A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations

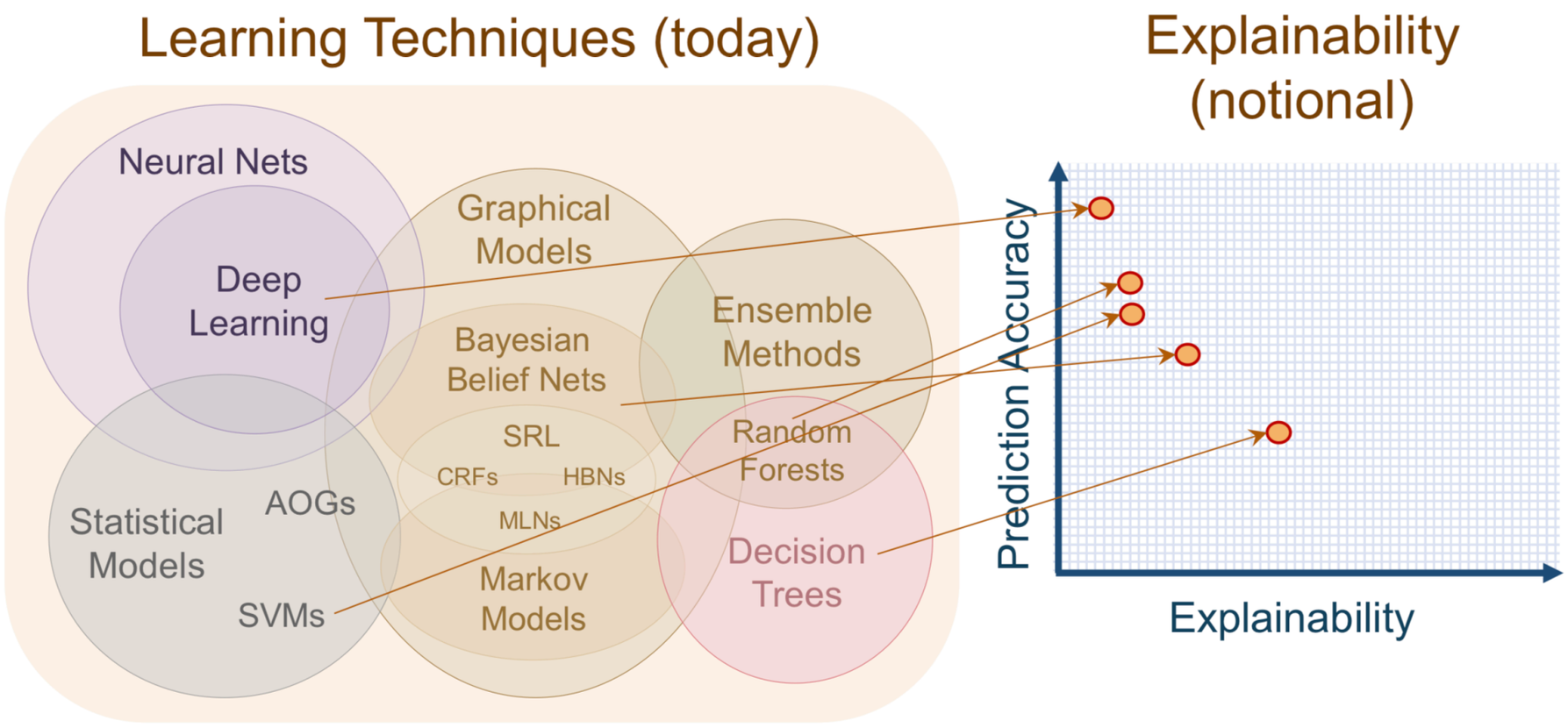
Weili Nie, Yang Zhang and Ankit B. Patel

Rice University
Baylor College of Medicine

Outline

- Motivation: Interpretability of deep learning
- Backpropagation-based visualizations
  - Formal definitions
  - Perplexing behaviors
- Theoretical explanations
  - Starting from simple: A three-layer random CNN
  - Extensions more complex cases
- Experiments
  - Impact of local connections
  - Impact of max-pooling and network depth
  - Average distance l_2 statistics
  - Adversarial attack on VGG
  - VGG with partly trained weights
- Conclusions

Motivation: Interpretability of deep learning



- Explainability versus performance in learning techniques [1]

[1] Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.

