

# Machine Learning

## Lecture 18: Explainable AI (I)

Nevin L. Zhang

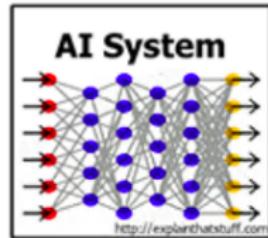
Department of Computer Science and Engineering  
The Hong Kong University of Science and Technology

This set of notes is based on internet resources and  
references listed at the end.

# Outline

- 1 Introduction
- 2 Pixel-Level Explanations
  - Pixel Sensitivity
  - Evaluation
- 3 Feature-Level Explanations
- 4 Concept-Level Explanations
- 5 Instance-Level Explanations

# What is XAI?



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

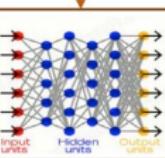
Cunning 2019

# What is XAI?

Today



Training Data



This is a cat  
( $p = .93$ )

Output



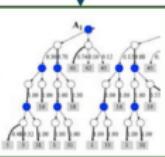
User with a Task

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow



Training Data



This is a cat:

- It has fur, whiskers, and claws.
- It has this feature:

Explanation Interface



User with a Task

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

Cunning 2019

# What is XAI?

Wikipedia:

- **Explainable AI (XAI)** refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts and users.
- It contrasts with the concept of the "black box" in machine learning where even their designers cannot explain why the AI arrived at a specific decision

# The Need for XAI: User Perspective

## Explanations foster trust and verifiability

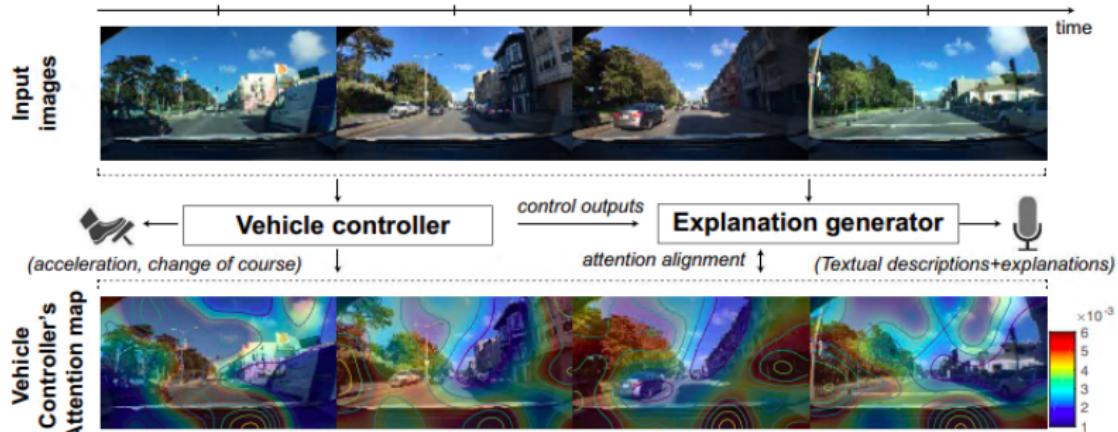
- Patients trust well-explained therapy.
- Doctors trust well-explained suggestions.



"Honey, drink the medicine."  
Man died later of poison.

# The Need for XAI: User Perspective

**Explanations foster trust and verifiability**



Example of textual descriptions + explanations:

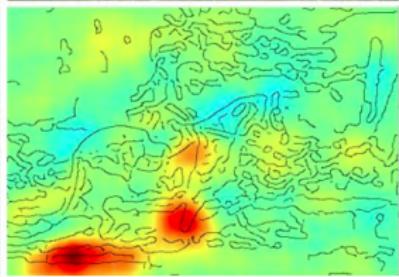
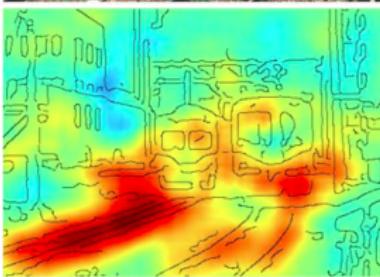
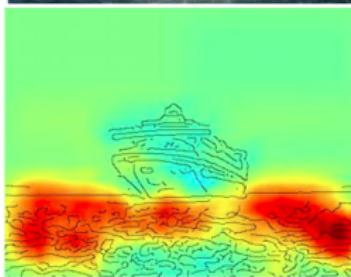
**Ours:** "The car is driving forward + because there are no other cars in its lane"

**Human annotator:** "The car heads down the street + because the street is clear."

Kim et al. 2018

# The Need for XAI: ML Expert Perspective

**Explanations help to determine if predication is based on the wrong reason (Clever Hans)**



Samek (2019)

# The Need for XAI: ML Expert Perspective

**Explanations help to determine if predication is based on the wrong reason (Clever Hans)**

Prediction probabilities



atheism

Posting  
0.15  
Host  
0.14  
NNTP  
0.11  
edu  
0.04  
have  
0.01  
There  
0.01

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.

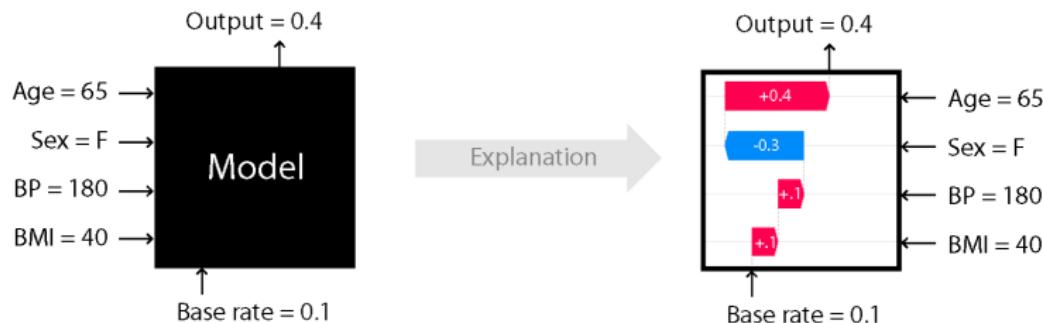
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

Rebeiro (2016)

