Evaluating XAI

CSEP 590B: Explainable AI

Ian Covert & Su-In Lee
University of Washington

Course announcements

- HW1 due last night
- HW 2 released today
  - Image explanations
  - XAI metrics

The course so far

- Focused on **feature importance** explanations

- Deep dive into the algorithms
  - Local and global methods
  - Removal-based methods
    - How to remove features
    - How to summarize influence
  - Propagation-based methods
    - Different ways to work with gradients

Now, zooming out

- Diverse algorithms, but all designed for one purpose: **identifying influential features**

- How can we test which methods do this best?

Questions to consider

- Do we need to know *a priori* what's important?
- Should explanations reflect what's important to the **model**, or what's important to **humans**?
- Are our performance metrics aligned with any specific explanation methods?

Setup

- Assume a model $f(x)$
  - Classifier with probability $f_y(x)$ for class y
- Assume an explanation algorithm
  - Local explanation (e.g., RISE)
  - Global explanation (e.g., permutation test)
  - Returns scores $a_i \in \mathbb{R}$ for each feature x_i

Today

- Section 1
  - Sanity checks 
  - Ground truth comparisons
- Section 2
  - Ablation metrics
  - Other criteria

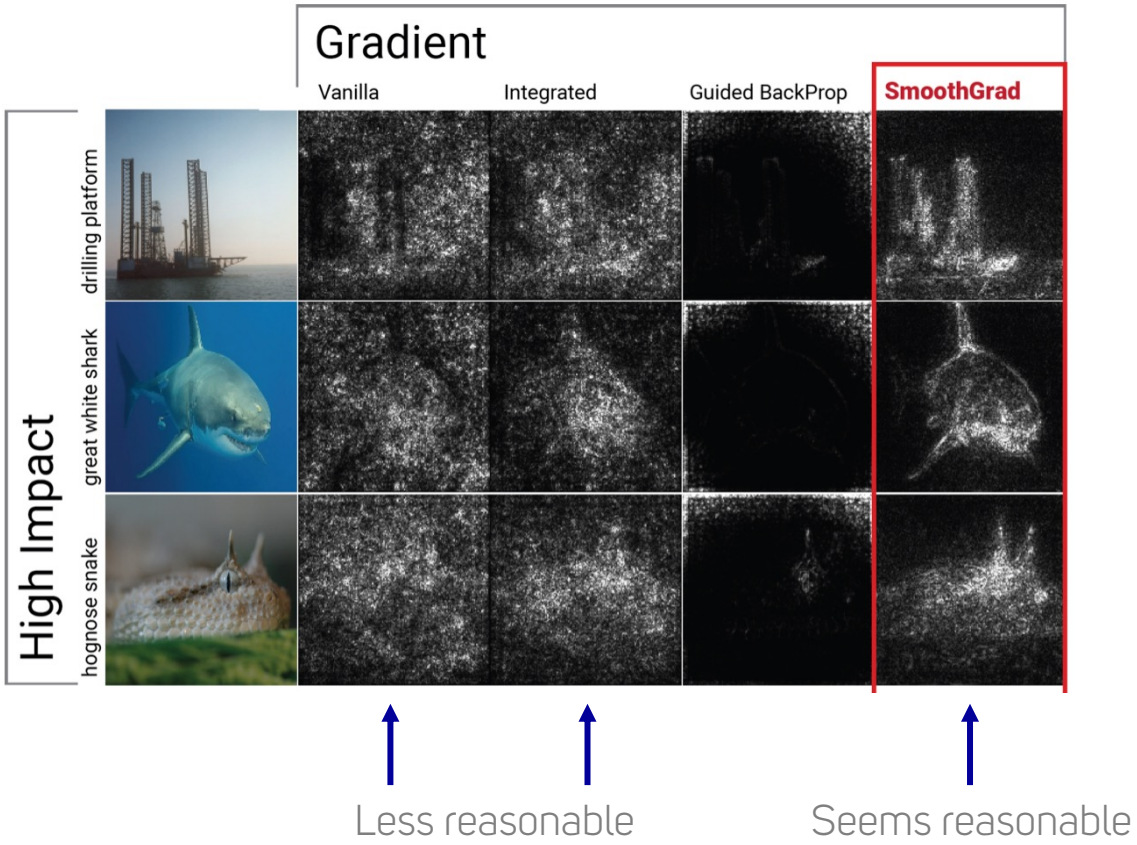
Sanity checks

- Sanity check = basic test to identify obvious issues
  - E.g., test a sorting algorithm with a small list, or a data structure with a few addition/deletion operations
- What are good sanity checks for an explanation algorithm?

Sanity checks for XAI

- Does the explanation make qualitative sense?
- Does it depend on the data?
- Does it depend on the model?

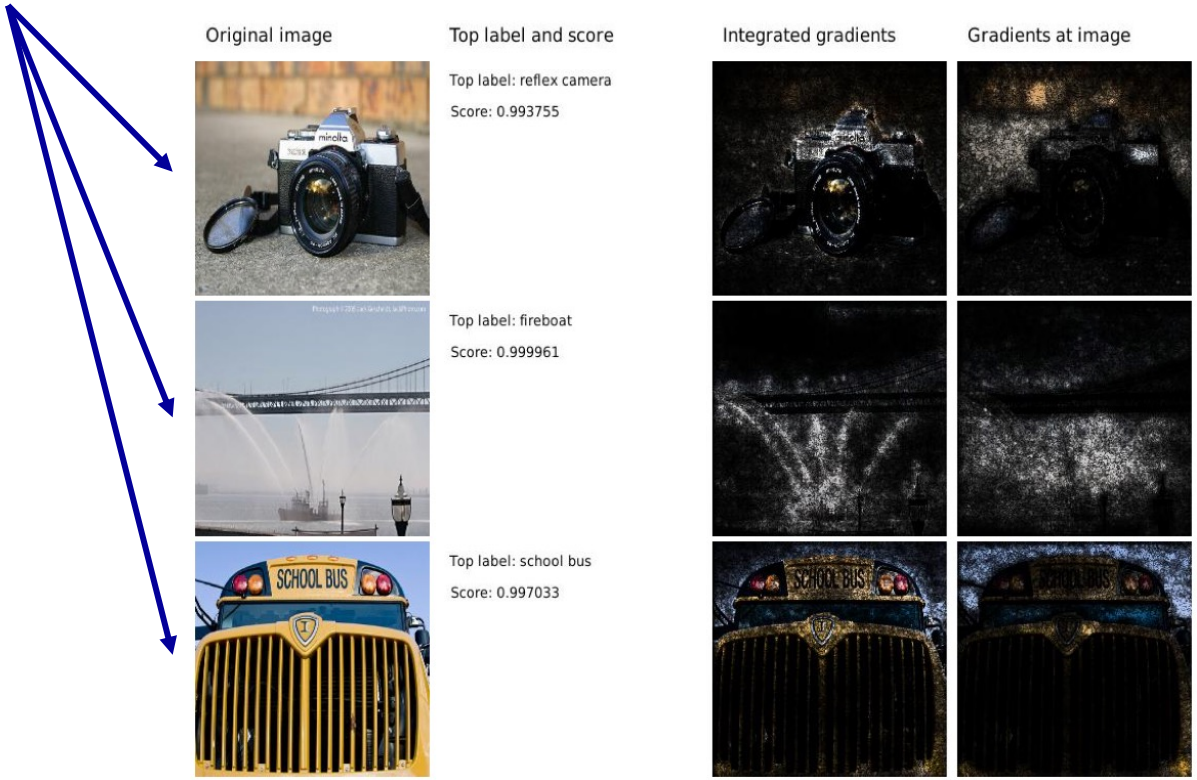
Qualitative evaluation



Smilkov et al., "SmoothGrad: Removing noise by adding noise" (2017)

