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

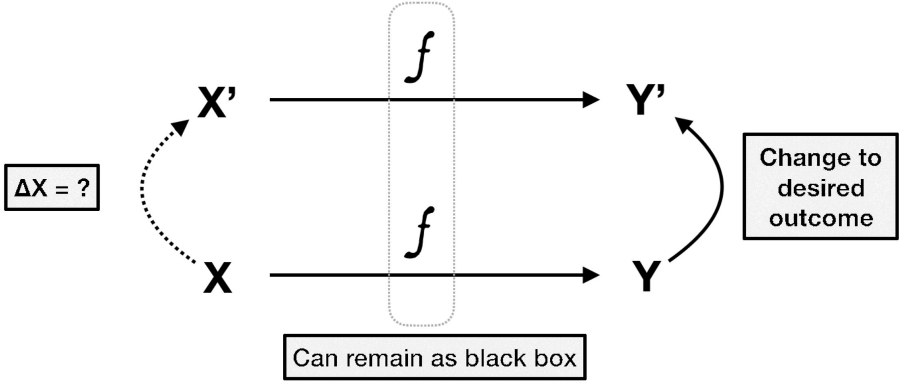
Recap

- Counterfactual Explanations

$$x' = arg\min_{x'} \lambda(\hat{f}(x') - y')^2 + d(x, x')$$

counterfactual example — desired outcome — distance function

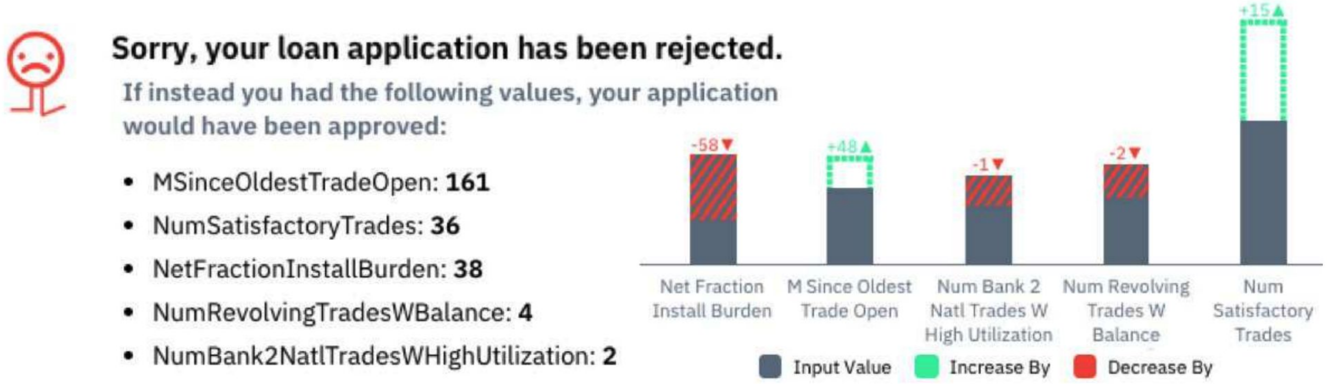
increase λ while $|\hat{f}(x') - y'| > \varepsilon$



Grath et al, 2018

Recap

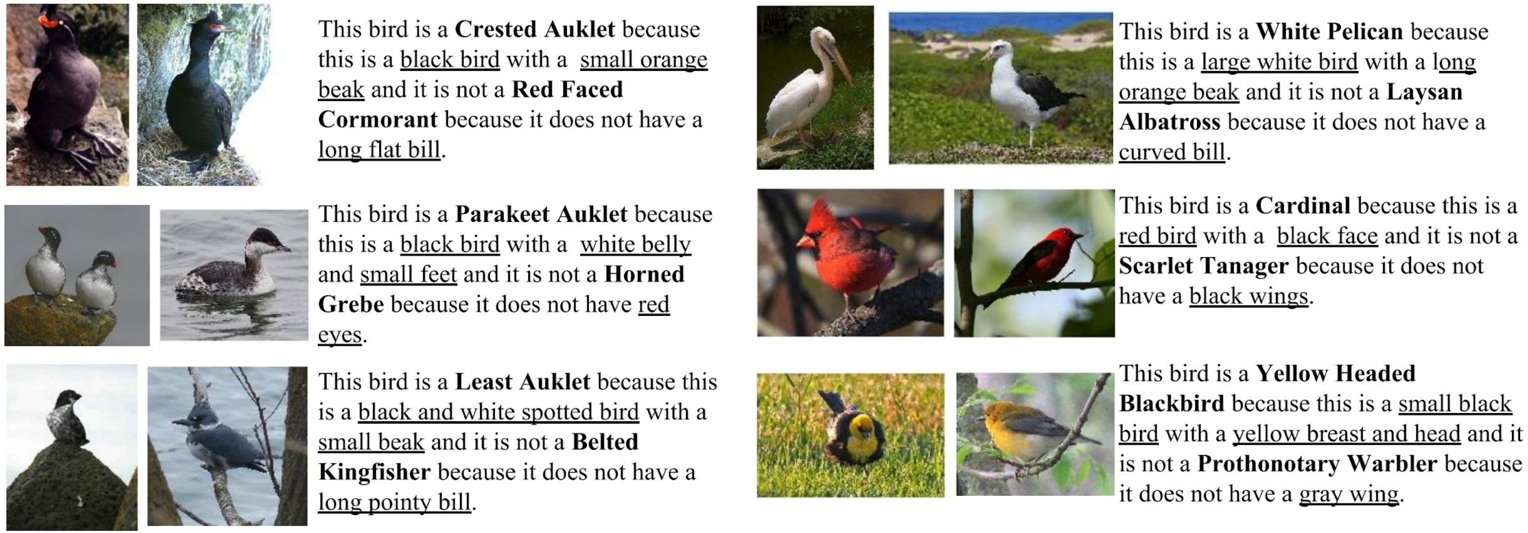
- Explaining Loan Decisions



[Grath et al, 2018]

Recap

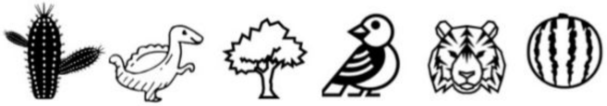
- Explaining Image Classifications



Hendricks et al , 2018

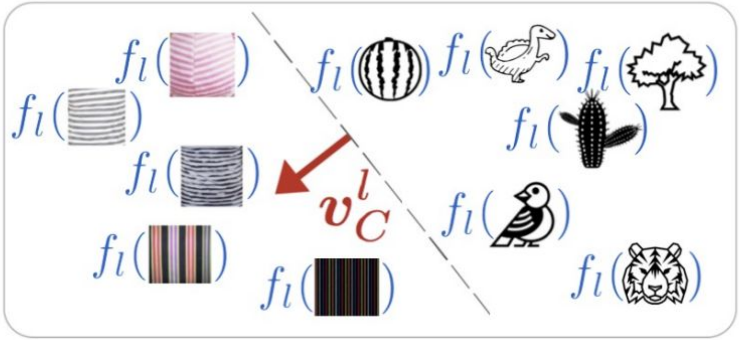
Recap

- Grouping Activation Layers Under the Same Concept (TCAV)



random samples

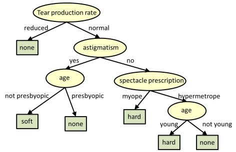
samples that represent a concept



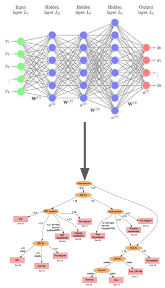
Kim et al, 2018

Recap

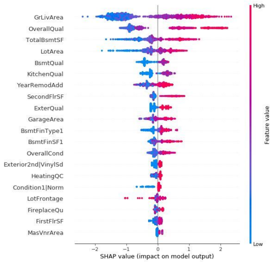
Lecture 3 Intrinsic Methods for Interpretability



