Contents and speakers

Overview of trustworthiness (Jindong Wang, 10min)

Robust machine learning
(Jindong Wang, 40min)

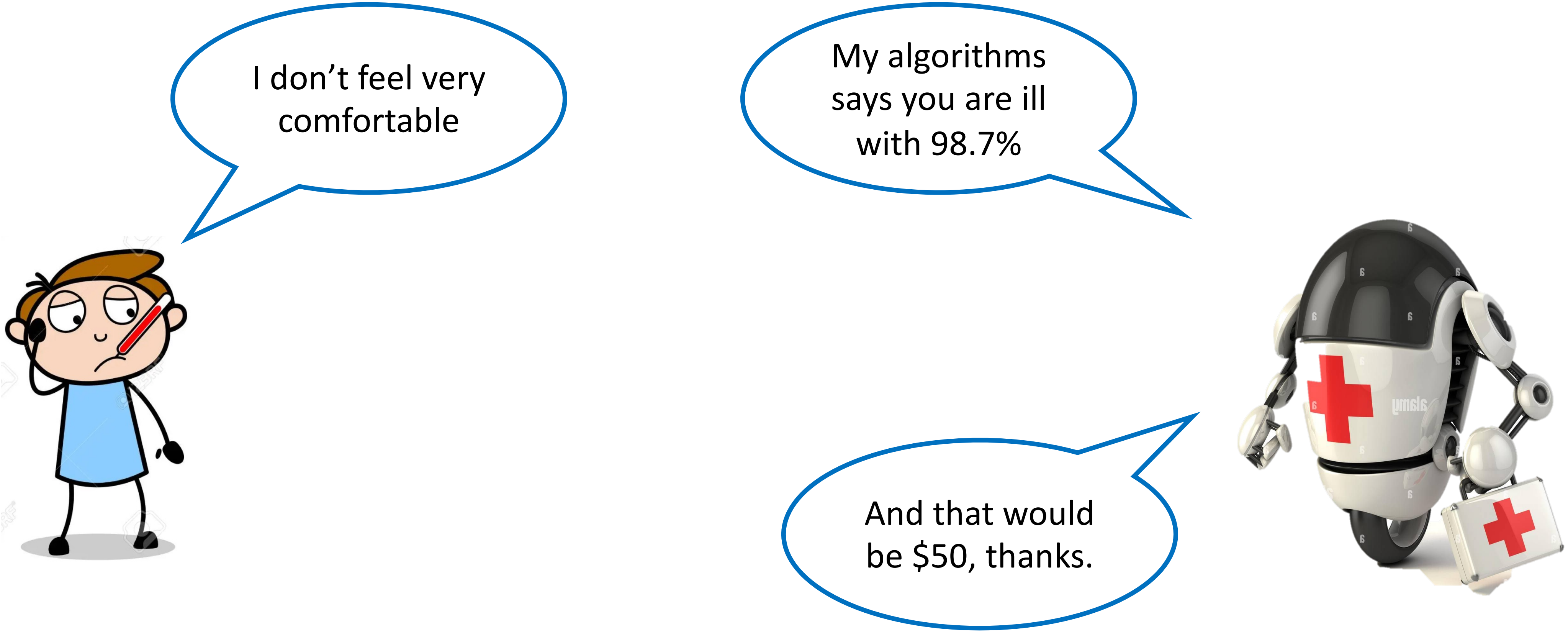
Out-of-distribution generalization
(Haohan Wang, 40min)

Interpretability
(Haohan Wang,
on behalf of Haoliang Li,
40min)

Trustworthiness in the era of large models (Jindong Wang, 40min)

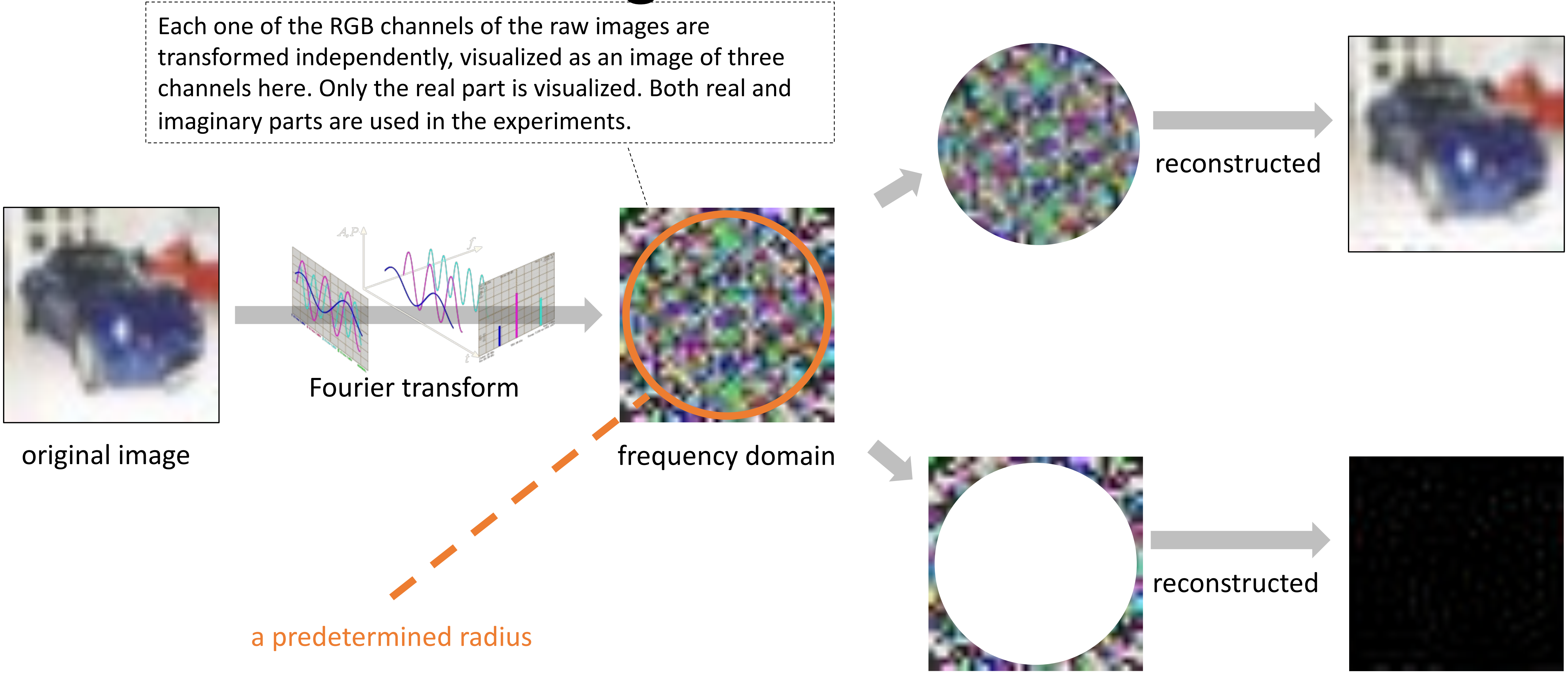
The Importance of Interpretability
From an application perspective

- In many applications, we need to know why



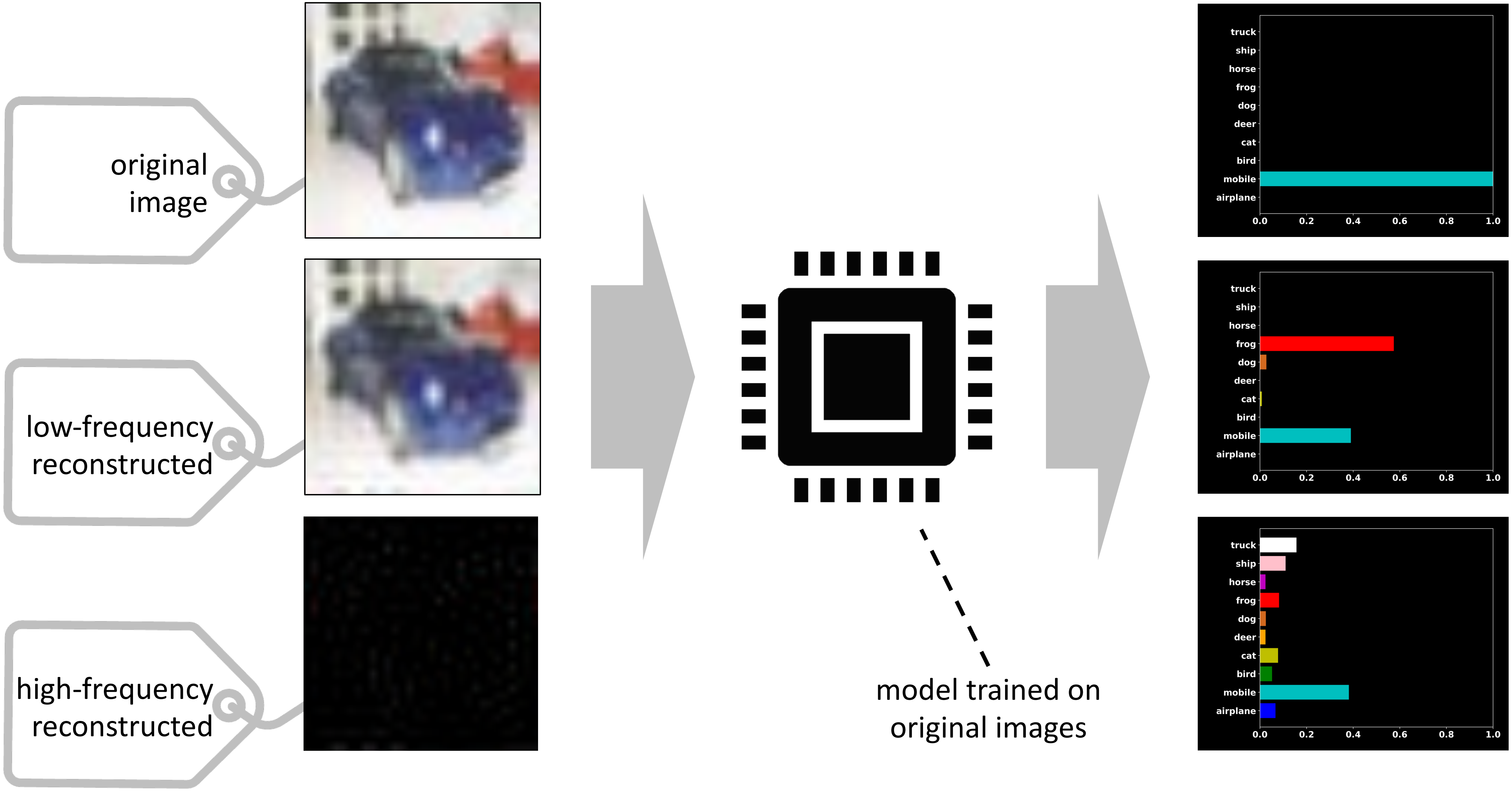
High/Low-frequency reconstructed Images

Wang, Haohan, Xindi Wu, Zeyi Huang, and Eric P. Xing. "High-frequency component helps explain the generalization of convolutional neural networks." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8684-8694. 2020.



