Testing 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 possible/impossible 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 techniques can reveal contingencies between feature pairs that are difficult to infer from their responses to natural images alone. We ultimate 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: ***?***

Introduction

What do you see in Figure 1? According to Dalle-3 (Betker et al., 2023), it's a depiction of a 'broccoli elephant'. One of the reasons we are so impressed by generative models like this is that 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.

This raises an important question when assessing features in discriminative models; namely, what features combinations 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. Appealing again to our perceptual intuition, I can perceive a black house, even if 'black' and 'house' have never co-occurred in the natural images I've seen before. However, can I perceive black milk? This is less clear, black milk may not look to me like milk at all. Lastly, it may be that some features *must* occur in combination, such as features for 'soccer ball' and 'curved'.

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 the outputs of 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 $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^2$. In the discrete case (where we usually consider 'combinations'), we may want to know if there exists some $\boldsymbol{x}$ for which $\boldsymbol{f}(\boldsymbol{x})$ yields $[1,1]$, $[1,0]$, or $[0,1]$. 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;

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

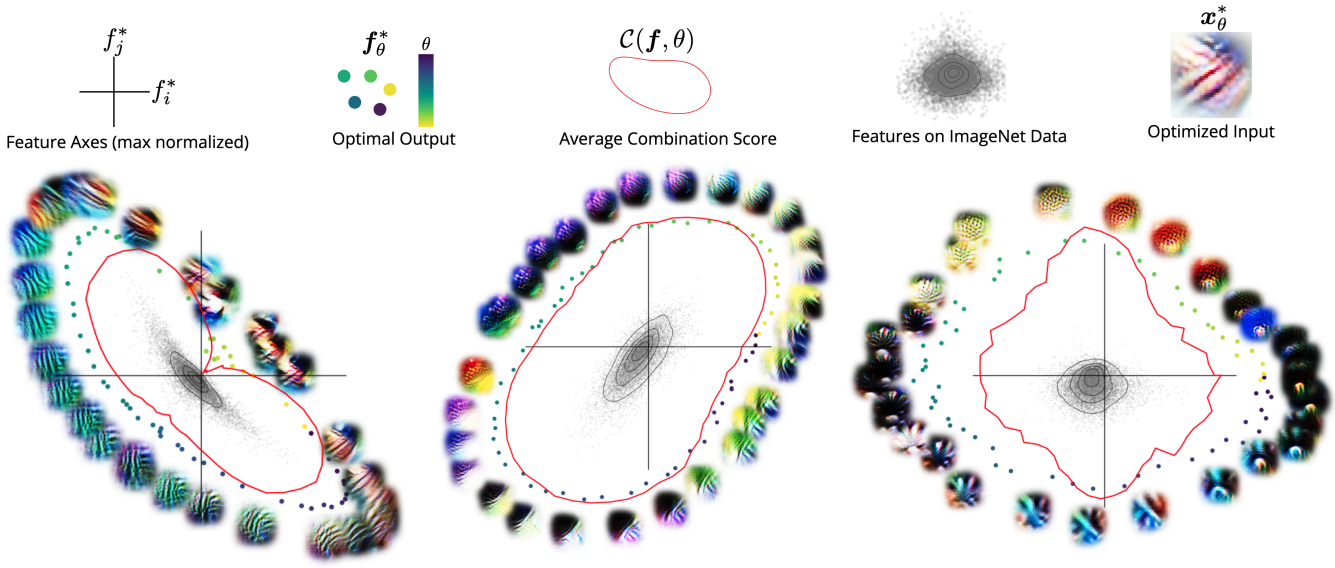

Figure 2: *?*

p is a hyperparameter that controls how much weight is to be placed on the cosine similarity (direction). Given this objective, we can attempt to optimize for image that yield arbitrary feature combinations;

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

We we can find through gradient descent, augmented with whatever additional feature visualization tricks we see fit.

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 incite. 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 those pairs of neurons that yield the highest, lowest, and closest to zero correlation in their respective layer. We next optimize conventional (activation maximization) feature visualizations for each neuron in the pair. We do this in order to define a *scale* with which to normalization the activations for each neuron. We are scaling activations by those values we get when the optimization process is unconstrained by direction. Finally, we optimize for

The text, tables and figures of a CCN submission can be no longer than two pages. If needed, references may extend onto an additional (third) page (in this example, references happen to fit into two pages).

