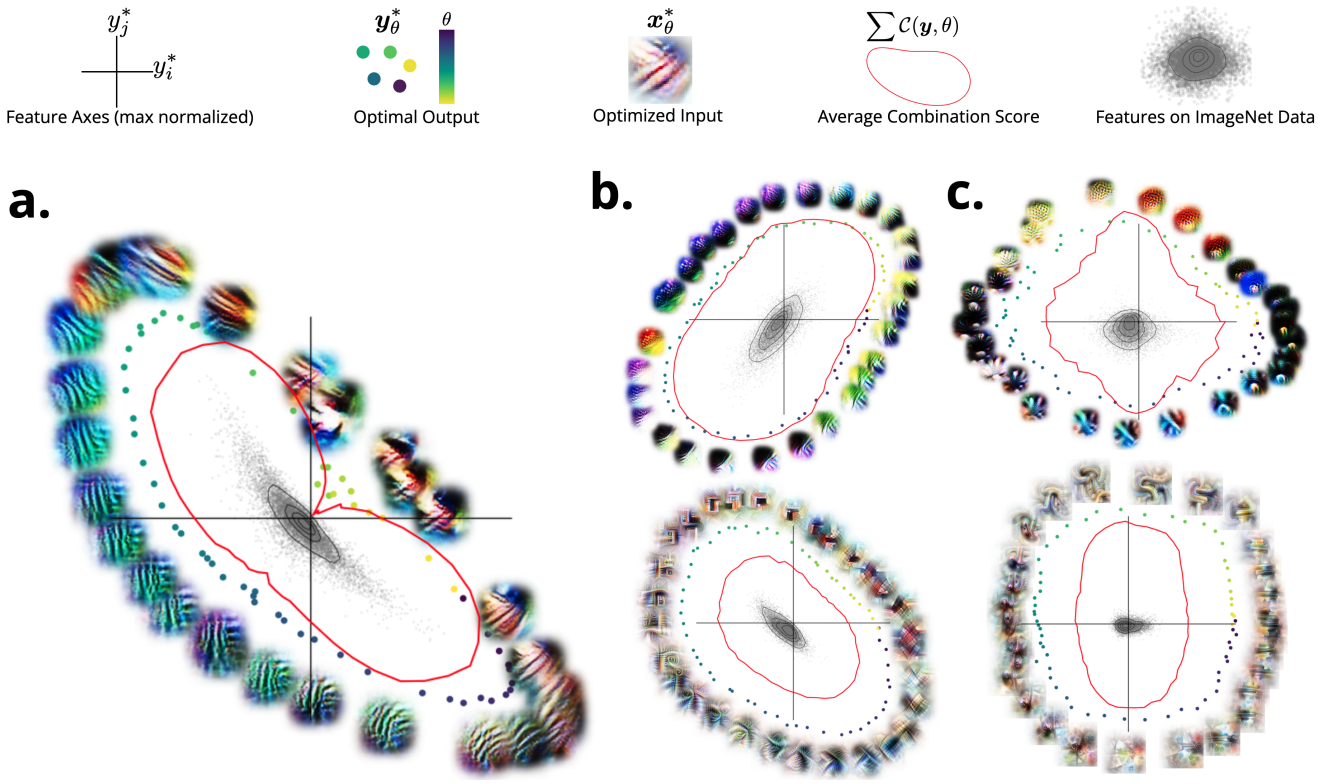

Figure 2: Pair-wise 'feature combination plots'

Experiment and Visualization

Here, we will test our optimization technique on pairs of latent InceptionV1 (Szegedy et al., 2015) neurons, and demonstrate our methods for visualizing the results to garner maximal insight. We begin by generating latent feature activations in response to a sample from the ImageNet (Deng et al., 2009) dataset. We do this to provide a criterion for selecting neural pairs to investigate. For example, in Figure 2. above, we visualizing pairs of neurons that yield the highest, lowest, and closest to zero correlation in their respective layer. We next optimize conventional feature visualizations (neuron-wise activation maximization) for each neuron in the pair, in order to define a *scale* with which to normalization y_i and y_j. We are effectively scaling the activations by those values we get when the optimization process is unconstrained by direction. Finally, we optimize $\mathcal{C}(\boldsymbol{f}(\boldsymbol{x}), \theta)$, targeting 60 distinct θ evenly spaced on $[0 - 2\pi]$. We repeated this procedure 5 times under different noise seeds for each feature pair tested.

Fig 2 shows the results of this experiment for several features, which we will explore in turn. First though, let's orient ourselves to parsing these feature combination plots in the general case. First, the axes convey the pairwise feature space $\mathcal{Y}$. As mentioned, each axis is scaled by the maximum y_i^* achieved with optimization, the tip of each axis corresponds to this value. The scatter plot shows for each θ that $\boldsymbol{y}_\theta^*$ which maximized $\mathcal{C}$ across the 5 seeds. For every other value of θ we also show the synthesized image $\boldsymbol{x}_\theta^*$, just beyond the $\boldsymbol{y}_\theta^*$ it induced. The grey point cloud shows the activations of $\boldsymbol{y}$ in response to ImageNet data, conveyed on the same scale. Finally, the red line conveys the average score of $\mathcal{C}(\boldsymbol{y}, \theta)$ across random seeds. We place this value on the plot using polar coordinates, drawing a red point at $(\sum \mathcal{C}, \theta)$, then connecting them in a loop. Importantly, $\mathcal{C}$ exists in a different space than $\mathcal{Y}$, while θ does not. This means, we are free to scale the red loop from the origin, but not rotate or translate it.

So, what can such plots actually tell us? Let's start with the example depicted in Figure 2.a, which conveys the most anti-correlated feature pair in layer 'conv2d2'. The feature pair is excited by gratings in opposite orientations, and positive co-activation only rarely happens in response to natural data. Additionally, we can see that co-activation is very difficult to optimize for, as indicated by the combination score (red loop) collapsing in the upper right quadrant. We find it curious that the quadrant which is asymmetrical with the others is precisely that which will bypass the relu in the next operation. It suggests this feature pair is made impossible through a learned constraint mechanism, and that co-occurrence would induce loss downstream. Importantly, our plots show that other feature pairs (Figure 2.b) in the model are anti-correlated but do not show this 'impossibility' signature. Finally, in Figure 2.c we see two pairs of features that each show 0 covariance in their activations to ImageNet, but disambiguate in their combination plots. In their limit of activation, it appears the first pair compete linearly; as one increases the other must decrease

proportionally in activation, tracing the contour of an L_1 norm. The second pair approximate a much higher p-norm, taking on all possible combinations with little trade-off in activation magnitude.

Examples like these may only scratch the surface of possible feature combination motifs. That said, the space of exploration is quite large - how should we select features for combination? We are excited to see what pairings the community comes up with.

References

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., ... others (2023). Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, *2*(3), 8.

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., & Olah, C. (2019). Activation atlas. *Distill*, *4*(3), e15.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Cvpr09.*

Fel, T., Boissin, T., Boutin, V., Picard, A., Novello, P., Colin, J., ... others (2023). Unlocking feature visualization for deeper networks with magnitude constrained optimization. *arXiv preprint arXiv:2306.06805*.

Mordvintsev, A., Pezzotti, N., Schubert, L., & Olah, C. (2018). Differentiable image parameterizations. *Distill*.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*. (https://distill.pub/2017/feature-visualization) doi: 10.23915/distill.00007

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).