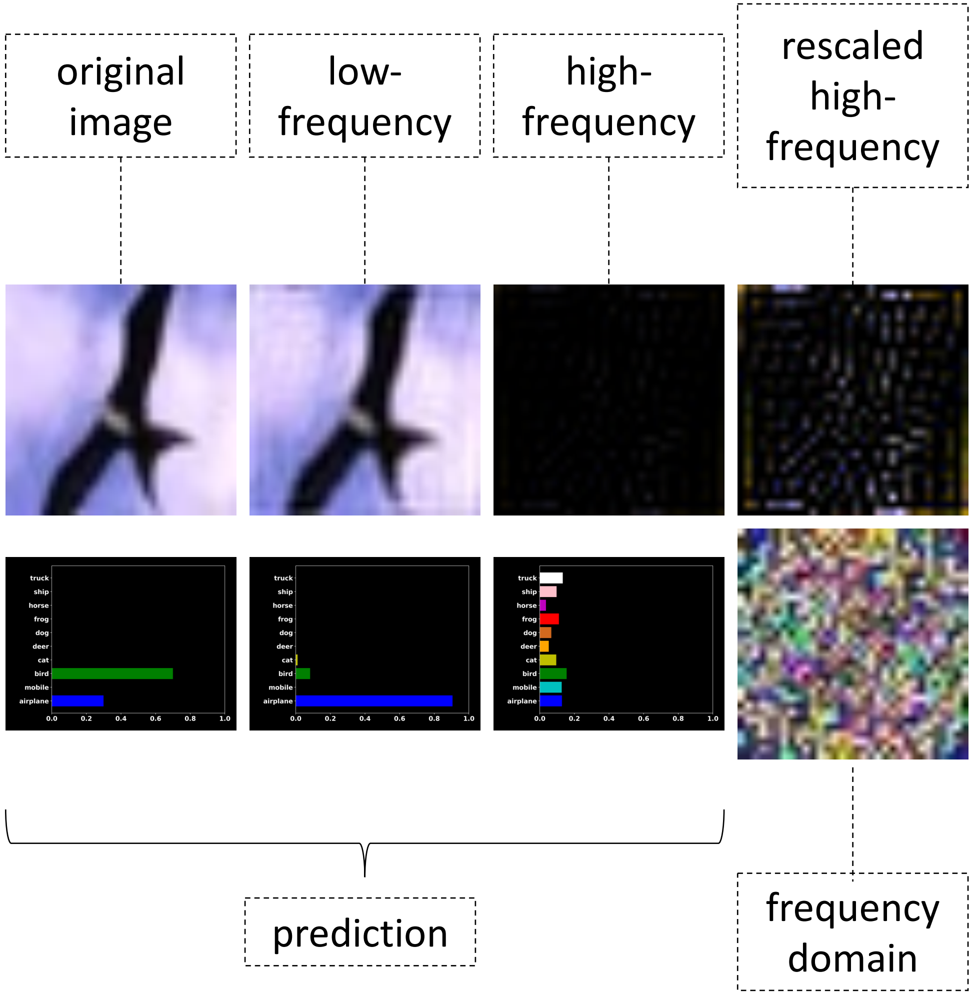
Misalignment between human and model

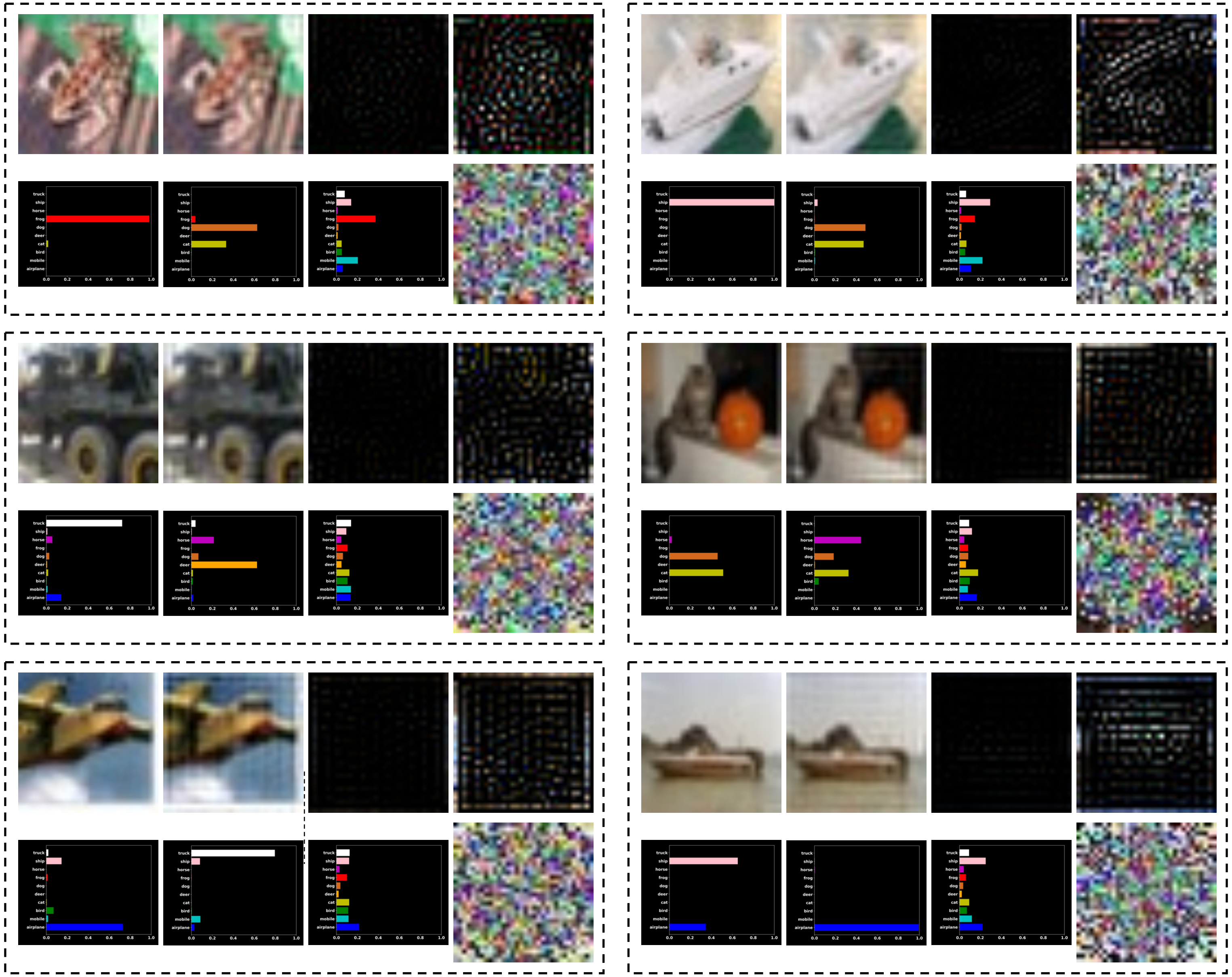
Wang, Haohan, Xindi Wu, Zeyi Huang, and Eric P. Xing. "High-frequency component helps explain the generalization of convolutional neural networks." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8684-8694. 2020.



Additional examples

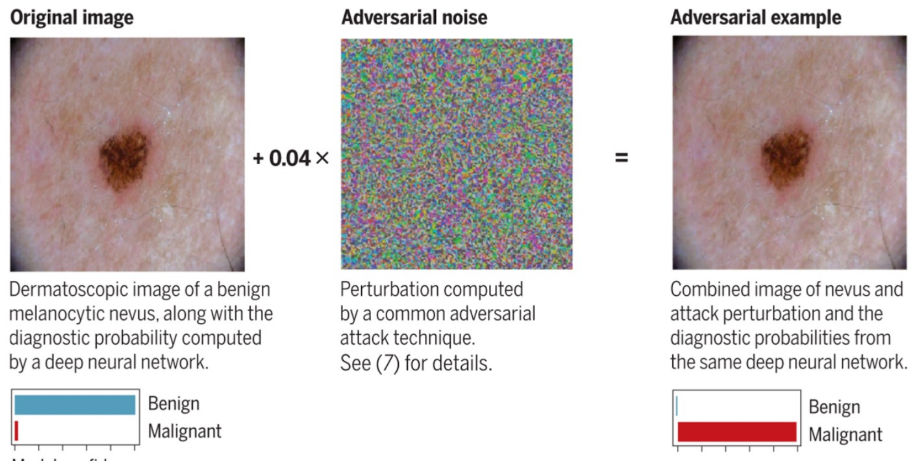
Wang, Haohan, Xindi Wu, Zeyi Huang, and Eric P. Xing. "High-frequency component helps explain the generalization of convolutional neural networks." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8684-8694. 2020.

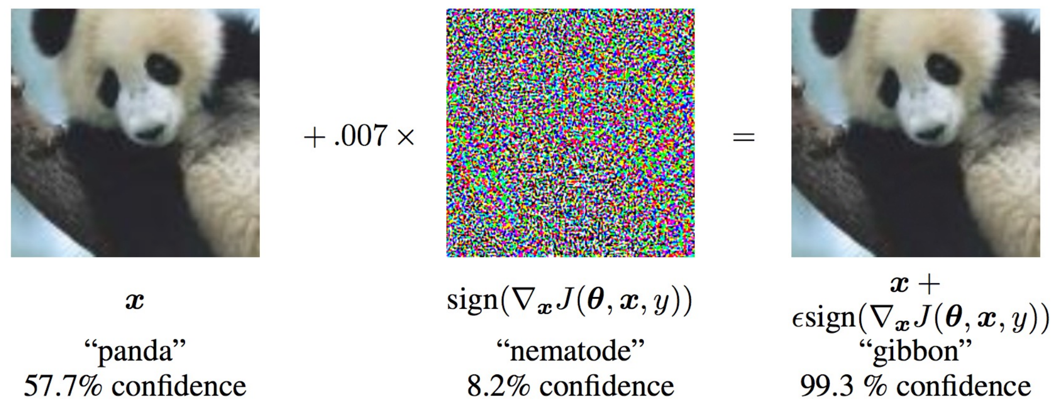




Motivation: Interpretability

- Adversarial Robustness seems to remind us about one in applying deep learning in practice:

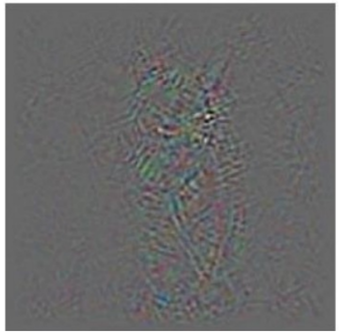




  - When a model is making prediction, the features it use can be quite arbitrary
    - So, we probably need to have a detailed look at the decision process

Backprop and Guided Backprop

- Directly using the gradient

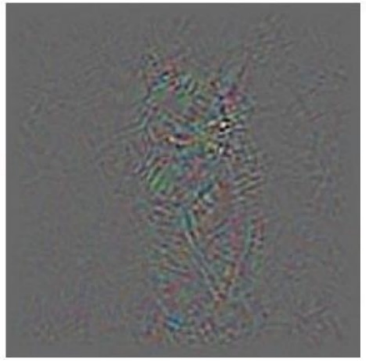


  - Include both positive values and negative values

Backprop and Guided Backprop

- Maybe we don’t care that much for the features that contribute negatively
  - How do we get rid of it?
    - Maybe just setting it to zeros

