Visual Explainability in Machine Learning

Lecture 3: Visual Explanations I

Ghassan AlRegib, PhD
Professor

Mohit Prabhushankar, PhD
Postdoctoral Fellow

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
School of Electrical and Computer Engineering
Georgia Institute of Technology
{alregib, mohit.p}@gatech.edu

Dec 5, 2023

Short Course Materials

Accessible Online

SCAN ME

https://alregib.ece.gatech.edu/sps-education-short-course/

{alregib, mohit.p}@gatech.edu

IEEE Signal Processing Society™

Title: Visual Explainability in Machine Learning

Presented by: ***Ghassan AlRegib, and Mohit Prabhushankar***

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)

School of Electrical and Computer Engineering

Georgia Institute of Technology, Atlanta, USA

https://alregib.ece.gatech.edu/

Short Course

Course Outline

Day 1: Define and Detail; Day 2: Evaluate; Day 3: Reassess

Day 1	Lecture 1: Why Explainability?	Lecture 2: What is Explainability?	Lectures 3 and 4: Popular Methods of Explainability
Day 2	Lecture 5: Evaluating Explainability	Lecture 6: Robustness: A proxy for Explainability	
Day 3	Lectures 7 and 8: Reassessing Explainability via Uncertainty and Concept vectors	Lectures 9 and 10: Causality and Generative AI	

Outline

Lecture 3: Visual Explanations I

- Human-centric Explanations
- Indirect Explanations
  - Visualizing filters
  - Visualizing activations
  - Visualizing Last layer Embedding
- Direct Explanations
  - Intervention-based visualizations
    - Saliency Maps
  - Gradient-based visualizations
    - Vanilla Backpropagation
    - Deconvolution Backpropagation
    - Guided Backpropagation
- Takeaways

Outline

Lecture 3: Visual Explanations I

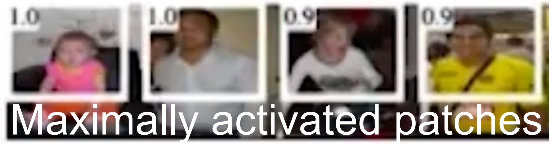
- Human-centric Explanations
- Indirect Explanations
  - Visualizing filters
  - Visualizing activations
  - Visualizing Last layer Embedding
- Direct Explanations
  - Intervention-based visualizations
    - Saliency Maps
  - Gradient-based visualizations
    - Vanilla Backpropagation
    - Deconvolution Backpropagation
    - Guided Backpropagation
- Takeaways

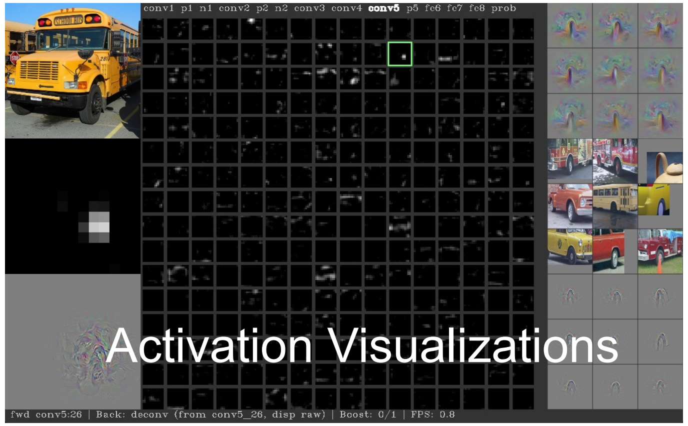
Explanations

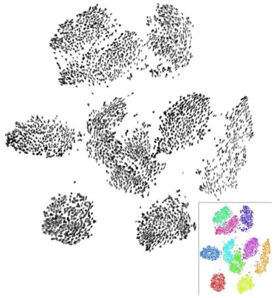
Human-centric Explanations

Explanations can be characterized based on the knowledge of the audience they cater to

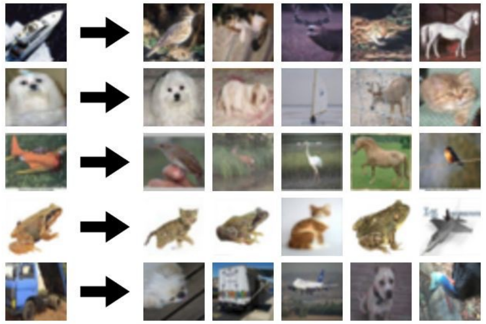
Lecture 3: Indirect and Direct Explanations





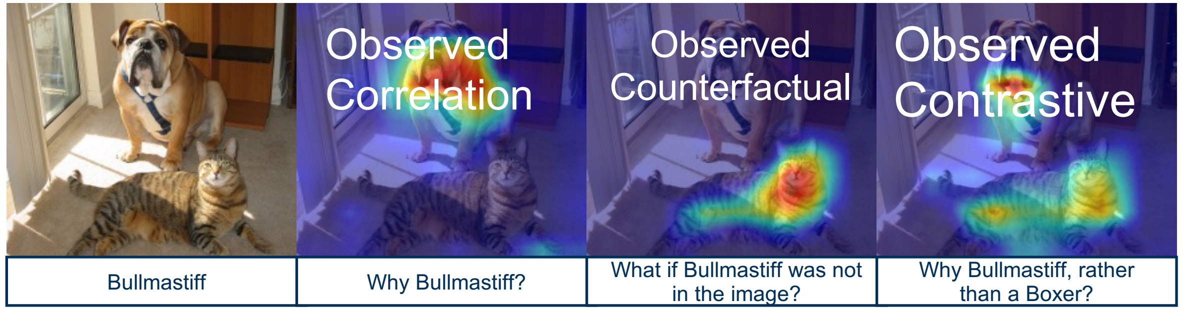


Dimensionality Reduction



Nearest Neighbor

Lecture 4: Targeted Explanations



AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Explanations

Indirect Explanations

Indirect explanations visually analyze network parameters and features and indirectly explain the output

- **Required knowledge to understand explanations:** Models, model parameters, and training data
- **Required Knowledge to obtain explanations:** Models, model parameters, and training data
- **Explanations audience:** Researchers and Engineers building the models
- **Explanatory chronology:** Initially, all Explainability techniques were indirect

	Definition		
Methods	Indirect	Direct	Targeted
Deconvolution [21]	✓	–	–
Inverted Representations [22]	✓	–	–
Guided-Backpropagation [18]	–	✓	–
SmoothGrad [17]	–	✓	–
LIME [39]	–	✓	–
CAM [24]	–	✓	–
Graph-CNN [23]	✓	–	–
GradCAM [12]	–	–	✓
TCAV [40]	–	✓	–
GradCAM++ [16]	–	–	✓
RISE [35]	–	✓	–
Causal-CAM [15]	–	–	✓
Counterfactual-CAM [12]	–	–	✓
Goyal et al. [26]	–	–	✓
CEM [29]	–	–	✓
Contrast-CAM [13]	–	–	✓
Contrastive reasoning [14]	–	–	✓

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Outline

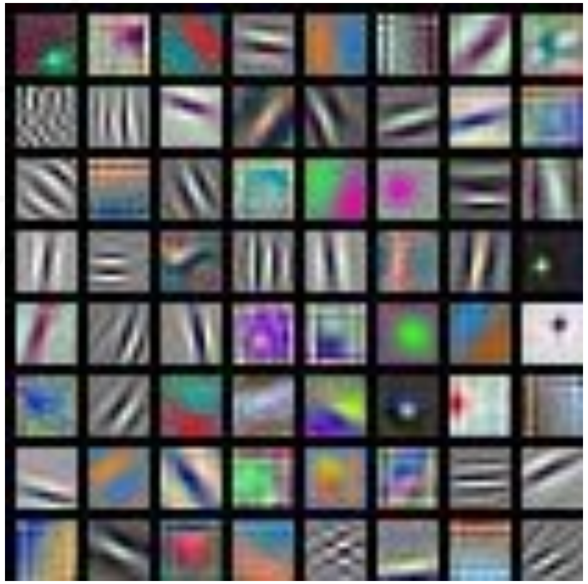
Lecture 3: Visual Explanations I

- Human-centric Explanations
- Indirect Explanations
  - Visualizing filters
  - Visualizing activations
  - Visualizing Last layer Embedding
- Direct Explanations
  - Intervention-based visualizations
    - Saliency Maps
  - Gradient-based visualizations
    - Vanilla Backpropagation
    - Deconvolution Backpropagation
    - Guided Backpropagation

Indirect Explanations

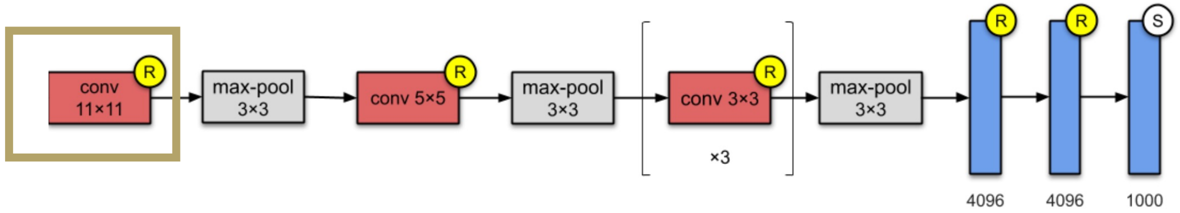
Visualizing Filters in the First Layer

Filters are looking for *low-level* oriented edges, color blobs, textures, background etc.



