

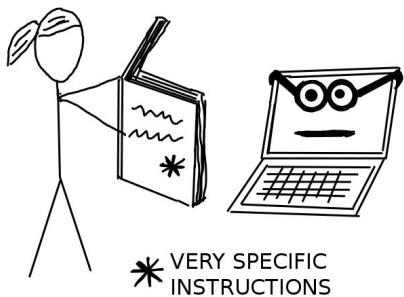


Interpretability in Machine Learning

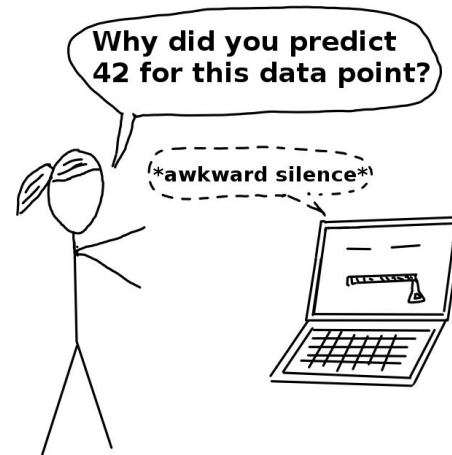


Interpretability

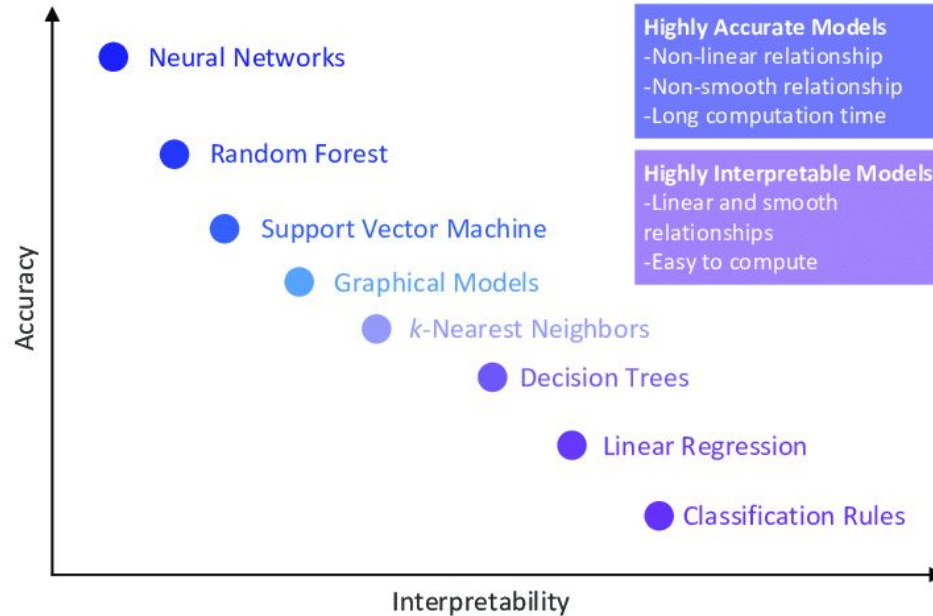
Without Machine Learning



With Machine Learning

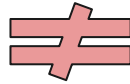


Interpretability



Interpretability vs Explainability

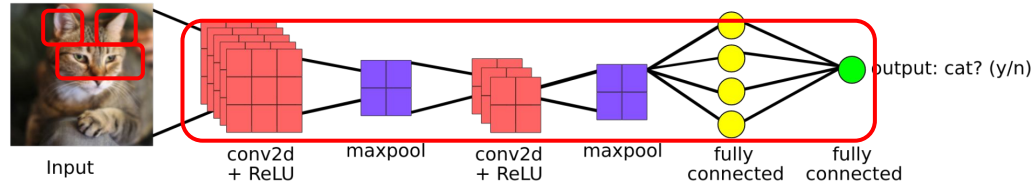
Interpretability



Explainability:

Understand the cause of a model decision.

Understand the internal mechanism leading to a model decision.



Interpretable Machine Learning:




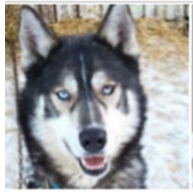


Methods and models that make the decisions taken by machine learning systems humanly understandable



Why do we want interpretability?

- Human nature
- Acceptance and trust
- Legal issues
- Fairness issues
- Improve human knowledge
- Debugging

Why do we want interpretability?

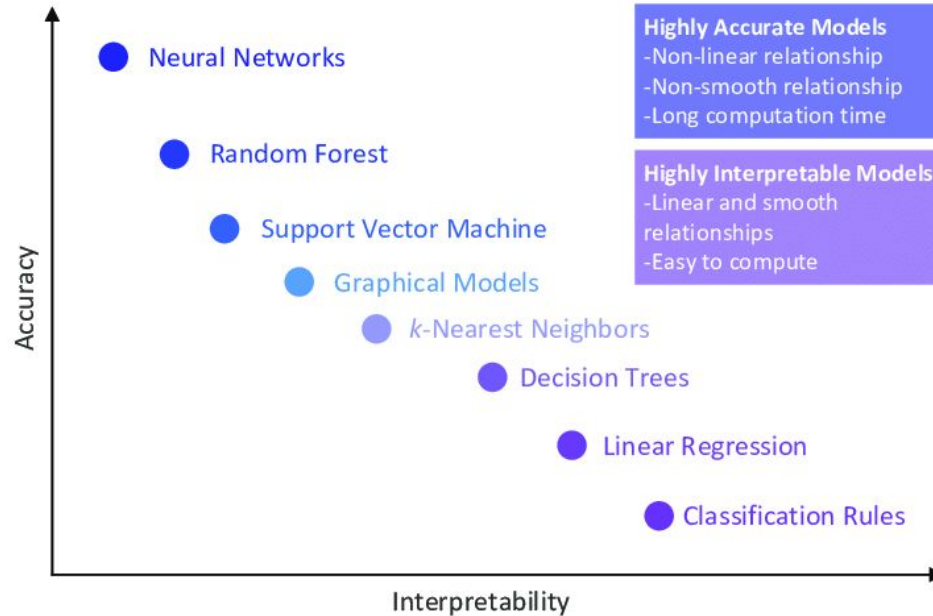
		
Predicted: wolf True: wolf	Predicted: husky True: husky	Predicted: wolf True: wolf
		
Predicted: wolf True: husky	Predicted: husky True: husky	Predicted: wolf True: wolf