The text of the paper should be formatted in two columns with an overall width of 7 inches (17.8 cm) and length of 9.25 inches (23.5 cm), with 0.25 inches between the columns. Leave two line spaces between the last author listed and the text of the paper. The left margin should be 0.75 inches and the top margin should be 1 inch. Use 10 point Modern with 12 point vertical spacing, unless otherwise specified.

The title should be in 14 point, bold, and centered. The title should be formatted with initial caps (the first letter of content words capitalized and the rest lower case). Each author's name should appear on a separate line, 11 point bold, and centered, with the author's email address in parentheses. Under each author's name list the author's affiliation and postal address in ordinary 10 point type.

Indent the first line of each paragraph by 1/8 inch (except for the first paragraph of a new section). Do not add extra vertical space between paragraphs.

First Level Headings

First level headings should be in 12 point, initial caps, bold and centered. Leave one line space above the heading and 1/4 line space below the heading.

Second Level Headings

Second level headings should be 11 point, initial caps, bold, and flush left. Leave one line space above the heading and 1/4 line space below the heading.

Third Level Headings Third level headings should be 10 point, initial caps, bold, and flush left. Leave one line space above the heading, but no space after the heading.

Formalities, Footnotes, and Floats

Use standard APA citation format. Citations within the text should include the author's last name and year. If the authors' names are included in the sentence, place only the year in parentheses, as in Newell and Simon (1972), but otherwise place the entire reference in parentheses with the authors and

year separated by a comma (Newell & Simon, 1972). List multiple references alphabetically and separate them by semicolons (Chalnick & Billman, 1988; Newell & Simon, 1972). Use the "et al." construction only after listing all the authors to a publication in an earlier reference and for citations with four or more authors.

Footnotes

Indicate footnotes with a number[1] in the text. Place the footnotes in 9 point type at the bottom of the column on which they appear. Precede the footnote block with a horizontal rule.[2]

Tables

Number tables consecutively. Place the table number and title (in 10 point) above the table with one line space above the caption and one line space below it, as in Table 1. You may float tables to the top or bottom of a column, or set wide tables across both columns.

Table 1: Sample table title.

Error type	Example
Take smaller	63 - 44 = 21
Always borrow	96 - 42 = 34
0 - N = N	70 - 47 = 37
0 - N = 0	70 - 47 = 30

Figures

Make sure that the artwork can be printed well (e.g. dark colors) and that the figures make understanding the paper easy. Number figures sequentially, placing the figure number and caption, in 10 point, after the figure with one line space above the caption and one line space below it, as in Figure 3. If necessary, leave extra white space at the bottom of the page to avoid splitting the figure and figure caption. You may float figures to the top or bottom of a column, or set wide figures across both columns.

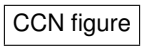

Figure 3: This is a figure.

Acknowledgments

Place acknowledgments (including funding information) in a section at the end of the paper.

References Instructions

Follow the APA Publication Manual for citation format, both within the text and in the reference list, with the following exceptions: (a) do not cite the page numbers of any book, including chapters in edited volumes; (b) use the same format for unpublished references as for published ones. Alphabetize references by the surnames of the authors, with single author entries preceding multiple author entries. Order references by the same authors by the year of publication, with the earliest first.

Use a first level section heading, "**References**", as shown below. Use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 1/8 inch. Below are example references for a conference paper, book chapter, journal article, dissertation, book, technical report, and edited volume, respectively.

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.

Chalnick, A., & Billman, D. (1988). Unsupervised learning of correlational structure. In *Proceedings of the tenth annual conference of the cognitive science society* (pp. 510–516). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Feigenbaum, E. A. (1963). The simulation of verbal learning behavior. In E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought.* New York: McGraw-Hill.

Hill, J. A. C. (1983). A computational model of language acquisition in the two-year old. *Cognition and Brain Theory*, *6*, 287–317.

Matlock, T. (2001). *How real is fictive motion?* Doctoral dissertation, Psychology Department, University of California, Santa Cruz.

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

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Ohlsson, S., & Langley, P. (1985). *Identifying solution paths in cognitive diagnosis* (Tech. Rep. No. CMU-RI-TR-85-2). Pittsburgh, PA: Carnegie Mellon University, The Robotics Institute.

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

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

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).

[1] Sample of the first footnote.

[2] Sample of the second footnote.