Backpropagation-based Visualizations

Formal Definitions

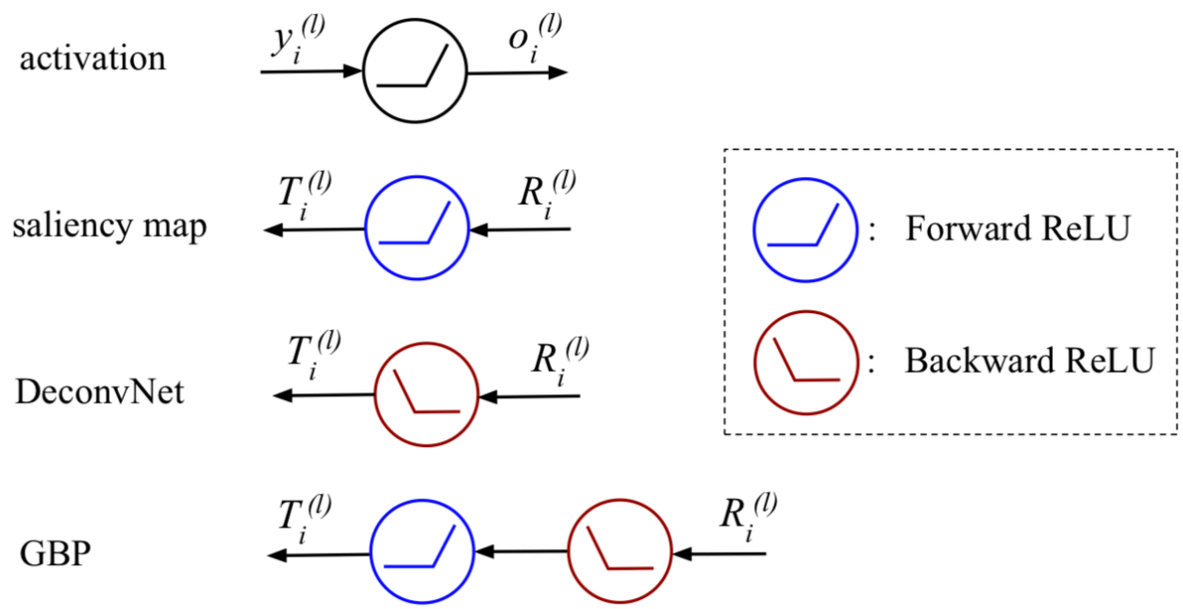
The backpropagation-based visualizations for propagating the output score back through the *i*-th ReLU activation in the *l*-th layer are defined as

$$T_i^{(l)} = h\left(R_i^{(l)}\right)\frac{\partial g\left(y_i^{(l)}\right)}{\partial y_i^{(l)}}$$