Taxonomy of Interpretability Methods

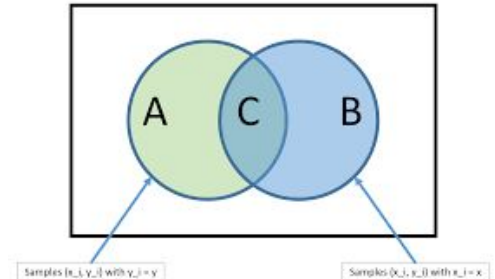
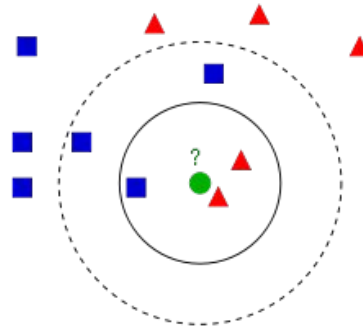
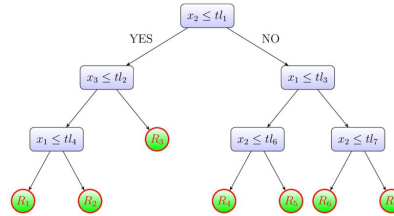
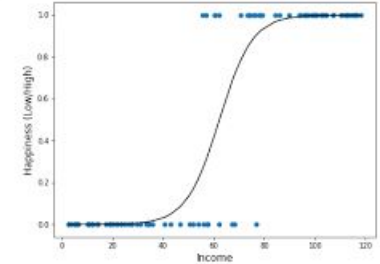
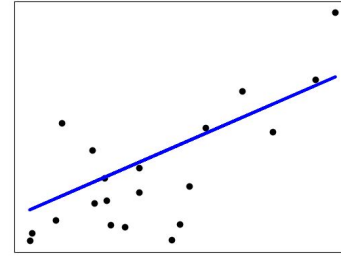
- Intrinsic or Post Hoc

Interpretability

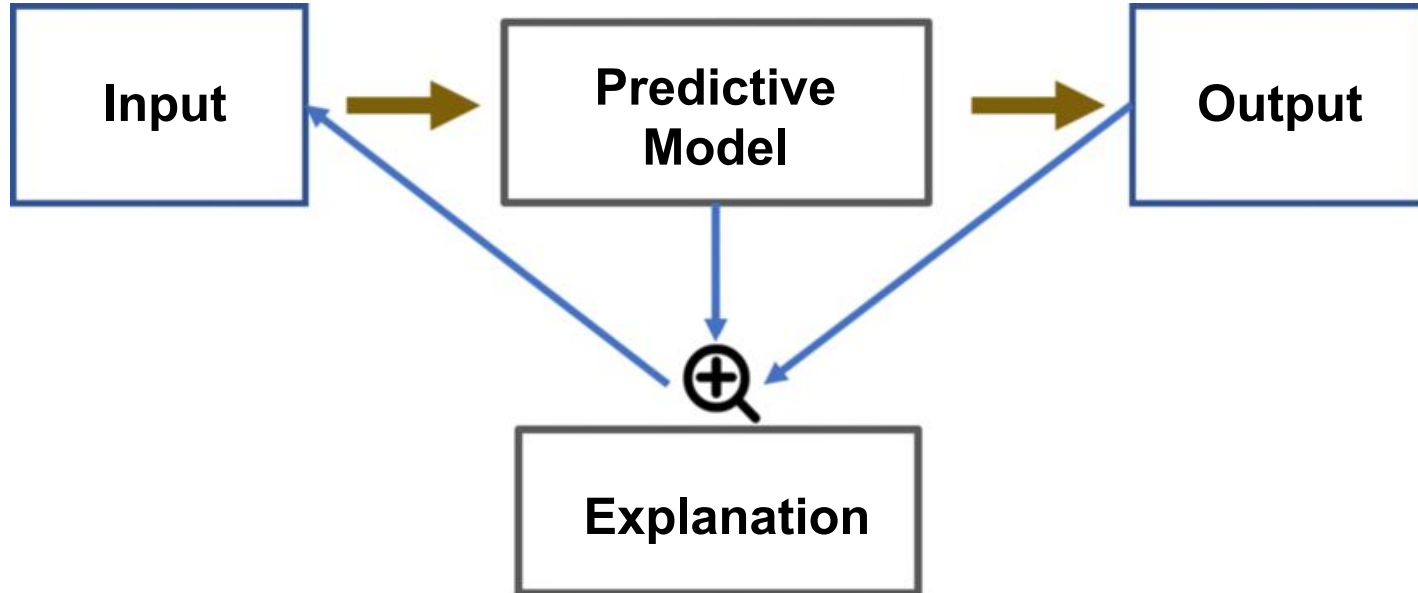


Intrinsic: Interpretable models

- Linear regression
- Logistic regression
- Decision trees
- Naïve Bayes
- K-nearest neighbors



Post Hoc

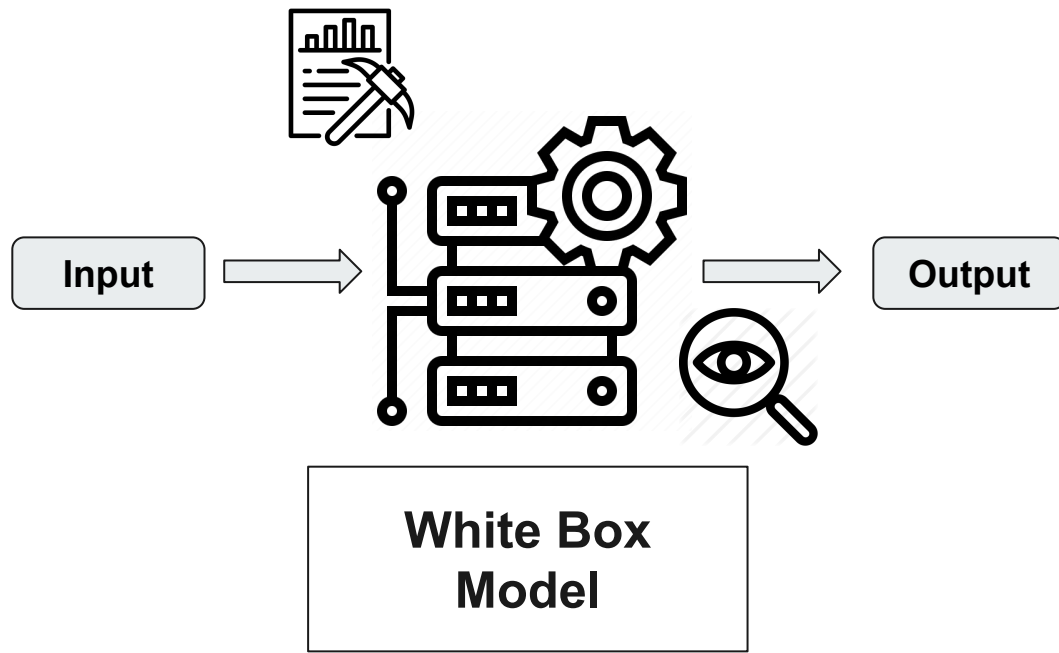




Taxonomy of Interpretability Methods

- Intrinsic or Post Hoc
- Model-specific or Model-agnostic

Model-Specific Methods



Model-Agnostic Methods



**Black Box
Model**



**Model-Agnostic
Methods**



Taxonomy of Interpretability Methods

- Intrinsic or Post Hoc
- Model-specific or Model-agnostic
- Local or global



Local vs Global Methods

Global Methods:

- Describe the average behavior
- Discover biases
- Useful for debugging at high level

Local Methods:

- Individual predictions insights
- Provide explanations to the final user
- Useful for debugging at the instance level

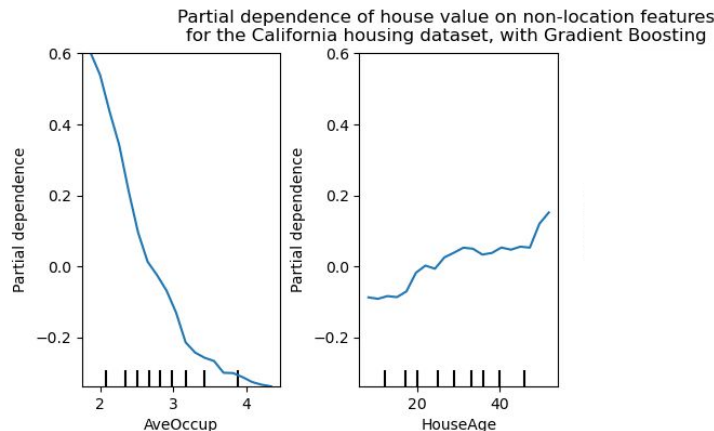


Model-agnostic methods

Partial Dependence Plot (PDP)

- Global Method
- Show the relationship between the target and a specific feature
- Monte Carlo method:

$$\hat{f}_S(x_S) = \frac{1}{n} \sum_{i=1}^n \hat{f}(x_S, x_C^{(i)})$$



Greedy function approximation: A gradient boosting machine.