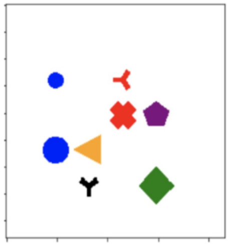
Post Hoc Methods for Interpretability



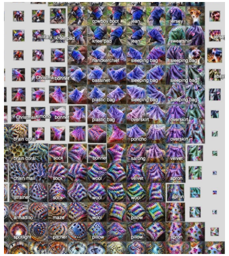
Lecture 4
Proxy Methods



Lecture 5
Feature Interaction



Lecture 6
Example Based
Methods



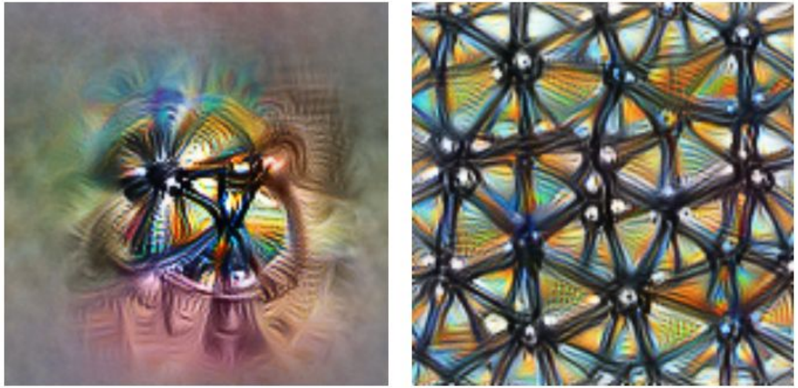
Lecture 7
Visualization Based
Methods*

*Some techniques are available only for deep neural networks

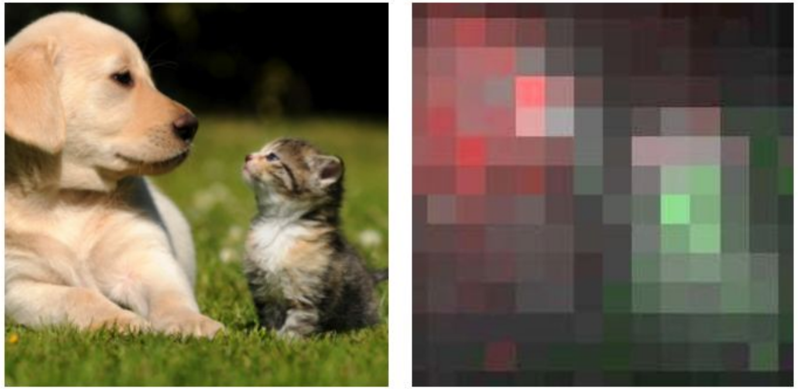
Outline

- 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



Feature Attribution

Visualize activations in neural networks

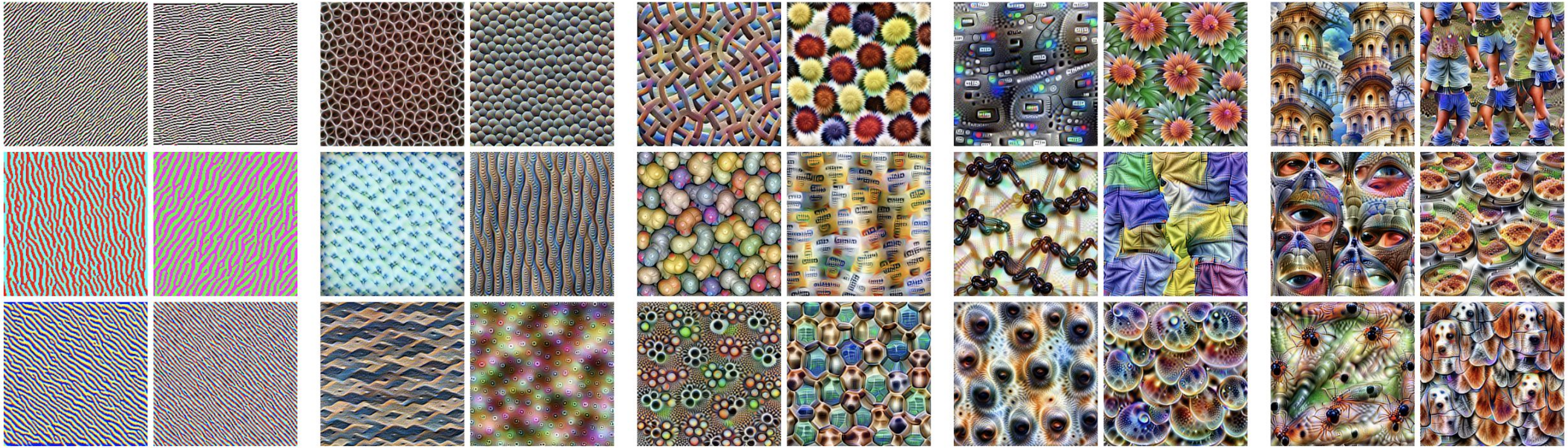
Explain decisions on the importance of specific inputs

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

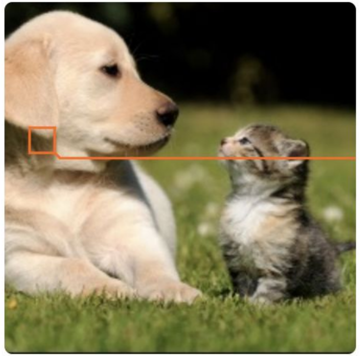


Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)

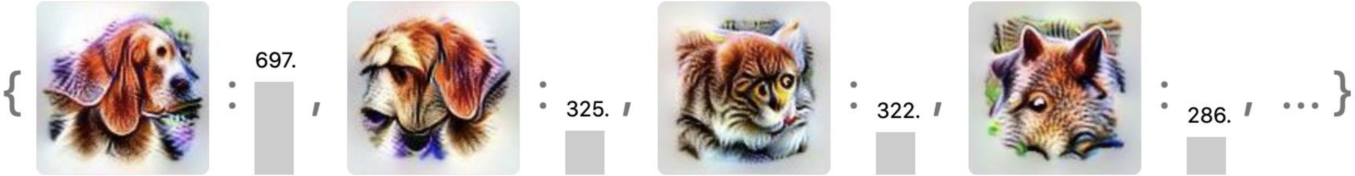
Visualizations of hidden layers of GoogLeNet

[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...