# The Need for XAI: Society Perspective

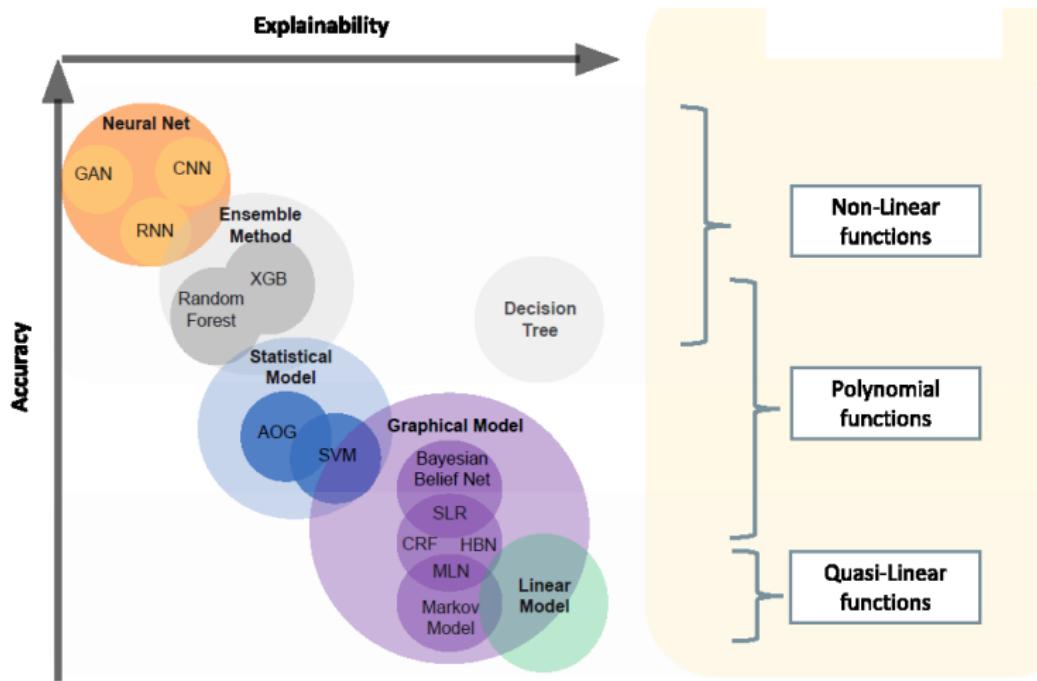
**Explanations are required by regulations for fairness and accountability**

- The EU's General Data Protection Regulation (GDPR) confers a right of explanation for all individuals to obtain **meaningful explanations of the logic involved** for automated decision making.

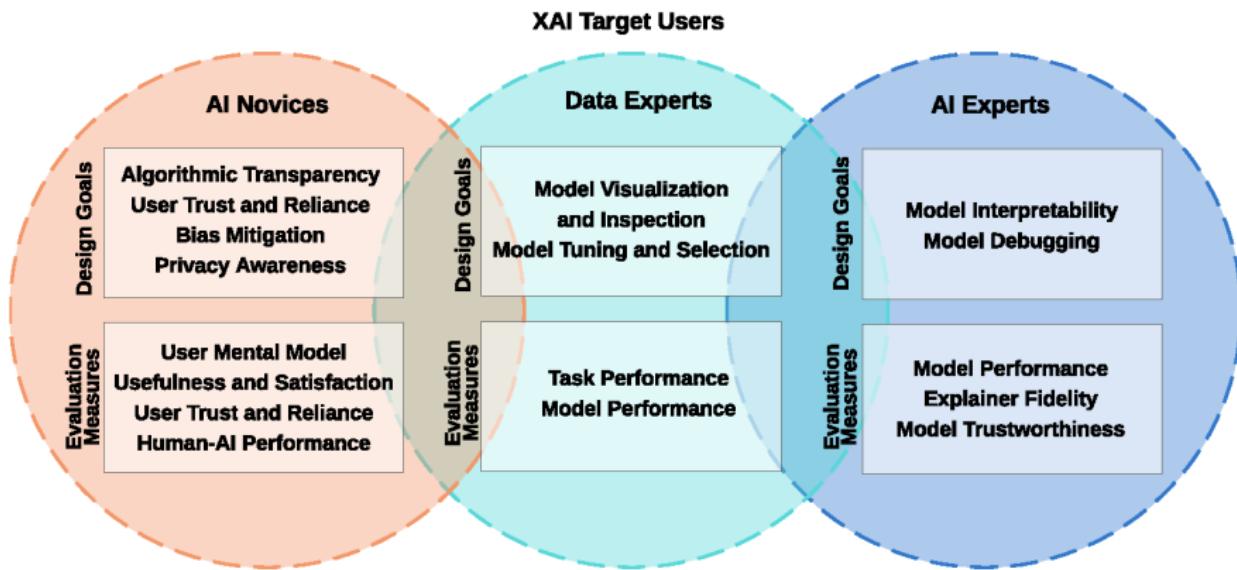


Lundberg (2019): Explaining interest rate of loan.

# The Interpretability and Accuracy Tradeoff



# Target Users of XAI



Mosheni et al. 2020

# Models to be explained

- **Image classifiers**
- **Tabular data classifiers**
- Large language models
- Reinforcement learning models
- Clustering algorithms
- ...

An XAI method is typically applicable to multiple models. We will focus on two tasks, image classification and tabular data classification.

# Types of Explanations

## Local vs Global explanations:

- Local XAI: Explains one particular prediction made by a model.
- Global XAI: Explains general behaviour of a model.

## Model-specific or model-agnostic:

- Model-agnostic XAI: Treats models as black-box.
- Model-specific XAI: Depends on the type of selected model

## Ante Hoc. vs Post Hoc.:

- Ante Hoc. XAI: Learn models that are interpretable.
- Post Hoc. XAI: Interpret models that are not interpretable by themselves.

# Outline

1 Introduction

2 Pixel-Level Explanations

- Pixel Sensitivity
- Evaluation

3 Feature-Level Explanations

4 Concept-Level Explanations

5 Instance-Level Explanations

# The Setup

$$\underbrace{[\![ \dots [\![} \xrightarrow{\bar{x}} [\![ \dots [\![ \xrightarrow{\bar{h}} \underbrace{[\![ \dots [\![}_{\bar{z}} \xrightarrow{\bar{w}^\top \bar{h} + \bar{b}} p(\bar{y} | \bar{z})$$

- An image  $\mathbf{x} = (x_1, \dots, x_D)^\top$  is fed to a DNN to produce a latent feature vector  $\mathbf{h}$ .
- An affine transformation is performed on  $\mathbf{h}$  to get a **logit vector**  $\mathbf{z} = (z_1, \dots, z_C)$ , which is used to define a probability distributions over the classes via softmax.
- Question: How important is a pixel  $x_i$  to the score  $z_c(\mathbf{x})$ ?
  - **Sensitivity:** How sensitive is the score  $z_c(\mathbf{x})$  to changes in  $x_i$ ?
  - **Attribution:** How much does  $x_i$  contribution to the score  $z_c(\mathbf{x})$ ?

# The Case of Linear Model

Let  $\mathbf{w}_c = (w_{c1}, \dots, w_{cD})$ . Suppose  $\mathbf{h} = \mathbf{x}$ . Then,

$$z_c = \mathbf{w}_c^\top \mathbf{x} + b_c,$$

- The sensitivity of  $x_i$  to  $z_c(\mathbf{x})$  is determined by

$$w_{ci} = \frac{\partial z_c}{\partial x_i}.$$

- The contribution of  $x_i$  to  $z_c(\mathbf{x})$  is

$$x_i w_{ci} = x_i \frac{\partial z_c}{\partial x_i}.$$

In words, it is  $\text{input} \times \text{gradient}$ .

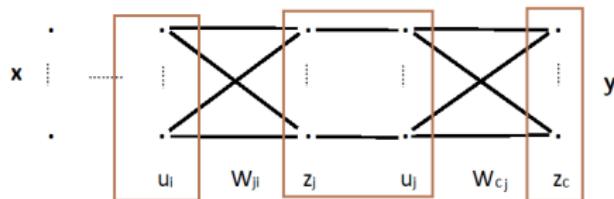
# Saliency Map

- In the general case, we can still determine the sensitivity of  $z_c$  to  $x_i$  using  $\frac{\partial z_c}{\partial x_i}$ .
- **Saliency Map** (*Simonyan et al. 2013*) is a way to visualize the gradients w.r.t all the pixel.



