Visualization Based Methods for Interpretability

Apr 29, 2020
Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning
Stanford University
Recap

- Counterfactual Explanations

\[ x' = \text{arg min}_{x'} \lambda (\hat{f}(x') - y')^2 + d(x, x') \]

increase $\lambda$ while $|\hat{f}(x') - y'| > \varepsilon$

Grath et al, 2018
Recap

- Explaining Loan Decisions

Grath et al, 2018
Recap

- Explaining Image Classifications

This bird is a **Crested Auklet** because this is a **black bird** with a **small orange beak** and it is not a **Red Faced Cormorant** because it does not have a **long flat bill**.

This bird is a **Parakeet Auklet** because this is a **black bird** with a **white belly** and **small feet** and it is not a **Horned Grebe** because it does not have **red eyes**.

This bird is a **Least Auklet** because this is a **black and white spotted bird** with a **small beak** and it is not a **Belted Kingfisher** because it does not have a **long pointy bill**.

This bird is a **White Pelican** because this is a **large white bird** with a **long orange beak** and it is not a **Laysan Albatross** because it does not have a **curved bill**.

This bird is a **Cardinal** because this is a **red bird** with a **black face** and it is not a **Scarlet Tanager** because it does not have a **black wings**.

This bird is a **Yellow Headed Blackbird** because this is a **small black bird** with a **yellow breast and head** and it is not a **Prothonotary Warbler** because it does not have a **gray wing**.

Hendricks et al., 2018
Recap

- Grouping Activation Layers Under the Same Concept (TCAV)

Kim et al, 2018
Recap

Lecture 3 Intrinsic Methods for Interpretability

Post Hoc Methods for Interpretability

Lecture 4
Proxy Methods

Lecture 5
Feature Interaction

Lecture 6
Example Based Methods

Lecture 7
Visualization Based Methods*

*Some techniques are available only for deep neural networks
Outline

- Visualization Based Methods
- Activation Visualization
  - Saliency Maps
  - GoogLeNet Activation Atlas
  - Interpretability via Activation Visualization
- Gradient Based Feature Attribution
  - Integrated Gradient
  - Baselines for Integrated Gradient
Visualization Based Methods

Activation Visualization

- Visualize activations in neural networks

Feature Attribution

- Explain decisions on the importance of specific inputs
  - e.g., LIME, SHAP, DeepLift, LRP
Outline

- Visualization Based Methods
- Activation Visualization
  - Saliency Maps
  - GoogLeNet Activation Atlas
  - Interpretability via Activation Visualization
- Gradient Based Feature Attribution
  - Integrated Gradient
  - Baselines for Integrated Gradient
Activation Visualization

Visualizations of hidden layers of GoogLeNet

Olah et al, 2017
Reasons for Activation Visualization

\[ a_{5,1} = [9.76, 0, 45.6, 33.1, 14.4, 0, 119.9, 84.9, 151.3, 5... \]
Class Specific Saliency Maps

- Generate a representative image for class $c$
  - Fix trained networks
  - Generate an input image that maximizes model probabilities

$$\arg \max_{I} S_c(I) - \lambda \|I\|^2_2$$

Logits of a trained model on class $c$

Deep Model

- class $c$

- class saliency map

Simonyan et al, 2014
Class Specific Saliency Maps

- ConvNet Trained on ILSVRC2013

Simonyan et al, 2014
Activation Visualization With a Prior

- Generating Saliency Maps With Prior Knowledge

$$\hat{y}^l = \arg \max_{y^l} (\Phi_h(G_l(y^l)) - \lambda \|y^l\|)$$

Nguyen et al, 2016
Activation Visualization With a Prior

- Using a Generator as a Prior

\[
\hat{y}^l = \arg \max_{y^l} (\Phi_h(G_l(y^l)) - \lambda\|y^l\|)
\]
Class Specific Activation Visualization With a Prior

AlexNet DNN trained on the MIT Places dataset

Nguyen et al, 2016
GoogLeNet

Szegedy et al, 2014
Deep Dream - Class Specific

Mordvintsev et al, 2015
Deep Dream - Class Specific

Mordvintsev et al, 2015
Deep Dream - Class Specific

Øygard, 2015
Visuzling Components of Neural Networks

- Neuron: $\text{layer}_n[x,y,z]$
- Channel: $\text{layer}_n[:,:,:]$  
- Layer/DeepDream: $\text{layer}_n[:,:,:]^2$
- Class Logits: $\text{pre}_\text{softmax}[k]$
- Class Probability: $\text{softmax}[k]$

Olah et al. 2017
Visualizations of Neuron Activations

Baseball—or stripes? 
mixed4a, Unit 6

Animal faces—or snouts? 
mixed4a, Unit 240

Clouds—or fluffiness? 
mixed4a, Unit 453

Buildings—or sky? 
mixed4a, Unit 492
Outline

● Visualization Based Methods

● Activation Visualization
  ○ Saliency Maps
  ○ GoogLeNet Activation Atlas
  ○ Interpretability via Activation Visualization

● Gradient Based Feature Attribution
  ○ Integrated Gradient
  ○ Baselines for Integrated Gradient
Model Explanation via Activation Visualizations

\[ a_{2,6} = [59.1, 95.6, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...] \]