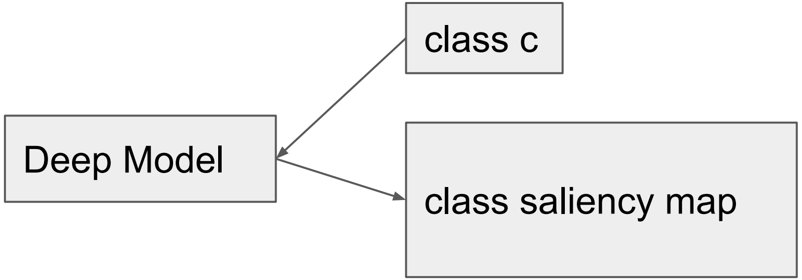


Class Specific Saliency Maps

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

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

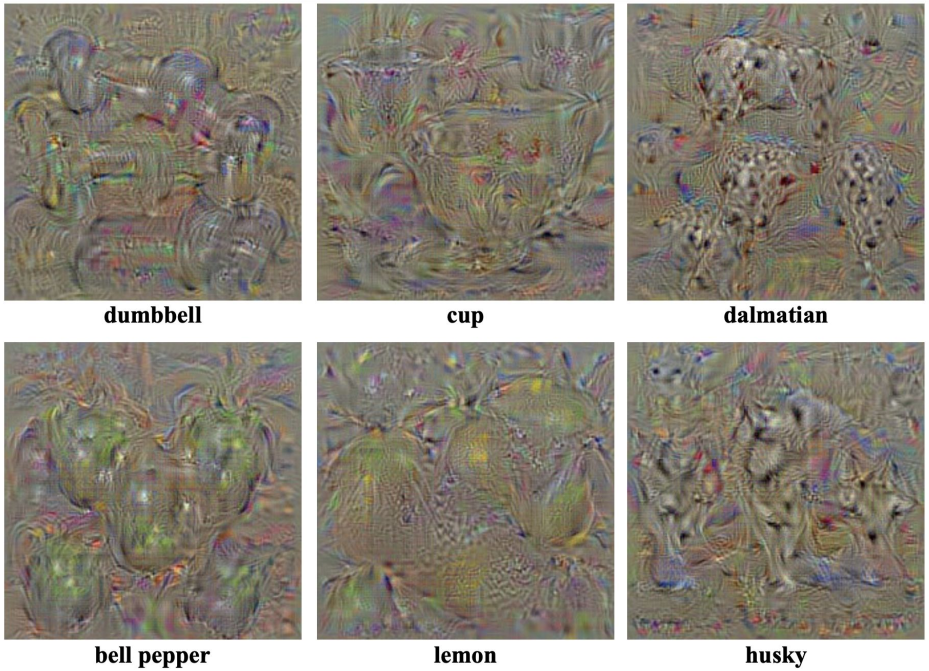
Logits of a trained model on class c



Simonyan et al, 2014

Class Specific Saliency Maps

- ConvNet Trained on ILSVRC2013



Simonyan et al, 2014

Activation Visualization With a Prior

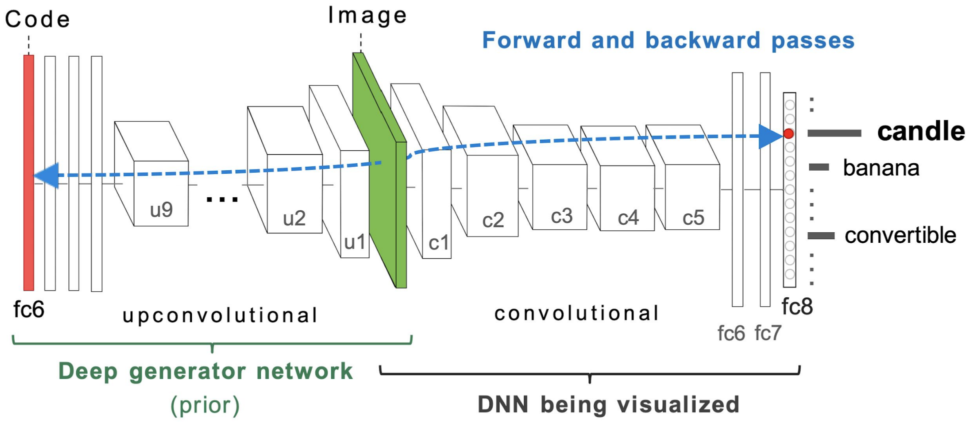
- Generating Saliency Maps With Prior Knowledge

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

Latent Code

Deep Model

Pretrained Generator

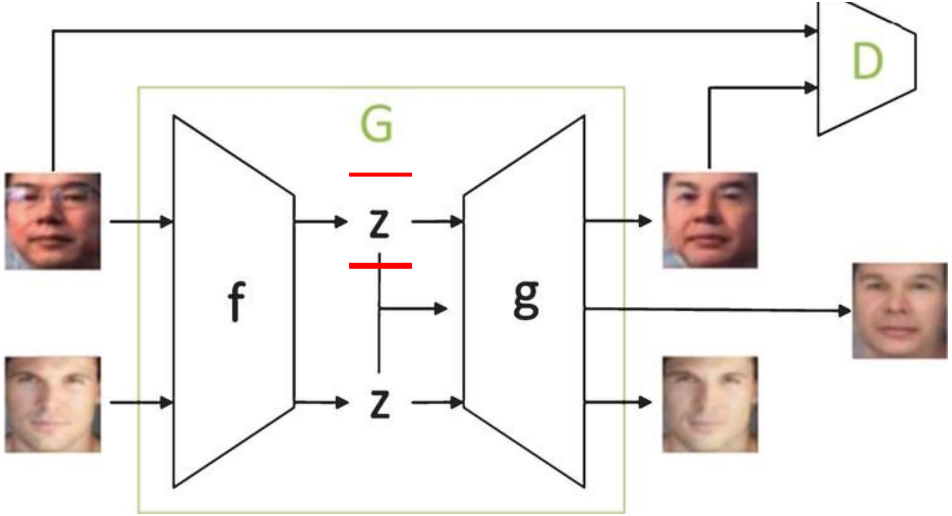


[Nguyen et al, 2016]