A saliency map highlights the pixels that have the largest impact on class score **if perturbed**.

# Computation of Saliency Map

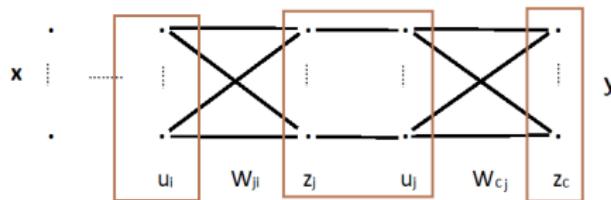


$$z_c = \sum_j u_j W_{cj}, \quad u_j = g(z_j), \quad z_j = \sum_i u_i W_{ji} \quad (\text{Bias ignored for simplicity.})$$

**Backpropagation during training revisited:** Training example  $(\mathbf{x}, y)$

- **Forward propagation:** Compute activations,  $\mathbf{z}$ , and loss  $L(\mathbf{z}, Y)$ .
- **Backward propagation:** (Purpose:  $\frac{\partial L}{\partial W_{ji}}$ )
  - For output unit (each class)  $c$ :  $\delta_c \leftarrow \frac{\partial L}{\partial z_c}$
  - For each unit  $j$  on second-last layer:  $\delta_j \leftarrow \frac{\partial u_j}{\partial z_j} \sum_c W_{cj} \delta_c$
  - For each unit  $i$  on third-last layer:  $\delta_i \leftarrow \frac{\partial u_i}{\partial z_i} \sum_j W_{ji} \delta_j$
- $\frac{\partial L}{\partial W_{ji}} \leftarrow u_i \delta_j$ , etc

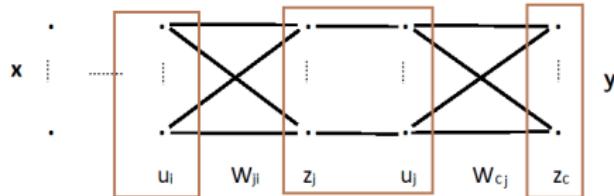
# Computation of Saliency Map



Explaining the score  $z_c(x)$  for input  $x$ :

- **Forward propagation:** Compute activations and  $z$ .
- **Backward propagation:** (Purpose:  $\frac{\partial z_c}{\partial x_i}$ )
  - For each unit  $j$  on second-last layer:  $\frac{\partial z_c}{\partial u_j} \leftarrow W_{cj}$
  - For each unit  $i$  on third-last layer :  $\frac{\partial z_c}{\partial u_i} \leftarrow \sum_j W_{ji} \frac{\partial u_j}{\partial z_j} \frac{\partial z_c}{\partial u_j}$
  - ...

# Computation of Saliency Map



**Notes on Backprop:**  $\frac{\partial z_c}{\partial u_i} \leftarrow \sum_j W_{ji} \frac{\partial u_j}{\partial z_j} \frac{\partial z_c}{\partial u_j}$  if  $u_j = g(z_j)$ ,  $z_j = \sum_i u_i W_{ji}$

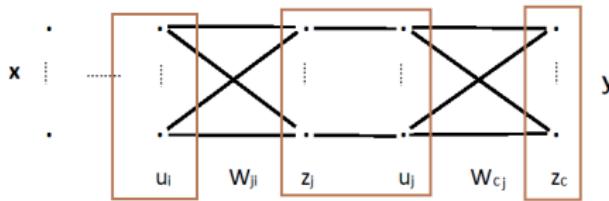
- In the case of max pooling:  $u_j = \max_i u_i$

$$\frac{\partial z_c}{\partial u_i} \leftarrow \begin{cases} \frac{\partial z_c}{\partial u_j} & \text{if } u_i = u_j \\ 0 & \text{if } u_i \neq u_j \end{cases}$$

$\frac{\partial z_c}{\partial u_j}$  is backpropagated along one of the connections. Forward activations act like "switches" for backprop.

- If unit  $j$  is a ReLU unit and  $z_j < 0$ , then  $u_j = 0$  and  $\frac{\partial u_j}{\partial z_j} = 0$ , and hence  $\frac{\partial z_c}{\partial u_j}$  is not backpropagated. Forward activations act like gates for backprop.

# Guided Backpropagation



**Vanilla Backprop:**  $\frac{\partial z_c}{\partial u_i} \leftarrow \sum_j W_{ji} \frac{\partial u_j}{\partial z_j} \frac{\partial z_c}{\partial u_j}$

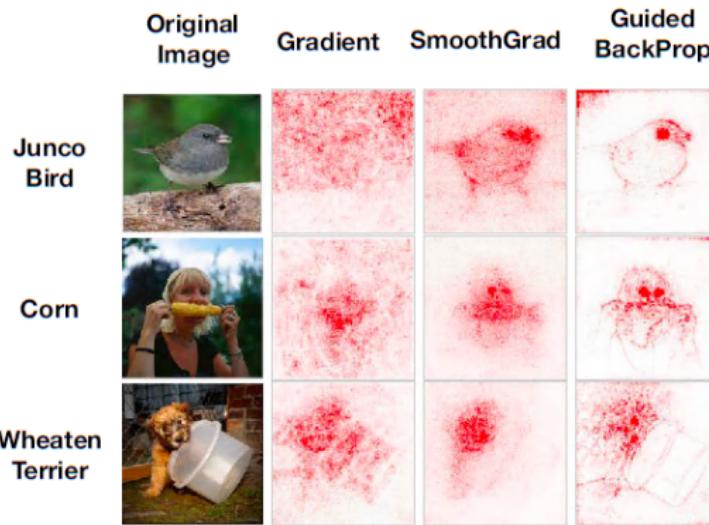
- If the gradient  $\frac{\partial z_c}{\partial u_j} < 0$ , then  $u_j (>= 0)$  contributes to  $z_c$  negatively.
- If we want to find the pixels the contribution to  $z_c$  positively, we can ignore negative gradients.
- This gives rise to **Guided Backpropagation** (Springenberg *et al.* 2014):

$$\frac{\partial z_c}{\partial u_i} \leftarrow \sum_j W_{ji} \frac{\partial u_j}{\partial z_j} \text{ReLU}\left(\frac{\partial z_c}{\partial u_j}\right).$$

ReLU is applied to gradients at all layers.