AlexNet:
64 x 3 x 11 x 11

Input Image:
3 x 224 x 224



AlexNet

Filters always extend the full depth of the input volume

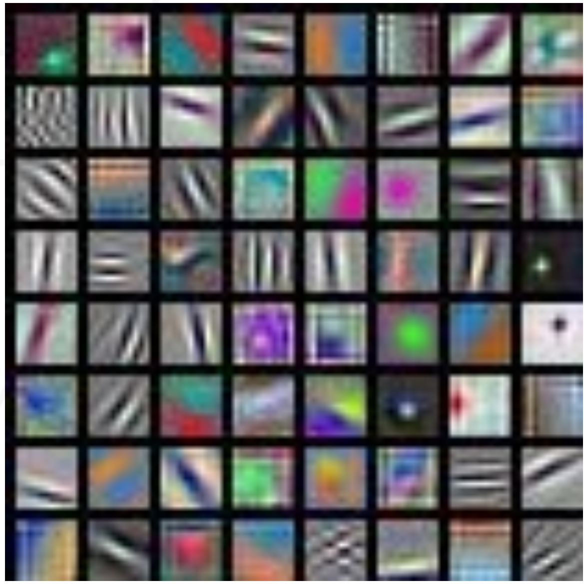
- 64 filters in the first convolutional layer
- Filter size: 11 x 11 x 3 (visualized as RGB images)

Krizhevsky et al. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012)

Indirect Explanations

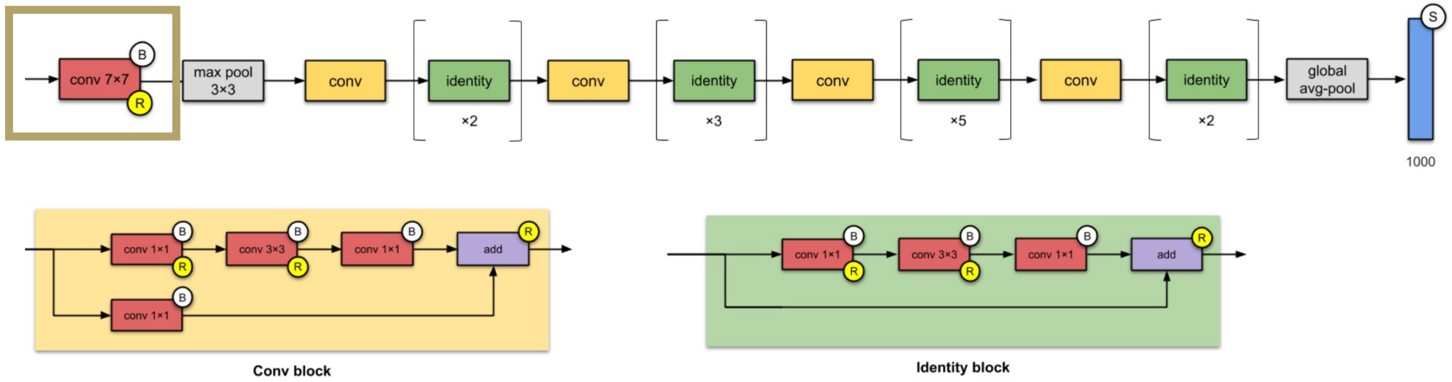
Visualizing Filters in the First Layer

Filters in the first convolutional layers across different architectures learn similar patterns



AlexNet:
64 x 3 x 11 x 11

ResNet-18:
64 x 3 x 7 x 7



ResNet

Krizhevsky et al. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012)

Indirect Explanations

Visualizing Filters in the Intermediate Layers

Filters in higher convolutional layers are not as interpretable as filters in the first layer

Visualizing the filters (raw weights)

Conv layer 2 weights
20 x 16 x 7 x 7
(visualize as 16 grayscale images)

Conv layer 3 weights
20 x 20 x 7 x 7
(visualize as 20 grayscale images)

Krizhevsky et al. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012)

Outline

Lecture 3: Visual Explanations I

- Human-centric Explanations
- Indirect Explanations
  - Visualizing filters
  - Visualizing activations
  - Visualizing Last layer Embedding
- Direct Explanations
  - Intervention-based visualizations
    - Saliency Maps
  - Gradient-based visualizations
    - Vanilla Backpropagation
    - Deconvolution Backpropagation
    - Guided Backpropagation
- Takeaways

Indirect Explanations

Visualizing Activations in the Intermediate Layers I

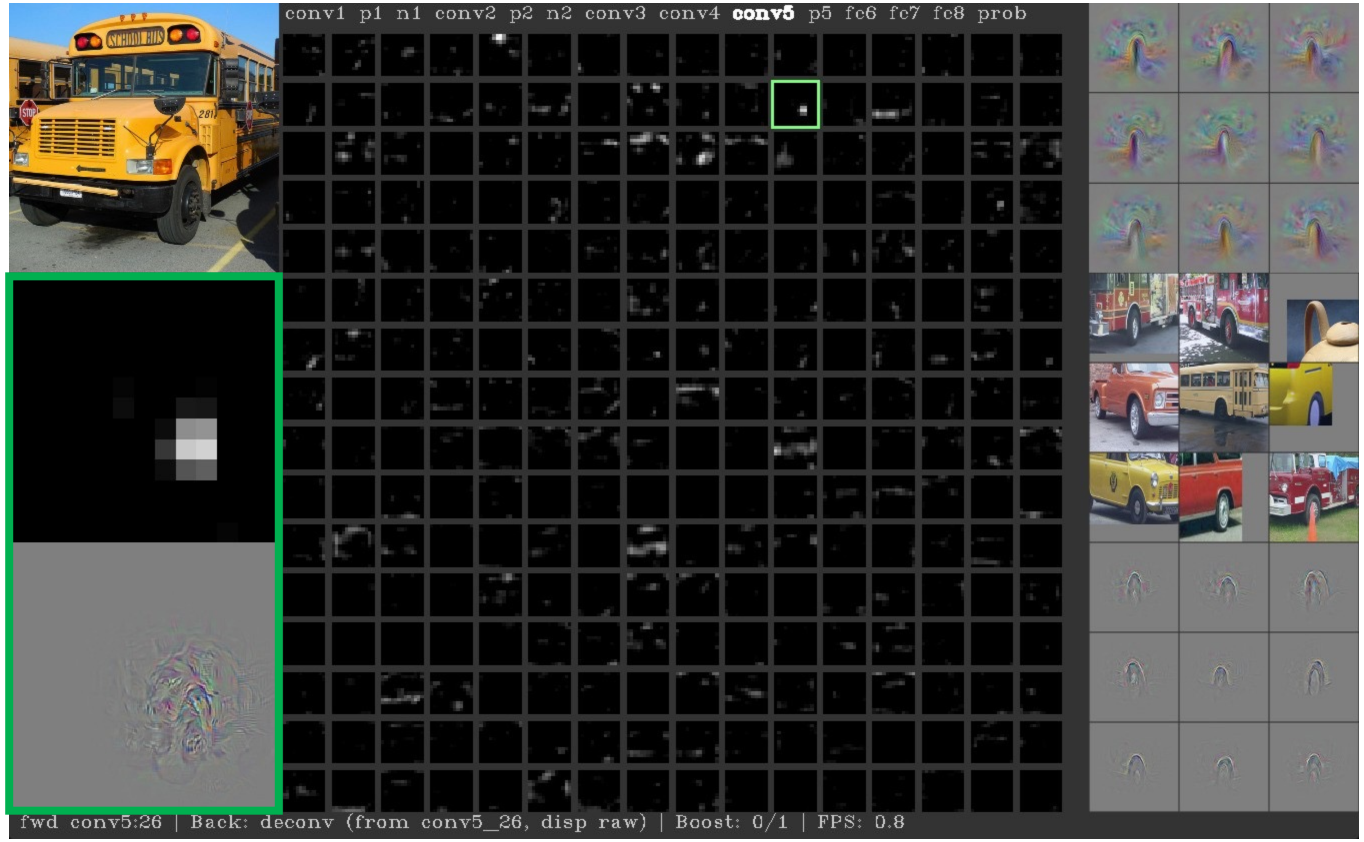
Higher layers are activated by *semantic concepts* rather than features

Intermediate layers:

- Weights: not very interpretable
- Activations: interpretable

The **filter in green box** is **activated when it sees a wheel**

However, it is irrational to explain billions of parameters by individual activation inspection



Conv 5 layer of AlexNet

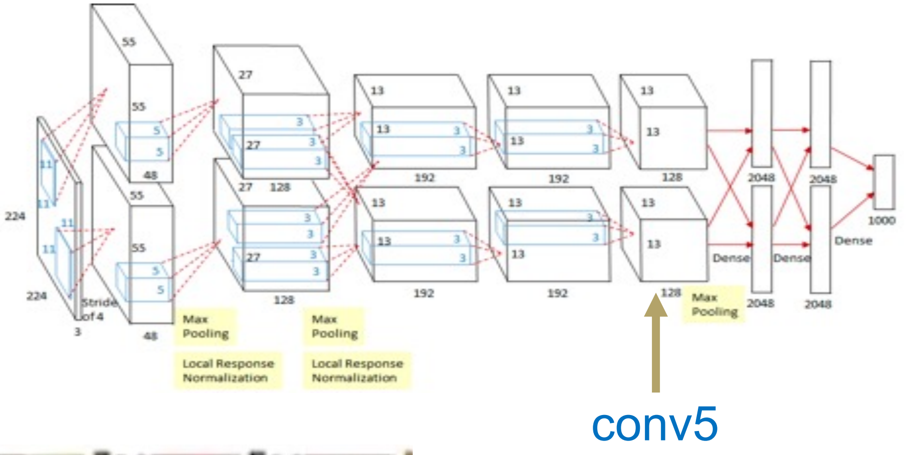
Yosinski et al, “Understanding Neural Networks Through Deep Visualization”, ICML DL Workshop 2014.

