Understanding Deep Networks via Extremal Perturbations and Smooth Masks

Ruth Fong†, Mandela Patrick†, Andrea Vedaldi†‡
†University of Oxford ‡Facebook AI Research  *equal contribution

Attribution
Where is the network “looking”? We optimize the mask \( \Phi \) on the blurred image \( x \) for a given area \( m \).

Perturbation analysis
We test a region counterfactually by blurring its complement and observing the change in the network’s response.

Formalization
We optimize the mask \( \Phi \) to maximize the response of the network on the blurred image \( x \) for a given area \( m \).

Algorithm
Pick area \( m \) and perform SGD to optimize:

\[
\arg\max \Phi \Phi(x(m) \otimes x) - \| \text{vec}(\Phi(x(m) \otimes x)) - r_x \|^2
\]

Results

- **Pointing Game**: Quantitative evaluation is difficult, but said, the method achieves competitive results on pointing game [6].

- **Sanity Checks**: Our method is sensitive to model weights. Below, model weights are progressively randomized [7].

Comparisons
Left: Extremal perturbations (ours) vs. meaningful perturbations [1]. We show that our method is more sensitive and stable than prior work.

Below: Comparison with related works. Left: Extremal perturbations (ours) vs. feature inversions (M & R). Right: “DIR” activations with feature inversions (M & R) to inspect which features are highlighted and minimized by our channel attribution mask.

Channel attribution
We learn channel-wise attribution masks to identify salient channels.

TorchRay
Interpretability library: github.com/facebookresearch/torchray

References