Optimizing for Possible Feature Combinations in Discriminative Vision Models

Chris Hamblin *(chrishamblin@fas.harvard.edu)*, Talia Konkle, George Alvarez

Problem: How can we distinguish *possible* feature combinations from correlations introduced by the data generating process?

- People can perceive *many* unnatural feature combinations
- Yet some feature combinations may be *impossible* to represent

black+house

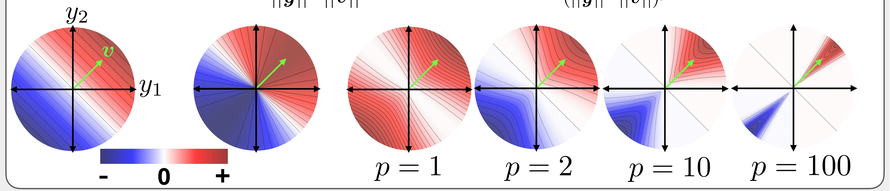
broccoli+elephant

pointy *curvy*

Solution: In a vision model, we can distinguish these cases by optimizing inputs for desired feature combinations.

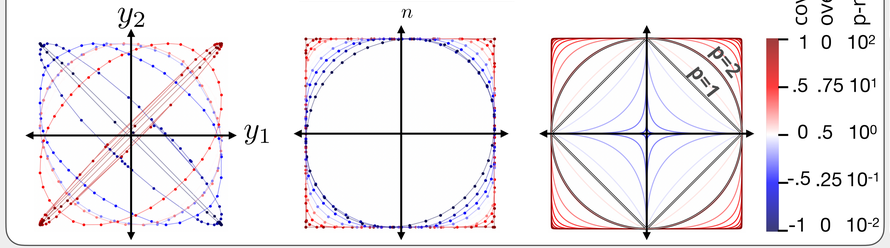
'Cosdot' Objective: Optimizes for both direction and magnitude

dot: $\boldsymbol{y} \cdot \boldsymbol{v}$ cosine: $\frac{\boldsymbol{y} \cdot \boldsymbol{v}}{||\boldsymbol{y}|| \cdot ||\boldsymbol{v}||}$ dot*cos: $\frac{(\boldsymbol{y} \cdot \boldsymbol{v})^{p+1}}{(||\boldsymbol{y}|| \cdot ||\boldsymbol{v}||)^{p}}$ *carter et al. (2019)*



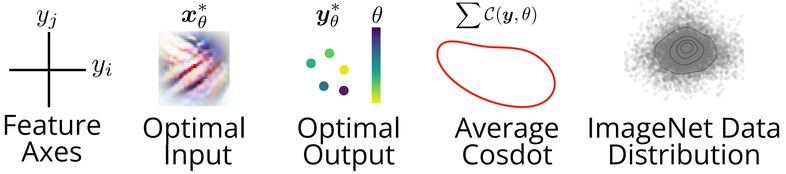
Toy Examples: What happens when we optimize 'cosdot' $\forall \boldsymbol{v}$ for a bounded linear model $\boldsymbol{y} = \mathbf{W}\sigma(\boldsymbol{x})$ *with* $\mathbf{W} \in \mathbb{S}^{2\times1000}$

varying: $Cov(W_1, W_2)$ $overlap(W_1, W_2)$ $\frac{|\{i : W_{1i} \neq 0 \wedge W_{2i} \neq 0\}|}{n}$ Overlap follows P-norm contours

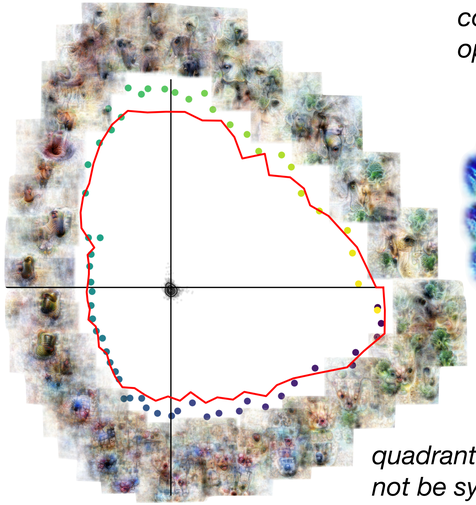


Experiments: We can perform the same experiment as in our toy examples on a *real* model (InceptionV1) by optimizing for neuron pairs. Using techniques from feature visualization *(Fel, et al. (2023))* guides the process towards perceptible optima

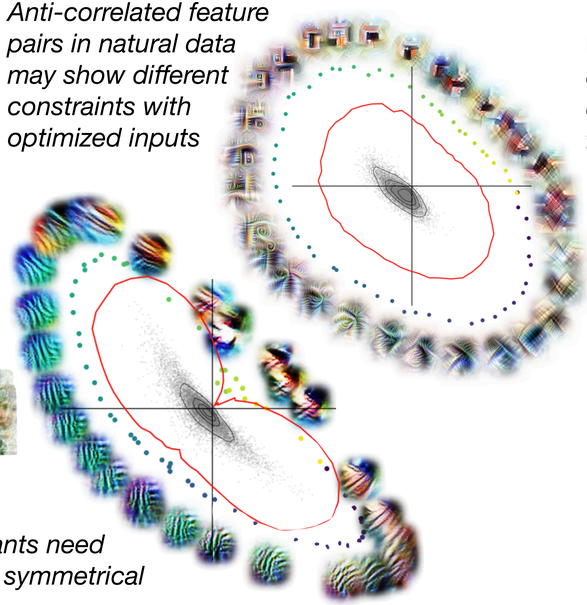
Legend



We could apply this technique to arbitrary features of interest, like the 'broccoli' and 'elephant' logits

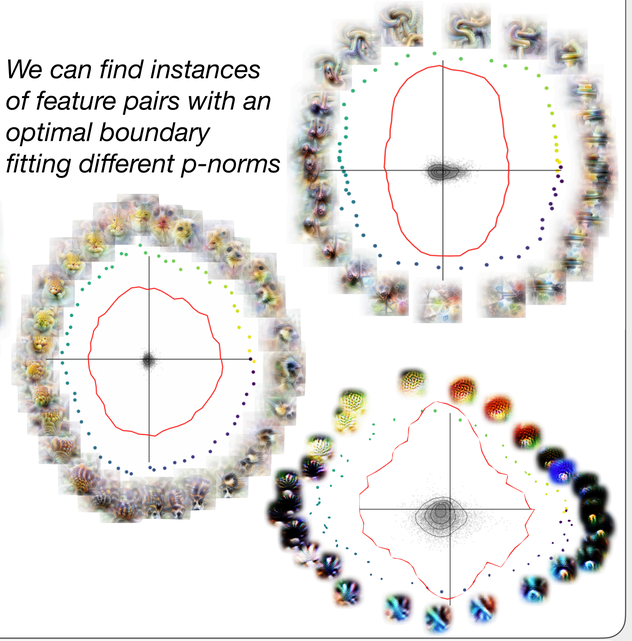


Anti-correlated feature pairs in natural data may show different constraints with optimized inputs



quadrants need not be symmetrical

We can find instances of feature pairs with an optimal boundary fitting different p-norms



Future Work & Open Questions:

- This analysis is slow; how do we aggregate summary statistics across pairs, and what should those stats be?
- Here we presuppose the 'features' correspond to neurons, thus superposition is an outstanding problem.
- How should we think about the 'unnaturalness' of these optimized images? Is it a blessing or a curse?
- How do we extend this analysis to more than 2 dimensions?