# Guided Backpropagation

In general, Guided Backprop produces sharper saliency maps than Vanilla Gradient

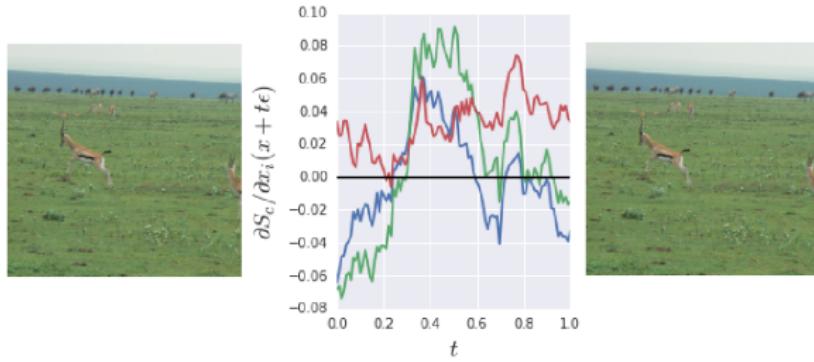


Adebayo *et al.* (2018)

It is a combination of Vanilla Gradient and **deconvNet** (Zeiler *et al.* 2014 ), which maps a neuron activation back to the input pixel space, showing what input pattern originally caused a given activation in the feature maps.

# Smoothgrad (Smilkov *et al.* 2017)

- Sensitivity maps are often visually noisy (see previous slide).
- One explanation:  $\frac{\partial z_c}{\partial x_i}$  may fluctuate sharply in response to small variations in input  $\mathbf{x}$ .



Two images  $\mathbf{x}$  and  $\mathbf{x} + \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, 0.01^2)$ . Middle plot shows  $\frac{\partial z_c}{\partial x_i}$  for  $\mathbf{x} + t\epsilon$  and a fixed pixel (divide by max over all pixels).

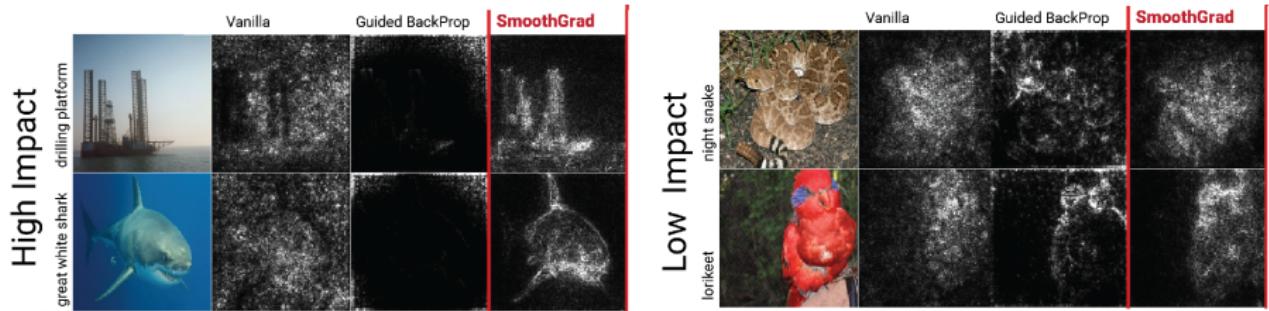
# Smoothgrad (Smilkov *et al.* 2017)

- For input  $\mathbf{x}$ , generate  $N$  copies by adding small perturbations:

$$\mathbf{x}_j = \mathbf{x} + \epsilon_j, \epsilon \sim \mathcal{N}(0, \sigma^2), j = 1, \dots, N.$$

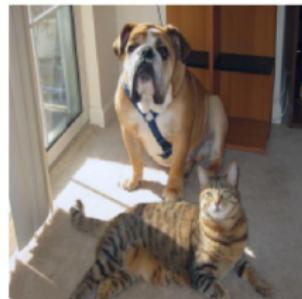
Take average of the corresponding gradients:

$$\text{smoothGrad}(z_c, x_i) = \frac{1}{N} \sum_{j=1}^N \frac{\partial z_c(\mathbf{x}_j)}{\partial x_i}$$



# Grad-CAM (Selvaraju *et al.* 2017)

- Unlike previous methods, **Grad-CAM (Gradient-weighted Class Activation Mapping)** is [class-discriminative](#): It localizes class in the image.



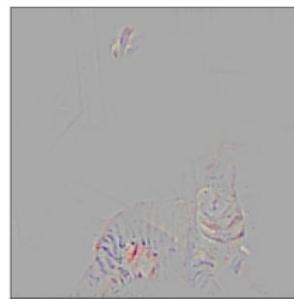
Original Image



Guided Backprop 'Cat'



Guided Backprop 'Dog'

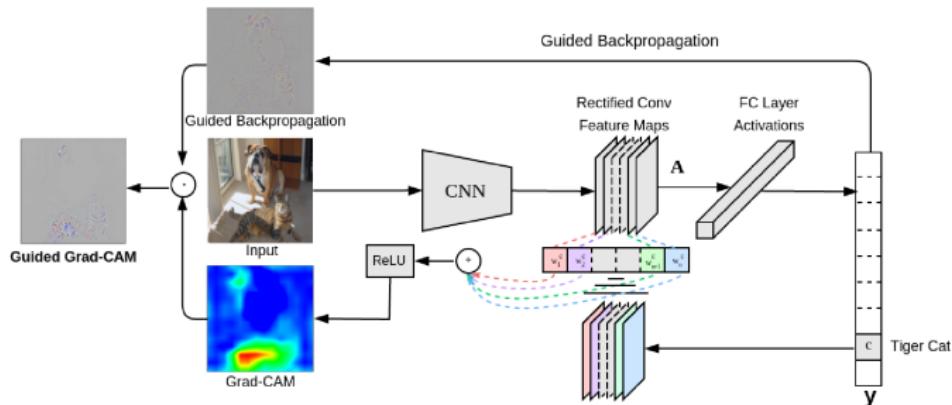


Guided Grad-CAM 'Cat'

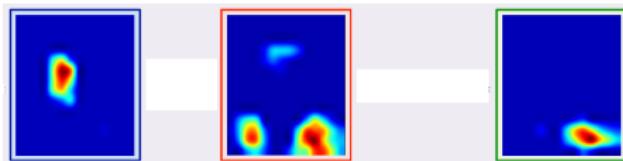


Guided Grad-CAM 'Dog'

# Grad-CAM (Selvaraju et al. 2017)



- Let  $A^k = [a_{ij}^k]$  be a **feature map** in the last convolutional layer for an input  $x$ . The activations are local.



# Grad-CAM (Selvaraju *et al.* 2017)

- The following quantity measures the “importance” of the feature map  $A^k$  is to the score  $z_c(\mathbf{x})$ :

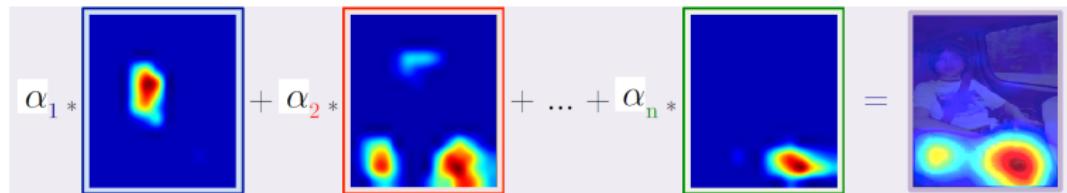
$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial z_c}{\partial a_{ij}^k},$$

where  $Z$  is the number of pixels in  $A^k$ . ( Global average pooling).