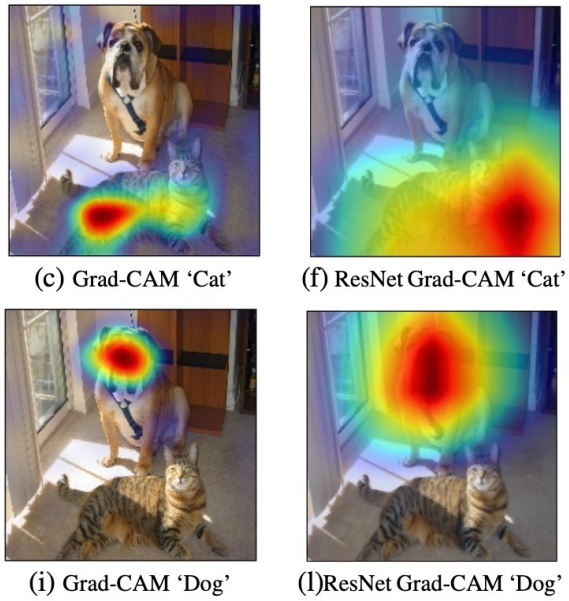
Data dependence

Clearly depends on the data

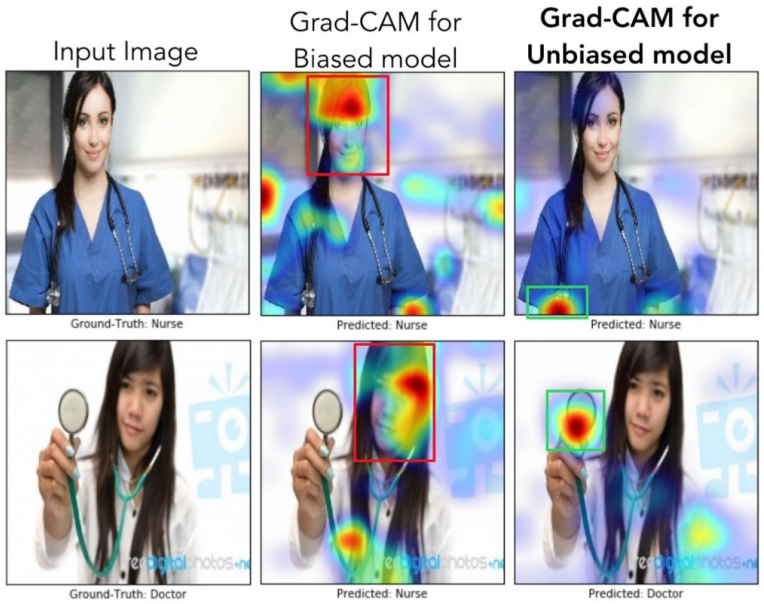


Sundararajan et al., "Axiomatic attribution for deep networks" (2017)

Model dependence



VGG-16 **ResNet-18**



Selvaraju et al., "Grad-CAM: Visual explanations from deep neural networks via gradient-based localization" (2017)

Randomization tests

- Scaled-up version of previous checks
  - Compare explanations after applying randomization
  - Either *model randomization* or *data randomization*

- Explanations should change significantly
  - Surprisingly, some methods don't change very much

Adebayo et al., "Sanity checks for saliency maps" (2018)

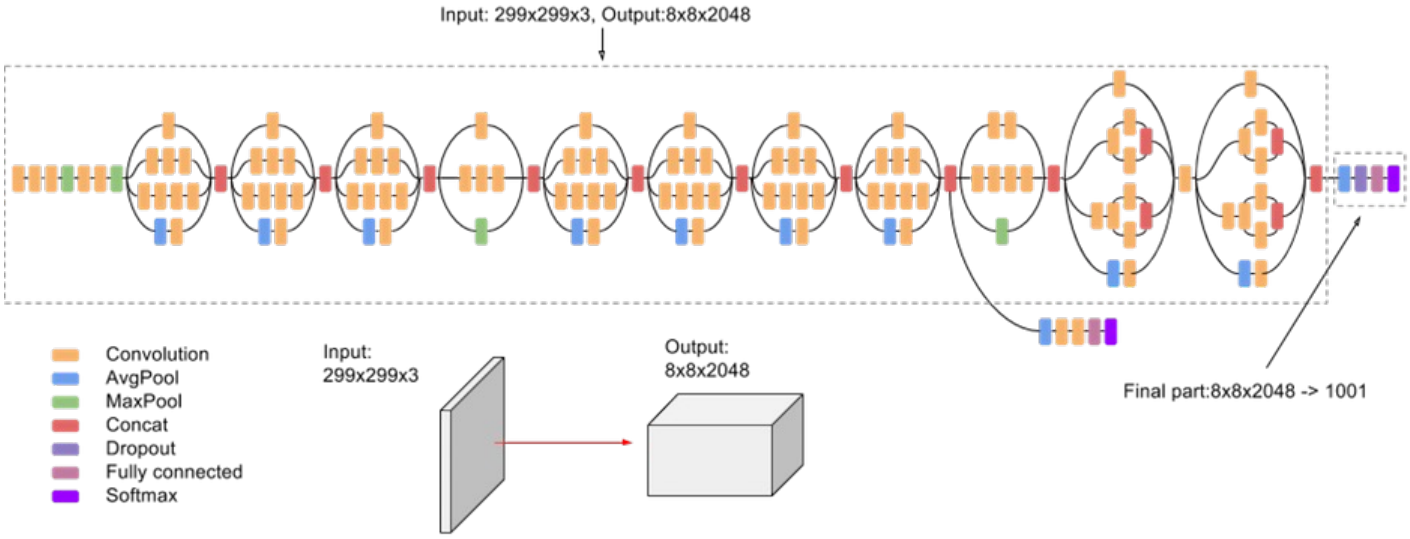
Model randomization

- Begin with a deep neural network
  - They use Inception-v3 architecture

- **Idea:** randomize parameters in specific layers
  - Begin with the final layer, then progressively randomize earlier layers (“cascading randomization”)

Model randomization

Inception-v3 architecture



Szegedy et al., "Rethinking the Inception architecture for computer vision" (2015)

Model randomization (cont.)