Indirect Explanations

Visualizing Activations in the Intermediate Layers II: Maximally Activating Patches

Visualize patterns in images that cause the maximum activations of certain neurons

- Maximally Activating Patches:
  - **Image patches** in the input that **cause the maximum activations** of certain filters
- Obtaining Maximally Activating Patches:
  - Pick activations in a layer
  - Feed forward images through the network, record values of the chosen channel
  - Visualize image patches that correspond to **maximal activation**



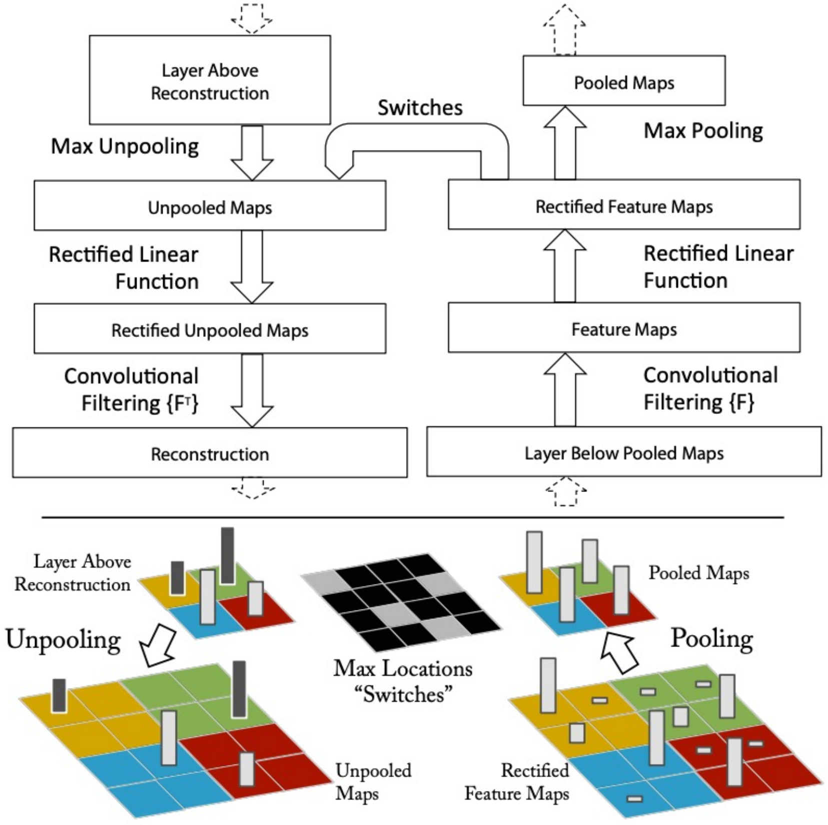
Each row corresponds to a particular neuron in conv5

Springenberg et al, “Striving for Simplicity: The All Convolutional Net”, *ICLR* Workshop 2015

Indirect Explanations

Visualizing Activations in the Intermediate Layers III: DeconvNet

Train a decoder network using activations from a given intermediate layer



Left: Deconvolution network, Right: Convolutional encoder

- **DeconvNet:** An additional deconvolution network is added to map features back into input space
- Instead of directly visualizing patches from the input images, reconstruct maximally activating patches

Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision. Springer, 2014, pp. 818–833

Outline

Lecture 3: Visual Explanations I

- Human-centric Explanations
- Indirect Explanations
  - Visualizing filters
  - Visualizing activations
  - Visualizing Last layer Embedding
- Direct Explanations
  - Intervention-based visualizations
    - Saliency Maps
  - Gradient-based visualizations
    - Vanilla Backpropagation
    - Deconvolution Backpropagation
    - Guided Backpropagation
- Takeaways

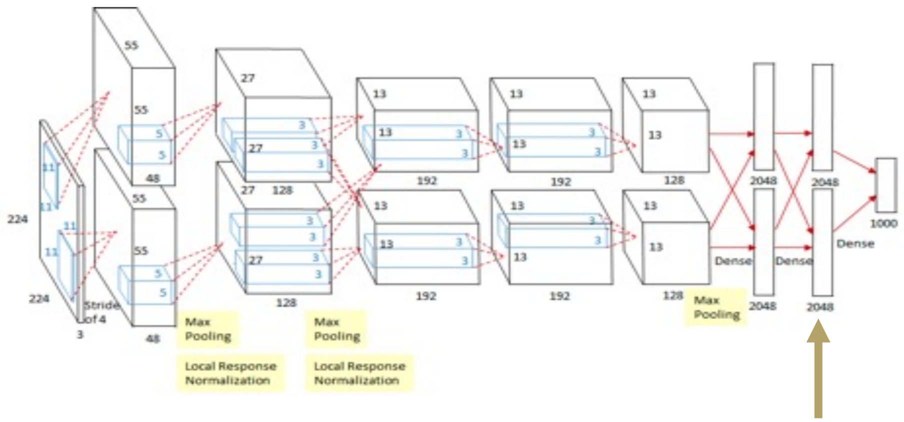
Indirect Explanations

Visualizing Last Layer Activations

Last layer activations consist of class-specific information

- We can group the images that have similar **class-specific information** by exploring **last layer activations**

- Last layer activations (embedding):
  - 4096-dimensional feature vector for an image (layer immediately before the classifier)
  - Representations of **entire input images** instead of specific patches
  - **Similar embeddings** correspond to **same classes of input images**



AlexNet fully-connected 7

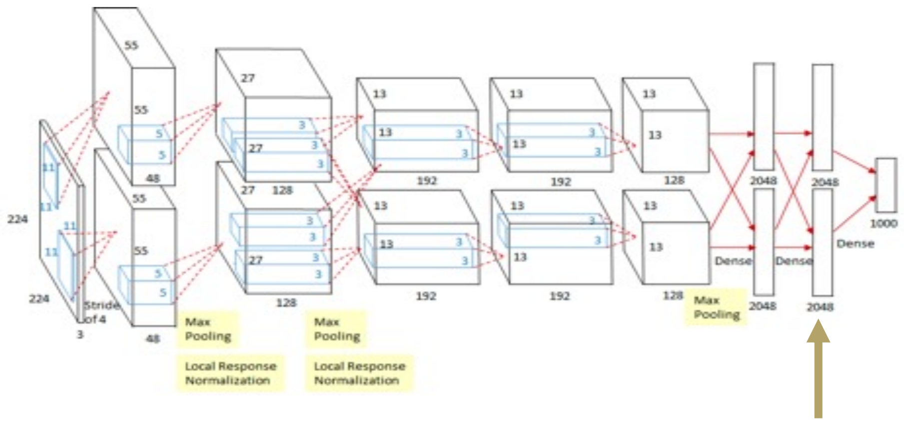
Indirect Explanations

Visualizing Last Layer Activations

Last layer activations consist of class-specific information

- Last layer activations (embedding):
  - 4096-dimensional feature vector for an image (layer immediately before the classifier)
  - Representations of *entire input images* instead of specific patches
  - Similar embeddings correspond to same classes of input images

- Feed forward images through the network, collect the final layer feature vectors
- **Visualize input images that have similar last layer embeddings**



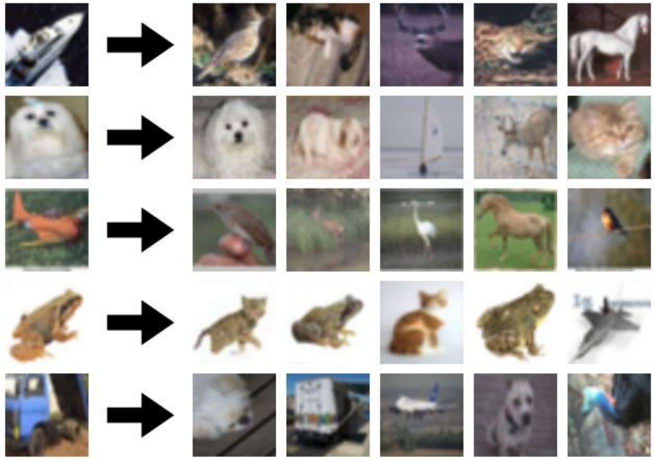
AlexNet   fully-connected 7

Indirect Explanations

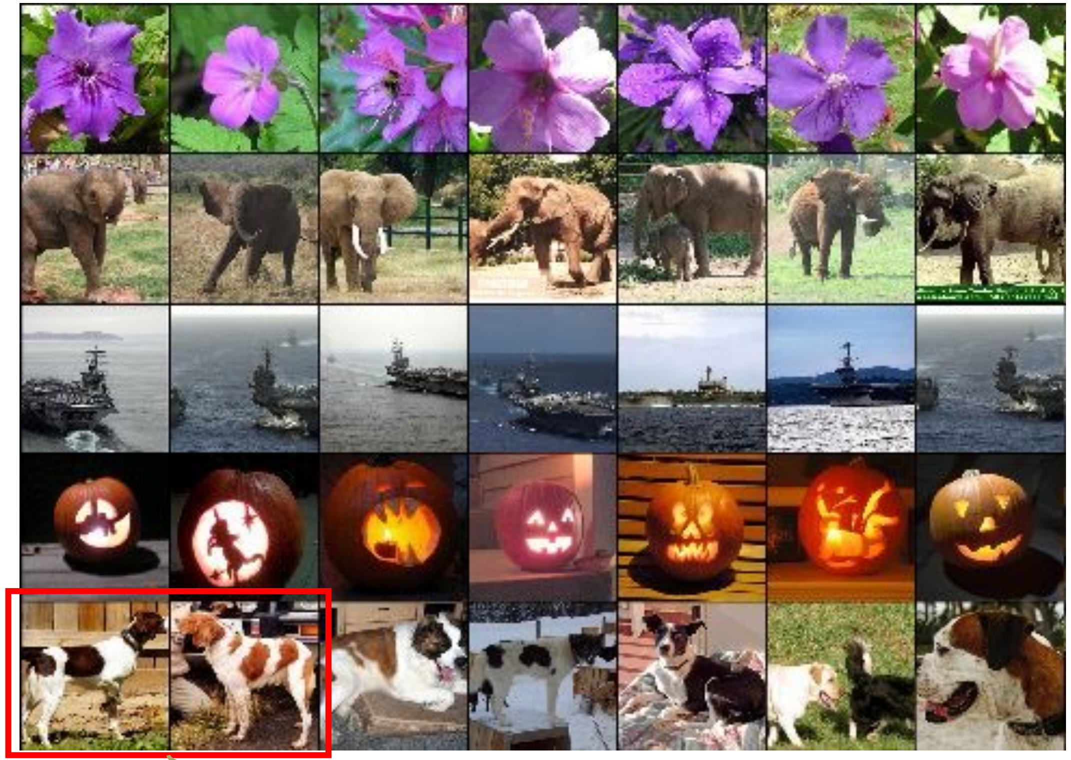
Visualizing Last Layer Activations I: Nearest Neighbor Samples

Explanations refer to retrieving the nearest neighbors (from train set) of given test image