Activation Visualization With a Prior

- Using a Generator as a Prior

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

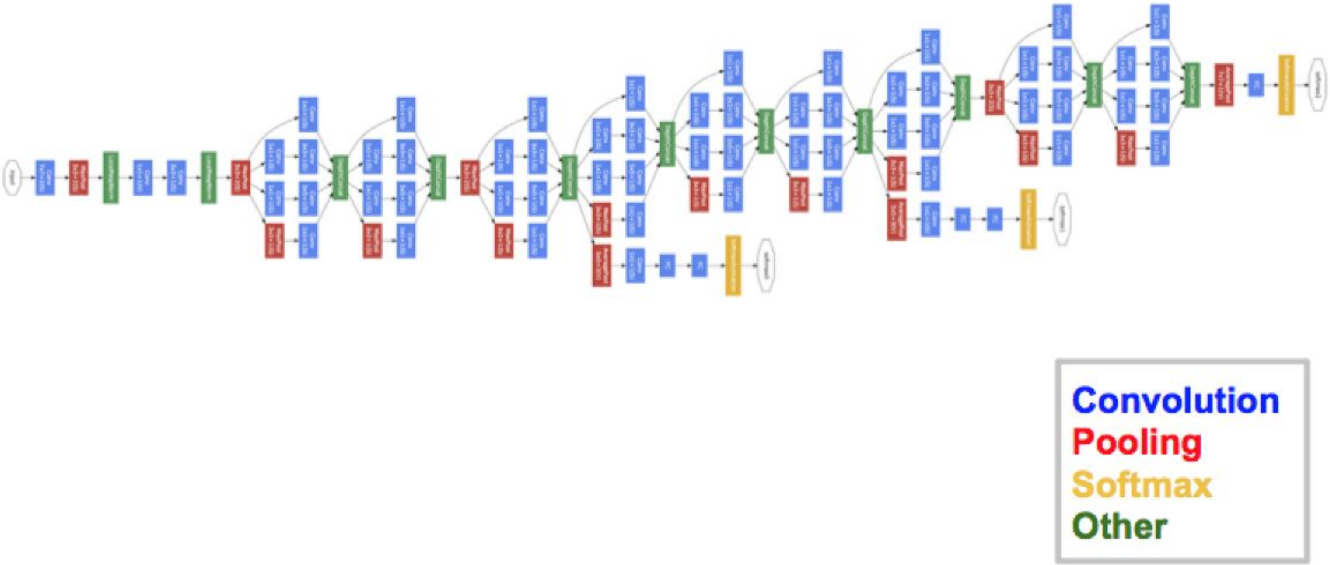


Class Specific Activation Visualization With a Prior

AlexNet DNN trained on the MIT Places dataset

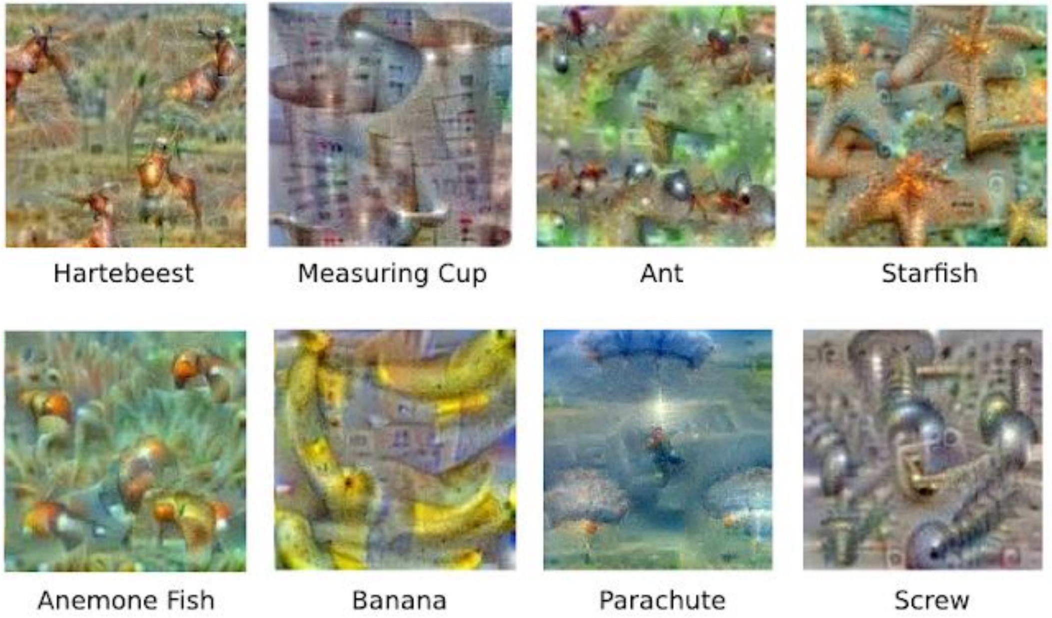
Nguyen et al, 2016

GoogLeNet



Szegedy et al, 2014

Deep Dream - Class Specific



Mordvintsev et al, 2015

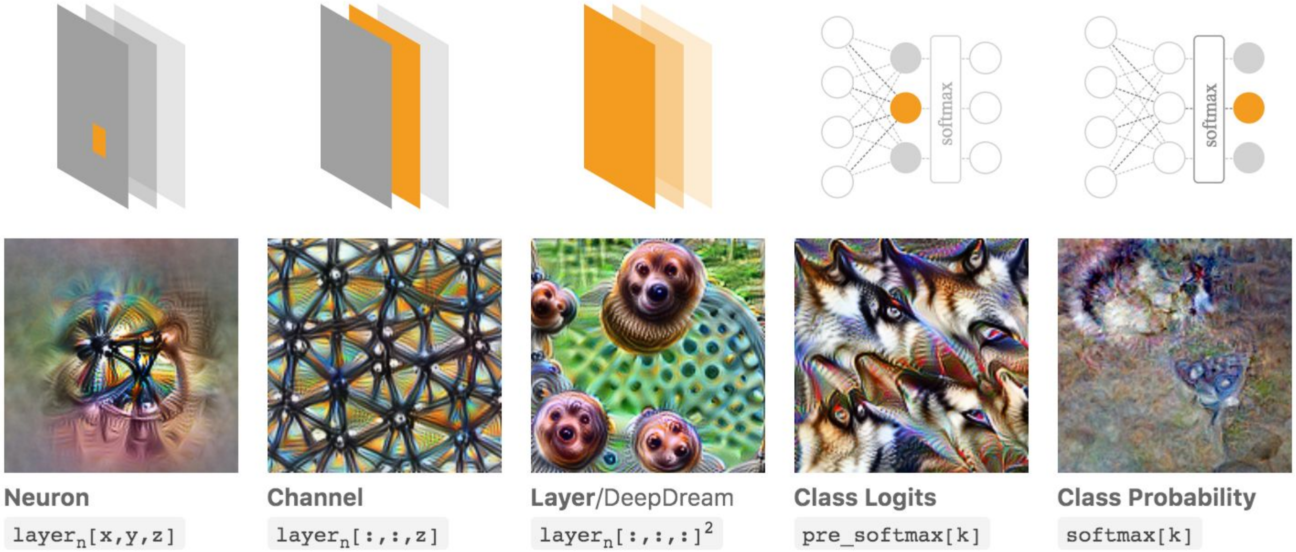
Deep Dream - Class Specific

Mordvintsev et al, 2015

Deep Dream - Class Specific

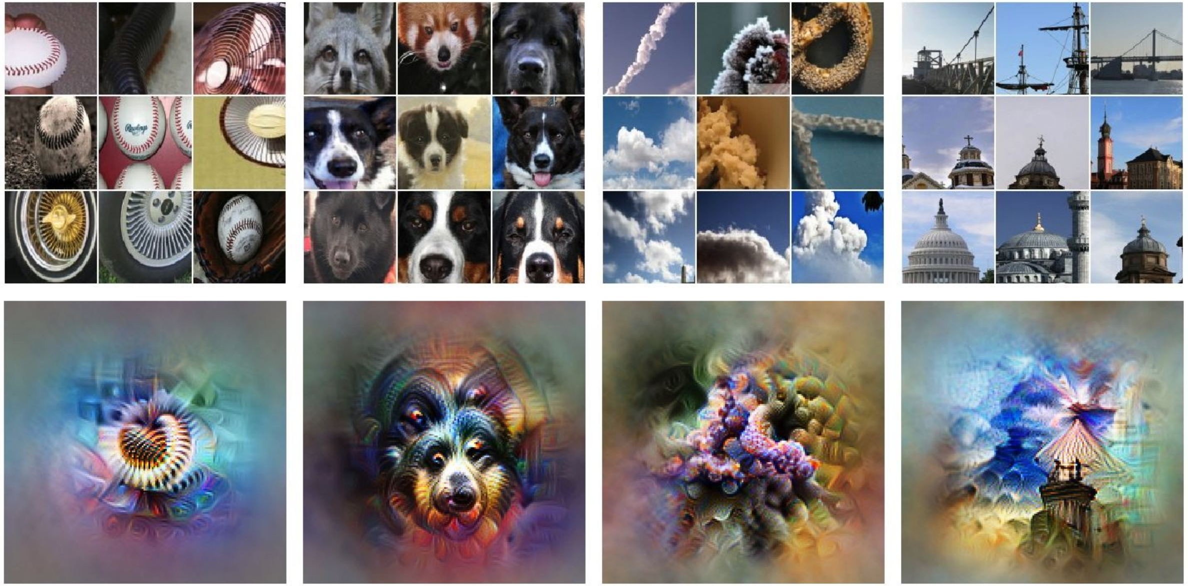
[Øygard, 2015]