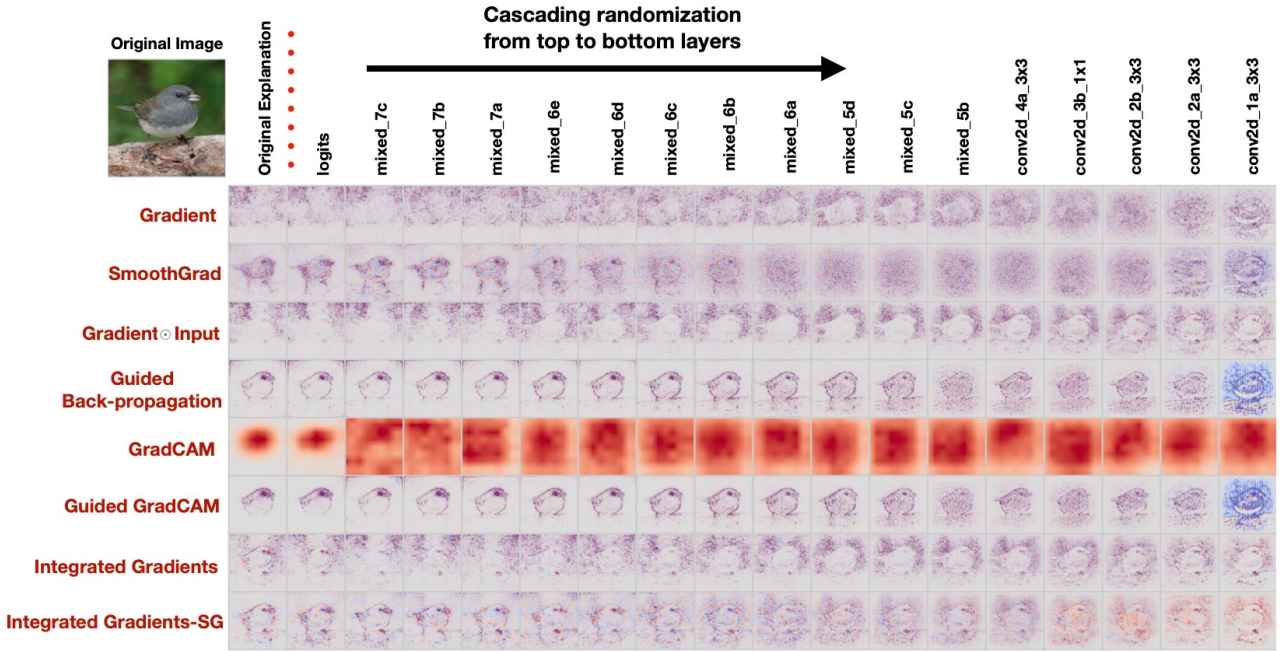
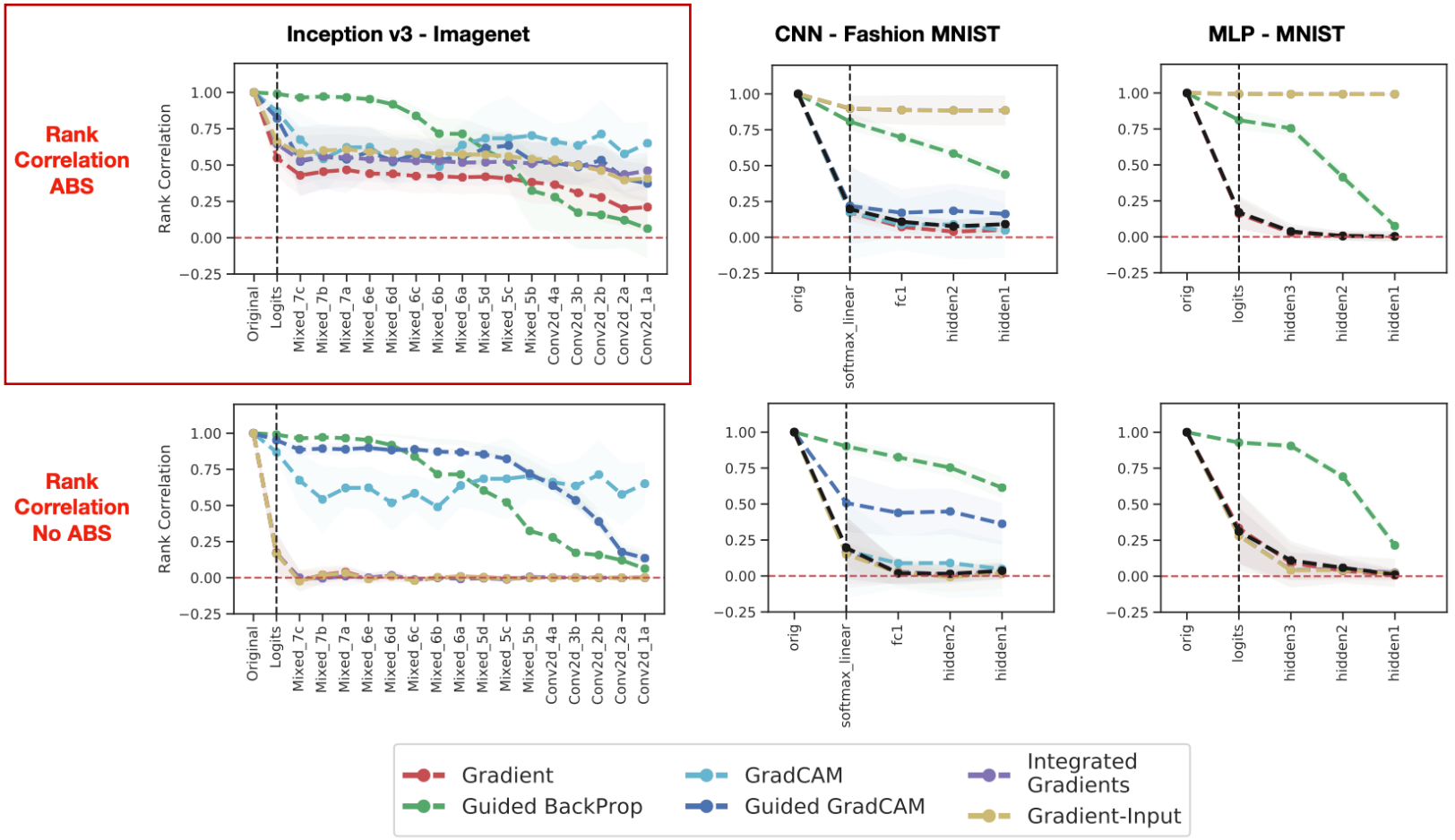


Figure 2: **Cascading randomization on Inception v3 (ImageNet).** Figure shows the original explanations (first column) for the Junco bird as well as the label for each explanation type. Progression from left to right indicates complete randomization of network weights (and other trainable variables) up to that 'block' inclusive. We show images for 17 blocks of randomization. Coordinate (Gradient, mixed_7b) shows the gradient explanation for the network in which the top layers starting from Logits up to mixed_7b have been reinitialized. The last column corresponds to a network with completely reinitialized weights. See Appendix for more examples.

Model randomization (cont.)



Data randomization

- **Idea:** retrain with randomized labels
  - Assign labels uniformly at random
  - New model should use different signals

Data randomization (cont.)