where the two functions $h(\cdot)$ and $g(\cdot)$ are defined as

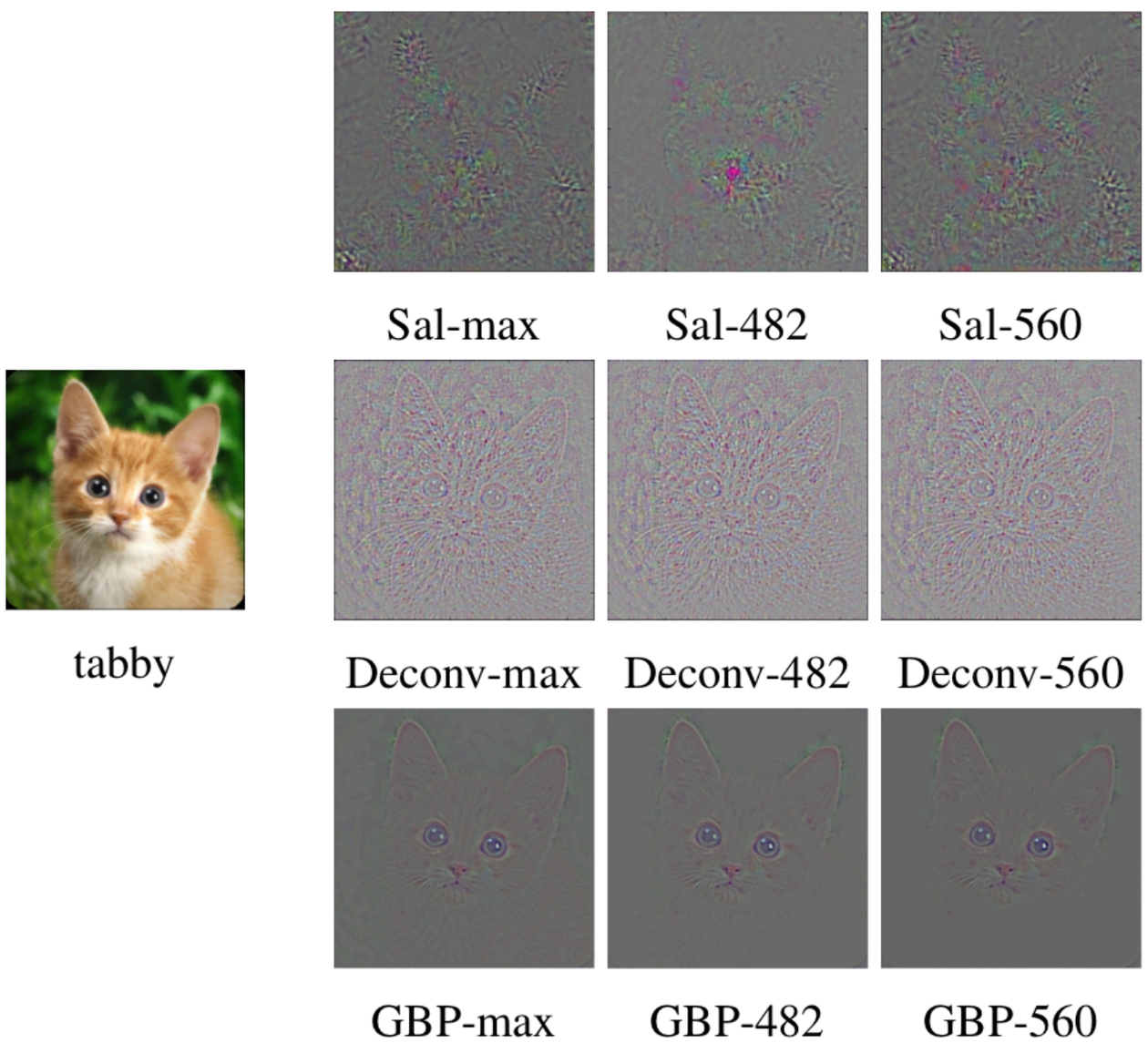
$$h(t) = \begin{cases} t & \text{for saliency map} \\ \sigma(t) & \text{for DeconvNet and GBP} \end{cases}$$

$$g(t) = \begin{cases} t & \text{for DeconvNet} \\ \sigma(t) & \text{for saliency map and GBP} \end{cases}$$



- Illustrations of saliency map, DeconvNet and GBP

Perplexing behaviors



- Saliency map, DeconvNet and GBP for the trained VGG-16 net.

Summary: DeconvNet and GBP are **more human-interpretable but less class-sensitive** than saliency map.