L2 Nearest neighbors in **pixel** space



Test image L2 Nearest neighbors in **feature** space



- The **features** of the two dogs **share L2 similarity** in feature space
- In image space, they are **not L2-similar** due to **horizontally flipped poses**

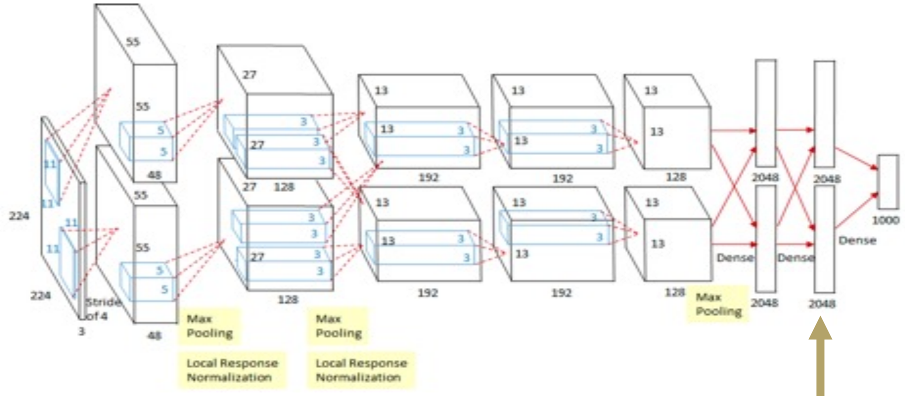
Indirect Explanations

Visualizing Last Layer Activations II: Dimensionality Reduction

Explanations refer to retrieving the nearest neighbors (from train set) of given test image

- Last layer embedding:
  - 4096-dimensional feature vector for an image

- Visualize the “feature space” by reducing dimensionality of feature vectors from 4096 to 2 dimensions
  - Each 2-dim feature correspond to an input image



fully-connected 7

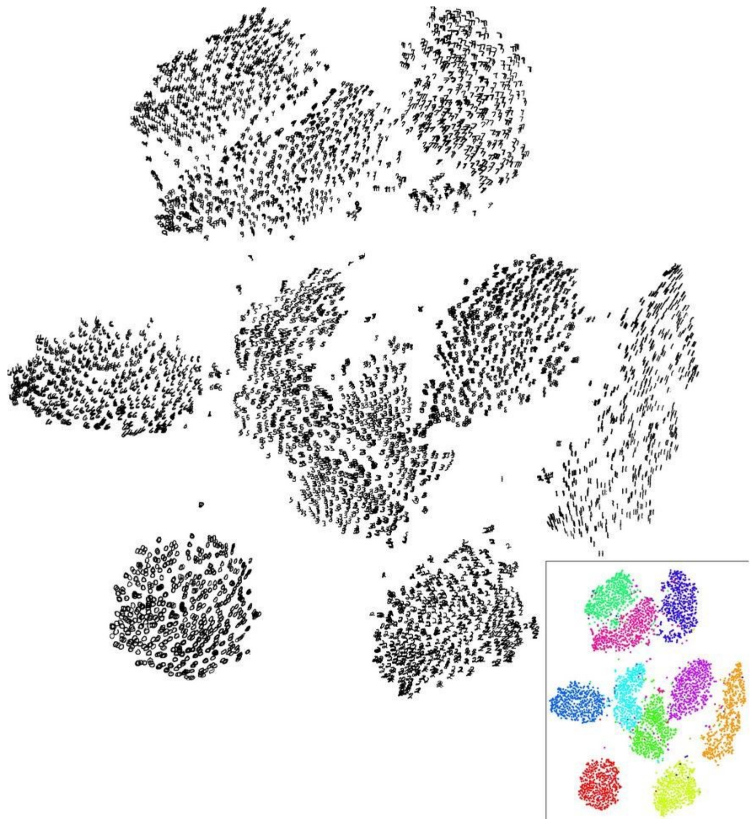
Van der Maaten and Hinton, “Visualizing Data using t-SNE”, JMLR 2008

Indirect Explanations

Visualizing Last Layer Activations II: Dimensionality Reduction

Explanations refer to retrieving the nearest neighbors (from train set) of given test image

- Last layer embedding:
  - 4096-dimensional feature vector for an image
- Dimensionality reduction using t-SNE (t-distributed stochastic neighbor embedding)
- Embed ***high-dimensional data*** points so that **locally, pairwise distances are conserved** i.e., similar classes end up in clusters, while dissimilar classes are separated



Van der Maaten and Hinton, “Visualizing Data using t-SNE”, JMLR 2008

Indirect Explanations

Summary

Indirect explanations require network knowledge from the humans interpreting the explanations

Indirect Explanations:

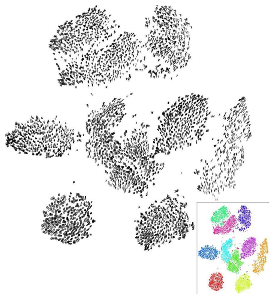
- **Visualize** weights (filters) in conv layers
- **Retrieve** maximally activating patches
- **Reconstruct** input images
- **Retrieve** nearest neighbor images in features space
- **Compute** last layer embeddings

Visualize, retrieve, reconstruct, and compute require the "technical know how"

filters

Maximally activating patches

Nearest neighbor images



Last layer embeddings

Outline

Lecture 3: Visual Explanations I

- Human-centric Explanations
- Indirect Explanations
  - Visualizing filters
  - Visualizing activations
  - Visualizing Last layer Embedding
- Direct Explanations
  - Intervention-based visualizations
    - Saliency Maps
  - Gradient-based visualizations
    - Vanilla Backpropagation
    - Deconvolution Backpropagation
    - Guided Backpropagation
- Takeaways

Explanations

Direct Explanations

Direct explanations highlight all regions in an image that lead to a decision

- **Required knowledge to understand explanations:** None
- **Required knowledge to obtain explanations:** Models, model parameters, and training data
- **Explanations audience:** Researchers and Engineers building the models
- **Explanatory chronology:** Most (existing) explanatory techniques are direct

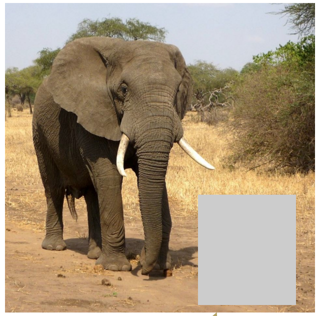
	Definition		
Methods	Indirect	Direct	Targeted
Deconvolution [21]	✓	–	–
Inverted Representations [22]	✓	–	–
Guided-Backpropagation [18]	–	✓	–
SmoothGrad [17]	–	✓	–
LIME [39]	–	✓	–
CAM [24]	–	✓	–
Graph-CNN [23]	✓	–	–
GradCAM [12]	–	–	✓
TCAV [40]	–	✓	–
GradCAM++ [16]	–	–	✓
RISE [35]	–	✓	–
Causal-CAM [15]	–	–	✓
Counterfactual-CAM [12]	–	–	✓
Goyal et al. [26]	–	–	✓
CEM [29]	–	–	✓
Contrast-CAM [13]	–	–	✓
Contrastive reasoning [14]	–	–	✓

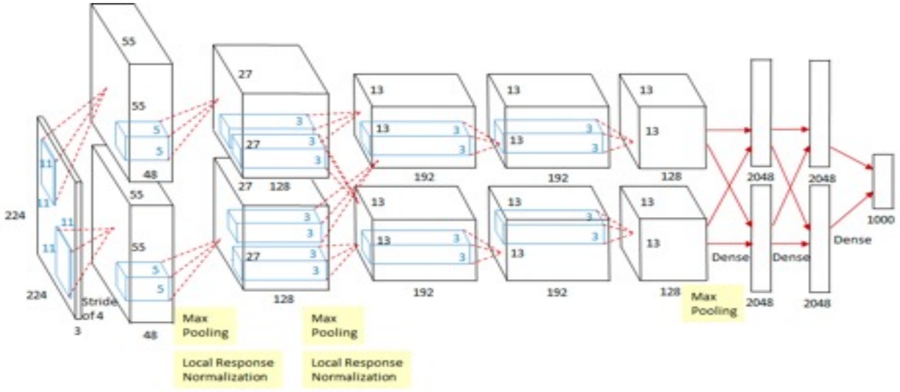
AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Direct Explanations

Saliency via Occlusion

Mask part of the image and check the change in predicted probabilities





P(elephant) = 0.95

A gray patch or patch of average pixel value of the dataset

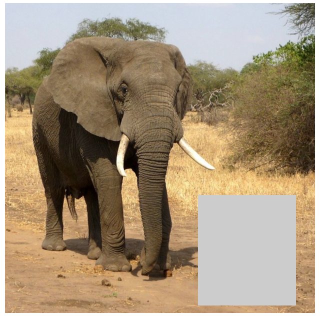
Note: <u>Not</u> a black patch because the input images are centered to zero in the preprocessing (More in lecture 5)

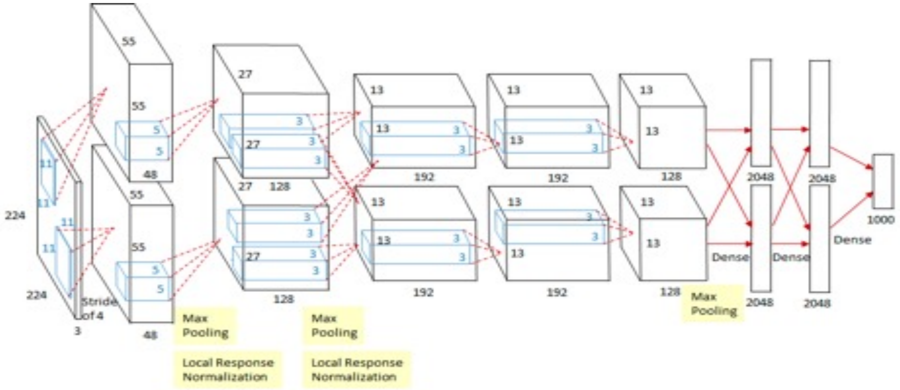
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Direct Explanations