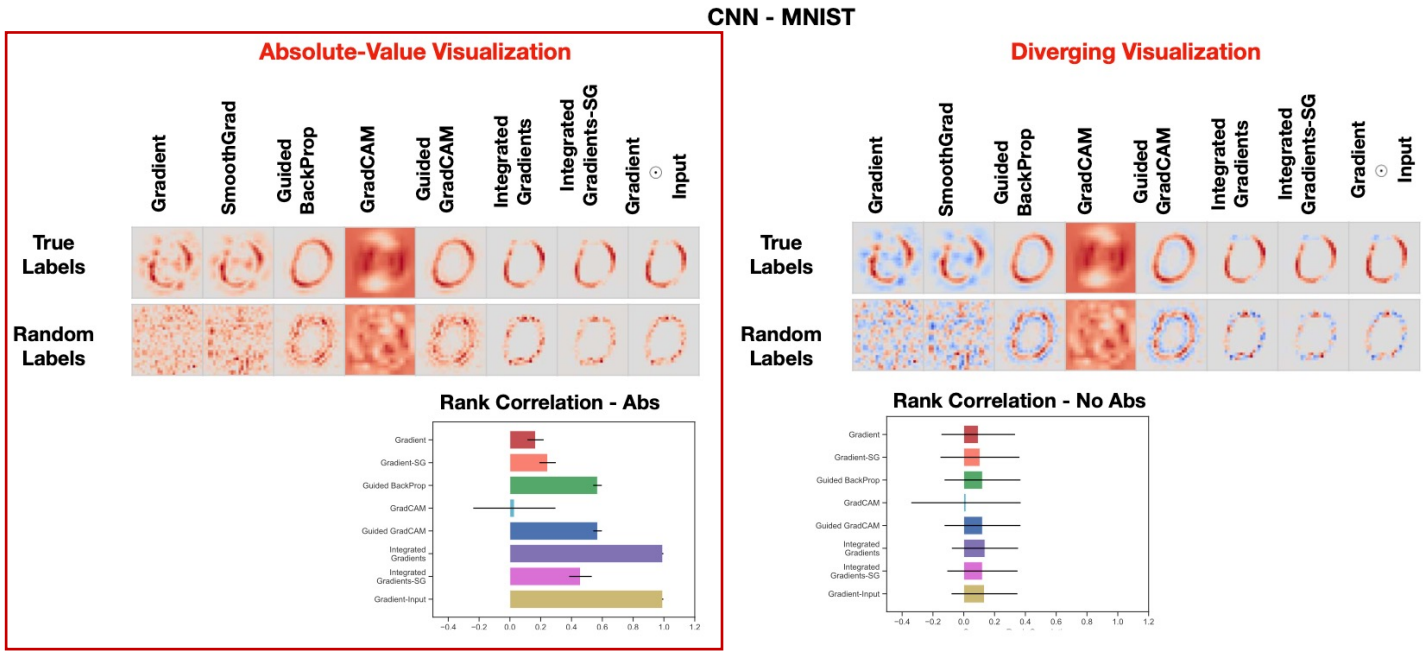


Figure 6: **Explanation for a true model vs. model trained on random labels. Top Left**: Absolute-value visualization of masks for digit 0 from the MNIST test set for a CNN. **Top Right**: Saliency masks for digit 0 from the MNIST test set for a CNN shown in diverging color. **Bottom Left**: Spearman rank correlation (with absolute values) bar graph for saliency methods. We compare the similarity of explanations derived from a model trained on random labels, and one trained on real labels. **Bottom Right**: Spearman rank correlation (without absolute values) bar graph for saliency methods for MLP. See appendix for corresponding figures for CNN, and MLP on Fashion MNIST.

Remarks

- **Pros:**
  - Sanity checks are simple, can rule out flawed methods
  - A first step before investing more time
- **Cons:**
  - Often not quantitative
  - Says little about an explanation’s correctness

Today

- Section 1
  - Sanity checks
  - Ground truth comparisons 
- Section 2
  - Ablation metrics
  - Other criteria

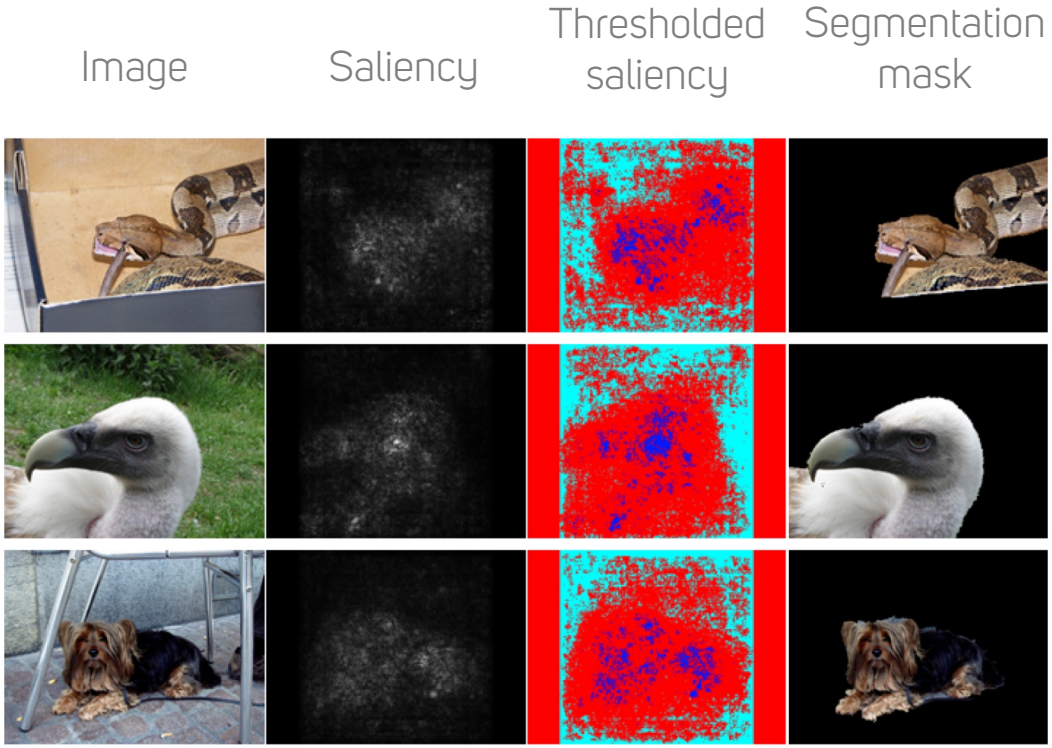
Ground truth importance

- Assume prior knowledge of “truly important” features
- Prior knowledge has various sources
  - Doctor annotations of medical images
  - Non-expert annotations of natural images
  - Genes with known role in disease (from survey of biology literature)
- Then, compare explanations to ground truth

Object localization

- Generate a bounding box from saliency map
- Then, compare to ground truth bounding box
  - Calculate area of overlap, count as correct localization if overlap exceeds threshold

Object localization (cont.)



Simonyan et al., "Deep inside convolutional networks: visualising image classification models and saliency maps" (2013)

Object localization (cont.)

- Generating bounding boxes is non-trivial
  - Can significantly affect the results
  - A simple approach:
    - Threshold saliency (e.g., at 50% quantile)
    - Find smallest bounding box containing salient features
- Simonyan et al. (2013) used a better approach
  - Inferred object and background colors using >95% and <30% salient features, did color segmentation
  - Strong results, despite using vanilla gradients

Object localization (cont.)

Localization errors can be low, despite models not being trained for localization ("weakly supervised")

Center	Grad [12]	Guid [13]	LRP [1]	CAM [20]	Exc [18]	Feed [2]	Mask [3]	This Work
46.3	41.7	42.0	57.8	48.1	39.0	38.7	43.1	**36.9**

Table 2: Localisation errors(%) on ImageNet validation set for popular weakly supervised methods. Error rates were taken from [3] which recalculated originally reported results using few different mask thresholding techniques and achieved slightly lower error rates. For a fair comparison, all the methods follow the same evaluation protocol of [2] and produce saliency maps for GoogLeNet classifier [15].