Theoretical explanations

- In a random three-layer CNN, we can approximate the backpropagation-based visualizations as

$$s_k\textbf{Sal}(x), s_k\textbf{Deconv}(x) \sim \mathcal{N}(0,I) \quad \textbf{(Theorem 1 \& 2)}$$
$$s_k\textbf{GBP}(x) \approx x$$

- GBP is approximately **recover the input** while saliency map and DeconvNet are **random noise**.

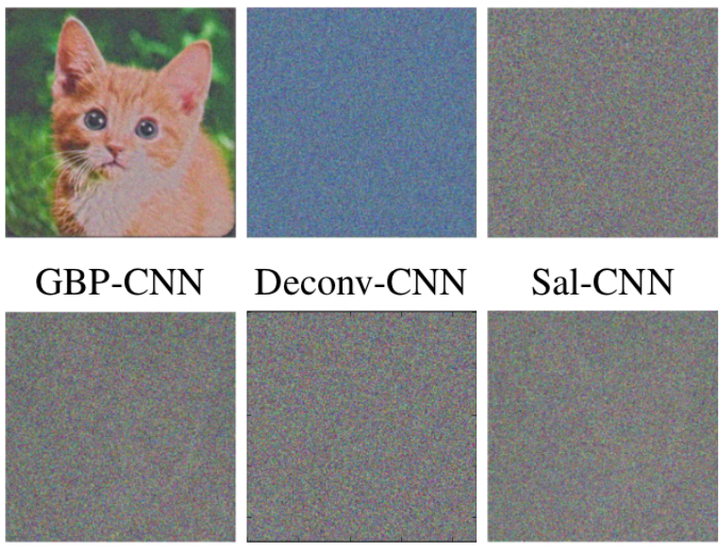
- Key insights: **Backward ReLU** and **local connections** are the two main causes for the input image recovery of GBP.

- By introducing the **max-pooling**, DeconvNet behaves approximately the same with GBP

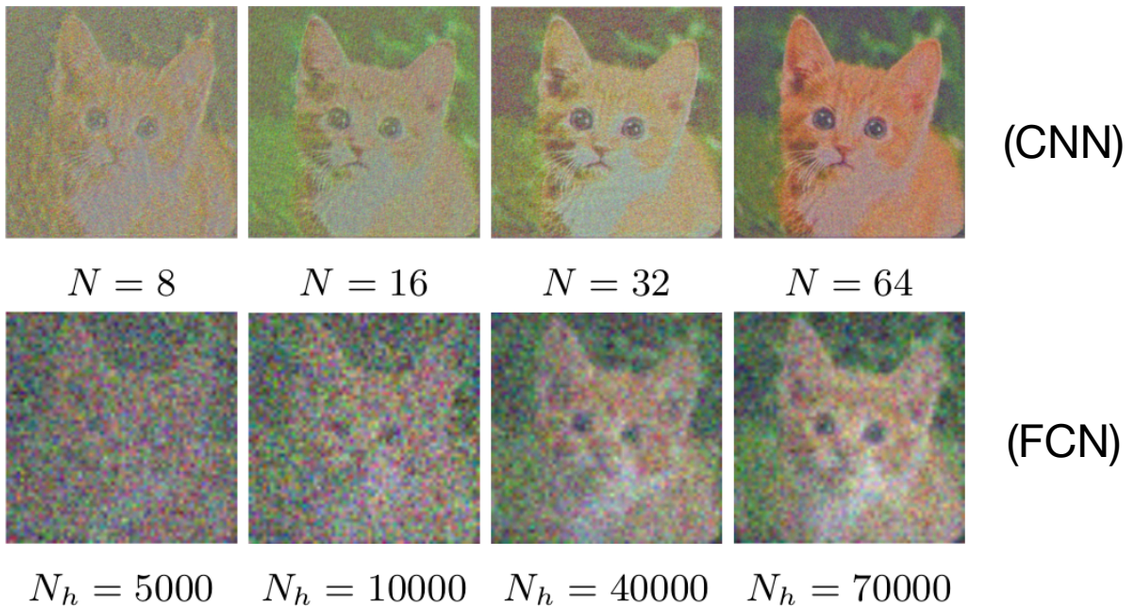
Experimental Results

Impact of local connections

- Compare visualizations in random three-layer CNNs vs. random three layer FCNs



- GBP, DeconvNet and Saliency map for fixed number of hidden filters (neurons)



