Visualizing and explaining neural networks

dumbbell
cup
dalmatian
Outline

• Basic visualization techniques
• Mapping activations back to the image
• Synthesizing images to maximize activation
• Saliency maps
• Explainability/interpretability
What’s happening inside the black box?

Visualize first-layer weights directly

“cat”
What’s happening inside the black box?

Not too helpful for subsequent layers

Features from a CIFAR10 network, via Stanford CS231n
What’s happening inside the black box?

Visualize maximally activating patches:
pick a unit; run many images through the network; visualize patches that produce the highest output values

“cat”
What’s happening inside the black box?

“cat”

Visualize maximally activating patches
What’s happening inside the black box?

What about FC layers?
Visualize nearest neighbor images according to activation vectors

Source: Stanford CS231n
What’s happening inside the black box?

Source: Andrej Karpathy

What about FC layers?
Fancy dimensionality reduction, e.g., t-SNE
What’s happening inside the black box?

Visualize activations for a particular image

Source
What’s happening inside the black box?

Visualize pixel values responsible for the activation

Visualize activations for a particular image

Source
What’s happening inside the black box?

Visualize activations for a particular image

Synthesize images to maximize activation

Source
Deep visualization toolbox

YouTube video

Overview

How to visualize parts of image responsible for the activation?

How to synthesize images to maximize activation?
Let’s take a single value in an intermediate feature map and propagate its gradient back to the original image pixels.

What does this tell us?
Mapping activations back to pixels

1. Forward an image through the network
2. Choose a feature map and an activation
3. Zero out all values except for the one of interest
4. Propagate that value back to the image
Mapping activations back to pixels

- Commonly used methods differ in how they treat the ReLU.

![Forward pass](image)

Propagating back negative gradients bad for visualization

---

Deconvnet visualization

AlexNet Layer 4

AlexNet Layer 5

M. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, ECCV 2014
Guided backpropagation visualization

Visualization by optimization

• How can we synthesize images that maximize activation of a given neuron?
• Basic approach: find image $x$ maximizing target activation $f(x)$ subject to natural image regularization penalty $R(x)$:

$$x^* = \arg \max_x f(x) - \lambda R(x)$$
Visualization by optimization

- Maximize $f(x) - \lambda R(x)$
  - $f(x)$ is score for a category before softmax
  - $R(x)$ is L2 regularization
  - Perform gradient ascent starting with zero image, add dataset mean to result

Visualization by optimization

• Alternative approach to regularization: at each step of gradient ascent, apply operator $r$ that regularizes the image:

$$x \leftarrow r \left( x + \eta \frac{\partial f}{\partial x} \right)$$

• Combination that gives good-looking results:
  • L2 decay
  • Gaussian blur (every few iterations)
  • Clip pixel values with small magnitude
  • Clip pixel values with small contribution to the activation (estimated by product of pixel value and gradient)

Visualization by optimization

- Example visualizations:

Visualization by optimization

- Example visualizations of intermediate features:

Multifaceted feature visualization

- Key idea: most neurons in high layers respond to a mix of different patterns or “facets”
- For coherent visualizations, zero in on individual facets

A. Nguyen, J. Yosinski, J. Clune, Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks, ICML workshop, 2016
Multifaceted feature visualization

- Key idea: most neurons in high layers respond to a mix of different patterns or “facets”
- For coherent visualizations, zero in on individual facets

Algorithm:
- Cluster FC activations of training images to identify facets
- For each facet, initialize optimization with mean image
- To attempt to produce image of a single object, use center-biased regularization (start with blurry image, gradually increase resolution and update center pixels more than edge pixels)

A. Nguyen, J. Yosinski, J. Clune, Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks, ICML workshop, 2016
Multifaceted feature visualization

A. Nguyen, J. Yosinski, J. Clune, Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks, ICML workshop, 2016
Google DeepDream

- Idea: adjust image to amplify existing activations

https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
Google DeepDream

- Idea: adjust image to amplify existing activations

https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
Google DeepDream

Choose an image and a layer in a CNN; repeat:
1. Forward: compute activations at chosen layer
2. Set gradient of chosen layer equal to its activation
   - Equivalent to maximizing $\sum_i f_i^2(x)$
3. Backward: Compute gradient w.r.t. image
4. Update image (with some tricks)

Source: Stanford CS231n

https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

https://deepdreamgenerator.com/
Saliency maps

• Which parts of the image played the most important role in the network’s decision?

Prediction: “car” 64%

Source: K. Saenko
“White box” saliency via gradients

- Back-propagate gradient of class score (before softmax) to the image, display max of absolute values across color channels

“White box” saliency via gradients

- Can be used for *weakly supervised* segmentation:

“Black box” saliency via masking

• Slide square occluder across image, see how class score changes

M. Zeiler and R. Fergus, *Visualizing and Understanding Convolutional Networks*, ECCV 2014
“Black box” saliency via masking

- Saliency of a class at a pixel is expected score for that class over all masks where the pixel is visible.

V. Petsiuk, A. Das, K. Saenko, RISE: Randomized Input Sampling for Explanation of Black-box Models, BMVC 2018
“Black box” saliency via masking

- Saliency of a class at a pixel is expected score for that class over all masks where the pixel is visible

V. Petsiuk, A. Das, K. Saenko, RISE: Randomized Input Sampling for Explanation of Black-box Models, BMVC 2018
“Black box” saliency via masking

- Application: detecting model/dataset bias

Prediction: “cow” 76%

Source: K. Saenko
“Black box” saliency via masking

- Application: detecting model/dataset bias

Baseline: *A man sitting at a desk with a laptop computer.*

Improved model: *A woman sitting in front of a laptop computer.*

“Black box” saliency via masking

• Application: detecting model/dataset bias

RISE applied to satellite image classification model shows that shadows have great influence on the model

Source: RISE poster
Quantifying interpretability of units

• From the very beginning, people have observed that many units in higher layers seem to fire on semantically meaningful concepts

• But how can we quantify this?

Figure: Zeiler & Fergus
Quantifying interpretability of units

- For a given unit, measure how much areas of high activation overlap semantic segmentations for a large set of visual concepts

Quantifying interpretability of units

- For a given unit, measure how much areas of high activation overlap semantic segmentations for a large set of visual concepts

Quantifying interpretability of units

Histogram of object detectors for Places AlexNet conv5 units
81/256 units with IoU > 0.04

conv5 unit 79        car (object)        IoU=0.13

conv5 unit 107       road (object)      IoU=0.15

Quantifying interpretability of units

Summary

- Basic visualization techniques
  - Showing weights, top activated patches, nearest neighbors
- Mapping activations back to the image
  - Deconvolution
  - Guided back-propagation
- Synthesizing images to maximize activation
  - Gradient ascent with natural image regularization
- Saliency maps
  - “White box” vs. “black box”
- Explainability/interpretability
  - Explaining network decisions, detecting bias
  - Quantifying interpretability of intermediate units