Dabkowski & Gal, "Real time image saliency for black box classifiers" (2017)

Pointing game

- A simpler localization task, no need to generate bounding boxes
- Check if explanation's most important pixel is within ground truth bounding box

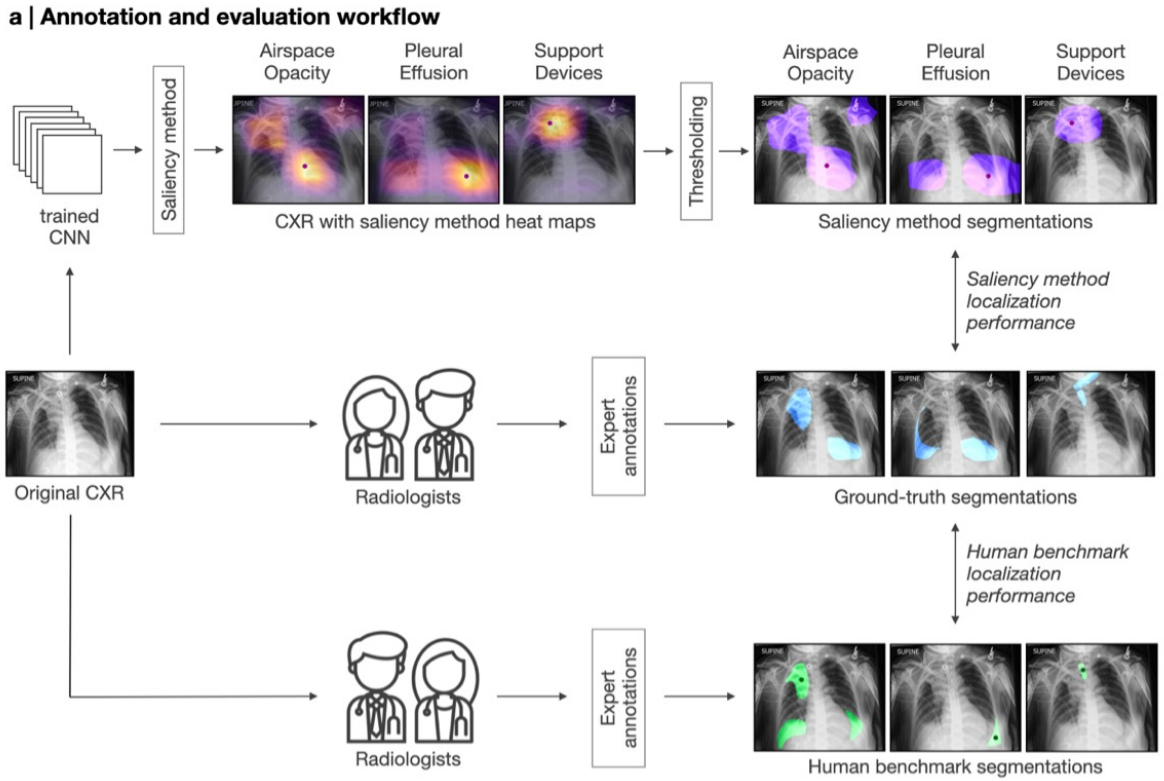
(Table from Petsiuk et al.)

Table 2: Mean accuracy (%) in the pointing game. Except for RISE, the rest require white-box model.

Base model	Dataset	AM [25]	Deconv [31]	CAM [33]	MWP [32]	c-MWP [32]	RISE
VGG16	VOC	76.00	75.50	-	76.90	80.00	**87.33** ± 0.49
	MSCOCO	37.10	38.60	-	39.50	49.60	**50.71** ± 0.10
Resnet50	VOC	65.80	73.00	**90.60**	80.90	89.20	88.94 ± 0.61
	MSCOCO	30.40	38.2	**58.4**	46.8	57.4	55.58 ± 0.51

Zhang et al., "Top-down neural attention by excitation backprop" (2016)

Localization in radiology



Saporta et al., "Benchmarking saliency methods for chest X-ray interpretation" (2021)

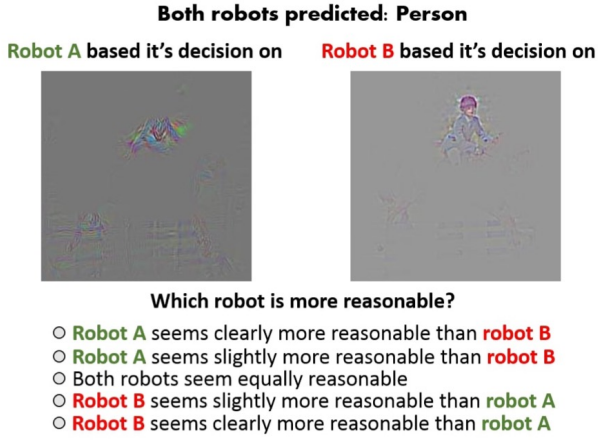
User studies

- Generate explanations using multiple methods, let humans decide which is best
  - Typically done on Mechanical Turk
- Different studies ask different questions
  - Which explanation is better, whether explanation indicates class, etc.

User studies

(a) Raw input image. Note that this is not a part of the tasks (b) and (c)

(b) AMT interface for evaluating the class-discriminative property



(c) AMT interface for evaluating if our visualizations instill trust in an end user

Fig. 5: AMT interfaces for evaluating different visualizations for class discrimination (b) and trustworthiness (c). Guided Grad-CAM outperforms baseline approaches (Guided-backprop and Deconvolution) showing that our visualizations are more class-discriminative and help humans place trust in a more accurate classifier.

Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization" (2017)

Synthetic datasets

- Synthetically generated data lets you control the ground truth

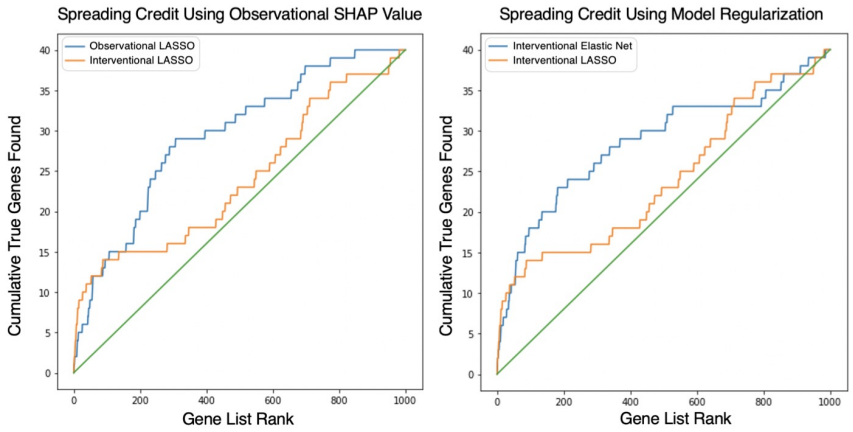


Figure 4. Left: When explaining a sparse model (Lasso regression), more true features are recovered when using the observational Shapley value to spread credit among correlated features than using the interventional Shapley value. Right: When using the interventional Shapley value, we recover more true features when the underlying model spreads credit among groups of correlated features (Elastic Net) than when the underlying model is sparse (Lasso).