- The **Grad-CAM heatmap** is computed as follows:

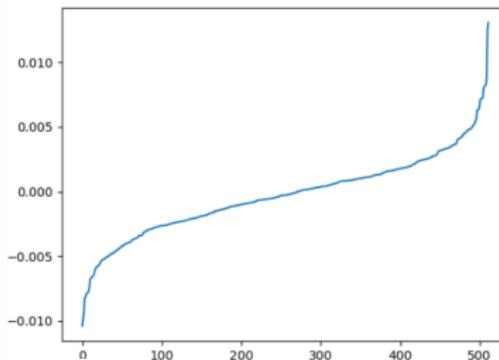
$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^K\right).$$

ReLU is used because we are only interested in features that have positive influence on  $z_c$ .



# Grad-CAM (Selvaraju *et al.* 2017)

Grad-CAM essentially combines only those feature maps that contribute positively to  $c$ , and hence is effective in localize it.



0: -0.010390018



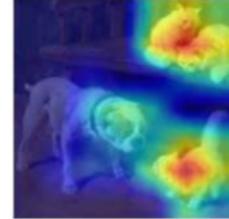
255: -0.00027679346



306: 0.00041336069



510: 0.012544246

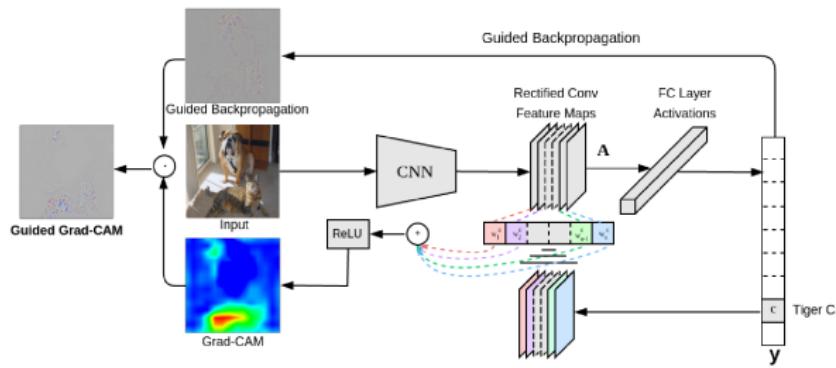
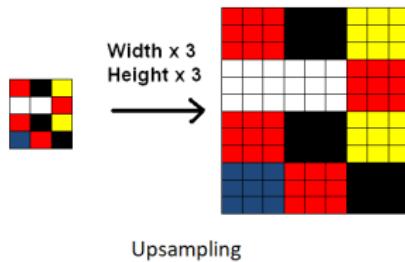


511: 0.013066023



# Guided Grad-CAM (Selvaraju *et al.* 2017)

- The Grad-CAM heatmap has smaller dimensions than the input.
- It is upsampled and multiplied with the saliency map by Guided BackProp.



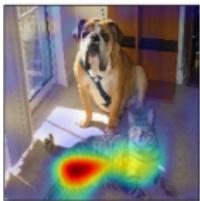
# Guided Grad-CAM (Selvaraju *et al.* 2017)



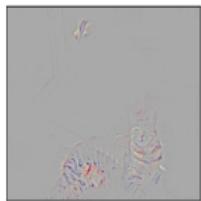
(a) Original Image



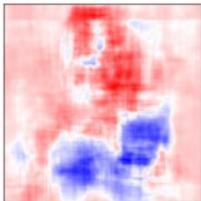
(b) Guided Backprop 'Cat'



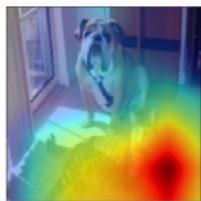
(c) Grad-CAM 'Cat'



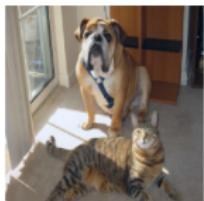
(d) Guided Grad-CAM 'Cat'



(e) Occlusion map 'Cat'



(f) ResNet Grad-CAM 'Cat'



(g) Original Image



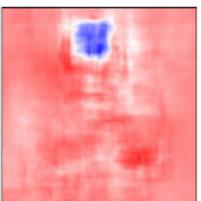
(h) Guided Backprop 'Dog'



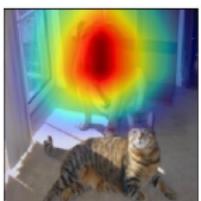
(i) Grad-CAM 'Dog'



(j) Guided Grad-CAM 'Dog'

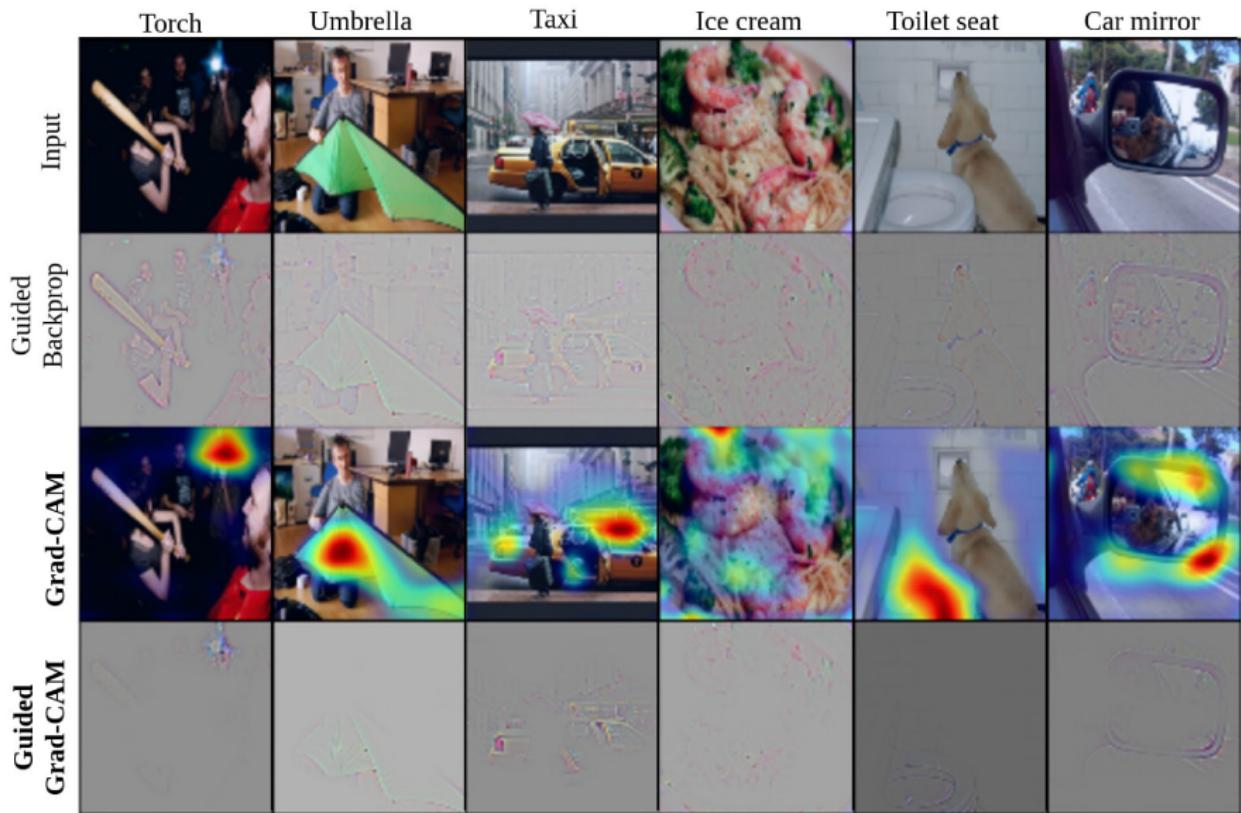


(k) Occlusion map 'Dog'