PDP-based Feature Importance

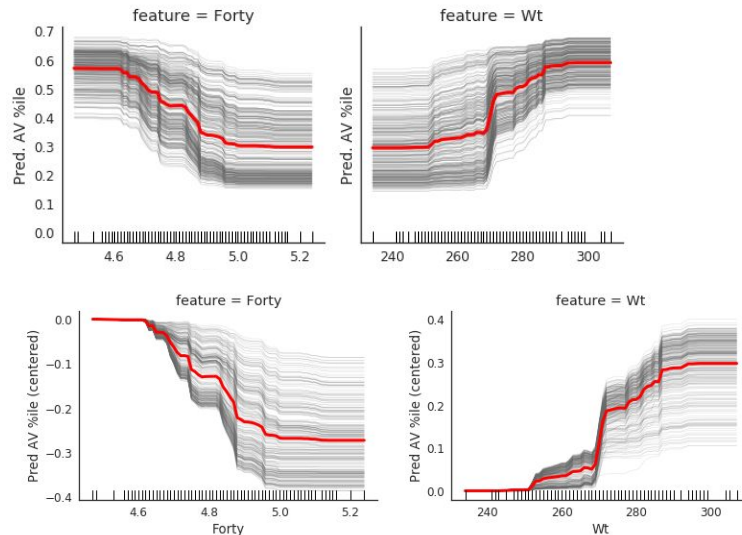
$$I(x_S) = \sqrt{\frac{1}{K-1} \sum_{k=1}^K \left(\hat{f}_S(x_S^{(k)}) - \frac{1}{K} \sum_{k=1}^K \hat{f}_S(x_S^{(k)}) \right)^2}$$

A Simple and Effective Model-Based Variable Importance Measure

[Brandon M. Greenwell](#), [Bradley C. Boehmke](#), [Andrew J. McCarthy](#)

Individual Conditional Expectation (ICE)

- Local Method
- Show the relationship between the target and a specific feature AND instance
- Centered ICE plots



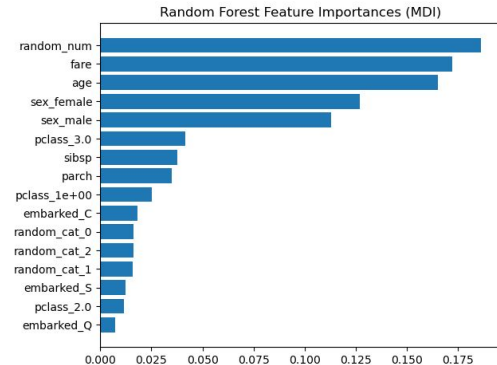
Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation

[Alex Goldstein](#), [Adam Kapelner](#), [Justin Bleich](#), [Emil Pitkin](#)

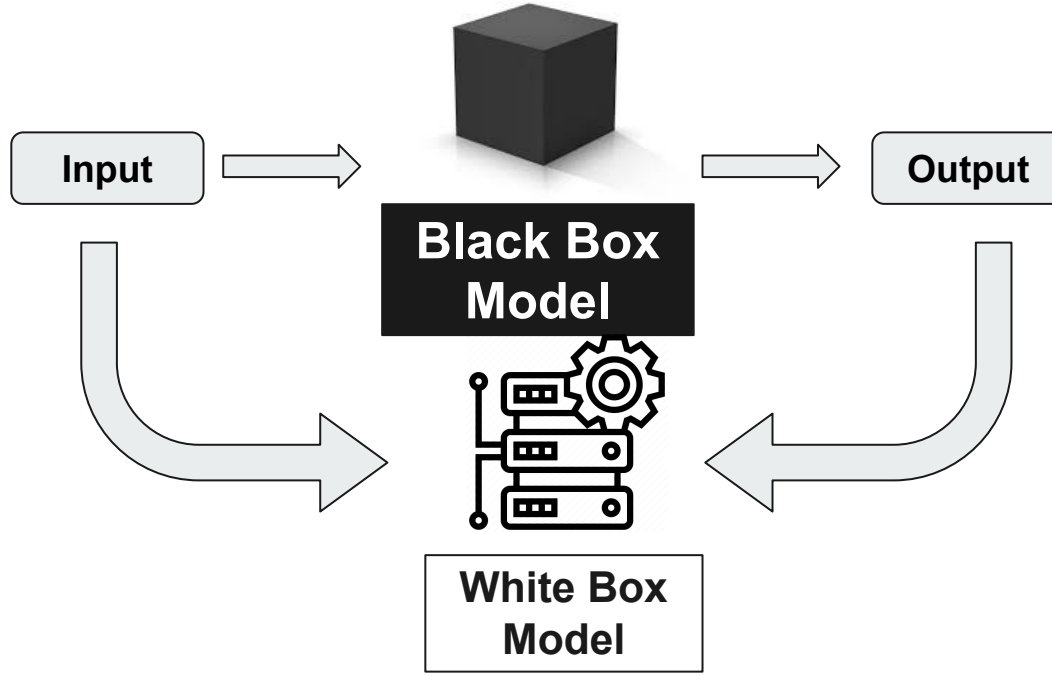
Permutation Feature Importance

- How much permuting a feature increases the prediction error?
- If error increases -> important feature
- If error decreases -> unimportant feature

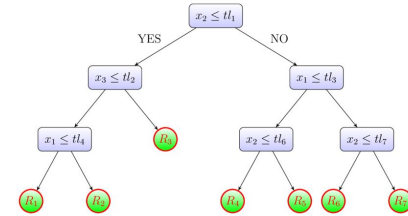
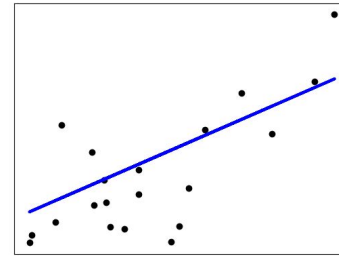
X_A	X_B	X_C	Y
<i>xa1</i>	<i>xb1</i>	<i>xc1</i>	<i>y1</i>
<i>xa2</i>	<i>xb2</i>	<i>xc2</i>	<i>y2</i>
<i>xa3</i>	<i>xb3</i>	<i>xc3</i>	<i>y3</i>
<i>xa4</i>	<i>xb4</i>	<i>xc4</i>	<i>y4</i>
<i>xa5</i>	<i>xb5</i>	<i>xc5</i>	<i>y5</i>
<i>xa6</i>	<i>xb6</i>	<i>xc6</i>	<i>y6</i>