Olah et al 2018
Visualizing Activations by Individual Neurons

Olah et al 2018
GoogLeNet Activation Atlas

https://distill.pub/2019/activation-atlas/

Carter et al, 2019
Visualizing Activations by Channel Activations

Olah et al 2018
Visualizing Activations by Neuron Groups
Visualizing Model Decisions

Olah et al 2018
Hands-On Session

● Can you explain the how GoogLeNet make decisions?
  ○ What Does the Network See?
  ○ How Are Concepts Assembled?
  ○ Making Things Human-Scale

https://distill.pub/2018/building-blocks/
Outline

- Visualization Based Methods
- Activation Visualization
  - Saliency Maps
  - GoogLeNet Activation Atlas
  - Interpretability via Activation Visualization
- Gradient Based Feature Attribution
  - Integrated Gradient
  - Baselines for Integrated Gradient
Feature Attribution

- Feature Attribution explains models by highlighting features
  - LIME, SHAP, LRP, DeepLift are feature attribution methods
- Gradient Based Feature Attribution
  - LRP, DeepLift
Implementation Invariance

- **Definition**
  - Attribution is Implementation Invariant if it is always identical for two functionally equivalent networks

[Images of LRP and DeepLift with '✗' symbols]

*Suvararajan et al 2017*
Integrated Gradients

- Feature Importance determined by the integral of gradients
  - $x$ - input
  - $x'$ - reference input
  - $F$ - black-box model

- Satisfies Implementation Invariance

\[
\text{IntegratedGrads}_i(x) \::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha
\]

\[
= (x_i - x'_i) \times \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i} \times \frac{1}{m}
\]

Riemann Approximation

Sundararajan et al 2017
Examples of Integrated Gradients

<table>
<thead>
<tr>
<th>Original image</th>
<th>Top label and score</th>
<th>Integrated gradients</th>
</tr>
</thead>
</table>
| ![Original image](image1.png) | **Top label: reflex camera**  
Score: 0.993755 | ![Integrated gradients](image2.png) |
| ![Original image](image3.png) | **Top label: fireboat**  
Score: 0.999961 | |
| ![Original image](image4.png) | **Top label: school bus**  
Score: 0.997033 | |
Examples of Integrated Gradients
Outline

● Visualization Based Methods
● Activation Visualization
  ○ Saliency Maps
  ○ GoogLeNet Activation Atlas
  ○ Interpretability via Activation Visualization
● Gradient Based Feature Attribution
  ○ Integrated Gradient
  ○ Baselines for Integrated Gradient
Choosing A Baseline Input $x'$

\[
\text{IntegratedGrads}_i(x) := (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha
\]

Sturmfels et al, 2020
Choosing A Baseline Input $x'$

\[ \text{IntegratedGrads}_i(x) := (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha \]

Sturmfels et al, 2020
Non-Constant Baselines

- Details will be eliminated where image and baseline has the same color
Expected Gradients

- Details will be eliminated where image and baseline has the same color
- Expected Gradients
  - Choosing Multiple Baselines $x^{i,j}$

\[
\hat{\phi}_i^{EG}(f, x; D) = \frac{1}{k} \sum_{j=1}^{k} (x_i - x^{i,j}_i) \times \frac{\delta f(x^{i,j} + \alpha^j(x - x^{i,j}))}{\delta x_i}
\]

Sturmfels et al, 2020
Smooth Gradients

$$\phi^S_i(f, x; N(\overline{0}, \sigma^2 I)) = \frac{1}{k} \sum_{j=1}^{k} (x + \epsilon^j) \times \frac{\delta f(x + \epsilon^j)}{\delta x_i}$$

Expected Gradients with Gaussian Noise

$$\hat{\phi}^E_i(f, x; D) = \frac{1}{k} \sum_{j=1}^{k} (x_i - x^*_{ij}) \times \frac{\delta f(x^*_{ij} + \alpha^j(x - x^*_{ij}))}{\delta x_i}$$

$$= \frac{1}{k} \sum_{j=1}^{k} \epsilon^j \times \frac{\delta f(x + (1 - \alpha^j)\epsilon^j)}{\delta x_i}$$

Daniel et al., 2017
Hands-On Session

- What do you think is the best baseline for each of the images?
  - Alternative Baseline Choices
  - The Gaussian Baseline
  - Expectations, and Connections to SmoothGrad
  - Using the Training Distribution (Optional)

[https://distill.pub/2020/attribution-baselines/](https://distill.pub/2020/attribution-baselines/)
Summary

- Activation Visualization
  - Visualize the hidden layers of neural networks
  - Generated using Saliency Maps
  - Powerful tools to explain ML models

- Gradient Based Feature Attribution
  - Integrated Gradients
Reading Assignments

- Kindermans, Pieter-Jan, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: PatternNet and PatternAttribution, ICLR 2018
- Zeiler, Matthew D., and Rob Fergus. Visualizing and understanding convolutional networks, ECCV 2014
- Dabkowski, Piotr, and Yarin Gal. Real time image saliency for black box classifiers, NeurIPS 2017
- Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, NeurIPS 2018
Next Lecture

Fairness Through Data/Prediction Manipulations