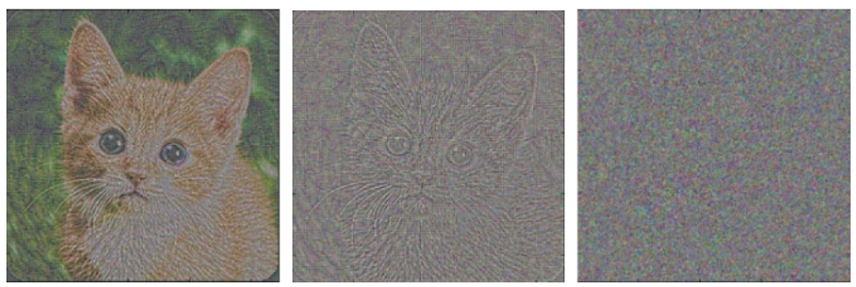
- GBP for different number of hidden filters (neurons)

Only GBP in the CNN can produce a human-interpretable visualization

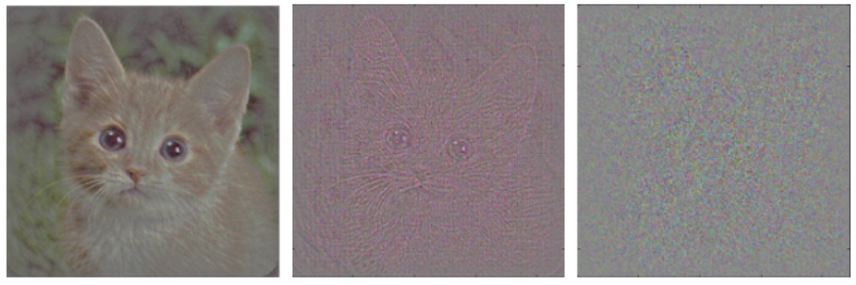
Local connections in CNNs contribute to the good visual quality of GBP

Impact of max-pooling and networks depth

- Add a max-pooling layer in the above random three-layer CNN and compare it with a random VGG-16 net



GBP-pool Deconv-pool Sal-pool



GBP-VGG Deconv-VGG Sal-VGG

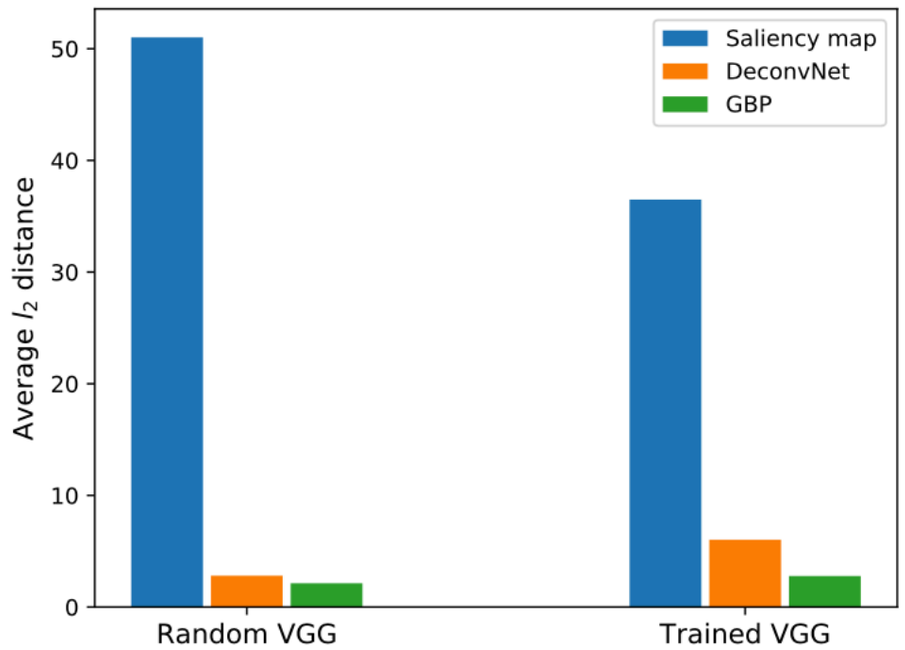
- GBP, DeconvNet and saliency map for CNN with max-pooling and VGG