Visuzling Components of Neural Networks



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

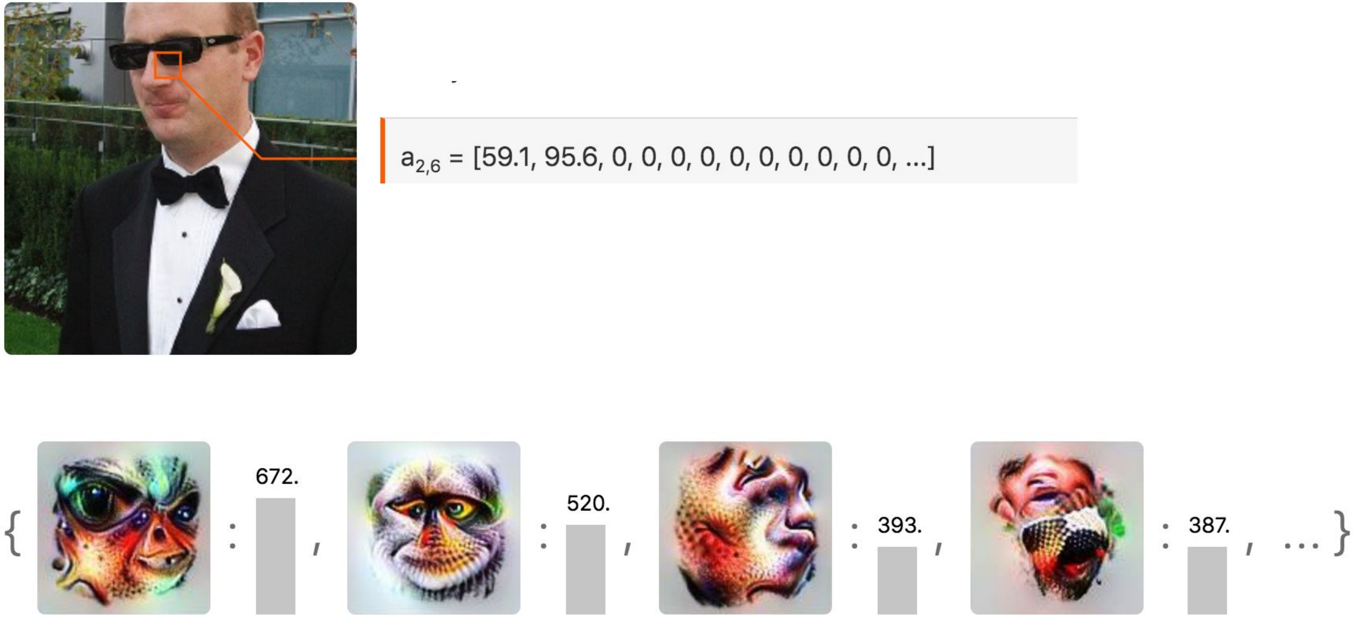
Buildings—or sky?
mixed4a, Unit 492

[Olah et al, 2017]

Outline

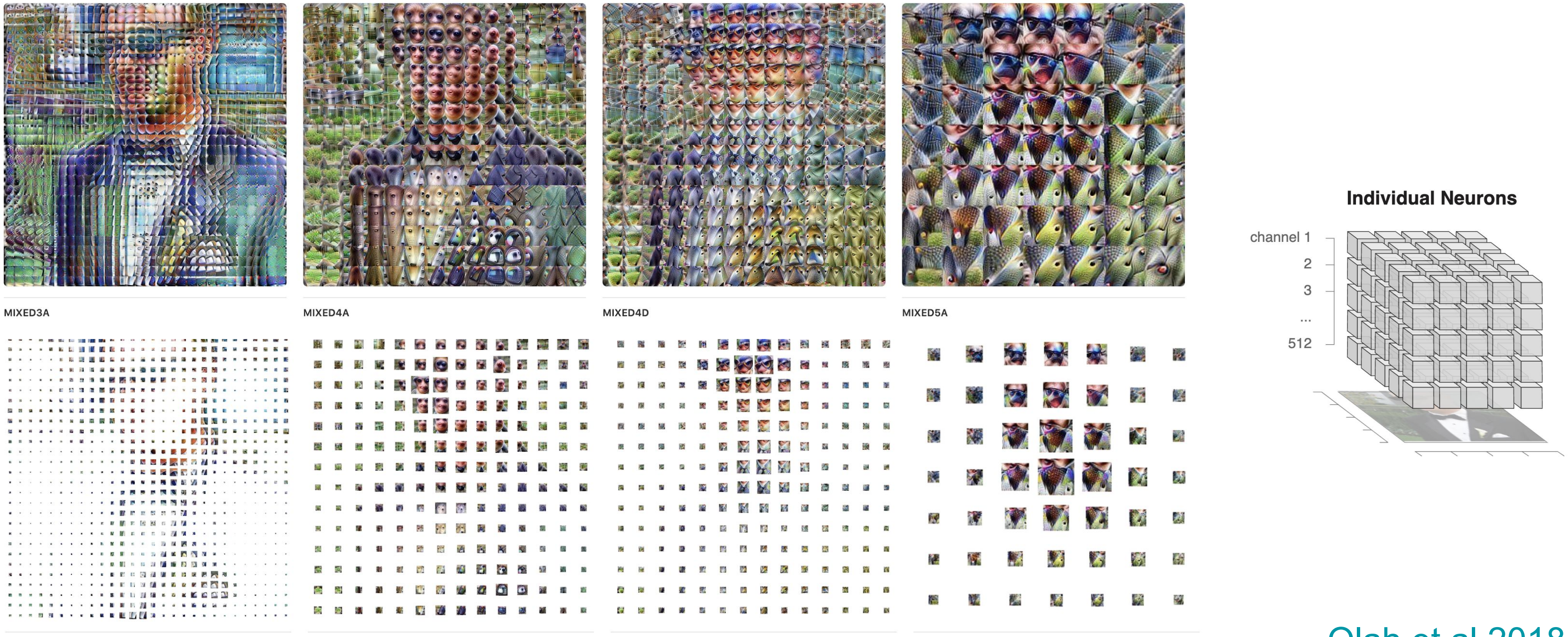
- Visualization Based Methods
- Activation Visualization
  - Saliency Maps
  - GoogLeNet Activation Atlas
  - Interpretability via Activation Visualization
- Gradient Based Feature Attribution
  - Integrated Gradient
  - Baselines for Integrated Gradient

Model Explanation via Activation Visualizations



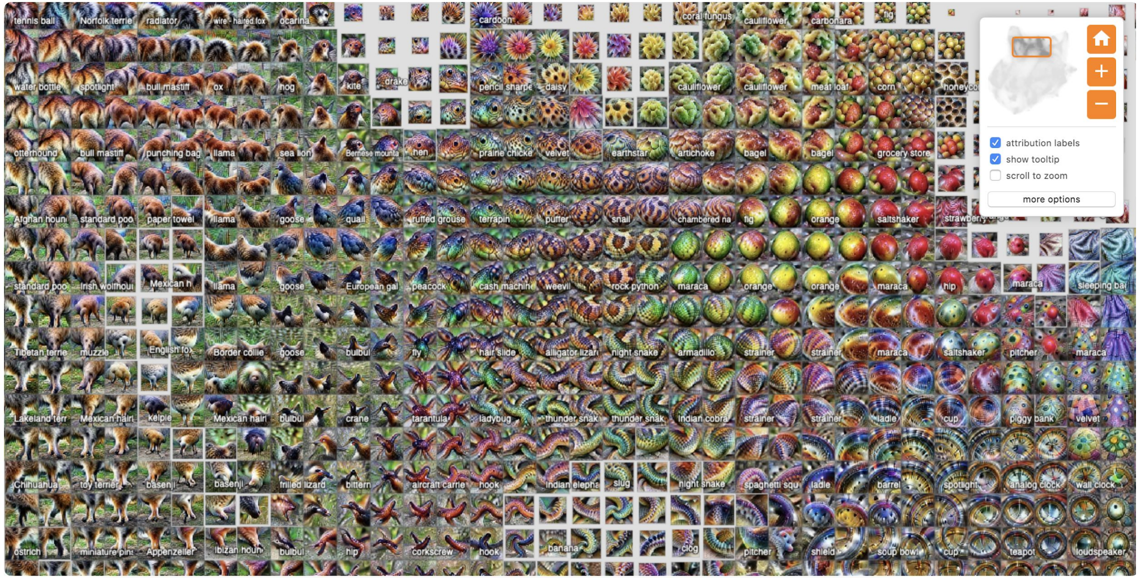
[Olah et al 2018]