Model Surrogates



- Global surrogate
- Local surrogate

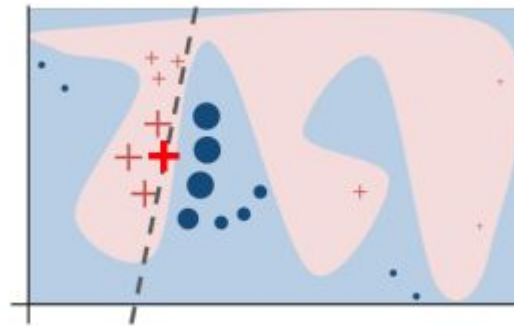


Local interpretable model-agnostic explanations (LIME)

Training local surrogate to explain individual predictions

Principle:

- Choose an instance
- Create local perturbations of the instance $\rightarrow X$
- Get the predictions of the black-box model on $X \rightarrow Y$
- Train a surrogate model on the (X,Y) dataset
(Using a weighted objective function according to the proximity with the original instance)

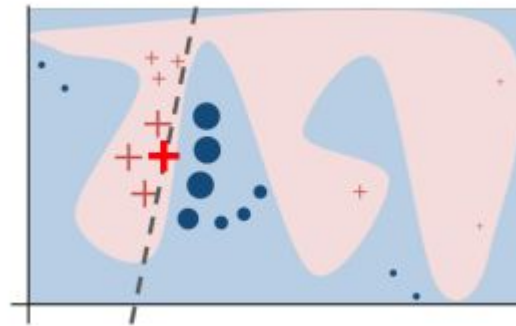


"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

[Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin](#)

Local interpretable model-agnostic explanations (LIME)

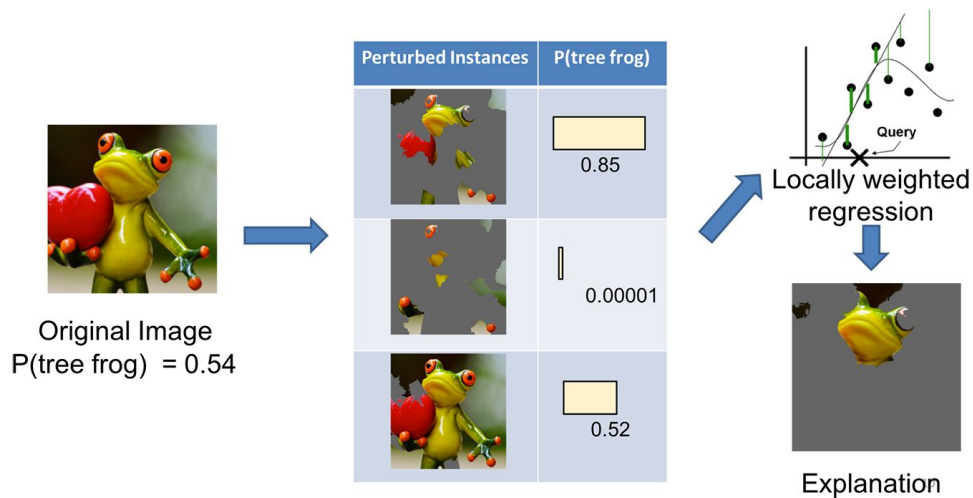
- How to generate perturbations?
- How to compute the distance to the original instance?
- Use several surrogates
- Fidelity measure
- Instability of the explanations



"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

[Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin](#)

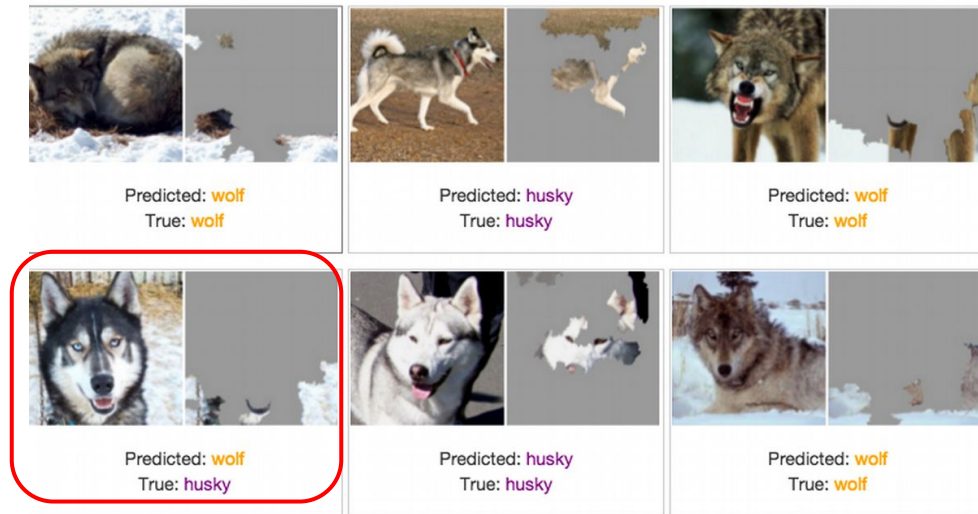
Local Surrogate (LIME)



"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

[Marco Tulio Ribeiro](#), [Sameer Singh](#), [Carlos Guestrin](#)

Local Surrogate (LIME)



"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

[Marco Tulio Ribeiro](#), [Sameer Singh](#), [Carlos Guestrin](#)

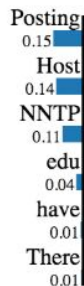
Local Surrogate (LIME)

Prediction probabilities



atheism

christian



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

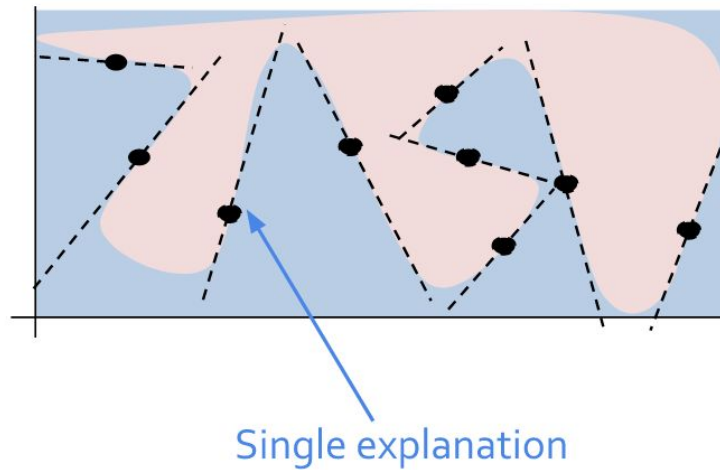
Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

[Marco Tulio Ribeiro](#), [Sameer Singh](#), [Carlos Guestrin](#)

SP-LIME



"Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016)

[Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin](#)