- Guided Backprop

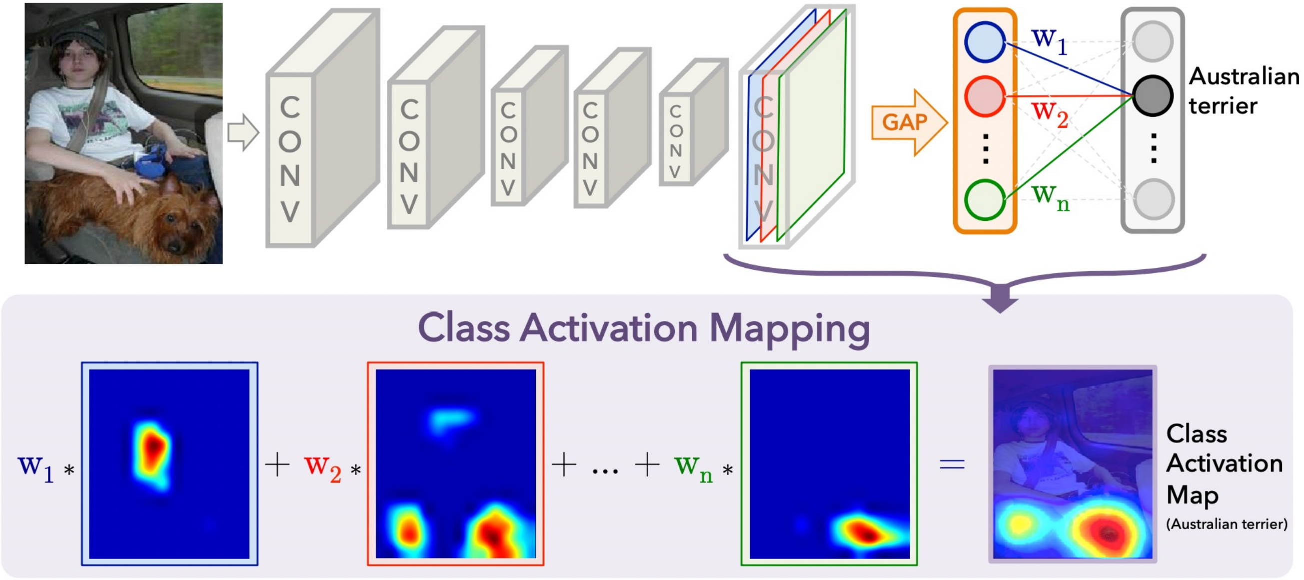


gradient

only positive gradient

Class Activation Map

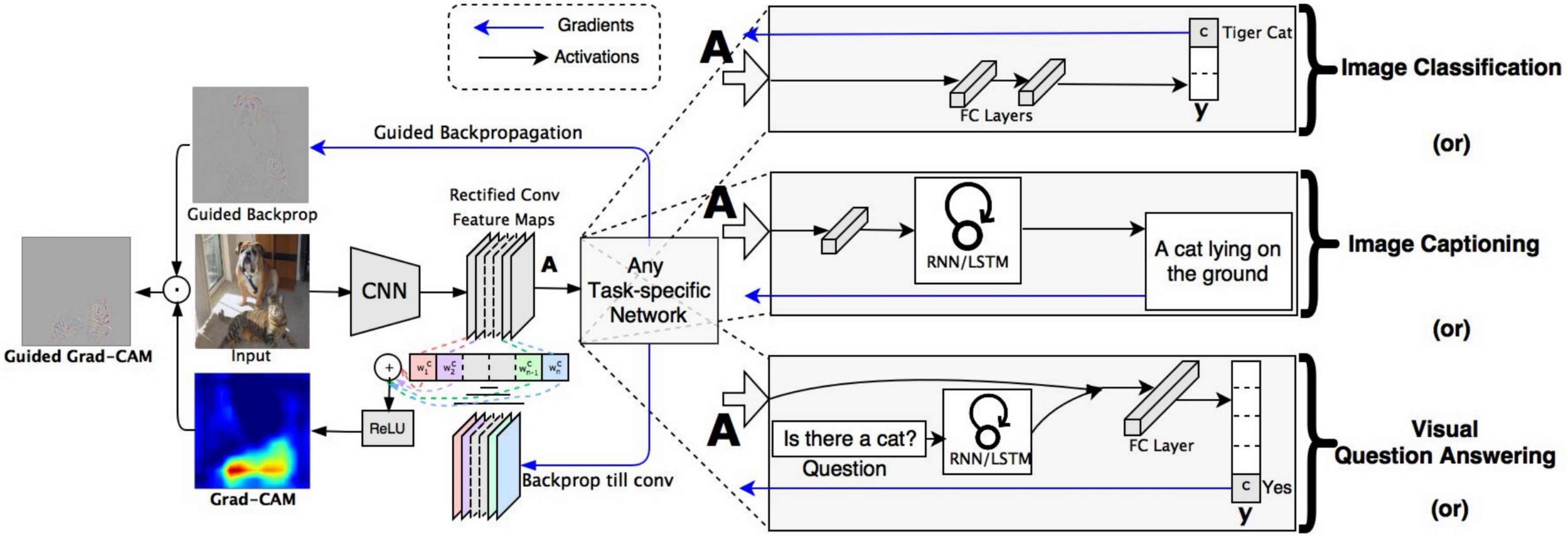
- Visualizing the decision for convolutional neural networks



- As the beginning work, the model has to follow certain structures

GradCam

- Connecting later layers to the convolutional layers with gradient



GradCam

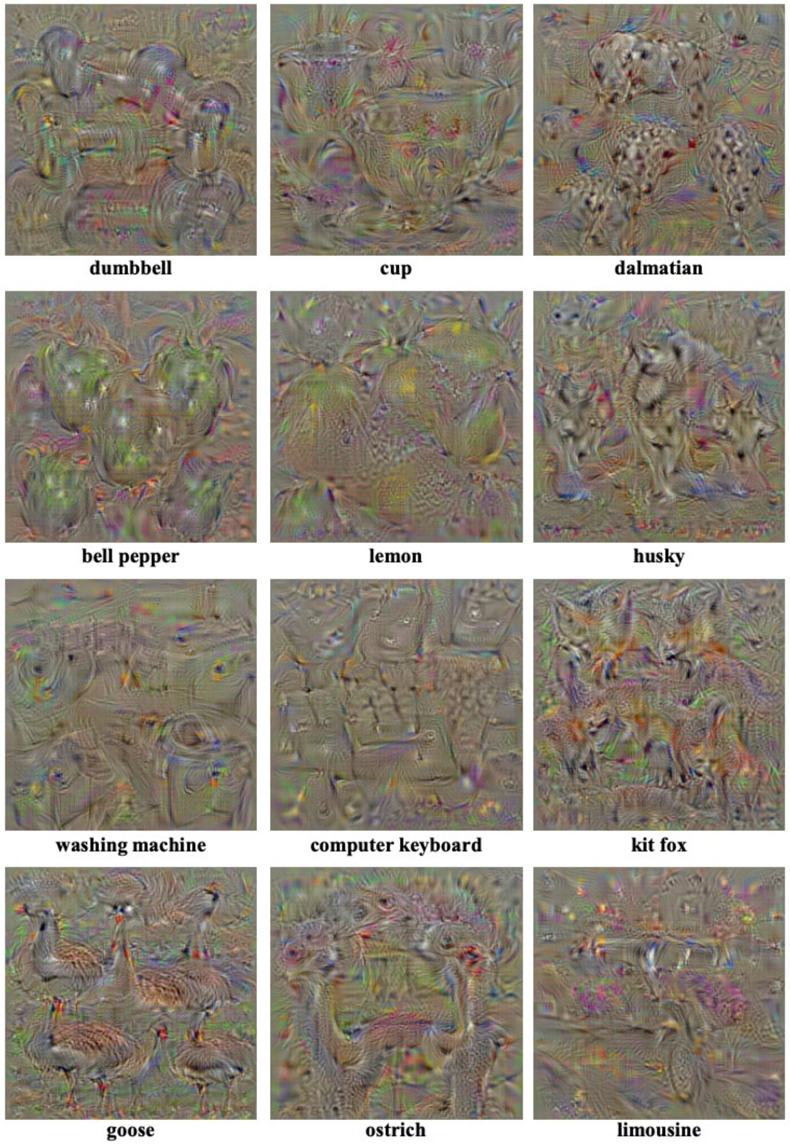
- Results Comparison

(a) Original Image (b) Guided Backprop ‘Cat’ (c) Grad-CAM ‘Cat’

(g) Original Image (h) Guided Backprop ‘Dog’ (i) Grad-CAM ‘Dog’

Model's understanding of certain labels

- Use to evaluate the model's understanding of the whole class
  - Given a (random) starting point
  - Update the samples following the gradient
  - Until the model predicts the resulting images with full confidence



Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

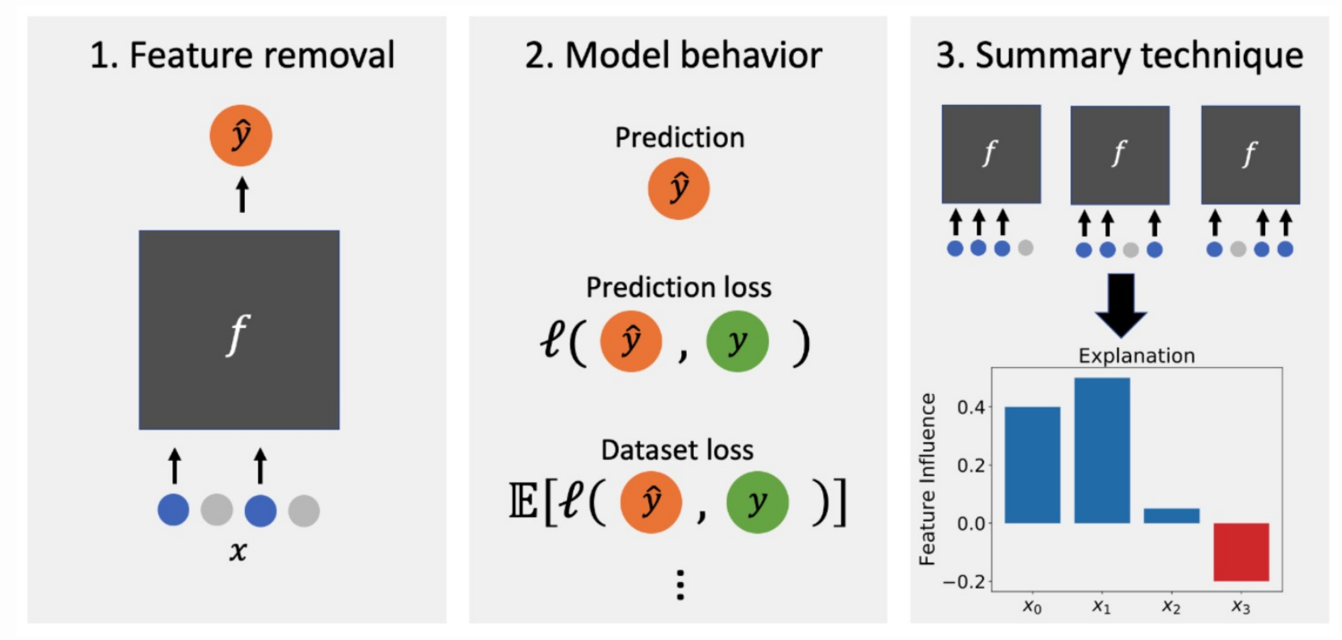
Explaining by Removing

- Explaining by Removing: A Unified Framework for Model Explanation
  - Many different papers in the community for interpretability

Method	Removal	Behavior	Summary
IME (2009)	Separate models	Prediction	Shapley value
IME (2010)	Marginalize (uniform)	Prediction	Shapley value
QII	Marginalize (marginals product)	Prediction	Shapley value
SHAP	Marginalize (conditional/marginal)	Prediction	Shapley value
KernelSHAP	Marginalize (marginal)	Prediction	Shapley value
TreeSHAP	Tree distribution	Prediction	Shapley value
LossSHAP	Marginalize (conditional)	Prediction loss	Shapley value
SAGE	Marginalize (conditional)	Dataset loss (label)	Shapley value
Shapley Net Effects	Separate models	Dataset loss (label)	Shapley value
Shapley Effects	Marginalize (conditional)	Dataset loss (output)	Shapley value
Permutation Test	Marginalize (marginal)	Dataset loss (label)	Remove individual
Conditional Perm. Test	Marginalize (conditional)	Dataset loss (label)	Remove individual
Feature Ablation (LOCO)	Separate models	Dataset loss (label)	Remove individual
Univariate Predictors	Separate models	Dataset loss (label)	Include individual
L2X	Missingness during training	Prediction mean loss	High-value subset
INVASE	Missingness during training	Prediction mean loss	High-value subset
LIME (Images)	Default values	Prediction	Linear model
LIME (Tabular)	Marginalize (replacement dist.)	Prediction	Linear model
PredDiff	Marginalize (conditional)	Prediction	Remove individual
Occlusion	Zeros	Prediction	Remove individual
CXPlain	Zeros	Prediction loss	Remove individual
RISE	Zeros	Prediction	Mean when included
MM	Default values	Prediction	Partitioned subsets
MIR	Extend pixel values	Prediction	High-value subset
MP	Blurring	Prediction	Low-value subset
EP	Blurring	Prediction	High-value subset
FIDO-CA	Generative model	Prediction	High-value subset