Saliency via Occlusion

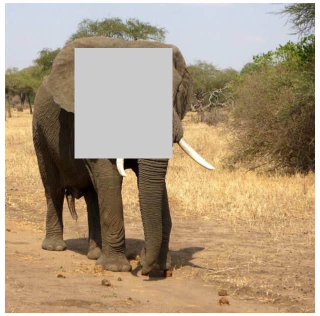
Mask part of the image and check the change in predicted probabilities

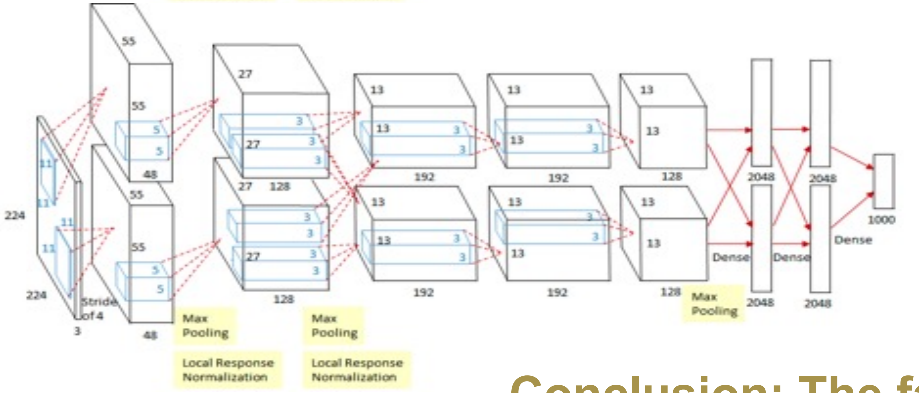




P(elephant) = 0.95

Decrease in probability (even with correct decision)





P(elephant) = 0.75

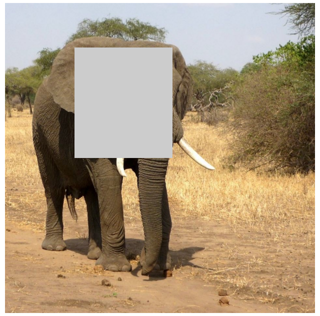
Conclusion: The face pixels affect the decisions more

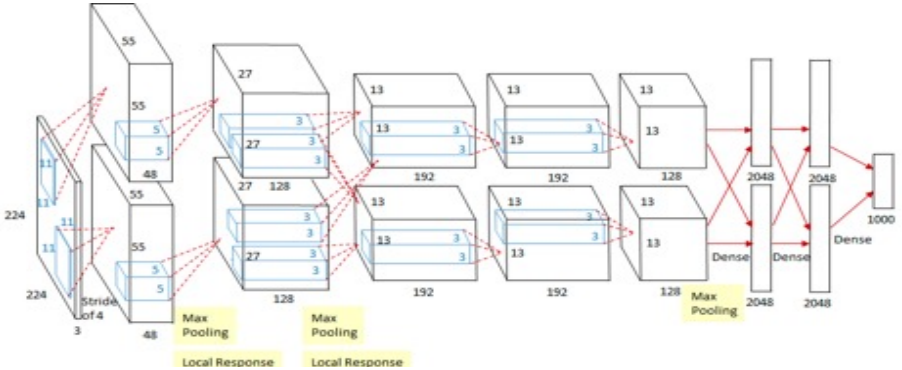
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Direct Explanations

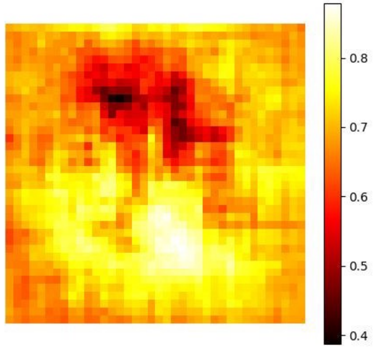
Saliency via Occlusion

Visualize the heatmap of pixels that cause decrease in probabilities when masked

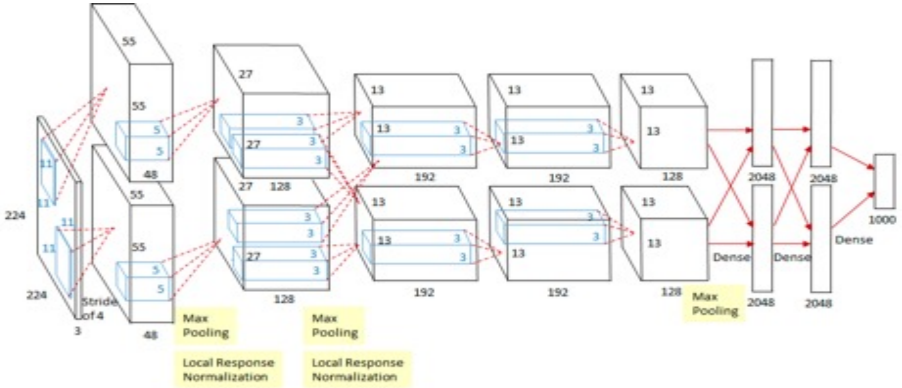




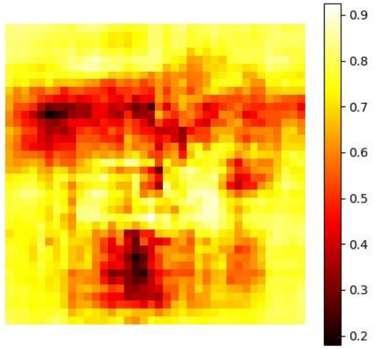
African elephant, Loxodonta africana



go-kart



go-kart

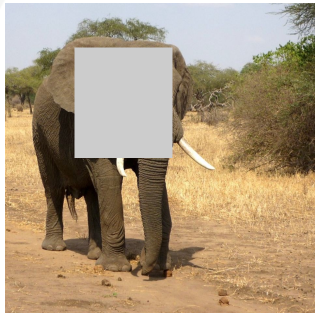


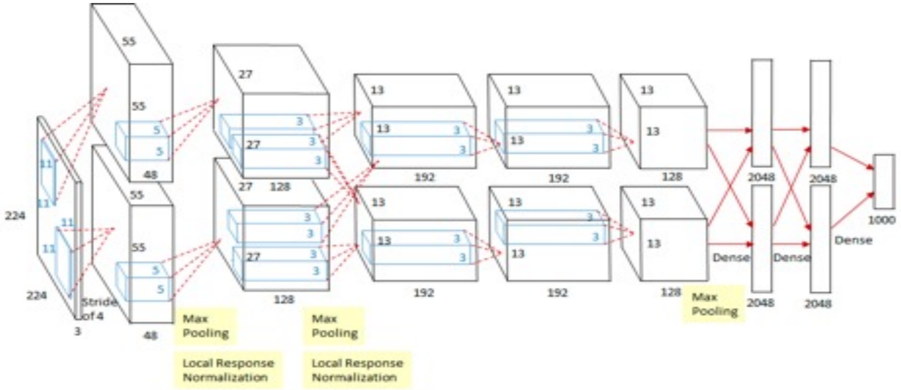
Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Direct Explanations

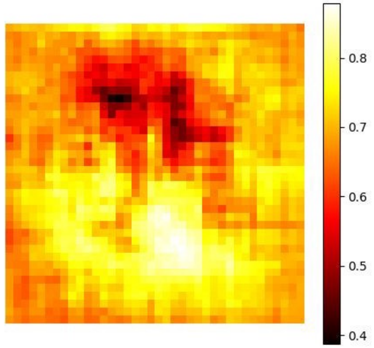
Saliency via Occlusion

Visualize the heatmap of pixels that cause decrease in probabilities when masked





African elephant, Loxodonta africana



Necessity property from Lecture 2: Features are said to be necessary if their deletion causes a misclassification

- **Saliency via Occlusion** is an approximation of necessity property and can objectively be evaluated as **”good”**
- However, the method is **computationally expensive**

Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

Outline

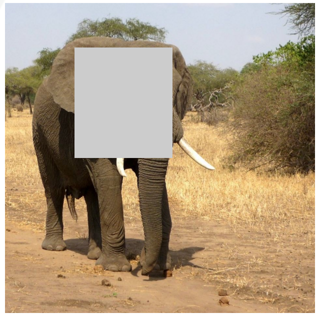
Lecture 3: Visual Explanations I

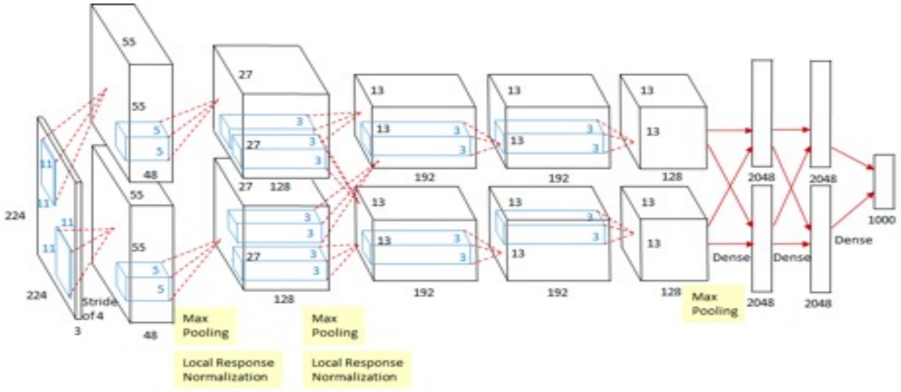
- Human-centric Explanations
- Indirect Explanations
  - Visualizing filters
  - Visualizing activations
  - Visualizing Last layer Embedding
- Direct Explanations
  - Intervention-based visualizations
    - Saliency Maps
  - Gradient-based visualizations
    - Vanilla Backpropagation
    - Deconvolution Backpropagation
    - Guided Backpropagation
- Takeaways

Direct Explanations

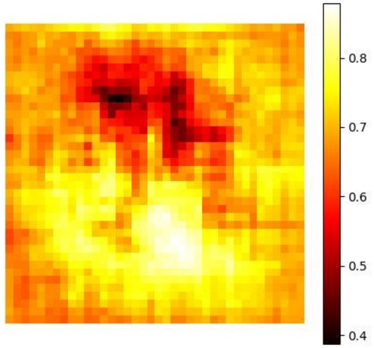
Saliency via Feature Importance

Finding alternatives to necessity property: Feature Importance





African elephant, Loxodonta africana



We define a new property called feature importance. A toy example:

- In logistic regression, for each feature x_i, a **weight** w_i represents its **importance**

$$P(y = 1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x} + b)$$
$$P(y = 0|\boldsymbol{x}) = 1 - \sigma(\boldsymbol{w}^T\boldsymbol{x} + b)$$

- We want to generate pixel saliency maps by deep models as feature importance maps

Direct Explanations

Saliency via Feature Importance

Saliency, approximated by gradients w.r.t. input, can be obtained via backpropagation

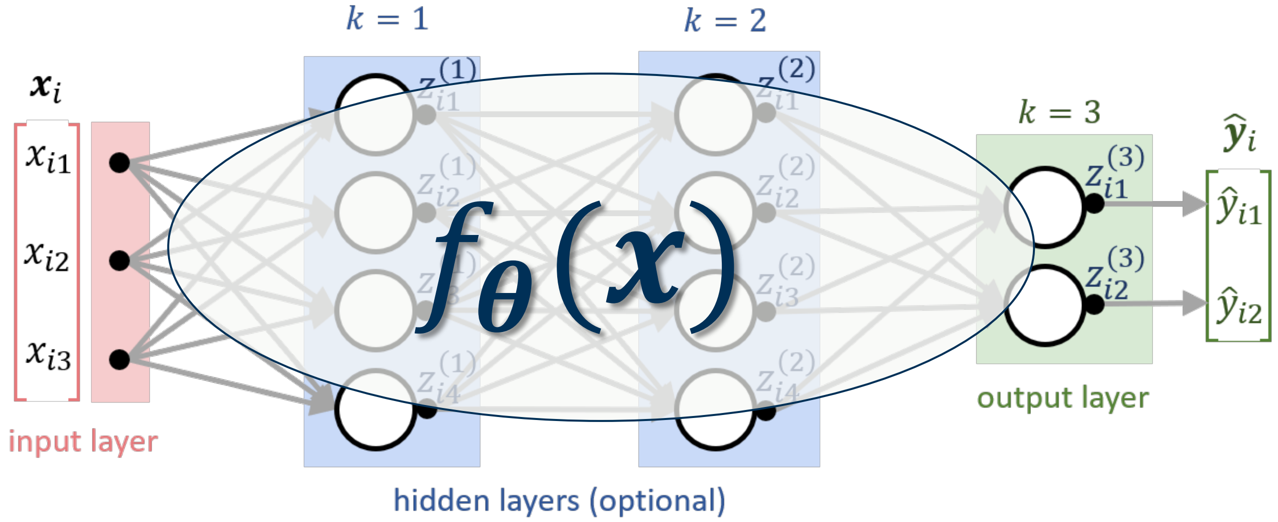
- *Highly non-linear mapping* function $f_{\theta}: \mathcal{X} \rightarrow \mathcal{Y}$:

$$\widehat{\boldsymbol{Y}} = \varphi\left(\varphi\left(\varphi\left(\boldsymbol{X}\left(\boldsymbol{W}^{(1)}\right)^{T} + \boldsymbol{o}(\boldsymbol{b}^{(1)})^{T}\right)\left(\boldsymbol{W}^{(2)}\right)^{T} + \boldsymbol{o}(\boldsymbol{b}^{(2)})^{T}\right)\left(\boldsymbol{W}^{(3)}\right)^{T} + \boldsymbol{o}(\boldsymbol{b}^{(3)})^{T}\right)$$

- Assume that we can 'linearize' the model using Taylor series

$$\widehat{\boldsymbol{Y}} \approx \boldsymbol{X}(\boldsymbol{W})^{T} + \boldsymbol{o}(\boldsymbol{b})^{T}$$

$$\boldsymbol{W} \approx \frac{\partial \widehat{\boldsymbol{Y}}}{\partial \boldsymbol{X}}$$



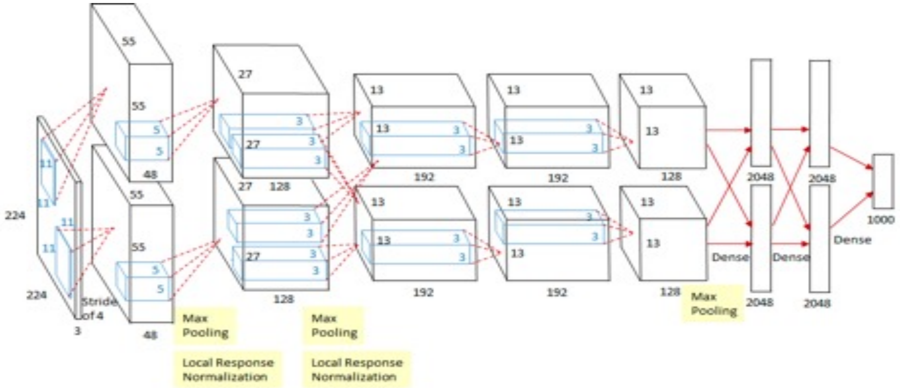
$$f(a)+\frac{f'(a)}{1!}(x-a)+\frac{f''(a)}{2!}(x-a)^2+\frac{f'''(a)}{3!}(x-a)^3+\cdots,$$

Direct Explanations

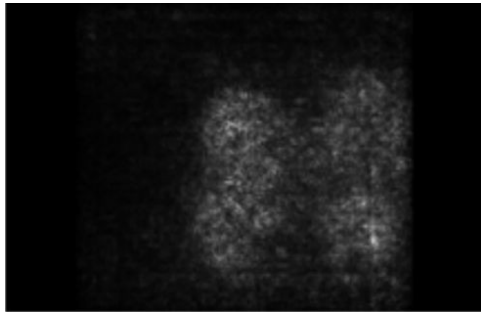
Gradient-based Saliency via Backpropagation

Saliency, approximated by gradients w.r.t. input, can be obtained via backpropagation

Forward pass: Compute probabilities



$$\hat{y} = \begin{bmatrix} 0.05 \\ 0.10 \\ \mathbf{0.85} \end{bmatrix}$$ Dog



Saliency map

$$\frac{\partial \hat{y}_c}{\partial X}$$

Backward pass: Compute gradients

Compute **gradient** of (unnormalized) class score **with respect to image pixels**

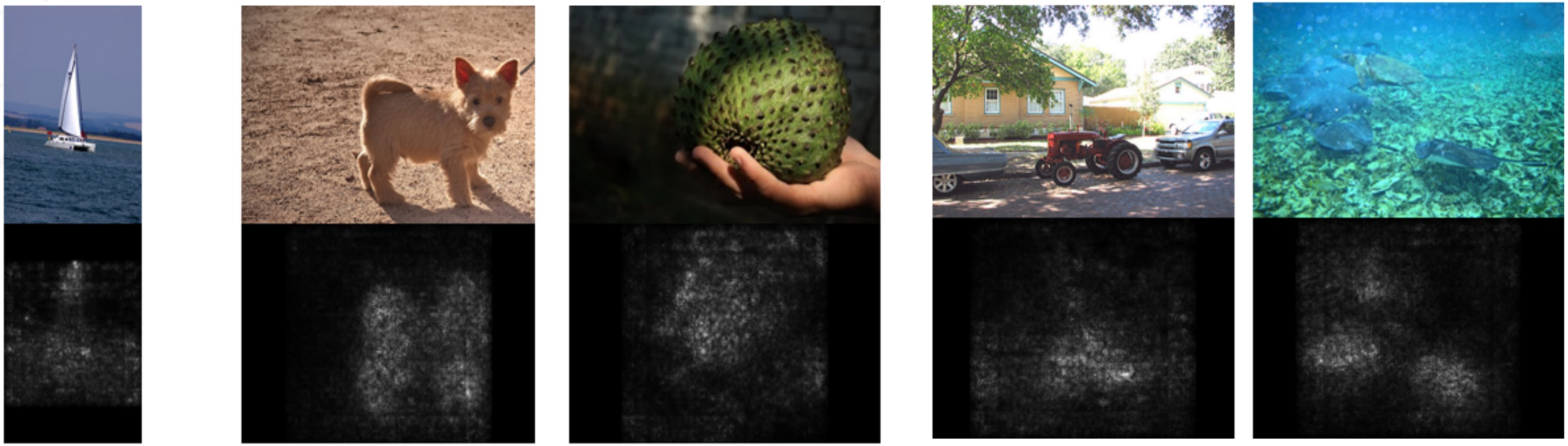
Then visualize the max of absolute value over RGB channels

Simonyan, Vedaldi, and Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014.