(l) ResNet Grad-CAM 'Dog'

# Guided Grad-CAM (Selvaraju *et al.* 2017)



# Guided Grad-CAM and Counterfactual Explanation

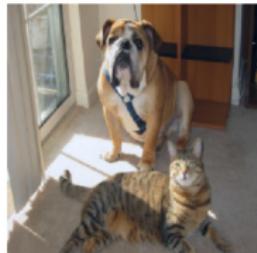
- To identify region that contributes **positively** to a classification:

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial z_c}{\partial a_{ij}^k}, \text{ReLU}\left(\sum_k \alpha_k^c A^K\right)$$

- To identify region that contributes **negatively** to a classification:

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} -\frac{\partial z_c}{\partial a_{ij}^k}, \text{ReLU}\left(\sum_k \alpha_k^c A^K\right)$$

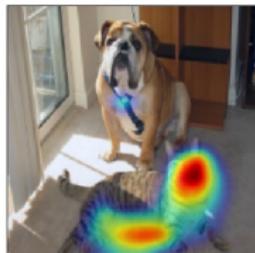
- Removing such negative region would make the classification more confident. The modified images are **counterfactual explanations**.



(a) Original Image



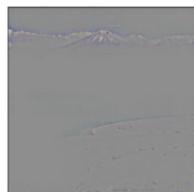
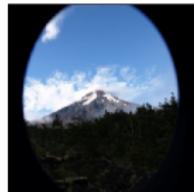
(b) Cat Counterfactual exp



(c) Dog Counterfactual exp

# Application of Guided Grad-CAM: Model Diagnosis

Seemingly unreasonable predictions have reasonable explanations.



Ground truth: volcano



Ground truth: volcano



Ground truth: beaker



Ground truth: coil



Predicted: sandbar



Predicted: car mirror



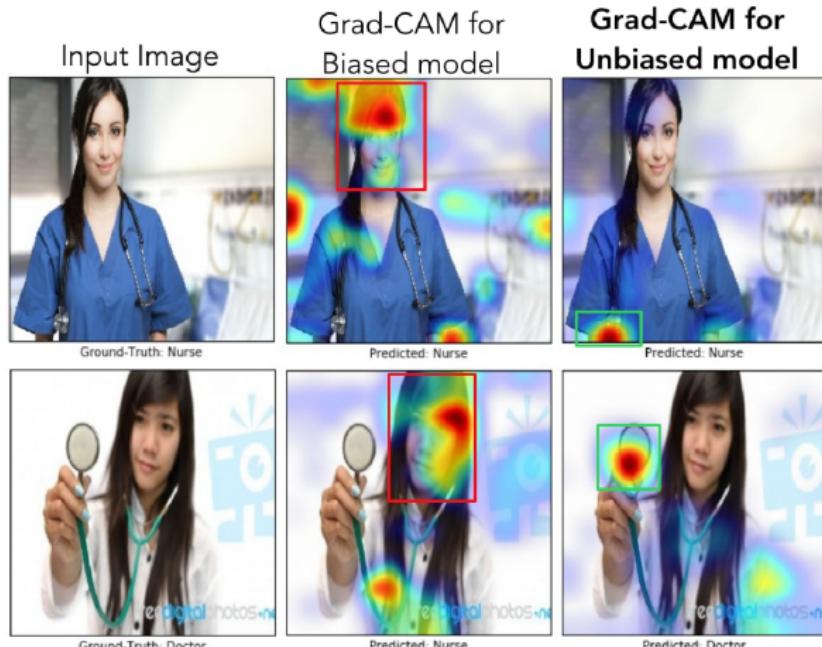
Predicted: syringe



Predicted: vine snake

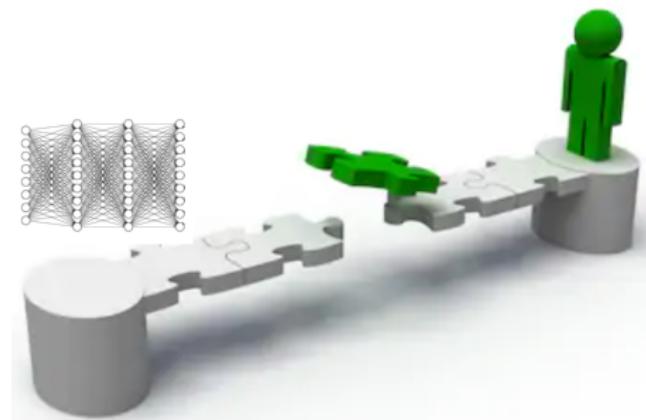
# Application of Guided Grad-CAM: Discovering Bias in Data

**Biased model** trained on model where gender strongly correlated with being doctor/nurse. **Unbiased model** trained on model where gender independent of being doctor/nurse.



# Faithfulness and Interpretability

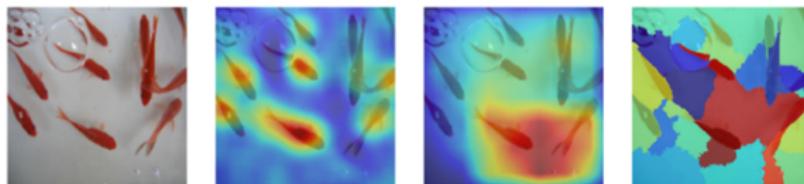
- XAI is a bridge between model and human.



- **Faithfulness:** How well it ties in with the **model**, revealing key evidence and reasoning that model uses for prediction.
- **Interpretability:** How well it is received by **human**, providing comprehensible and meaningful information, and allowing a good understanding of model behavior.