Visualizing Activations by Individual Neurons



Olah et al 2018

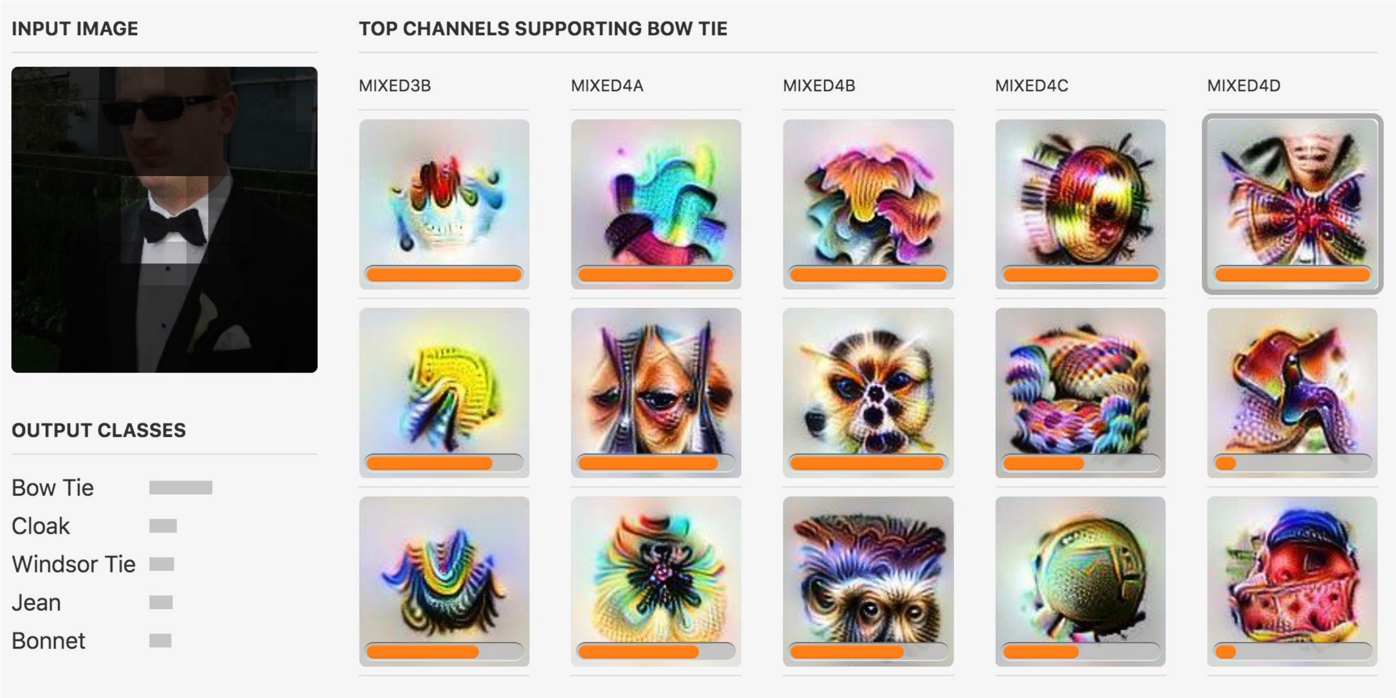
GoogLeNet Activation Atlas

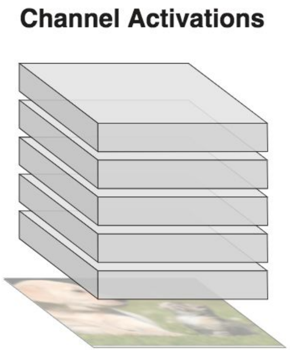


https://distill.pub/2019/activation-atlas/

[Carter et al, 2019]()

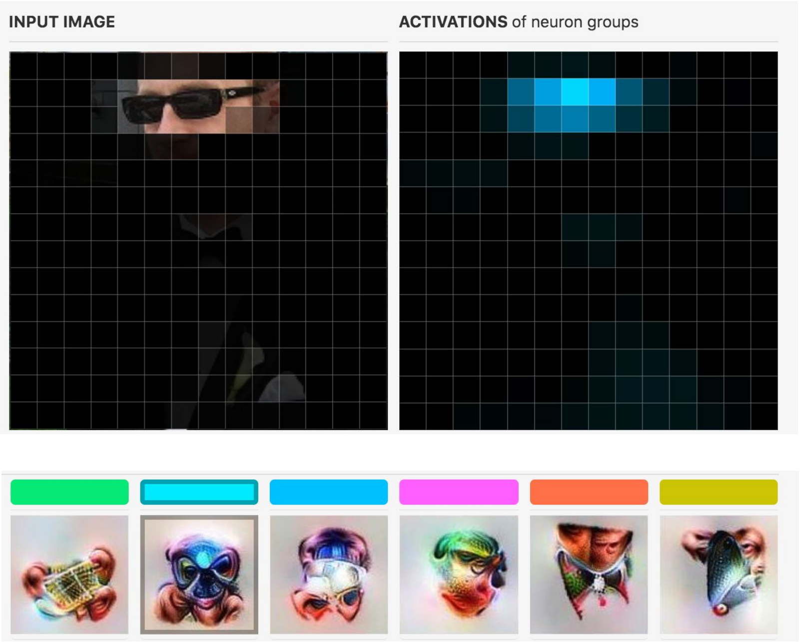
Visualizing Activations by Channel Activations

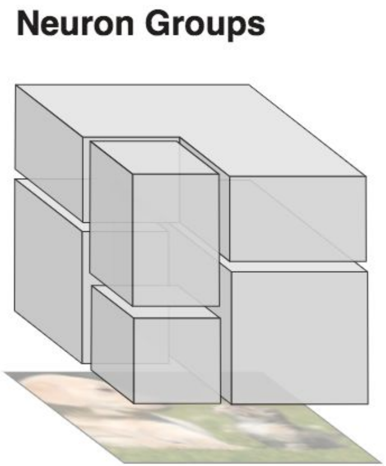




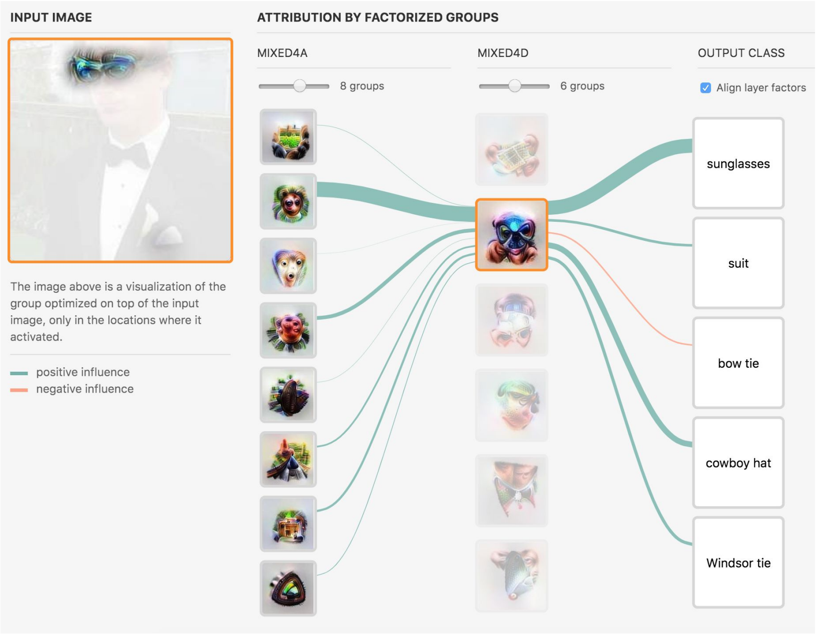
[Olah et al 2018](#)

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

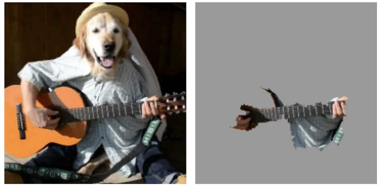
https://distill.pub/2018/building-blocks/

Outline

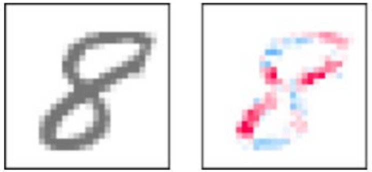
- Visualization Based Methods
- Activation Visualization
    - Saliency Maps
    - GoogLeNet Activation Atlas
    - Interpretability via Activation Visualization
- Gradient Based Feature Attribution
    - Integrated Gradient
    - Baselines for Integrated Gradient