Direct Explanations

Gradient-based Saliency via Backpropagation

Saliency, approximated by gradients w.r.t. input, can be obtained via backpropagation

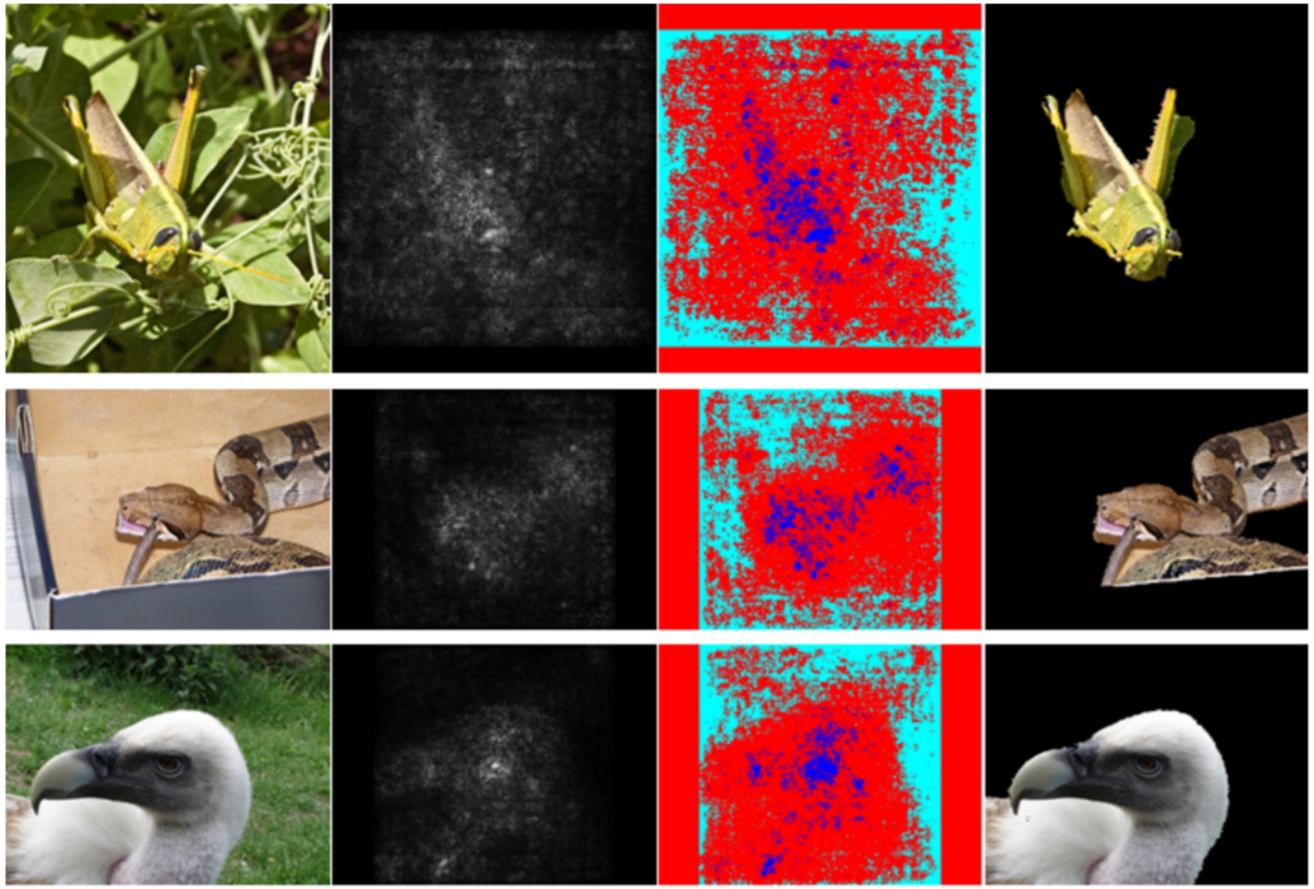


Simonyan, Vedaldi, and Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014.

Direct Explanations

Gradient-based Saliency via Backpropagation

Saliency maps can be used to help unsupervised semantic segmentation



Note: The network is trained only for classification. But it is **sensitive** to the all class-related visual **regions/features** in images

Simonyan, Vedaldi, and Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014.

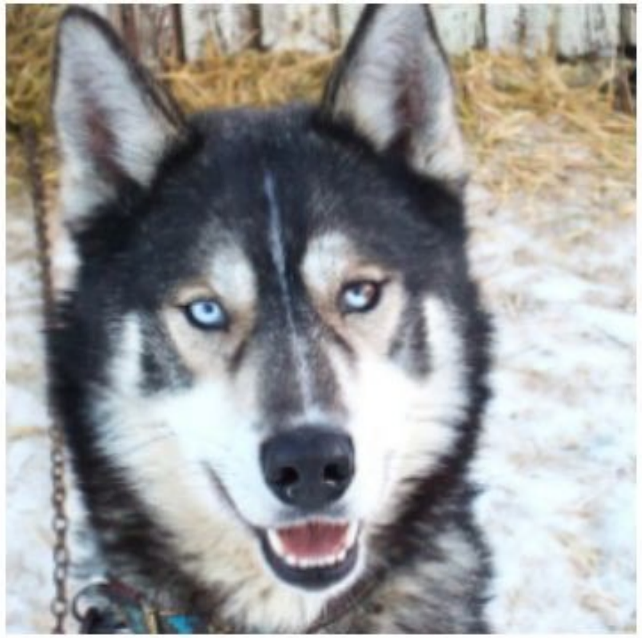
Direct Explanations

Gradient-based Saliency via Backpropagation

Saliency Maps can find biases

When all training wolf images have snow, network may use these snow pixels as salient regions for prediction

Wolf vs. dog classifier is actually a snow vs. no-snow classifier



(a) Husky classified as wolf

(b) Explanation

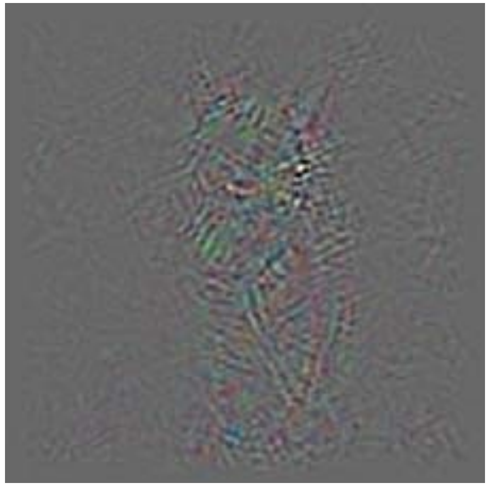
snow pixels as salient regions

Ribeiro et al, ““Why Should I Trust You?” Explaining the Predictions of Any Classifier”, ACM KDD 2016

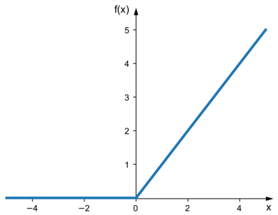
Direct Explanations

Vanilla Backpropagation

Saliency map by vanilla backprop



Method: Backpropagate by performing all the operations of the network (Unpooling, Filtering...). For ReLU non-linearities, **only pass gradients to regions of positive activations**



$$h^{l+1} = \max\{0, h^l\}$$

Forward pass

$$h^l \quad \begin{bmatrix} 1 & -1 & 5 \\ 2 & -5 & -7 \\ -3 & 2 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 5 \\ 2 & 0 & 0 \\ 0 & 2 & 4 \end{bmatrix} \quad h^{l+1}$$

$$\frac{\partial L}{\partial h^l} = [\![h^l > 0]\!] \frac{\partial L}{\partial h^{l+1}}$$

Backward pass: backpropagation

$$\begin{bmatrix} -2 & 0 & -1 \\ 6 & 0 & 0 \\ 0 & -1 & 3 \end{bmatrix} \leftarrow \begin{bmatrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{bmatrix} \quad \frac{\partial L}{\partial h^{l+1}}$$

Gradients from the later layer

positive activations in the previous layer

Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015

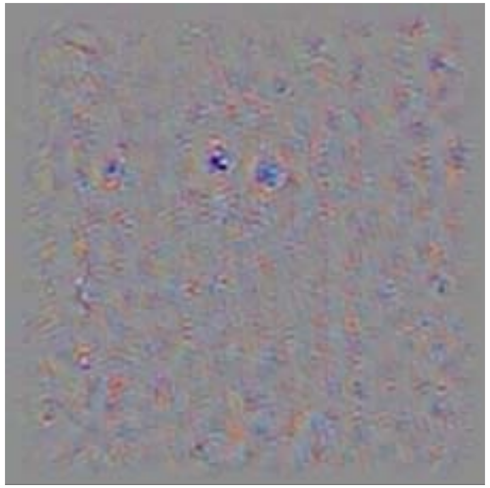
Direct Explanations

Deconvnet Backpropagation

The way DeconvNet Backpropagation handles the ReLU non-linearities is different as they propose to **only propagate positive gradient**

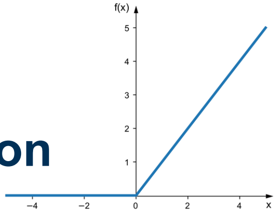
Rectifying the backpropagation empirically produce better saliency visualizations

Saliency map by deconv backprop



Cleaner saliency map

We can think of **Deconvnet** as **rectified gradients propagation**



$$\frac{\partial L}{\partial h^l} = [[\frac{\partial L}{\partial h^{l+1}} > 0]] \frac{\partial L}{\partial h^{l+1}}$$

Backward pass: "deconvnet"

0	3	0
6	0	1
2	0	3

←

-2	3	-1
6	-3	1
2	-1	3

$\frac{\partial L}{\partial h^{l+1}}$

Gradients from the later layer

positive gradient in the **later layer**

Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015

Direct Explanations

Guided Backpropagation

Guided backpropagation propose to **propagate positive gradient and rectified by positive activations**

Non-intuitive approach but **empirically** produce better saliency visualizations

Saliency map by Guided backpropagation

Cleaner saliency map

$$h^{l+1} = \max\{0, h^l\}$$

Forward pass

h^l

1	-1	5
2	-5	-7
-3	2	4

→

h^{l+1}

1	0	5
2	0	0
0	2	4

$$\frac{\partial L}{\partial h^l} = [[h^l > 0]][[\frac{\partial L}{\partial h^{l+1}} > 0]]\frac{\partial L}{\partial h^{l+1}}$$

Backward pass: *guided backpropagation*

0	0	0
6	0	0
0	0	3

←

$\frac{\partial L}{\partial h^{l+1}}$

-2	3	-1
6	-3	1
2	-1	3

positive activations in the previous layer

positive gradient in the later layer

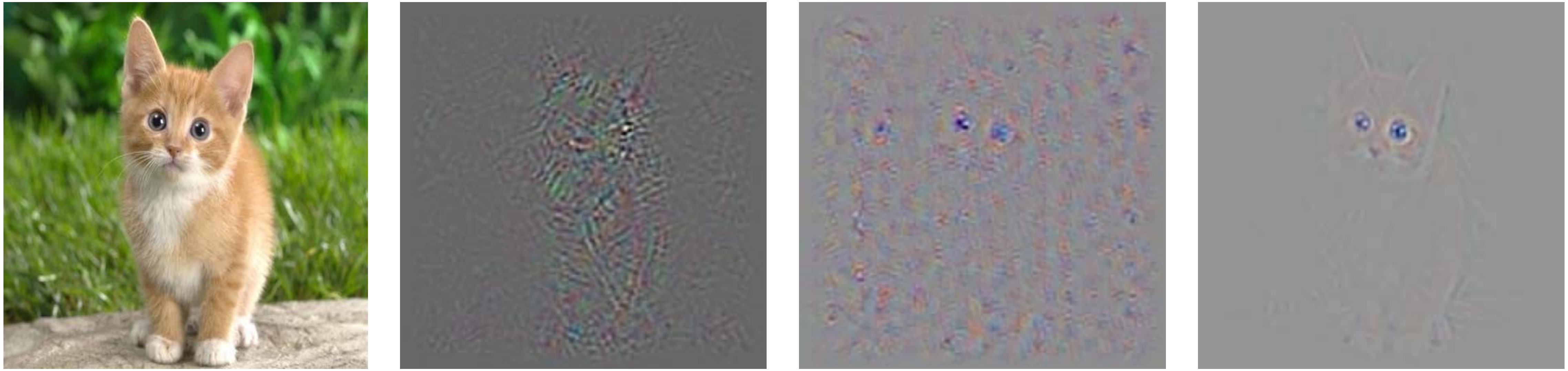
Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015