Explaining by Removing

- The methods typically have three core steps
  - How/what to remove the features
  - What to observe
  - Summarizing the observation to users



Explaining by Removing

- Feature Removal
  - Zero-ing features: $F(x, S) = f(x_S, 0)$
  - Setting features to a default value r: $F(x, S) = f(x_S, r_{\bar{S}})$
  - Sampling from a conditional generative model $\sim p_G(X_{\bar{S}} | X_S)$: $F(x, S) = f(x_S, \tilde{x}_{\bar{S}})$
  - Marginalizing with condition: $F(x, S) = \mathbb{E}[f(x) | X_S = x_S]$
  - Marginalizing with marginal: $F(x, S) = \mathbb{E}[f(x_S, X_{\bar{S}})]$

Model Behavior

- What to observe after removing the features

  - at the prediction level *(local explanations)*: Given an input $x \in \mathcal{X}$, study $F(x, S)$, that is how removed features are impacting a prediction higher or lower;
  - at the prediction loss level *(local explanations)*: Given an input x and its true label y, study $-\ell(F(x, S), y)$, that is how some features are making the prediction more or less correct.
  - the average prediction loss *(local explanations)*: Given an input x and the label's conditional distribution $p(Y|X = x)$, study $-\mathbb{E}_{p(Y|X=x)}[\ell(F(x, S), Y)]$, that is how a certain set of features can correctly predict what could have occured on average. Can be useful with uncertain labels.
  - dataset loss wrt label *(global explanations)*: How much the model's performance degrades when different features are removed, i.e. $-\mathbb{E}_{XY}[\ell(F(X_S), Y)]$
  - dataset loss wrt output *(global explanations)*: What are the features' influence on the model output (rather than on the model performance), i.e. $-\mathbb{E}_X[\ell(F(X_S), F(X))]$

Summary Techniques

- To provide a concise summary of the information we obtained
  - Examples:
    - Feature attribution:
      - Give every feature a score
    - Feature section
      - Select a subset of features

So, now, we should possess the knowledge of hundreds of model explanation methods

Explaining by Removing

- Summary of the techniques by “Explaining by Removing” at 2020

Summary technique

Feature removal	Feature attribution					Feature selection		
	Remove individual	Include individual	Mean when included	Shapley value	Additive model	High value subset	Low value subset	Partitioned subsets
Zeros	Occlusion CXPlain		RISE					MM
Default values					LIME (images)			
Extend pixels						MIR		
Blurring						EP	MP	
Generative model						FIDO-CA		
Marginalize (replacement distribution)					LIME (tabular)			
Marginalize (uniform)				IME 2010				
Marginalize (marginals product)				QII				
Marginalize (marginal)	Permutation test			SHAP KernelSHAP				
Marginalize (conditional)	PredDiff Conditional perm. test			SHAP SAGE LossSHAP Shapley Effects				
Tree distribution				TreeSHAP				
Surrogate model						L2X REAL-X		
Missingness during training						INVASE		
Separate models	Feature ablation	Univariate predictors		Shapley Net Effects IME 2009 SPVIM				

Model behavior ■ Prediction ■ Prediction loss ■ Prediction mean loss ■ Dataset loss ■ Prediction loss (output) ■ Dataset loss (output)

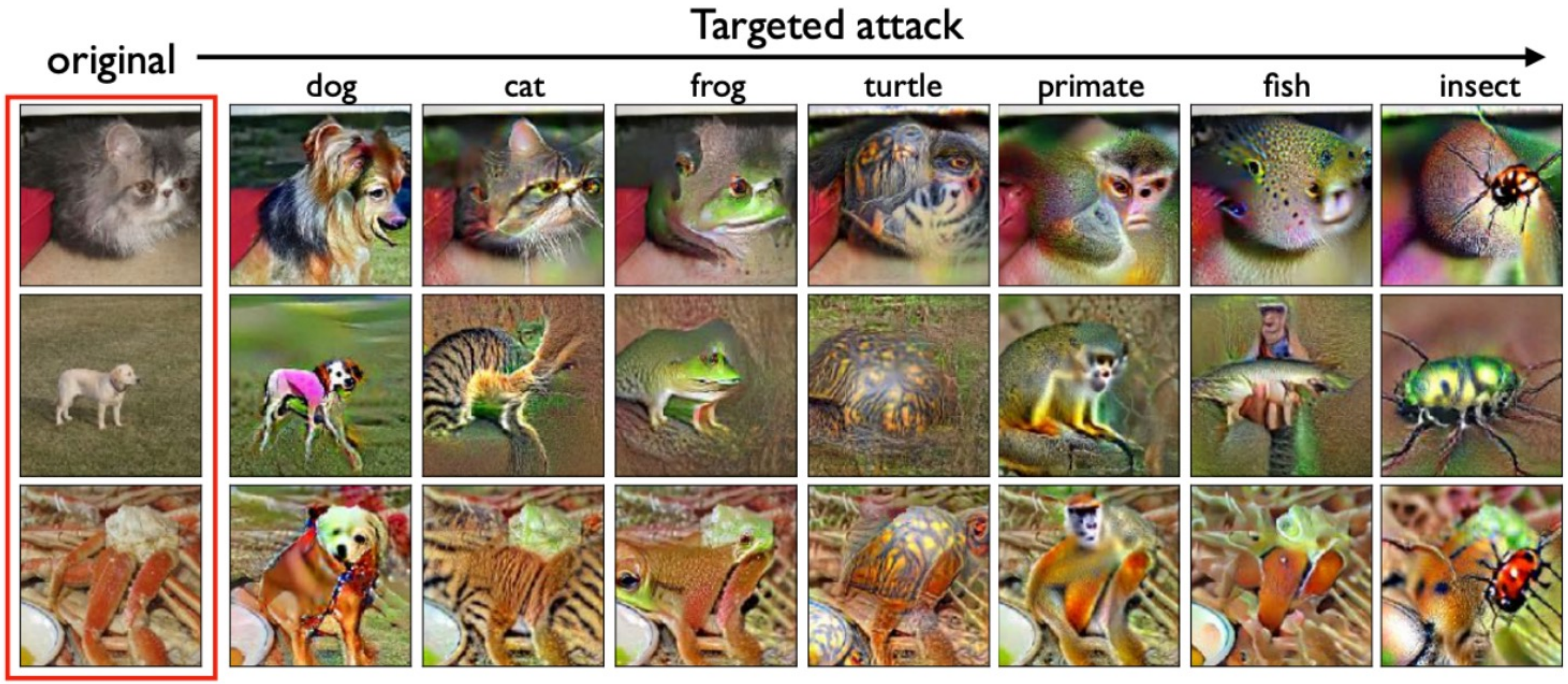
Just in case you want to try a submission in this field but not sure about what to do :D

An Interesting Question

Do we need better interpretability methods or better model?

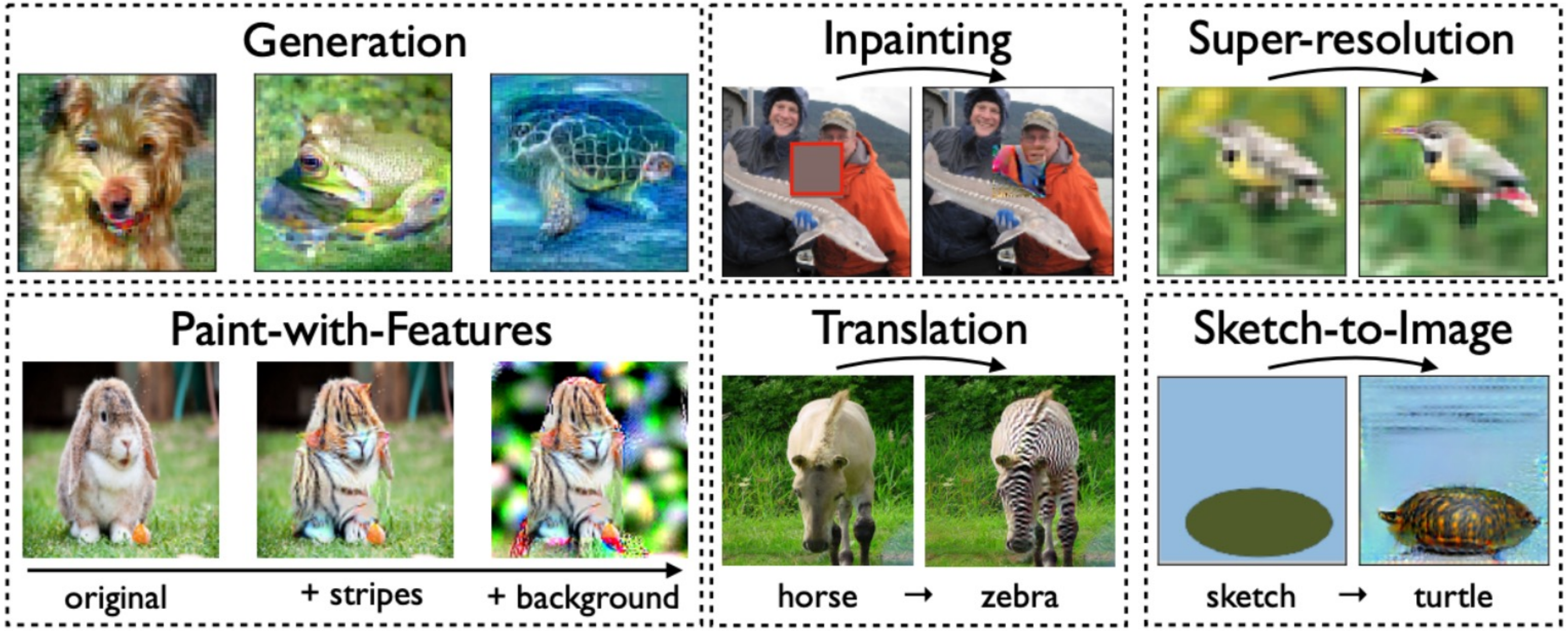
Adversarially Robust Models

- Adversarially robust model might learn perceptually aligned representations
  - (Santurkar et al. 2019)