Feature Attribution

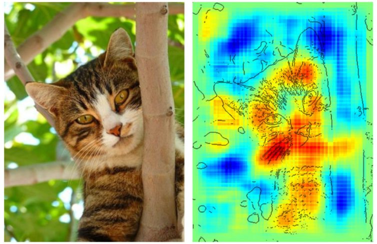
- Feature Attribution explains models by highlighting features
  - LIME, SHAP, LRP, DeepLift are feature attribution methods
- Gradient Based Feature Attribution
  - LRP, DeepLift



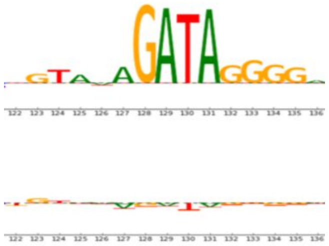
LIME



SHAP



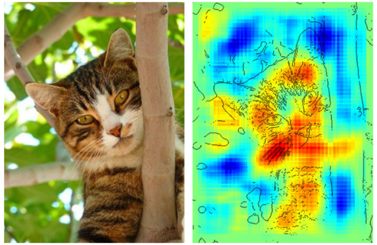
LRP



DeepLift

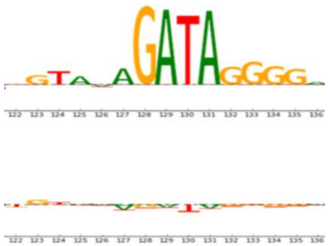
Implementation Invariance

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



LRP

✗



DeepLift

✗

Sundararajan 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

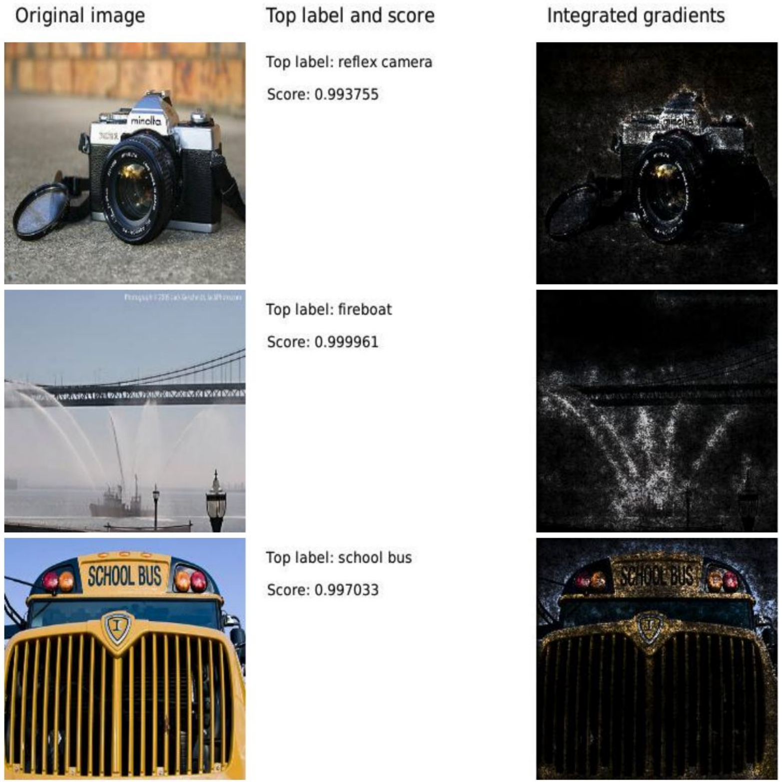
$$\mathsf{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i}\, d\alpha$$