Chen et al., "True to the model or true to the data?" (2020)

Challenges with ground truth

- Prior knowledge comes from humans
  - Difficult to obtain extra annotations
  - Reflects current understanding of the world
    - Penalizes models for using new, legitimate signals
- Not always derived from experts
  - Doctor annotations are probably trustworthy
  - Mechanical Turk users are less reliable

Jointly testing model and explanation

- For best results, we require two things:
  1. Explanations that correctly identify a model's dependencies
  2. A model that depends on the "correct" signals
     - Cannot use shortcuts or confounders (e.g., image background)

- **Problem:** poor results may be due to the model
  - Ground truth metrics don't directly test the explanation

A mathematical view

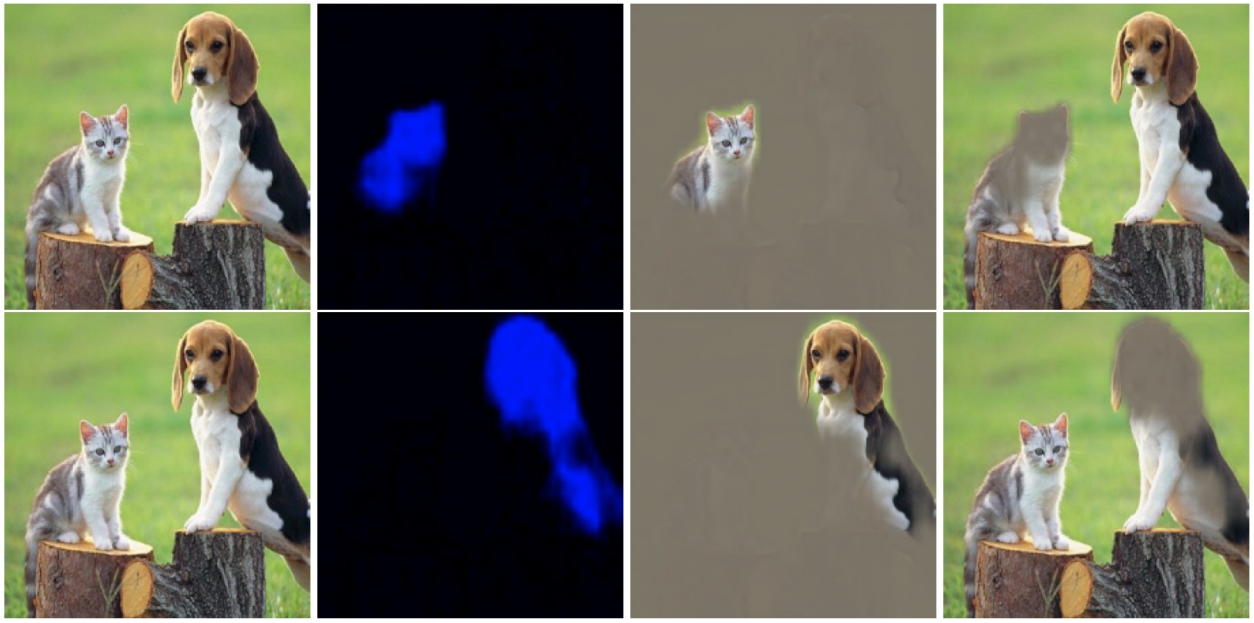
- Consider a classification problem, let $\mathrm{p}(y \mid x)$ be the **true** conditional probability
  - Assume an input x and label y where $\mathrm{p}(y \mid x) \approx 1$
  - Assume we can examine $\mathrm{p}(y \mid x_S)$ for all $S \subseteq \{1, \ldots, d\}$

- Ideally, the "truly important" features x_S should satisfy:

$$\mathrm{p}(y \mid x_S) \approx 1 \quad \longleftarrow \text{Sufficient}$$

$$\text{Necessary} \longrightarrow \quad \mathrm{p}(y \mid x_{\bar{S}}) \approx 0$$

A mathematical view (cont.)



(a) Input Image (b) Generated saliency map (c) Image multiplied by the mask (d) Image multiplied by inverted mask

Figure 1: An example of explanations produced by our model. The top row shows the explanation for the "Egyptian cat" while the bottom row shows the explanation for the "Beagle". Note that produced explanations can precisely both highlight and remove the selected object from the image.

Dabkowski & Gal, "Real time image saliency for black-box classifiers" (2017)

A mathematical view (cont.)

- Assume that $f_y(x) = \mathrm{p}(\, y \mid x \,)$
  - This is the implicit goal of model training

- Then, assume we can marginalize out features with their **conditional distribution**:

$$\mathbb{E}_{x_{\bar{S}} \mid x_S}\left[f_y(x)\right] = \mathrm{p}(\, y \mid x_S \,)$$

- This suggests that we can use removal-based methods to identify correct features x_S

Remarks

- **Pros:**
  - Ground truth metrics reflect the goal of XAI in some use cases: identifying true relationships in the data
- **Cons:**
  - Obtaining ground truth is difficult, imperfect
  - For good results, need a correct explanation and a correct model

Today

- Section 1
  - Sanity checks
  - Ground truth comparisons
  - **10 min break**
- Section 2
  - Ablation metrics
  - Other criteria

Evaluating XAI (continued)

CSEP 590B: Explainable AI

Ian Covert & Su-In Lee
University of Washington

Today

- Section 1
  - Sanity checks
  - Ground truth comparisons
- Section 2
  - Ablation metrics 
  - Other criteria

Ablation metrics

- Assume we can evaluate models with held-out features
- Importance values suggest how the prediction should change
  - Remove important features → prediction should change significantly

- **Idea:** test if explanations predict behavior with held-out features

Insertion/deletion

- Rank features x_i by importance a_i
- **Insertion:** add features, starting with the most important
  - Prediction should go up quickly
- **Deletion:** remove features, starting with most important
  - Prediction should drop quickly

Petsiuk et al., "RISE: Randomized input sampling for explanation of black-box models" (2018)

Insertion/deletion (cont.)

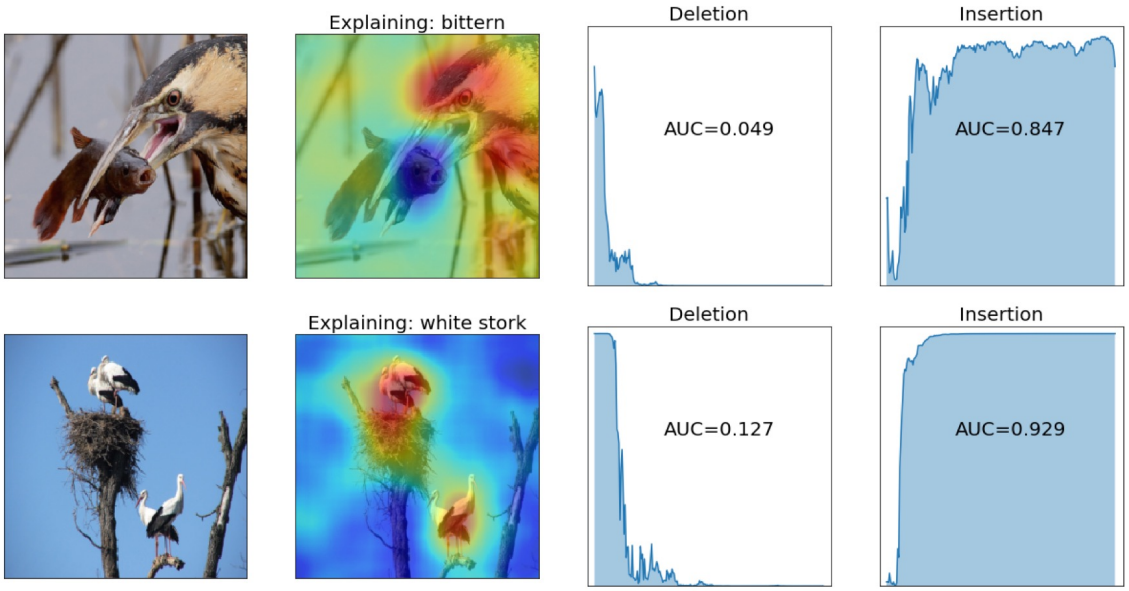


Figure 4: RISE-generated importance maps (second column) for two representative images (first column) with deletion (third column) and insertion (fourth column) curves.