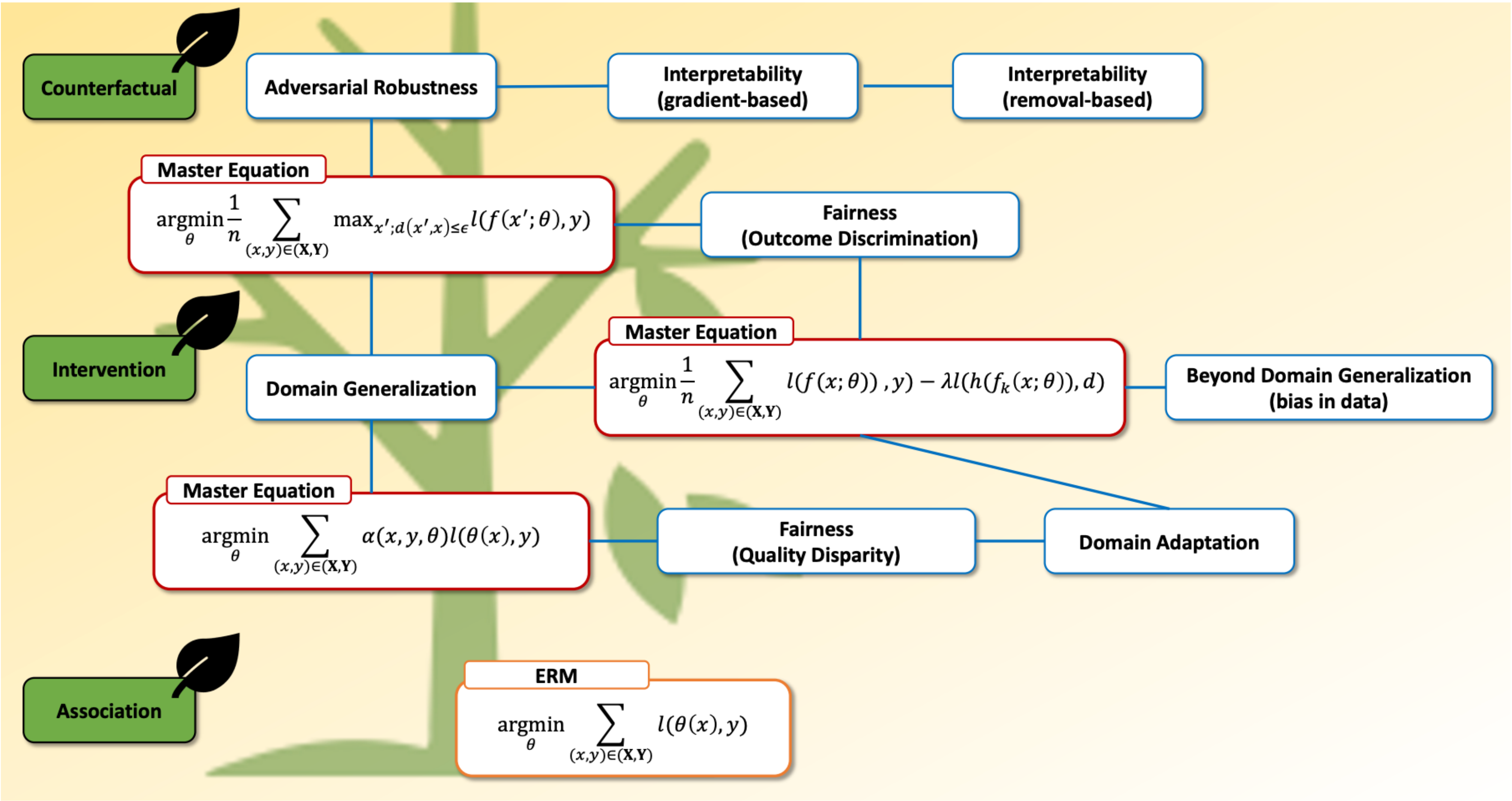


Adversarially Robust Models

- Adversarially robust model might learn perceptually aligned representations
  - (Santurkar et al. 2019)

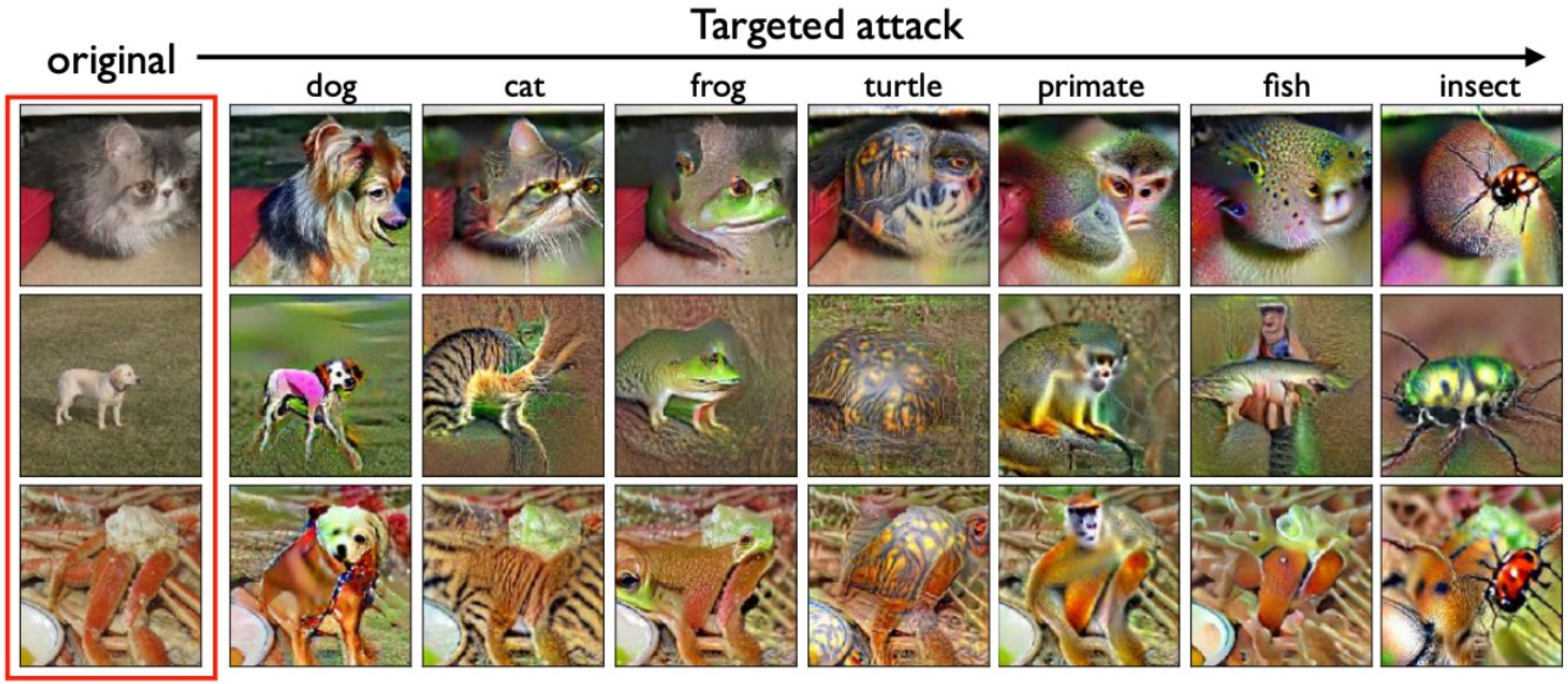


Potential connection between interpretability and adversarial robustness



Adversarially Robust Models

- With this connection, let's look at these results again
  - It seems these results are just supposed to happen



Following Slides are from Haoliang Li

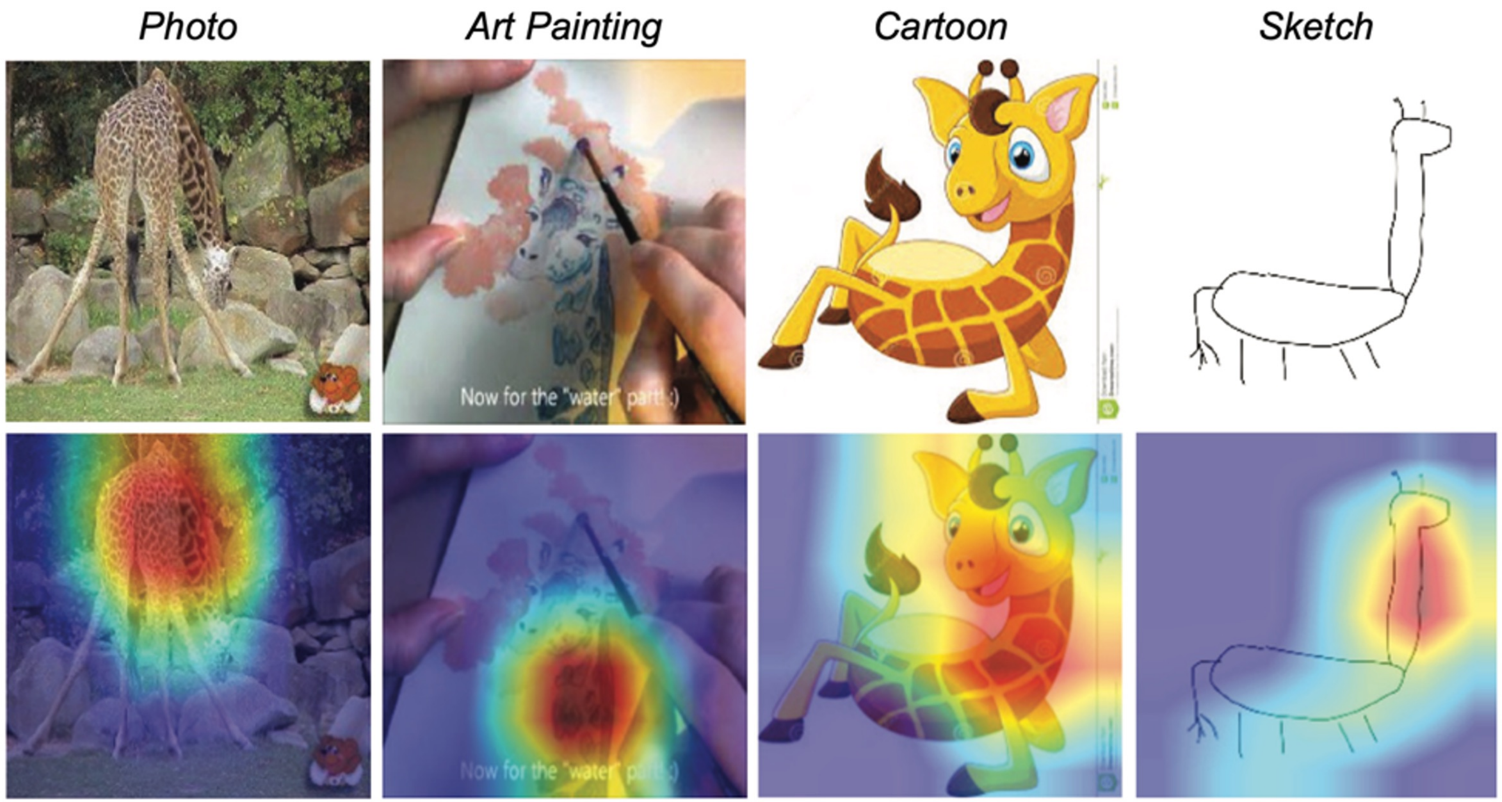
Interpretability via attention explanations

Shared causality across domains.

But can we believe such appealing visualizations?