# Local Faithfulness

- If input pixels are deemed important, removing them should cause a drop in class probability for the given example (local).
- Local faithfulness via perturbation (Samek et al. 2016)
  - Sort pixels by importance
  - Perturb them one by one (replace them by random values)
  - Measure drop in probability of true class

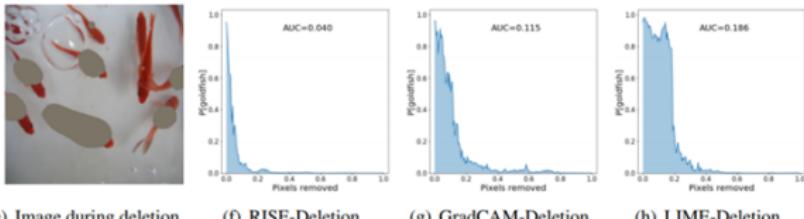


(a) Input

(b) RISE (ours)

(c) GradCAM

(d) LIME



(e) Image during deletion

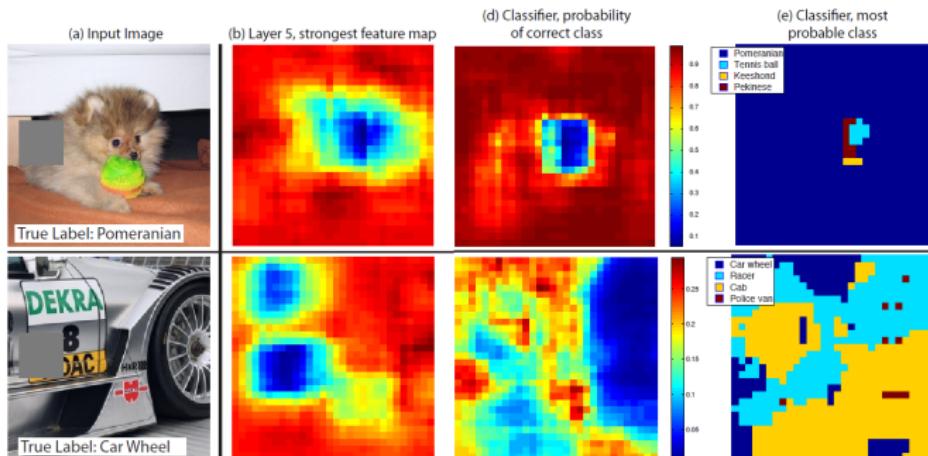
(f) RISE-Deletion

(g) GradCAM-Deletion

(h) LIME-Deletion

# Early XAI method: Occlusion Map (Zeiler et al. 2014)

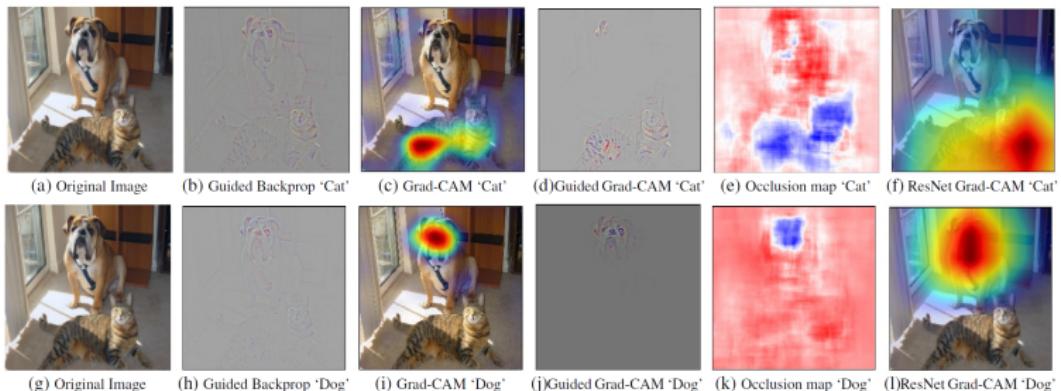
- **A model-agnostic explanation method:** Systematically occlude different portions of the input image with a grey square, and monitor the output of the classifier.



(e): the most probable label as a function of occluder position. E.g. in the 1st row, for most locations it is “pomeranian”, but if the dog’s face is obscured but not the ball, then it predicts “tennis ball”.

- Too expensive, but can be used as a baseline for evaluation.

# Local Faithfulness via Occlusion (Selvaraju *et al.* 2019)



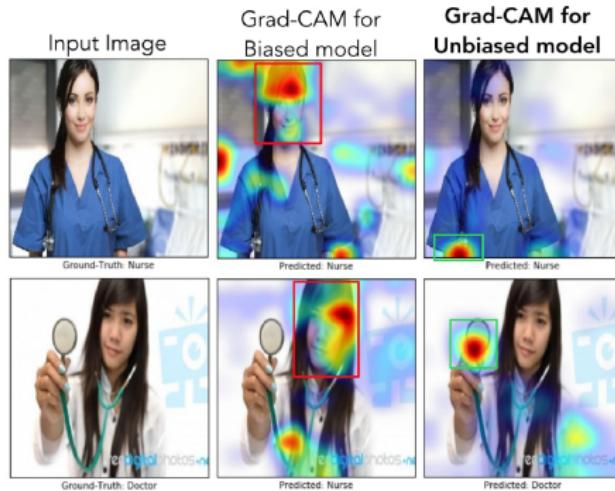
- Interestingly, patches which change the CNN score are also patches to which Grad-CAM and Guided Grad-CAM assign high intensity.
- Rank correlation ([https://en.wikipedia.org/wiki/Rank\\_correlation](https://en.wikipedia.org/wiki/Rank_correlation))

Grad-CAM	Guided Grad-CAM	Guided Backprop	c-MWP	CAM
0.254	0.261	0.168	0.220	0.208

Averaged over 2510 images in the PASCAL 2007 val set.

# Interpretability via Localization (Selvaraju *et al.* 2019)

- Given an image, we first obtain class predictions from our network and then generate Grad-CAM maps for each of the predicted classes and binarize them with a threshold of 15% of the max intensity.
- This results in connected segments of pixels and we draw a bounding box around the single largest segment.



# Interpretability via Localization (Selvaraju *et al.* 2019))

- The ImageNet location challenge has ground-truth bounding boxes, which are provided by **human**.
- Heatmap by **XAI** is interpretable if it matches the ground truth bounding boxes.

## ILSVRC Image Localization (LOC) Task



Correct



Bad localization



Bad classification



# Interpretability via Localization (Selvaraju *et al.* 2019)

- The ImageNet location challenge has ground-truth bounding boxes, which are provided by **human**.
- Heatmap by **XAI** is interpretable if it matches the ground truth bounding boxes.

		Classification		Localization	
		Top-1	Top-5	Top-1	Top-5
VGG-16	Backprop [51]	30.38	10.89	61.12	51.46
	c-MWP [58]	30.38	10.89	70.92	63.04
	Grad-CAM (ours)	30.38	10.89	<b>56.51</b>	46.41
	CAM [59]	33.40	12.20	57.20	<b>45.14</b>
AlexNet	c-MWP [58]	44.2	20.8	92.6	89.2
	Grad-CAM (ours)	44.2	20.8	68.3	56.6
GoogleNet	Grad-CAM (ours)	31.9	11.3	60.09	49.34
	CAM [59]	31.9	11.3	60.09	49.34

Table 1: Classification and localization error % on ILSVRC-15 val

# Interpretability via Human Studies Selvaraju *et al.* (2019)



**What do you see?**

**Your options:**

- Horse
- Person

How well can heatmap help user predict behavior of model?

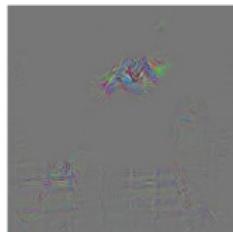
- Heatmaps shown to workers on Amazon Mechanical Turk.
- If **human** can correctly identify the class predicted by **model** from **heatmap**, then the **heatmap** is meaningful to **human** and faithful to the **model**.

Guided Grad-CAM	Guided Backprop	Deconv Grad-cam	Deconv
0.61	0.44	0.60	0.53

# Interpretability via Human Studies Selvaraju *et al.* (2019)

Both robots predicted: Person

Robot A based it's decision on



Robot B based it's decision on



Which robot is more reasonable?