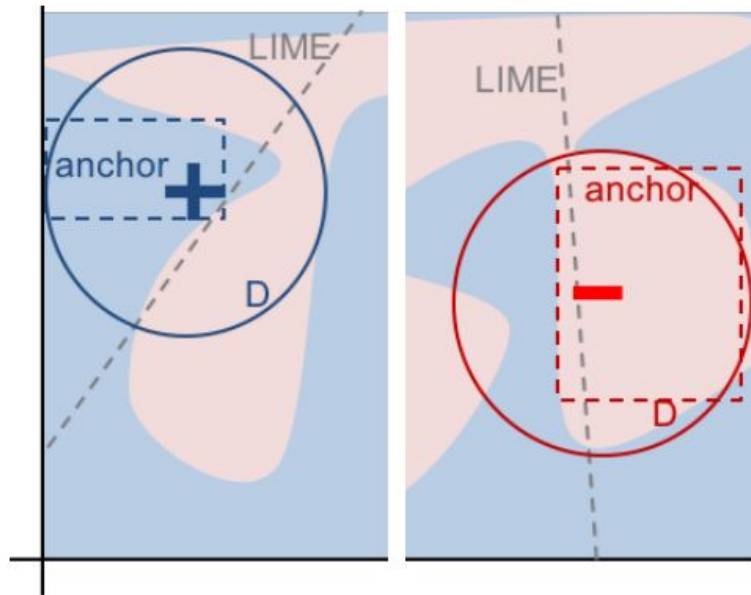
Scoped Rules (Anchors)

- Local explanation
- Set of IF-THEN rules (Anchors)
- Precision
- Coverage
- Multi-armed bandit KL-LUCB algorithm

IF HEIGHT < 60cm
AND WIDTH < 70cm
THEN PREDICT Class = Cat
WITH PRECISION 97% AND COVERAGE 15%

Anchors: High-Precision Model-Agnostic Explanations (2018)

[Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin](#)



Scoped Rules (Anchors)

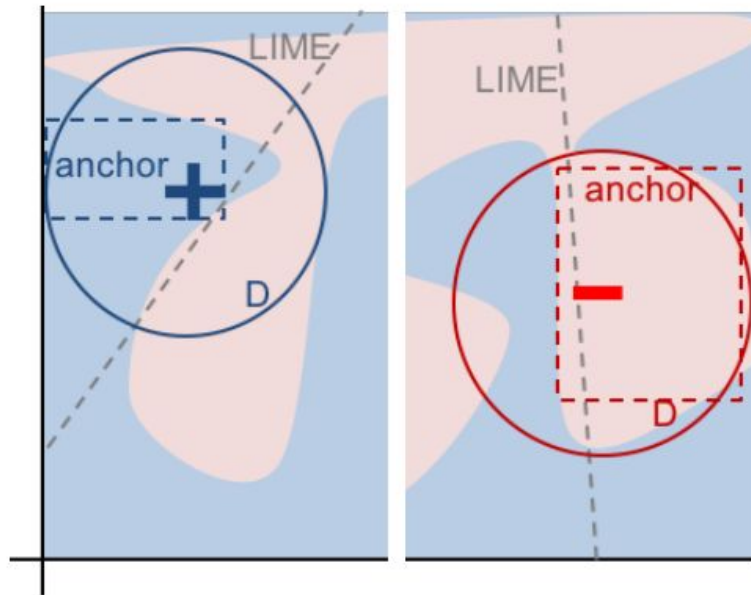
- Local explanation
- Set of IF-THEN rules (Anchors)
- Precision
- Coverage

IF HEIGHT < 60cm
AND WIDTH < 70cm
THEN PREDICT Class = Cat

With Precision=95%
And Coverage=35%

Anchors: High-Precision Model-Agnostic Explanations (2018)

[Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin](#)





SHAP (SHapley Additive exPlanations)

Shapley values:

- Game theory
- Collaborative multi-player game
- Fair repartition of a collaborative game winnings
- Measure individual contributions

A Unified Approach to Interpreting Model Predictions (2017)

[SM Lundberg, SI Lee](#)



SHAP (SHapley Additive exPlanations)

- Features = Players
- Winnings = Model's predictions
- Coalitions
- Local instance to explain
- Local perturbations of coalitions
- Estimate 'winnings' with a linear model
- Proximity based on SHAP kernel:

$$\pi_x(x') = \frac{(M-1)}{\binom{M}{|x'|} |x'| (M - |x'|)}$$

X	X ₁ =7.9	X ₂ =10.8	...	X _M =1.5
Coalition	1	1	...	1

X'	X ₁ =7.9	X ₂ =9.9	...	X _M =2.4
Coalition	1	0	...	0



Counterfactual Explanations

- How to change the model's prediction on a specific instance
- Find the smallest change that changes the prediction to a predefined output.

Example: Loan application

Instance: (male, 30 years old, salary 1500€, 5 years professional experience, 2 kids, divorced,) -> rejected

-> Increase salary to 2000 €

-> Increase professional experience to 6 years

Counterfactual explanation desired property:

- Be as similar as possible to the original instance
- Change as few features as possible
- A plausible example

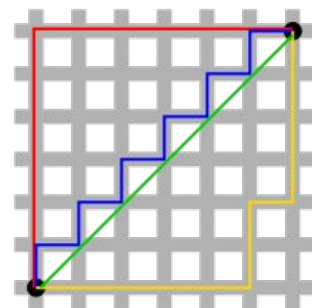
How to generate counterfactuals?

Counterfactual Explanations

Minimize: $L(x, x', y', \lambda) = \lambda \cdot \left(\hat{f}(x') - y' \right)^2 + d(x, x')$

Using scaled Manhattan distance:
$$d(x, x') = \sum_{j=1}^p \frac{|x_j - x'_j|}{MAD_j}$$

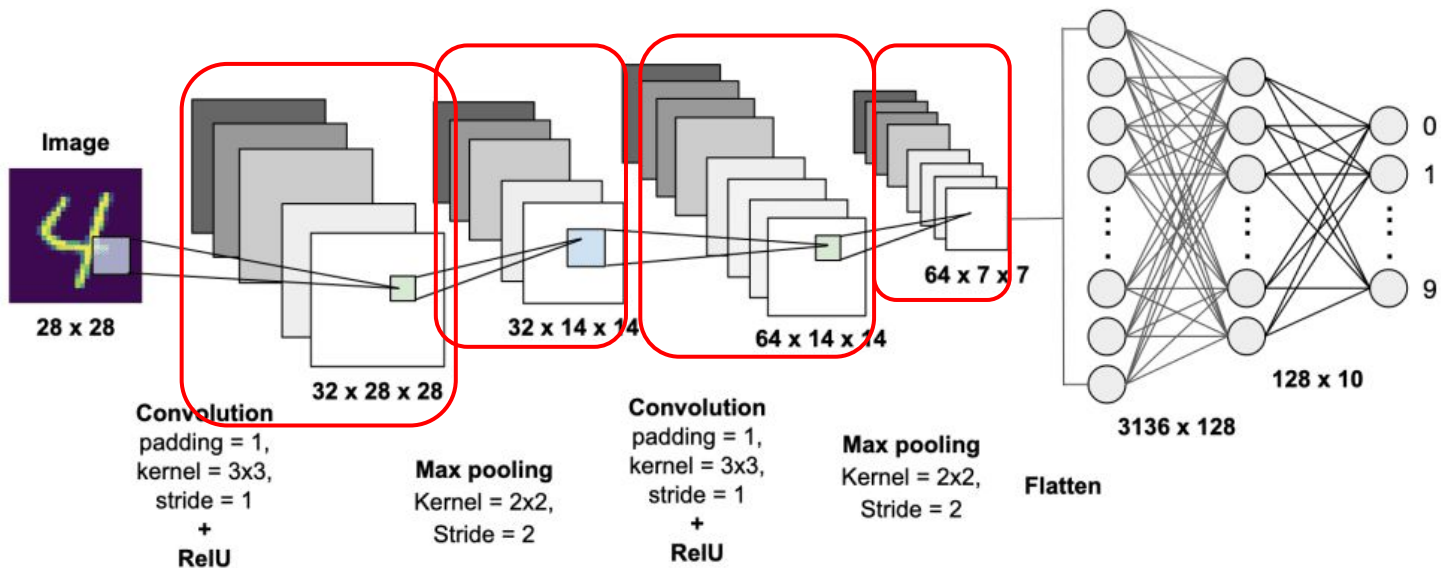
With median absolute deviation: $MAD_j = \text{median}_{i \in \{1, \dots, n\}} (|x_{i,j} - \text{median}_{l \in \{1, \dots, n\}} (x_{l,j})|)$