Direct Explanations

Guided vs Deconvnet vs Vanilla Backpropagation

Guided Backpropagation tends to be “cleanest”

Backprop | Deconv | Guided Backprop



Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015

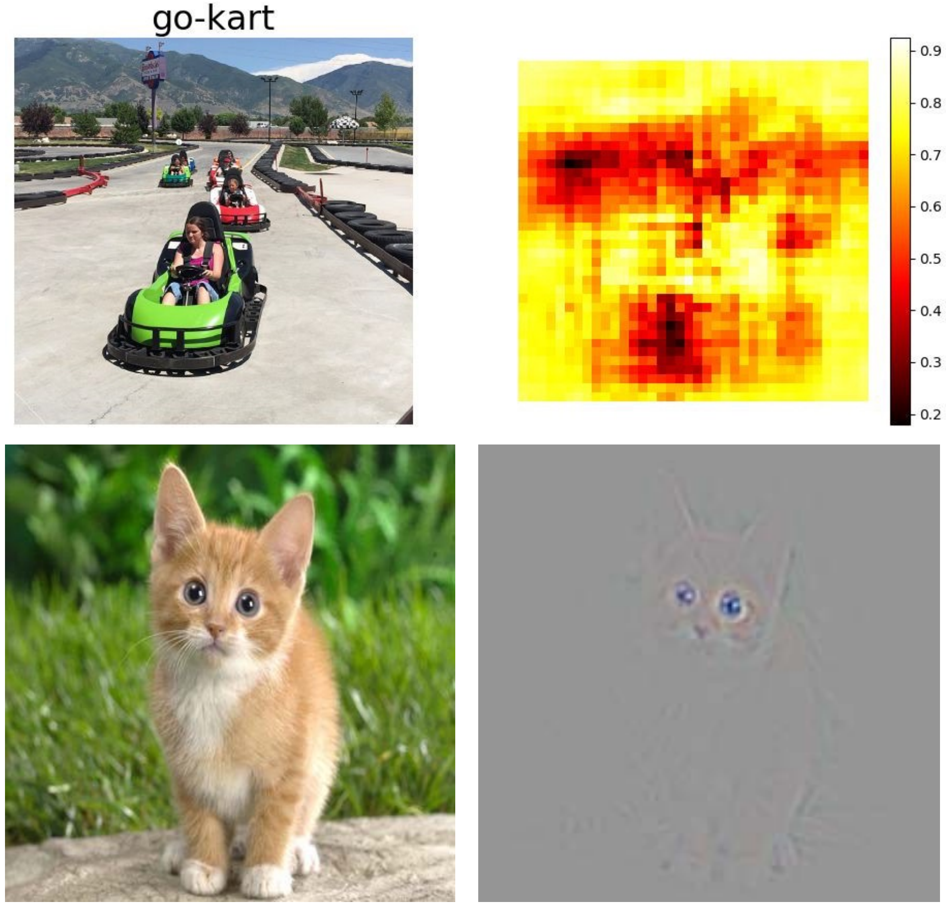
Direct Explanations

Summary

Direct explanations highlight all regions in an image that lead to a decision

- **Intervention-based:**
  - perturbing pixels and see how the decision change
  - Computationally expensive
- **Gradient-based:**
  - approximates feature importance by backpropagation
  - computationally efficient

However, direct explanations assume no knowledge from the audience either about the network or the data



Direct Explanations

Shortcomings in Guided Backpropagation

However, Guided Backpropagation explanations are not class-discriminative

GB explanation for “airliner”

GB explanation for “bus”

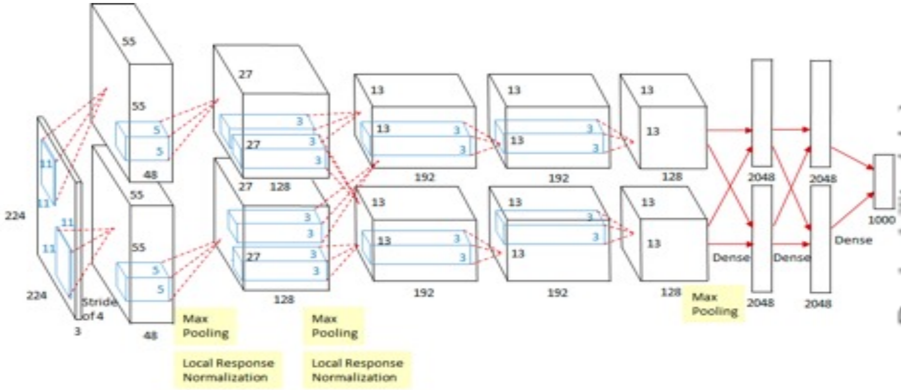
Nie, et al. "A theoretical explanation for perplexing behaviors of backpropagation-based visualizations." International Conference on Machine Learning. PMLR, 2018.

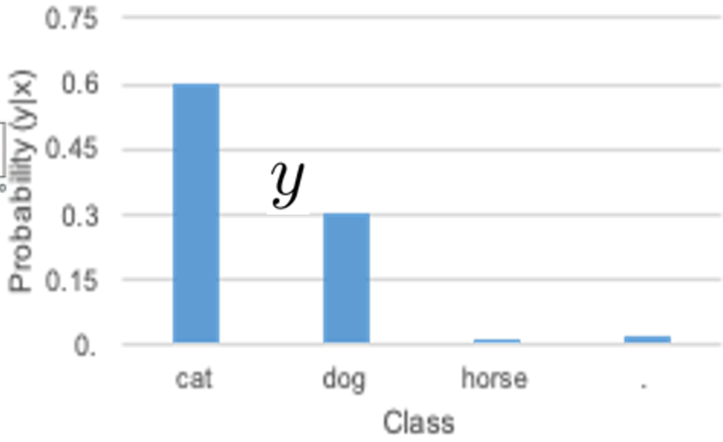
Direct Explanations

Shortcomings in Guided Backpropagation

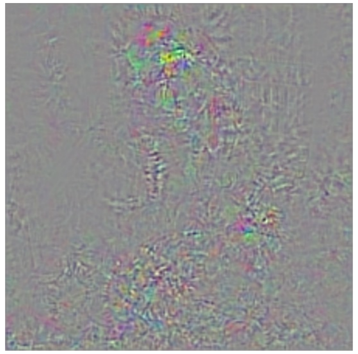
Guided Backpropagation does not explain decisions; They reconstruct inputs

x

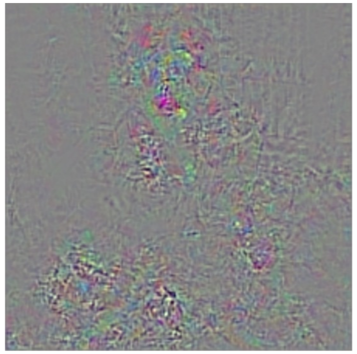




backprop



Backprop for \`cat’



Backprop for \`dog’

$$w_c = \frac{\partial y_c}{\partial x}\Big|_{x=x_0}$$

Guided Backprop for \`cat’

Guided Backprop for \`dog’

Nie, et al. "A theoretical explanation for perplexing behaviors of backpropagation-based visualizations." International Conference on Machine Learning. PMLR, 2018.

Direct Explanations

Shortcomings in Direct Explanations

Direct explanations highlight all pixels that lead to decision making; However, they provide no mechanism to choose targeted pixels based on class discriminability

Why Bullmastiff?

Why Tigercat?

We need Targeted Explanations!

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.

Takeaways

Takeaways from Lecture 3

- There are **no "one size fits all" explanations** and techniques
- **Indirect explanations requires knowledge of networks and data**
  - They are only accessible to a few
- **Direct explanations** place **no constraints** on knowledge of the audience
  - Saliency via occlusion is a direct but computationally expensive explanation
  - **Backpropagation assigns importance scores to pixels**
  - Rectification can be performed on gradients to obtain deconvolution and guided backpropagation
- Guided backpropagation provides the cleanest explanations
  - **However, it only reconstructs salient regions of the image without providing class-specific information**

References

Lecture 3: Visual Explanations I

- AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine* 39.4 (2022): 59-72.
- Krizhevsky et al. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems 25 (2012)
- Yosinski et al, "Understanding Neural Networks Through Deep Visualization", ICML DL Workshop 2014.
- Springenberg et al, "Striving for Simplicity: The All Convolutional Net", *ICLR* Workshop 2015
- Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in European conference on computer vision. Springer, 2014, pp. 818–833
- Van der Maaten and Hinton, "Visualizing Data using t-SNE", JMLR 2008
- Zeiler and Fergus, "Visualizing and Understanding Convolutional Networks", ECCV 2014
- Simonyan, Vedaldi, and Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", ICLR Workshop 2014.
- Ribeiro et al, ""Why Should I Trust You?" Explaining the Predictions of Any Classifier", ACM KDD 2016
- Springenberg, Dosovitskiy, et al., Striving for Simplicity: The all convolutional net, 2015
- Nie, et al. "A theoretical explanation for perplexing behaviors of backpropagation-based visualizations." International Conference on Machine Learning. PMLR, 2018.