Petsiuk et al., "RISE: Randomized input sampling for explanation of black-box models" (2018)

Insertion/deletion (cont.)

Table 1: Comparative evaluation in terms of deletion (lower is better) and insertion (higher is better) scores on ImageNet dataset. Except for Grad-CAM, the rest are black-box explanation models.

Method	ResNet50		VGG16	
	Deletion	Insertion	Deletion	Insertion
Grad-CAM [23]	0.1232	0.6766	0.1087	0.6149
Sliding window [31]	0.1421	0.6618	0.1158	0.5917
LIME [21]	0.1217	0.6940	0.1014	0.6167
RISE (ours)	**0.1076** $\pm$ **0.0005**	**0.7267** $\pm$ **0.0006**	**0.0980** $\pm$ **0.0025**	**0.6663** $\pm$ **0.0014**

Petsiuk et al., "RISE: Randomized input sampling for explanation of black-box models" (2018)

Many possible variations

- Measure different model behaviors
  - Prediction probability
  - Log-probability, log-odds
  - Accuracy

← Should not make a big difference

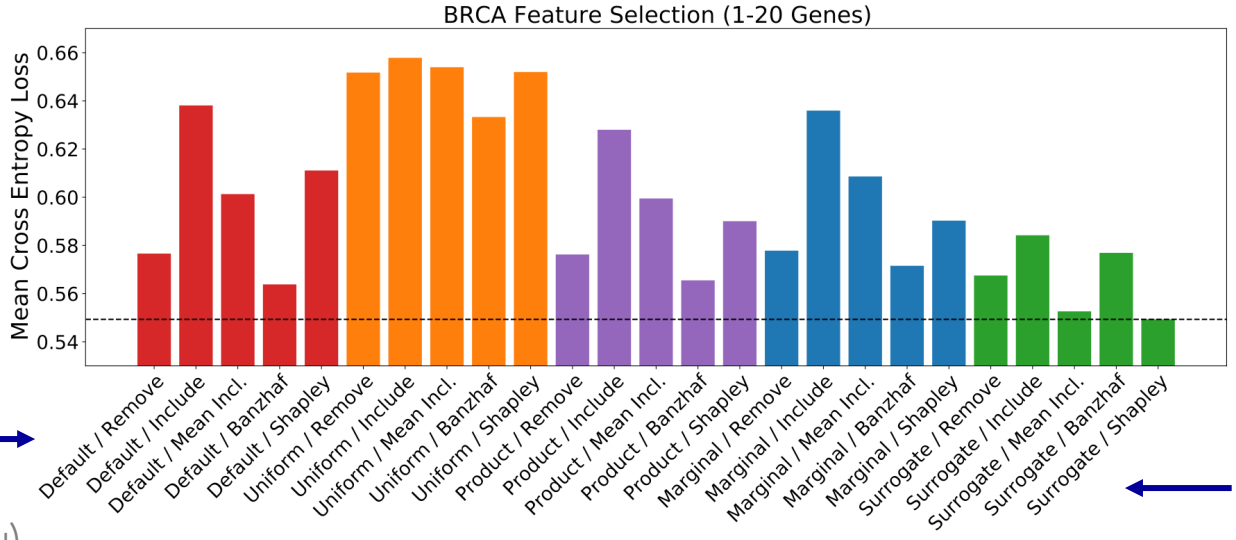
- Remove features differently
  - Zeros
  - Random noise
  - Sampled values from dataset

← Can make a big difference

Feature selection variation

- Can we apply the same idea to evaluate global explanations?
- Retrain models with **most** (**least**) important features
  - Should observe **high** (**low**) accuracy

Feature selection variation (cont.)



Grid of global explanations (removal/summary)

Best-performing method

Figure 12: Feature selection results for BRCA subtype classification when using top genes identified by each global explanation. Each bar represents the average loss for models trained using 1-20 top genes (lower is better).

Covert et al., "Explaining by removing: a unified framework for model explanation" (2020)

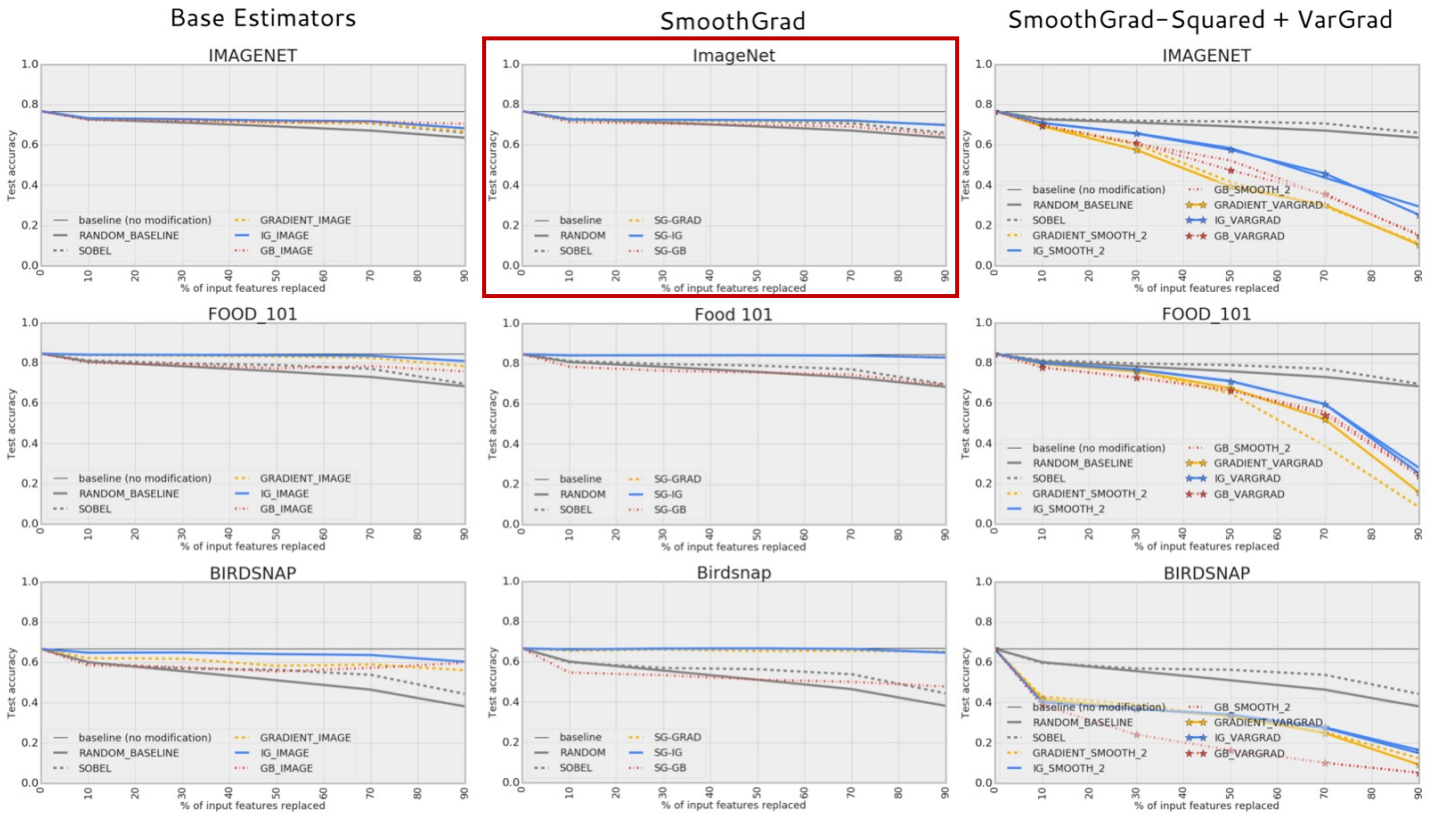
Remove and retrain (ROAR)