Input

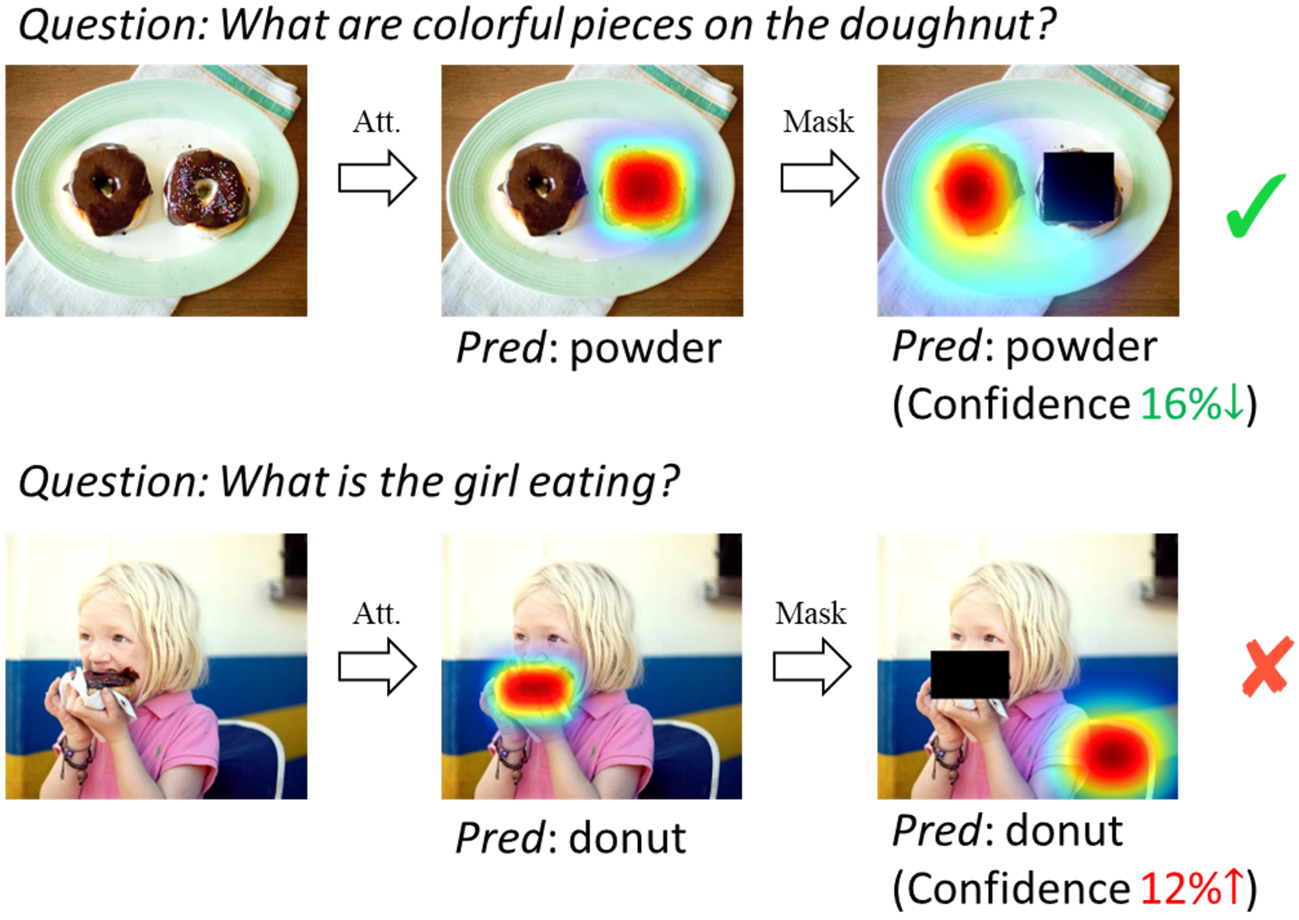
Explanation



- Kim et al. SelfReg: Self-supervised Contrastive Regularization for Domain Generalization. ICCV 2021.

Interpretability via attention explanations

Can existing explanation methods faithfully represent model decisions?



No! Don't be misled by appealing visualizations!

- Liu et.al. Rethinking Attention-Model Explainability through Faithfulness Violation Test, ICML 2022. https://arxiv.org/abs/2201.12114

Interpretability via attention explanations

The reason is simple:

- Recall the formulation of attention mechansims

$$\boldsymbol{A} = \mathrm{softmax}(\frac{\boldsymbol{Q} \cdot \boldsymbol{K}^T}{\sqrt{d_h}})$$

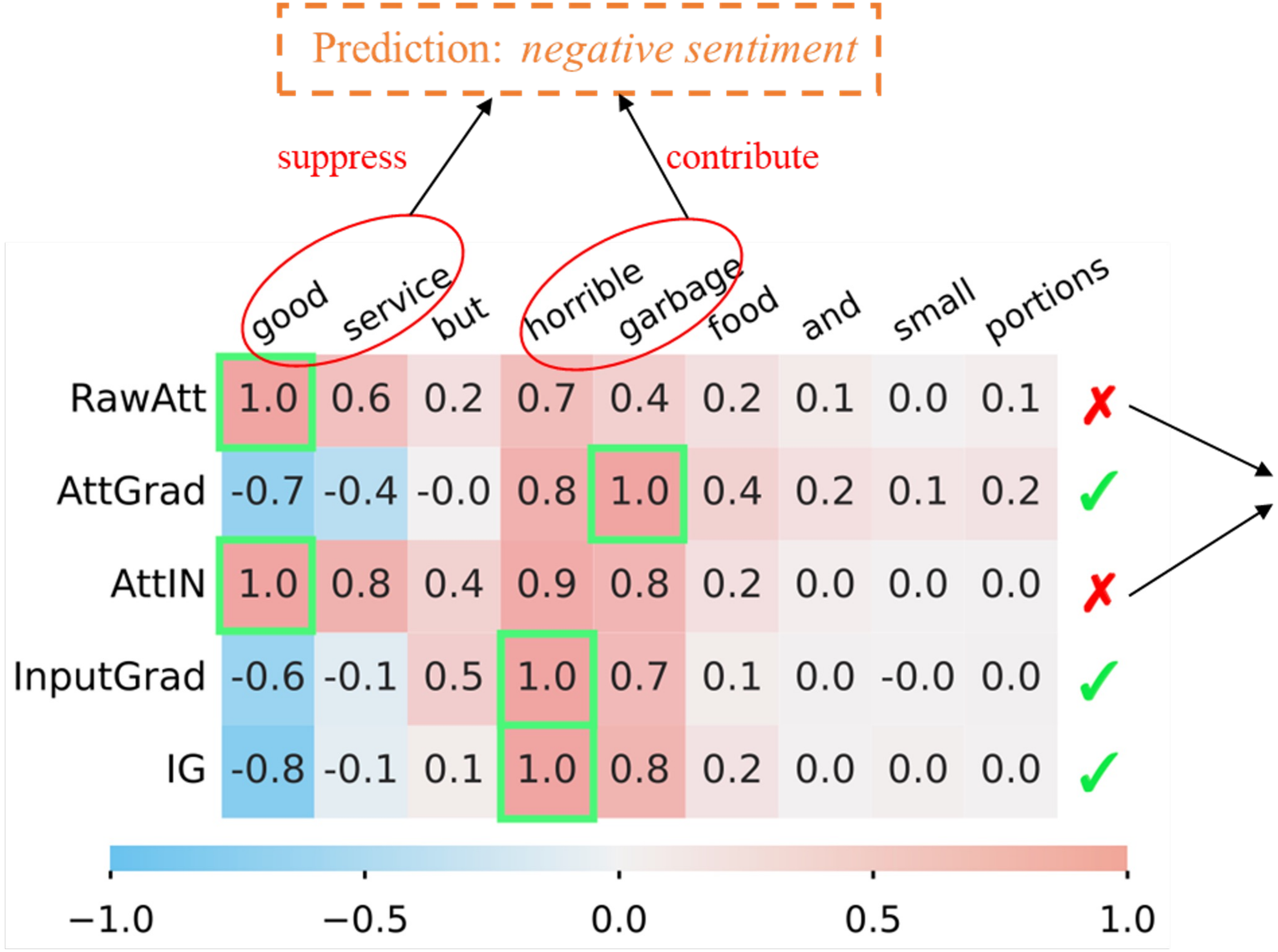
$$\boldsymbol{O} = \boldsymbol{A} \cdot \boldsymbol{V}$$

(Vaswani et al., 2017)

1. Attention weights are always non-negative since they are the output of softmax function.
2. So they cannot differentiate the direction of feature impacts, *i.e.*, impact polarity

Interpretability via attention explanations

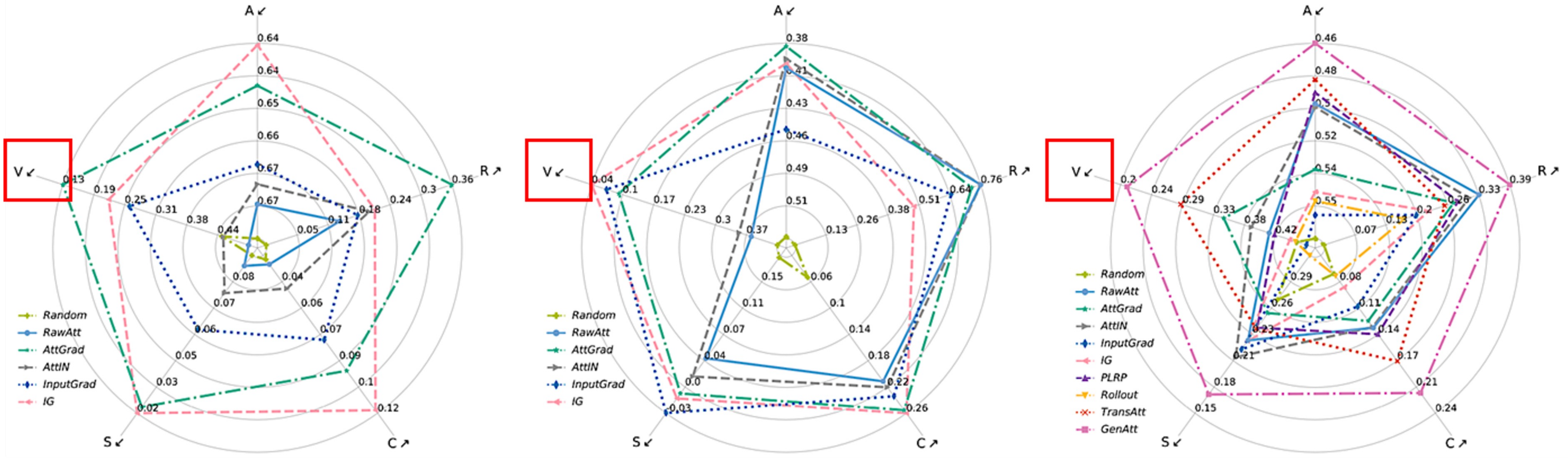
This heavily degrades the faithfulness of attention explanations:



Because they cannot differentiate the impact polarity!!

Interpretability via attention explanations

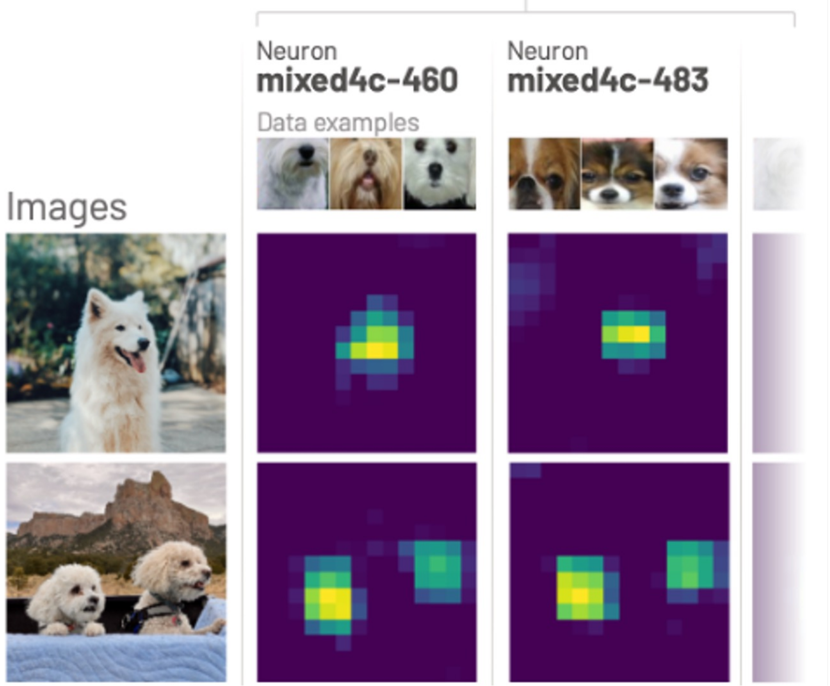
Most explanation methods fail to suffer from the polarity consistency issue.