- Robot A seems clearly more reasonable than robot B
- Robot A seems slightly more reasonable than robot B
- Both robots seem equally reasonable
- Robot B seems slightly more reasonable than robot A
- Robot B seems clearly more reasonable than robot A

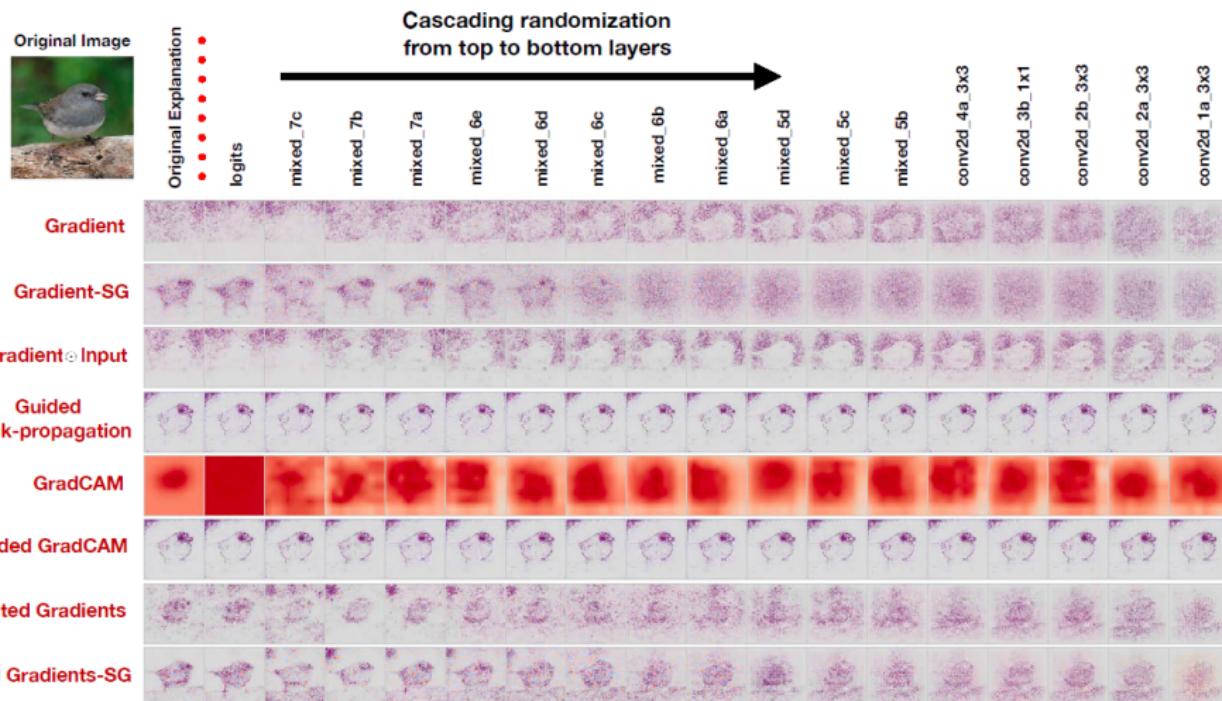
How well can heatmap help user tell the differences of two models?

- Both AlexNet and VGG-16 predict the same class. XAI method used to generate heatmaps for the two models. Turkers asked to rate which model is more **trust**worthy: -2, -1, 0, 1, 2.
- Results show that Guided Grad-Cam can help the turkers understand the behaviours of the models better.

Guided Grad-CAM	Guided Backprop
1.27	1.0

# Sanity Checks for Saliency Maps Adebayo *et al.* (2018)

Some saliency methods give similar results even when many weights are [randomly re-initialized](#). (Inception V3)



# Parameter Sensitivity Adebayo *et al.* (2018)

- **Reason:** When applied to a random convolution, saliency methods seem to act like edge detectors. (There is some theory for this.)
- **Implications:**
  - Confirmation bias: Human observer might think the edges are important to the model, but they are more or less independent of the model.
  - Explanations that do not depend on model parameters might still depend on the model architecture and thus provide some useful information about the prior incorporated in the model architecture.
  - However, in this case, the explanation method should only be used for tasks where we believe that knowledge of the model architecture on its own is sufficient for giving useful explanations.

# Outline

1 Introduction

2 Pixel-Level Explanations

- Pixel Sensitivity
- Evaluation

3 Feature-Level Explanations

4 Concept-Level Explanations

5 Instance-Level Explanations

# Outline

- 1 Introduction
- 2 Pixel-Level Explanations
  - Pixel Sensitivity
  - Evaluation
- 3 Feature-Level Explanations
- 4 Concept-Level Explanations
- 5 Instance-Level Explanations

# Outline

- 1 Introduction
- 2 Pixel-Level Explanations
  - Pixel Sensitivity
  - Evaluation
- 3 Feature-Level Explanations
- 4 Concept-Level Explanations
- 5 Instance-Level Explanations

# References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In Advances in Neural Information Processing Systems (pp. 9505-9515).
- Alvarez-Melis, David, and Tommi S. Jaakkola. "On the robustness of interpretability methods." arXiv preprint arXiv:1806.08049 (2018).
- Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015).
- David Cunning (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. <https://www.youtube.com/watch?v=nX-4ClxWXYg>
- F. Lecue *et al.* (2020). XAI Tutorial, AAAI.
- S. Lundberg (2019). <https://shap.readthedocs.io/en/latest/>.
- Kim, J., Rohrbach, A., Darrell, T., Canny, J., & Akata, Z. (2018). Textual explanations for self-driving vehicles. In Proceedings of the European conference on computer vision (ECCV) (pp. 563-578).
- Christoph Molnar (2020): Interpretable Machine Learning.  
<https://christophm.github.io/interpretable-ml-book/index.html>.
- Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. arXiv preprint arXiv:1811.11839v4.

# References

- M. T. Ribeiro (2016). <https://homes.cs.washington.edu/~marcotcr/blog/lime/>
- Samek, Wojciech, et al. "Evaluating the visualization of what a deep neural network has learned." *IEEE transactions on neural networks and learning systems* 28.11 (2016): 2660-2673.
- Samek, W. (2019). Explainable AI: interpreting, explaining and visualizing deep learning (Vol. 11700). Springer Nature.
- W. Samek (2019): Meta-Explanations, Interpretable Clustering & Other Recent Developments. [http://xai.unist.ac.kr/static/img/event/ICCV\\_2019\\_VXAI\\_Samek\\_Talk.pdf](http://xai.unist.ac.kr/static/img/event/ICCV_2019_VXAI_Samek_Talk.pdf)
- Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

# References

- Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825.
- Shrikumar, Avanti, et al. "DeepLIFT: Learning important features through propagating activation differences." arXiv preprint <https://arxiv.org/pdf/1704.02685.pdf> (2019).
- Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806.
- Sundararajan, M., Taly, A., & Yan, Q. (2017, August). Axiomatic attribution for deep networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 3319-3328). JMLR. org.
- Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.
- Zhang, Jianming, et al. "Top-down neural attention by excitation backprop." International Journal of Computer Vision 126.10 (2018): 1084-1102.