DeconvNet now becomes human interpretable via the max-pooling

Three visualizations change little after increasing the network depth

Average l_2 distance statistics

- Calculate the l_2 distance of two visualization results given two different class logits for each of the 10K images from ImageNet and then take the average

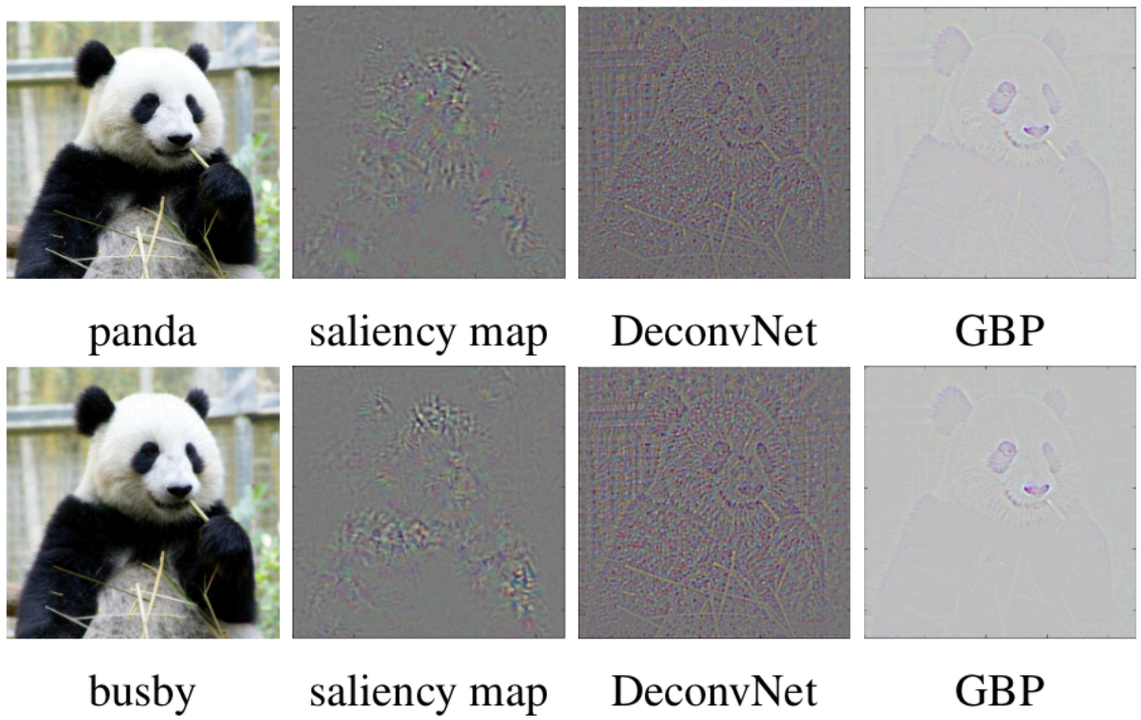


- Average L2 distance in a random VGG vs. a trained VGG

The average L2 distance of saliency map is much larger than that of both GBP and DeconvNet in either a random VGG or a trained VGG

Adversarial Attack on VGG

- Generate an adversarial example via FGSM and then compare visualizations of the original image and its adversarial counterpart