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

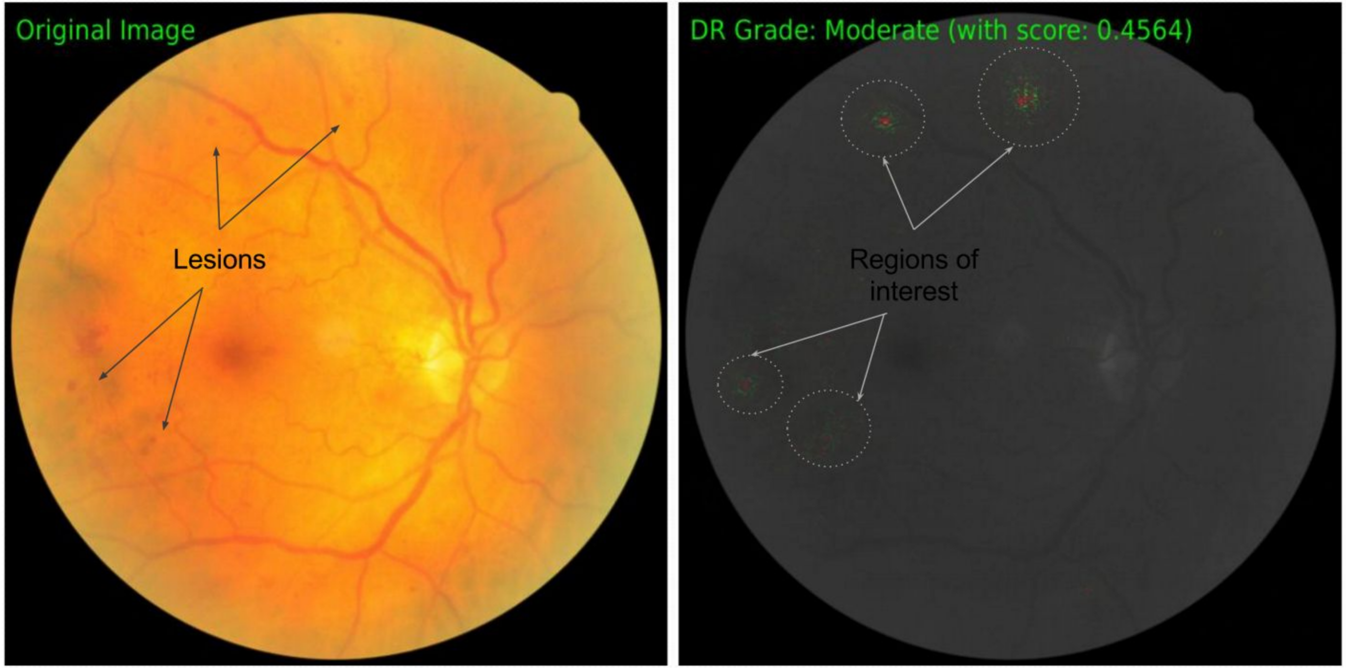
Riemman Approximation

Sundararajan et al 2017

Examples of Integrated Gradients



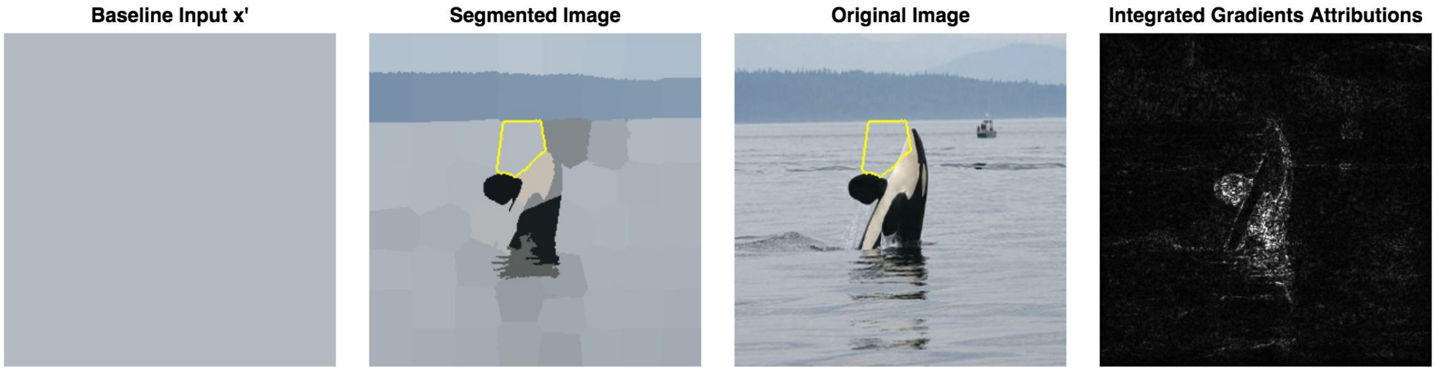
Examples of Integrated Gradients



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

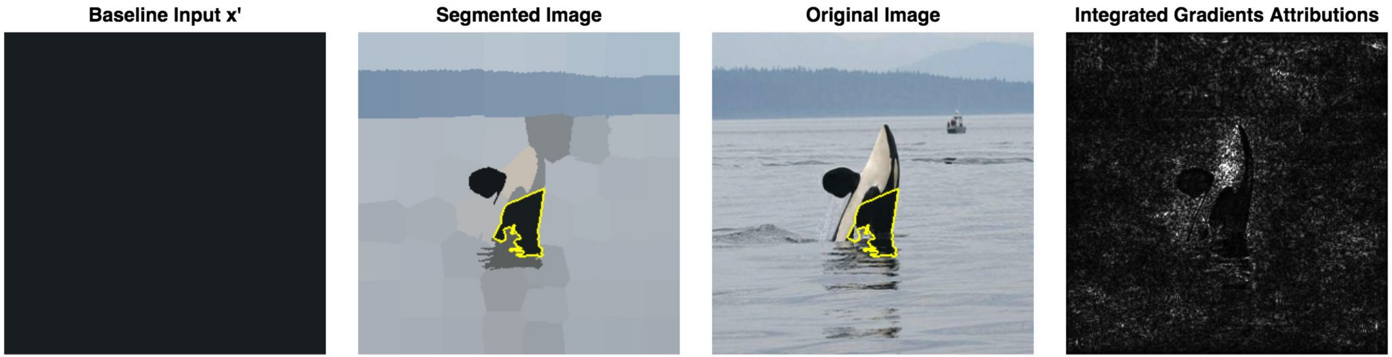
Choosing A Baseline Input x'



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

Sturmfels et al, 2020

Choosing A Baseline Input x'

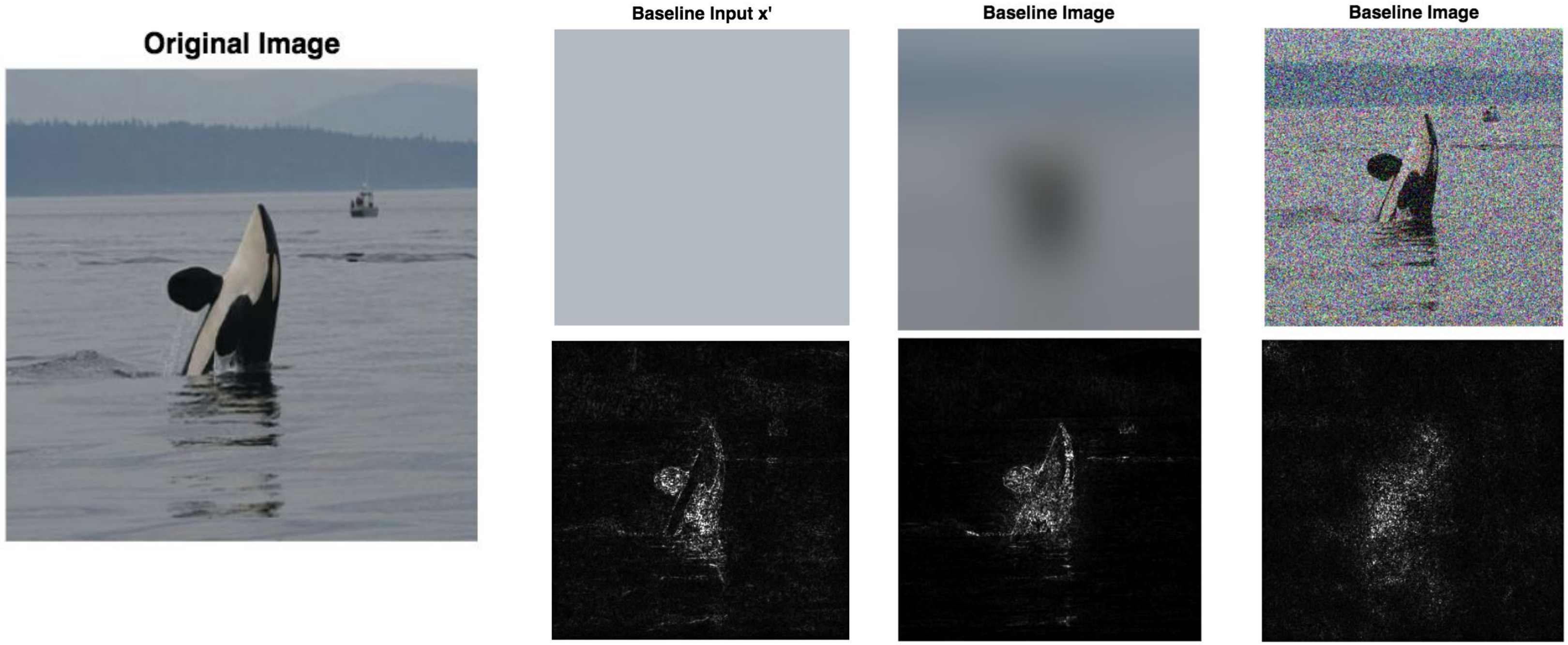


$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

Sturmfels et al, 2020

Non-Constant Baselines

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



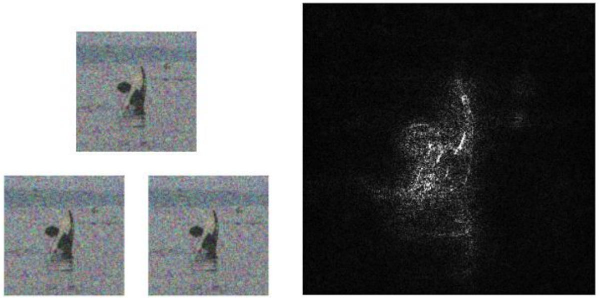
Constant Blur Noise

Expected Gradients

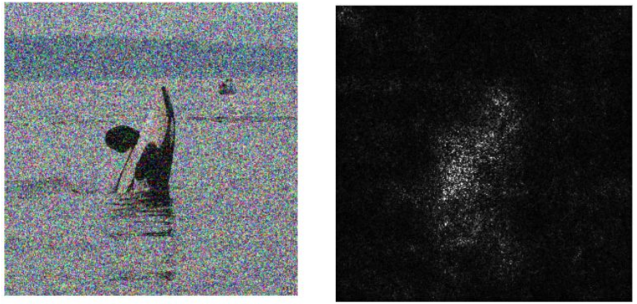
- Details will be eliminated where image and baseline has the same color
- Expected Gradients
  - Choosing Multiple Baselines x'^{j}

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

Baseline Images



Baseline Image



[Sturmfels et al, 2020]

Smooth Gradients

Smooth Gradients

$$\phi_i^{SG}(f, x; N(\bar{0}, \sigma^2 I)) = \frac{1}{k}\sum_{j=1}^{k}(x + \epsilon_\sigma^j) \times \frac{\delta f(x + \epsilon_\sigma^j)}{\delta x_i}$$

Expected Gradients
with Gaussian Noise

$$\begin{aligned}\hat{\phi}_i^{EG}(f, x; D) &= \frac{1}{k}\sum_{j=1}^{k}(x_i - x'^j_i) \times \frac{\delta f(x'^j + \alpha^j(x - x'^j))}{\delta x_i} \\ &= \frac{1}{k}\sum_{j=1}^{k}\epsilon_\sigma^j \times \frac{\delta f(x + (1 - \alpha^j)\epsilon_\sigma^j)}{\delta x_i}\end{aligned}$$

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/

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
- Melis, David Alvarez, and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks, NeurIPS 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, NuerIPS 2018

Next Lecture

Fairness Through Data/Prediction Manipulations