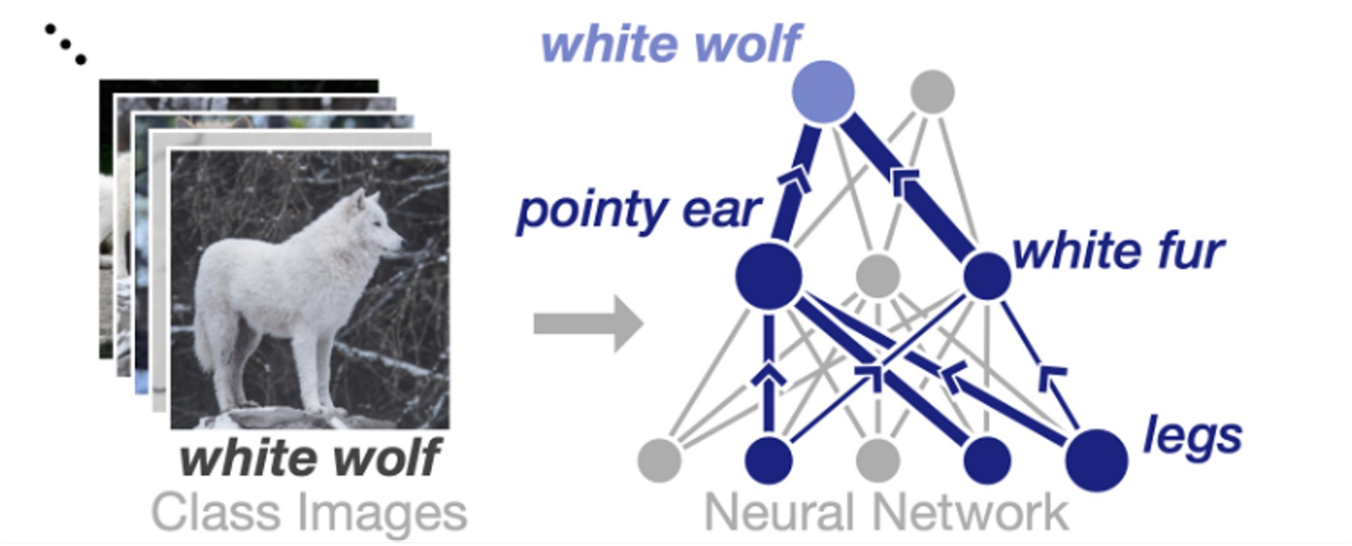


(a) LSTM+DotAtt on QQP dataset (b) BUTD on VQA 2.0 dataset (c) LXMERT on GQA dataset

Interpretability via neuron explanations

What did a single unit (neuron) learn?

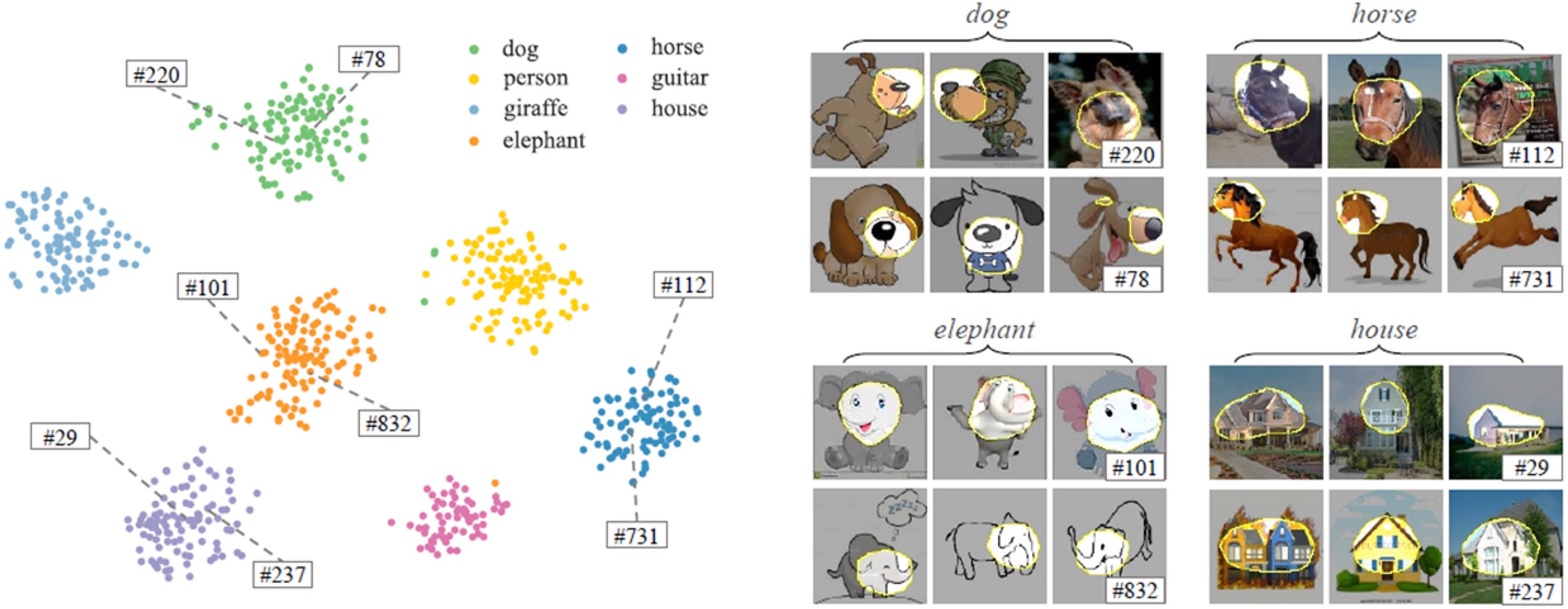




- H. Park, et al. "NeuroCartography: Scalable Automatic Visual Summarization of Concepts in Deep Neural Networks ," TVCG, 2021.
- F. Hohman, et al. "SUMMIT: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations ," TVCG, 2019.

Interpretability via neuron explanations

Take the last convolutional layer in ResNet as an example,

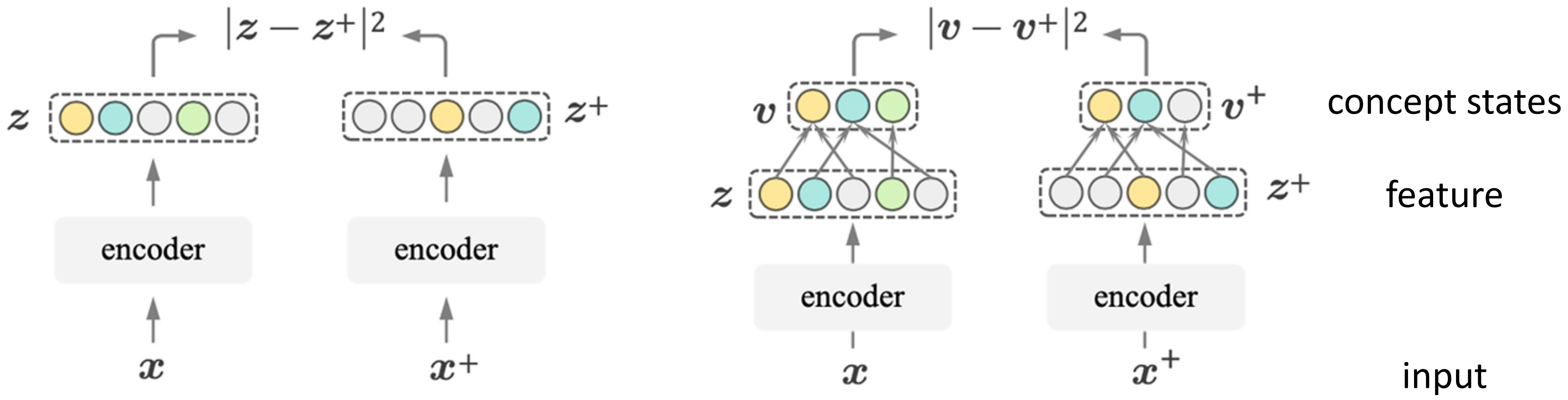


(a) Top activated neurons under different predictions. (b) Visual concepts shared by different neurons.

- Liu, Y., Tian, C. X., Li, H., & Wang, S. (2022). Generalization Beyond Feature Alignment: Concept Activation-Guided Contrastive Learning. arXiv preprint arXiv:2211.06843.

Interpretability via neuron explanations

We proposed concept-level contrast (CoCo) to learn features beyond conventional feature-level contrast.

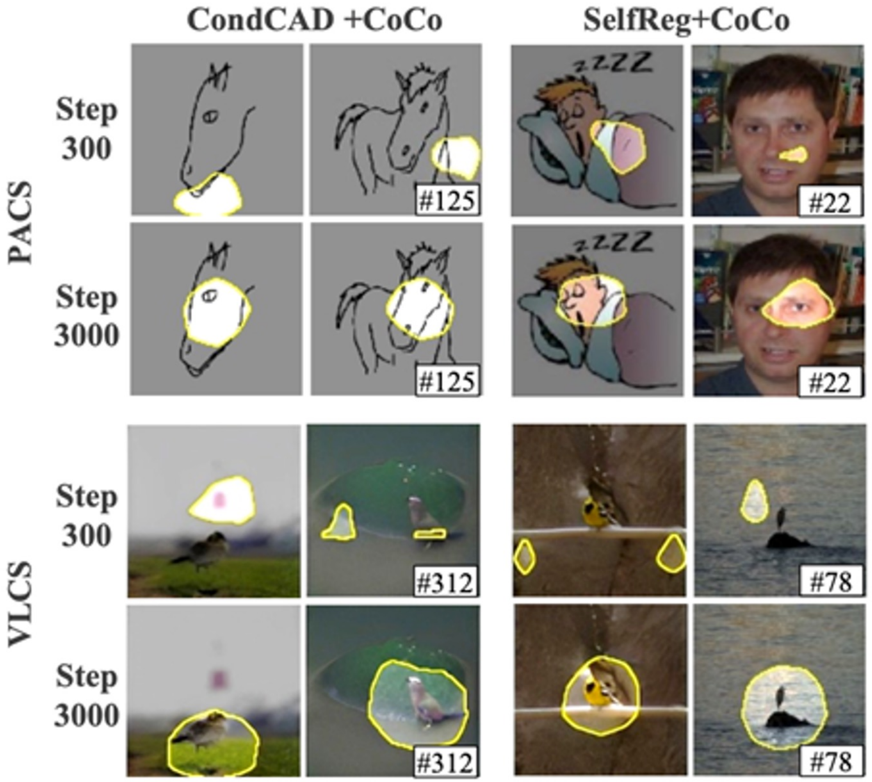


(a) Element-wise Feature Contrast (b) Concept Contrast (CoCo)

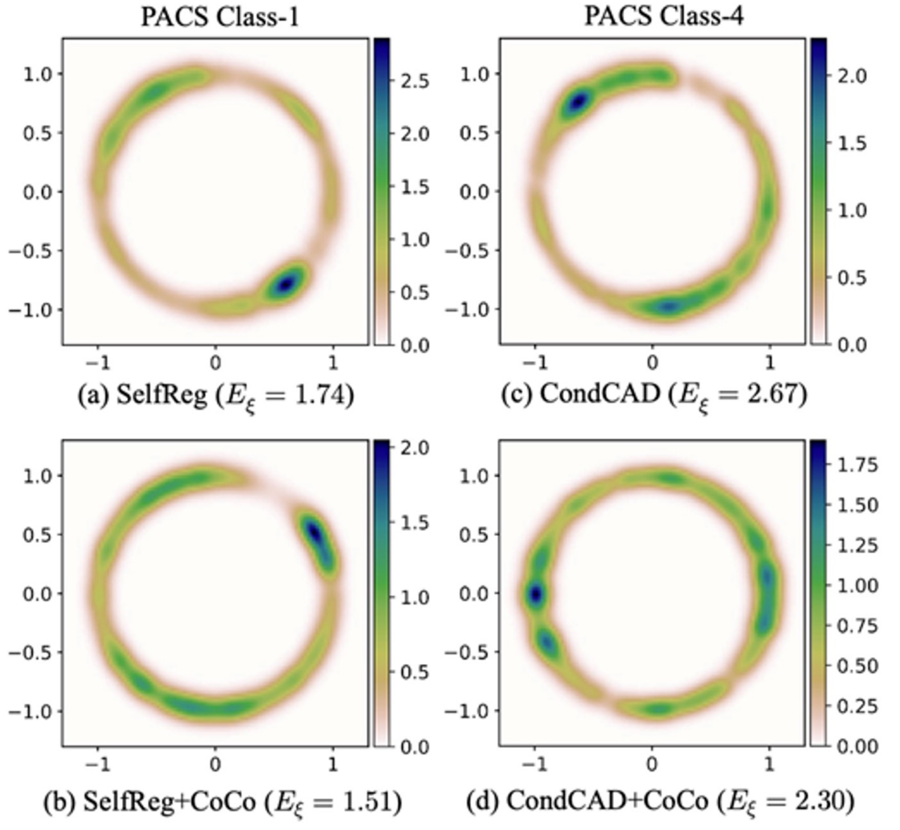
- Liu, Y., Tian, C. X., Li, H., & Wang, S. (2022). Generalization Beyond Feature Alignment: Concept Activation-Guided Contrastive Learning. arXiv preprint arXiv:2211.06843.