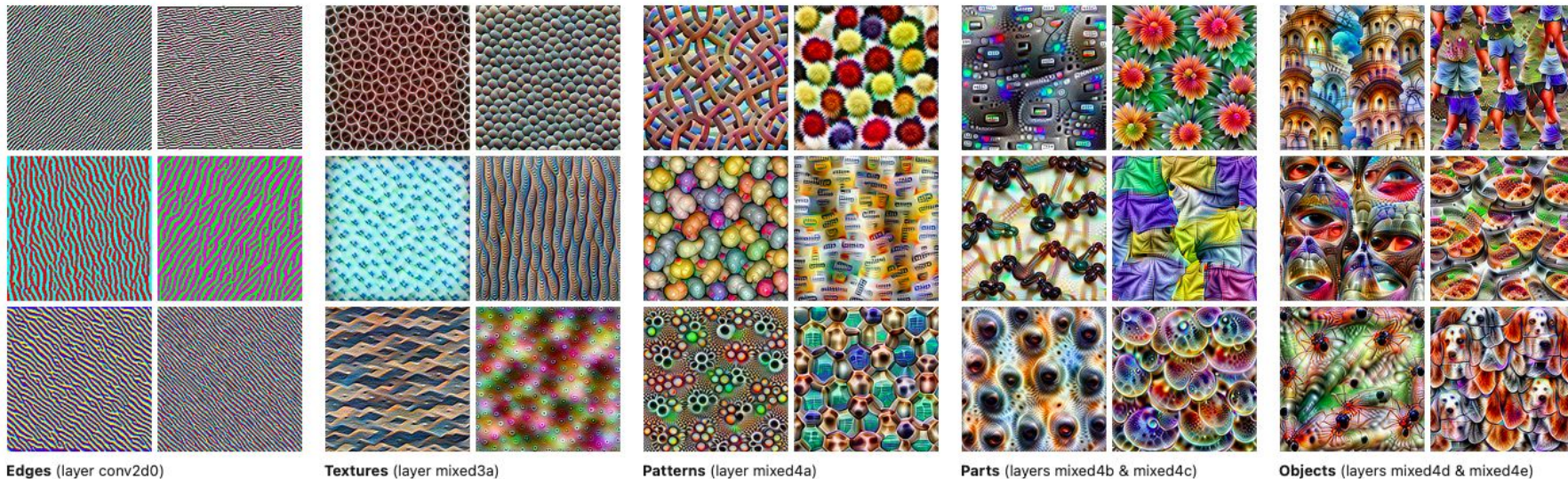


Model-specific methods (Neural networks)

Feature visualization



Feature visualization



Feature Visualization How neural networks build up their understanding of images

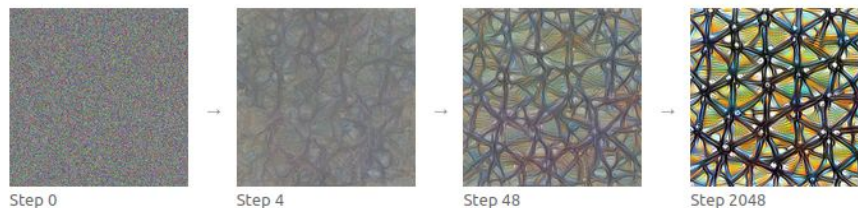
[Chris Olah](#), [Alexander Mordvintsev](#) and [Ludwig Schubert](#)

Feature visualization

- Neuron
- Layer Channel
- Layer

For a fixed network:

- Find the image that would maximize unit output using



Feature Visualization How neural networks build up their understanding of images

[Chris Olah](#), [Alexander Mordvintsev](#) and [Ludwig Schubert](#)



Network Dissection

Link feature map channels with human concepts

Principle:

- Define a large set of concepts
- Find units responding to these pre-defined concepts
- Measure the Activation-concept alignment

Network Dissection: Quantifying Interpretability of Deep Visual Representations (2017)

[David Bau](#), [Bolei Zhou](#), [Aditya Khosla](#), [Aude Oliva](#), [Antonio Torralba](#)

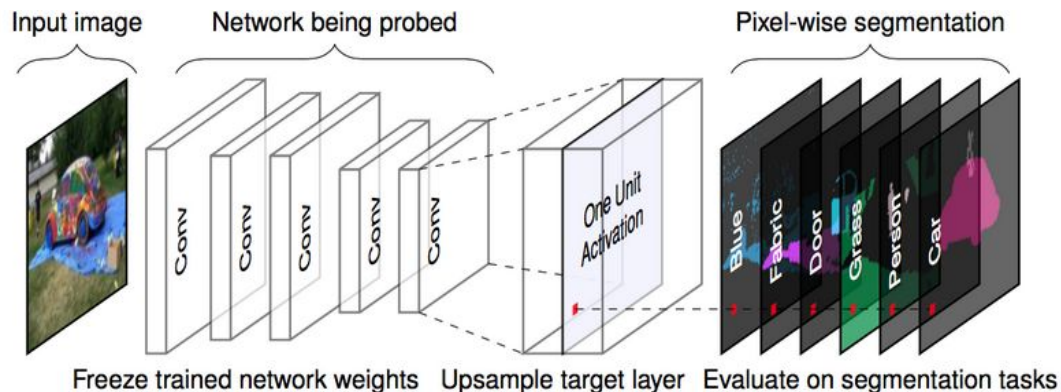
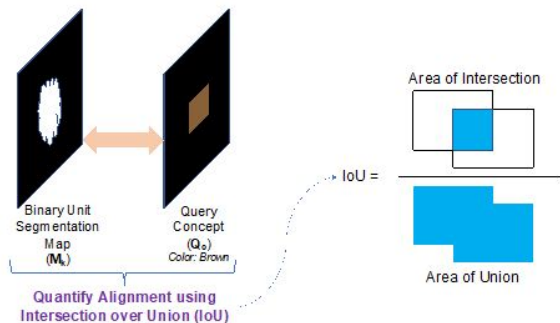
Network Dissection