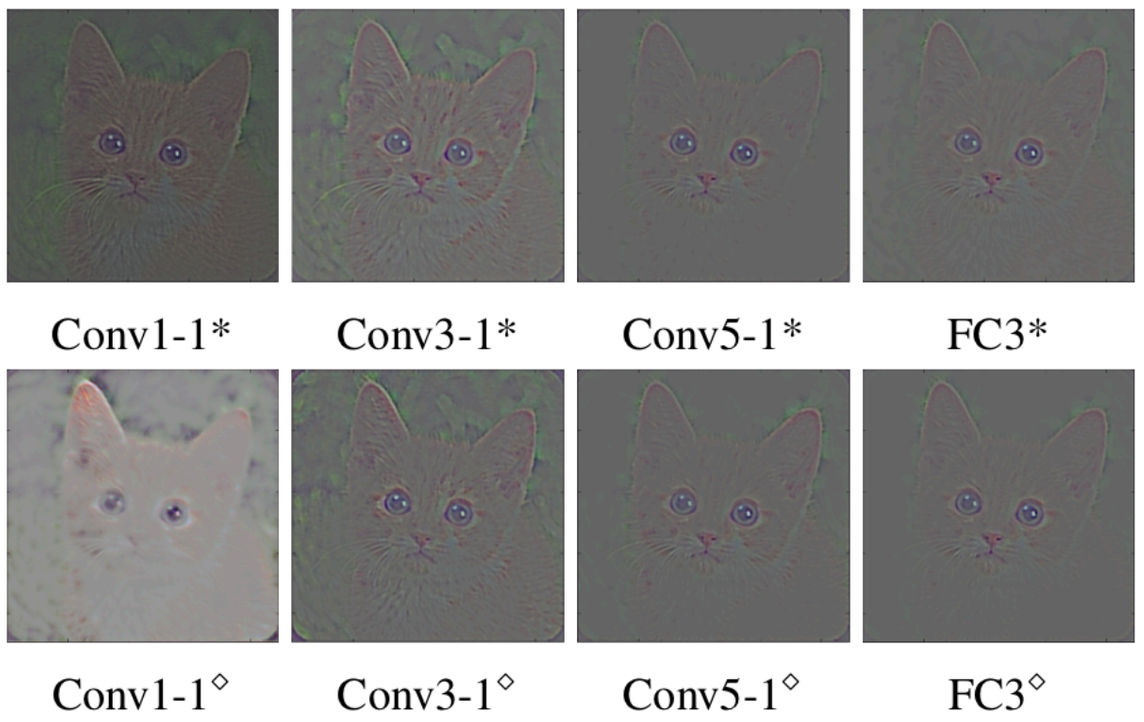


- The original image “panda” vs. its adversarial counterpart “busby” in a trained VGG-16 net

Saliency map changes significantly whereas GBP and DeconvNet remain almost unchanged after replacing “panda" by “busby”

VGG with Partly Trained Weights

- Investigate the contributions of different layers in the trained VGG-16 net to the visual differences of GBP between trained and random nets



- Load trained weights **up to** the indexed layer and leave the later layers randomly initialized vs. load trained weights **except for** the indexed layer is randomly initialized instead.

It is the convolutional layers rather than the dense layers that account for filtering out image patches in GBP

The earlier convolutional layer has more important impact in the GBP visualization than the later convolutional layer

Conclusions

- Proposed a theoretical explanation for perplexing behaviors of backpropagation-based visualizations, which reveals:

    1. Unlike saliency map, both GBP and DeconvNet are essentially doing (partial) image recovery, which is unrelated to the network decisions

    2. It is the backward ReLU, used by both GBP and DeconvNet, along with the local connections in CNNs, that is responsible for human-interpretable visualizations.

    3. DeconvNet also relies on the max-pooling to recover the input

- Extensive experiments are provided that support the theoretical analysis

- Come and see our poster tonight **06:15 - 09:00 PM** in **Hall B #19**