Visualizing and Understanding

First layer visualized filters show 유의미한 info.

Last layer : Nearest Neighbors

- pixel space에서 nearest neighbors는 부정확
- L2 nearest neighbors in feature space (ex. Last layer)는 상당히 정확 (원쪽 바라보는 코끼리, 오른쪽 바라보는 코끼리 동일시)

Dimensionality Reduction (4096 → 2)

- simple algorithm : Principle Component Analysis (PCA : 주성분분석)
- more complex : t-SNE (t-distributed Stochastic Neighbor Embeddings)

Visualizing activation map — available

Maximally Activating Patches

- pick a layer and a channel
- run many images through the network, record values of chosen channel
- visualize image patches that correspond to maximal activations

Occlusion Experiments

- mask part of the image before feeding to CNN, draw heatmap of probability at each mask location

Saliency Maps

- Compute gradient of class score with respect to image pixels

Intermediate Features via backprop

Gradient Ascent : generate a synthetic img. that maximally activates a neuron

<——> Gradient Descent

Fooling Images / Adversarial Examples

(1) Start from an arbitrary img.

(2) Pick an arbitrary class

(3) Modify the image to maximize the class

(4) Repeat until network is fooled

Deep Dream : Amplify existing features

Feature Inversion

Texture Synthesis

Neural Texture Synthesis : Gram Matrix

- Reconstructing texture from higher layers recovers larger features from the input texture

Neural Style Transfer

Content Image + Style Image = Style Transfer