Interpretability via neuron explanations

With CoCo, concept evolution happens



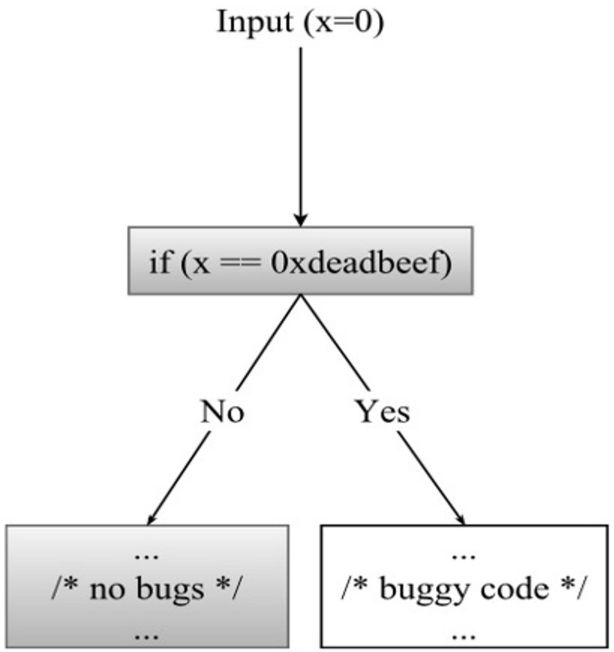
Diversified features



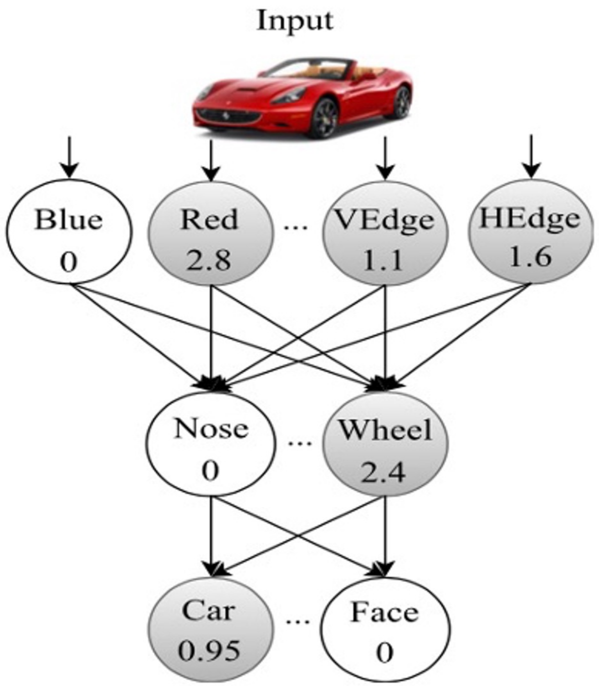
Interpretability via neuron explanations

Code / Logic not covered during testing -> Bugs may hide there

This could also happen in neural networks!



Neurons are not covered?



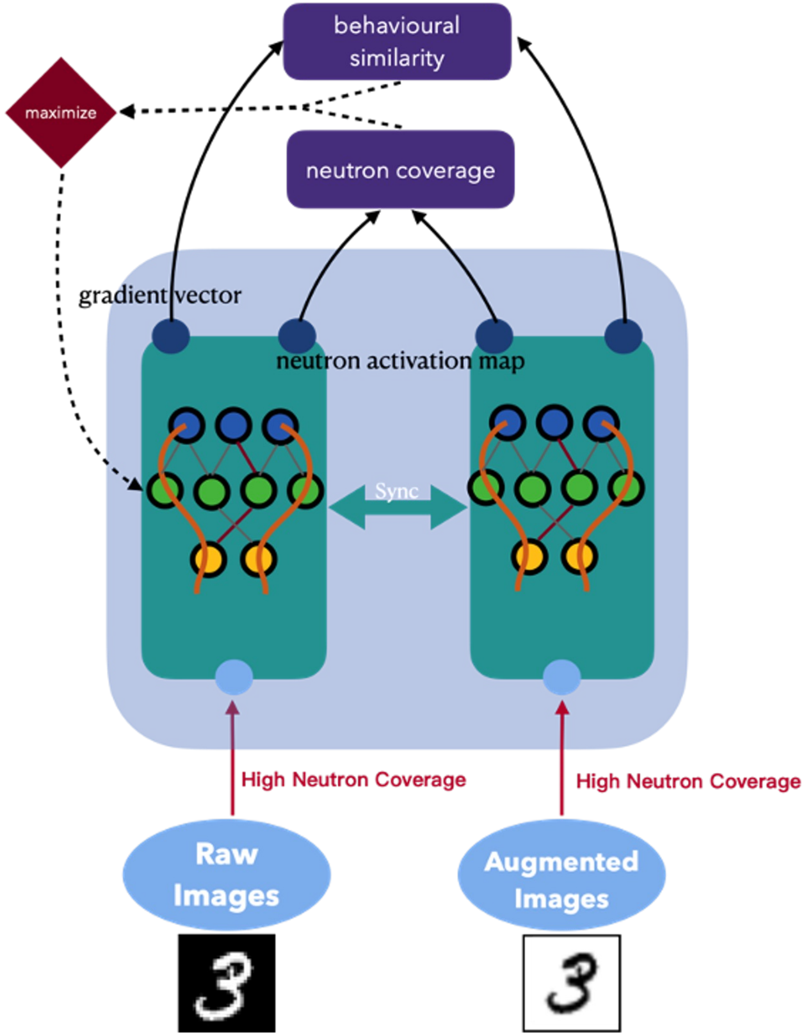
DeepXplore: Automated Whitebox Testing of Deep Learning Systems. Best paper in SOSP'17

Interpretability via neuron explanations

We proposed NCDG to actively activate the inactive neurons during training with the neuron coverage maximization loss.

If a neuron is inactive
(always low-value output)
during the **WHOLE** training process.

Once it gets activated (outputs high-value)
during the evaluation,
errors may happen.

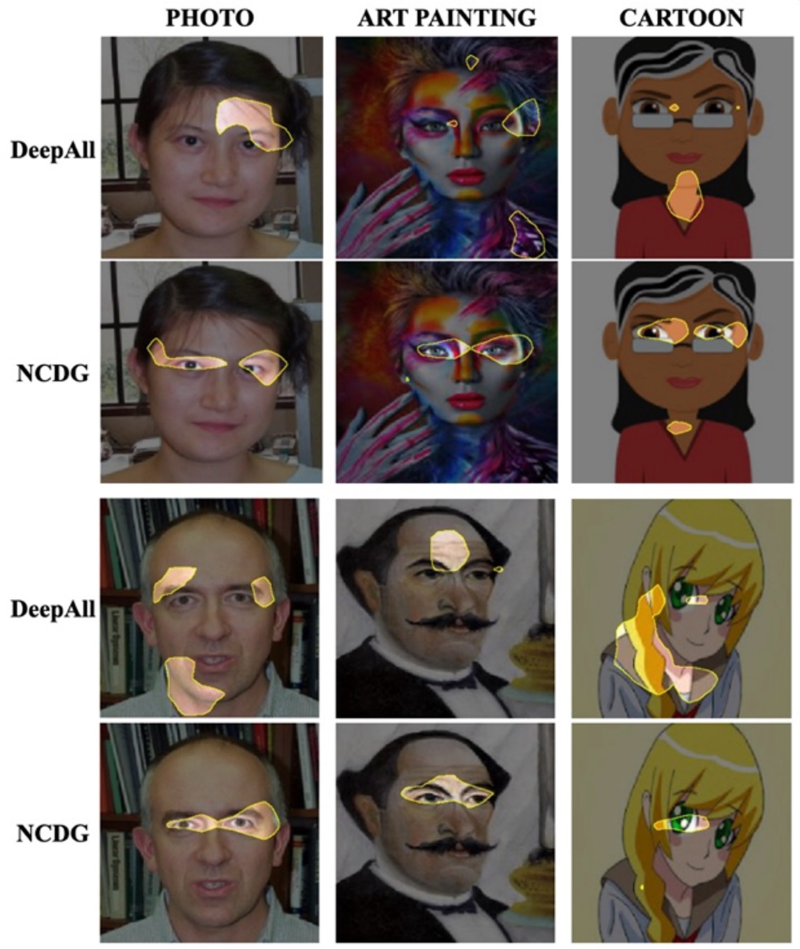


Tian, Chris Xing, et al. "Neuron Coverage-Guided Domain Generalization." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

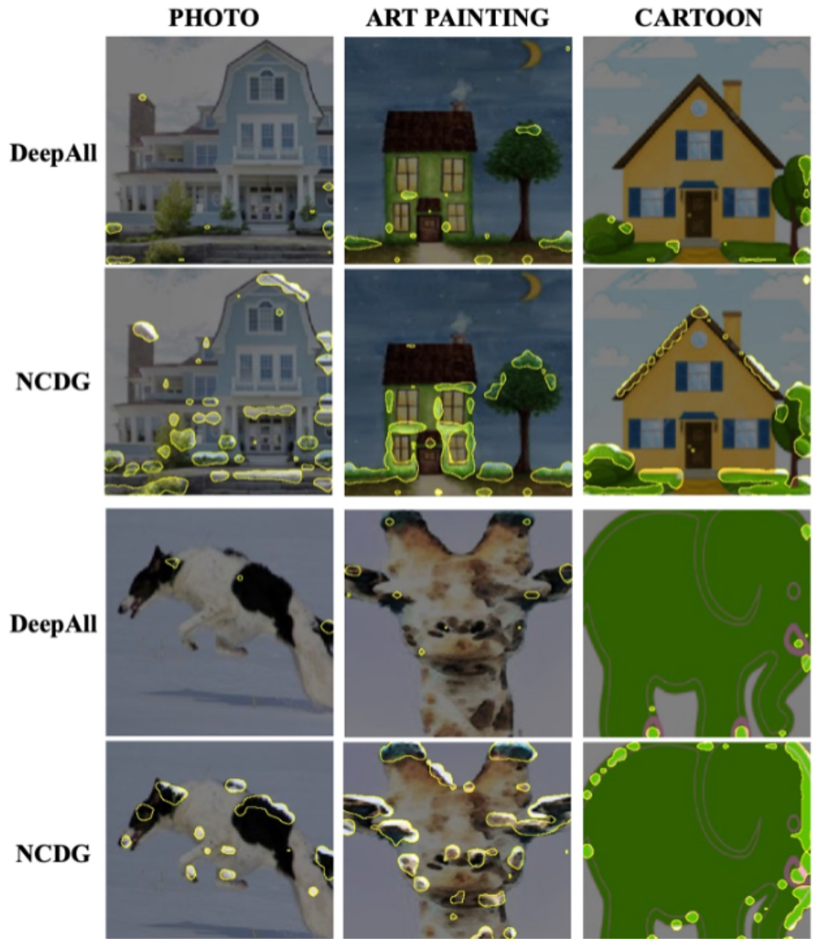
Interpretability via neuron explanations

Neuron Dissection before & after (being activated)

Model trained on PACS-Sketch domain



ResNet-18 block 3 unit 170



ResNet-18 block 2 unit 20