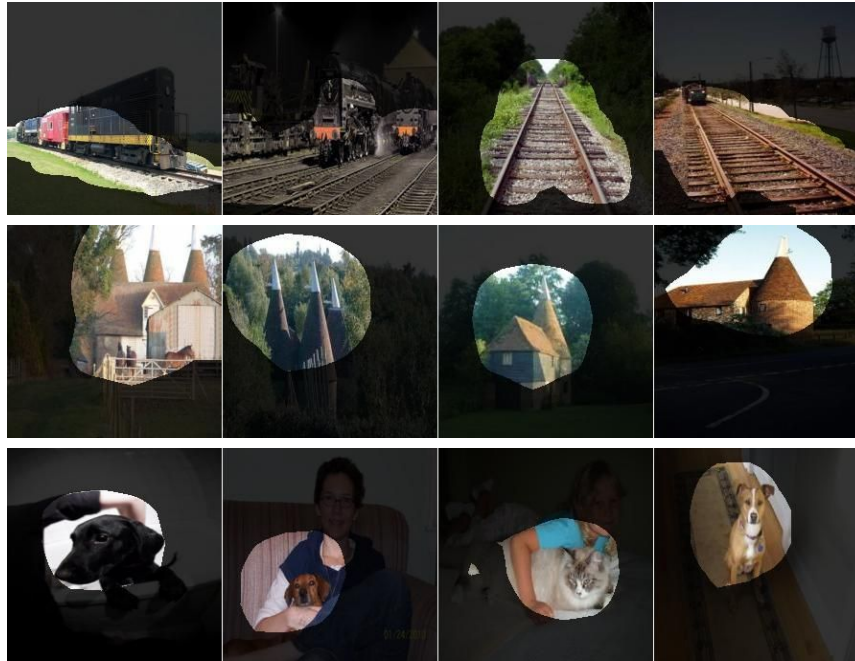
Network Dissection

For every channel:

- Compute the distribution of the feature map pixels output on the entire concept dataset
- Compute the 0.995-quantile
- For every image create a mask of feature map pixels > 0.995 -quantile
- Scale the feature map output
- Compute IoU



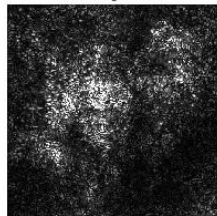
Network Dissection



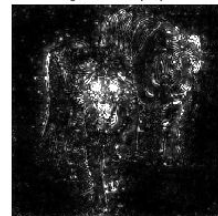
Saliency Maps



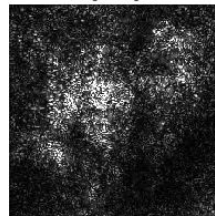
vanilla gradient



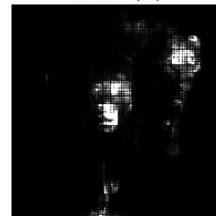
guided backprop



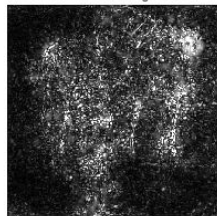
integrated grad



visual backprop



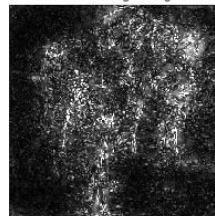
smoothed vanilla gradient



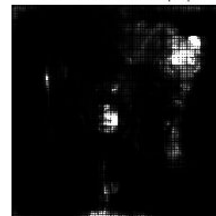
smoothed guided backprop



smoothed integrated grad



smoothed visual backprop





Saliency Maps

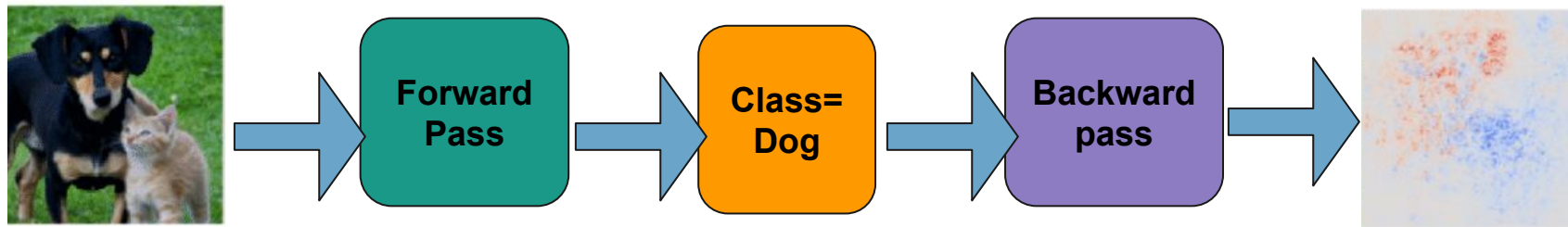
Visual attribution methods

Perturbations

Gradients

Activations

Vanilla gradient Saliency Maps



Deep inside convolutional networks: Visualising image classification models and saliency maps.”(2013)

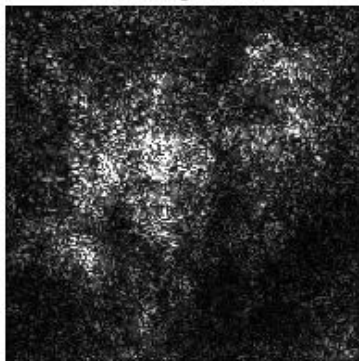
[Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman](#)

Gradient Saliency Maps

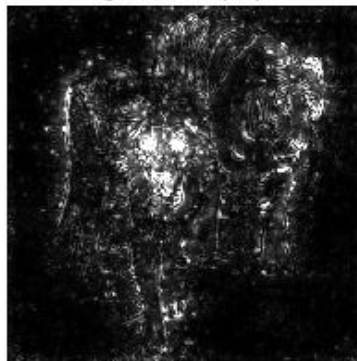
- Guided Backpropagation:
Striving for Simplicity: The All Convolutional Net (2014)
[Jost Tobias Springenberg](#), [Alexey Dosovitskiy](#), [Thomas Brox](#), [Martin Riedmiller](#)
- Smooth Grad:
SmoothGrad: removing noise by adding noise (2017)
[Daniel Smilkov](#), [Nikhil Thorat](#), [Been Kim](#), [Fernanda Viégas](#), [Martin Wattenberg](#)



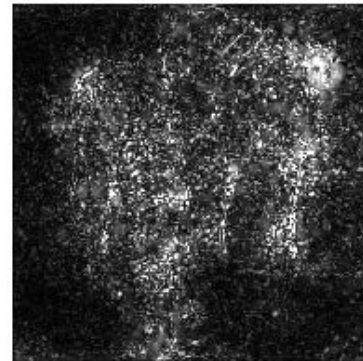
vanilla gradient



guided backprop



smoothed vanilla gradient





Saliency Maps

Visual attribution methods

Perturbations

Gradients

Activations

Grad-cam

- Gradient-weighted Class Activation Map
- Gradient back-propagated to the last convolutional layer
- Produce coarse localization map



Original Image



Grad-CAM 'Cat'



Grad-CAM 'Dog'

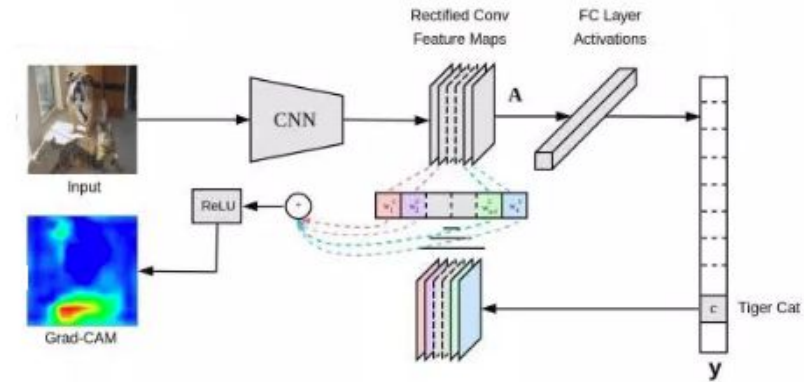
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

[Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra](#)

Grad-cam

Principle:

- Compute feature maps A^k with forward pass



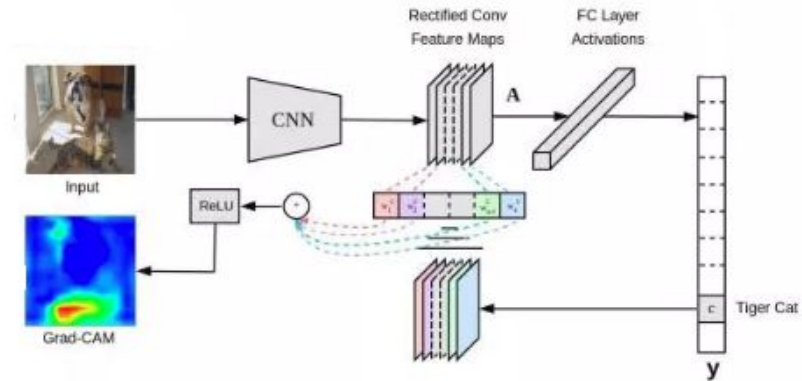
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

[Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra](#)

Grad-cam

Principle:

- Compute feature maps A^k with forward pass
- Backpropagate gradients of class c to the last convolutional layer



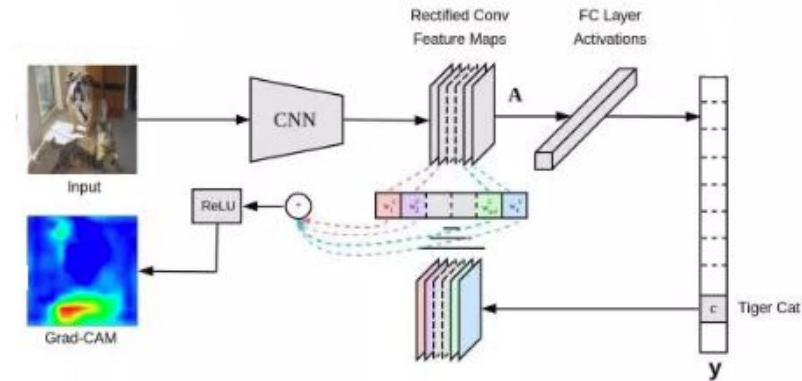
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

[Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra](#)

Grad-cam

Principle:

- Compute feature maps A^k with forward pass
- Backpropagate gradients of class C to the last convolutional layer
- Average gradients : $\alpha_k^c = \overbrace{\frac{1}{Z} \sum_i \sum_j}^{\text{global average pooling}} \frac{\partial y^c}{\partial A_{ij}^k}$



Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

[Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra](#)

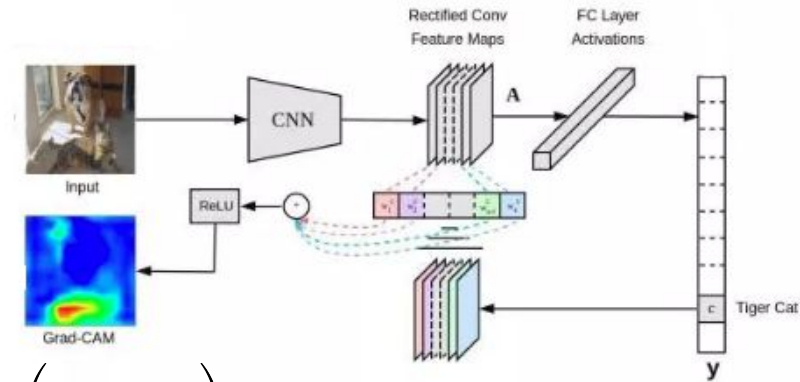
Grad-cam

Principle:

- Compute feature maps A^k with forward pass
- Backpropagate gradients of class C to the last convolutional layer

- Average gradients : $\alpha_k^c = \overbrace{\frac{1}{Z} \sum_i \sum_j}^{\text{global average pooling}} \frac{\partial y^c}{\partial A_{ij}^k}$

- Compute Grad-CAM heatmap: $L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right)$



Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

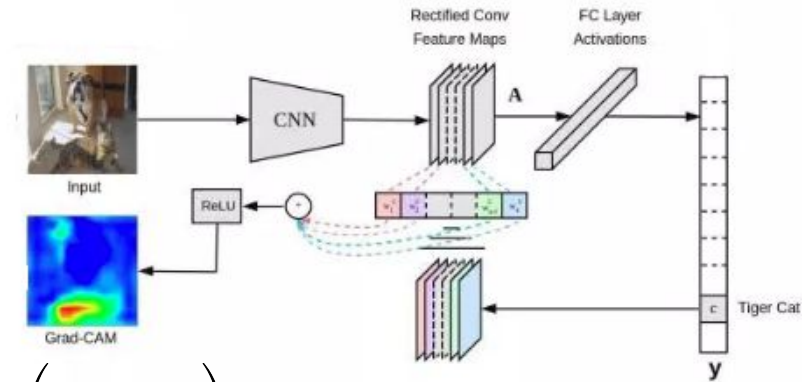
[Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra](#)

Grad-cam

Principle:

- Compute feature maps A^k with forward pass
- Backpropagate gradients of class C to the last convolutional layer
- Average gradients : $\alpha_k^c = \overbrace{\frac{1}{Z} \sum_i \sum_j}^{\text{global average pooling}} \frac{\partial y^c}{\partial A_{ij}^k}$

- Compute Grad-CAM heatmap: $L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right)$
- Upsample the heatmap



Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization (2016)

[Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra](#)

Recap:

Intrinsic:

- Linear models
- Decision trees
- Naive Bayes
- K-nn
- ...

Post Hoc:

Model agnostic

Global:

- PDP plots
- Permutation Feature Importance
- ...

Local:

- ICE plots
- LIME
- SHAP
- Anchors
- Counterfactuals

Model Specific

Global:

- Features visualization
- Network dissection

Local:

- Gradient saliency maps: (vanilla gradient, guided backprop, smooth-grad...)
- Activation saliency maps: Grad-CAM



References

Molnar, Christoph. “Interpretable machine learning. A Guide for Making Black Box Models Explainable”, 2019.
<https://christophm.github.io/interpretable-ml-book/>.

Ajay Thampi: [Interpretable AI](#)

- [Greedy function approximation: A gradient boosting machine.\(2001\)](#)
- [A Simple and Effective Model-Based Variable Importance Measure.\(2018\)](#)
- [Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation](#)
- [Random Forests \(2001\)](#)
- ["Why Should I Trust You?": Explaining the Predictions of Any Classifier \(2016\)](#)
- [Anchors: High-Precision Model-Agnostic Explanations \(2018\)](#)
- [A Unified Approach to Interpreting Model Predictions \(2017\)](#)
- [Counterfactual explanations without opening the black box: Automated decisions and the GDPR\(2017\)](#)
- [Feature VisualizationHow neural networks build up their understanding of images\(2018\)](#)
- [Network Dissection: Quantifying Interpretability of Deep Visual Representations \(2017\)](#)
- [Deep inside convolutional networks: Visualising image classification models and saliency maps.”\(2013\)](#)
- [Striving for Simplicity: The All Convolutional Net \(2014\)](#)
- [SmoothGrad: removing noise by adding noise \(2017\)](#)
- [Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization \(2017\)](#)