- Models are not made to handle missing features

- **Idea:** retrain with top features missing, test if accuracy drops
  - Mask important features
  - Retrain model with masked inputs
  - Measure the drop in accuracy

Hooker et al., "A benchmark for interpretability methods in deep neural networks" (2019)

ROAR (cont.)



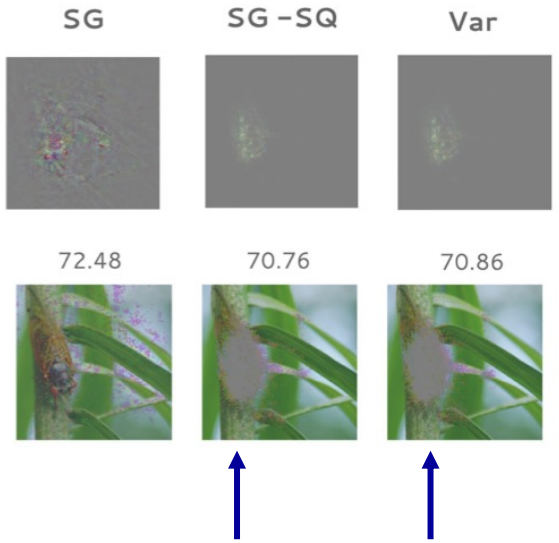
Hooker et al., "A benchmark for interpretability methods in deep neural networks" (2019)

ROAR problems

- Retraining many models is costly
- Does not test explanation's correctness *for the original model*
- Training with masking encourages use of confounders, yields inflated accuracy
  - 63% ImageNet accuracy with 90% of features masked is suspiciously high

ROAR problems (cont.)

- Information leakage problem
  - Masking is not random
  - Removed features can indicate class label



Given background and shape/size, likely an insect

Limitations

- So far, focused on importance *rankings*
  - Invariant to addition/multiplication by a constant
  - Invariant to any change that preserves ordering

- How to test the importance scores $a_i \in \mathbb{R}$ more precisely?

Additive proxy metrics

- Many methods have scores a_i that sum to the prediction (IntGrad, LRP)
- Some are explicitly designed as additive proxies for the model (LIME, SHAP)

- **Idea:** test accuracy of importance scores as additive proxy

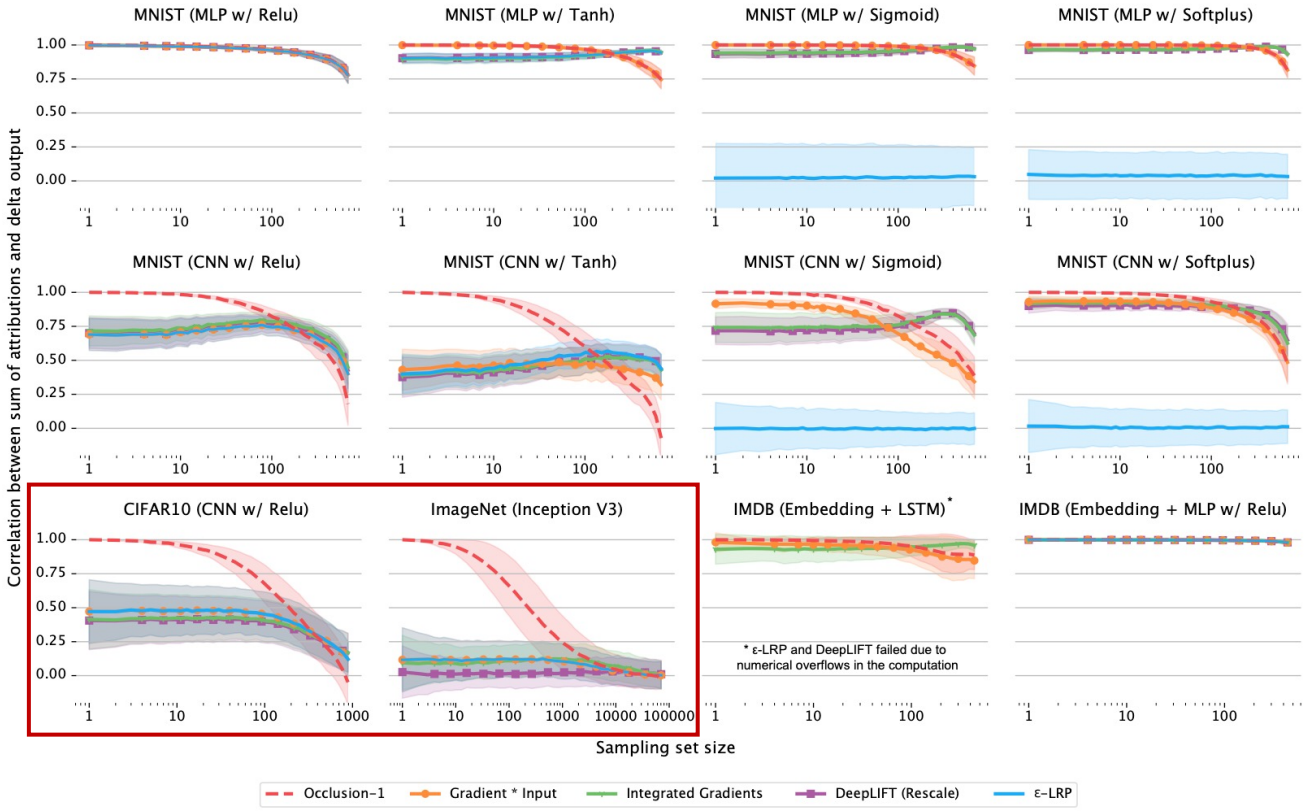
Sensitivity-n

- Test the proxy's correlation for random subsets with fixed cardinality
  - Uniform distribution over $S \subseteq \{1, \ldots, d\}$ with $|S| = n$

$$\text{Corr}\left(f_y(x_S), \sum_{i \in S} a_i\right)$$

Ancona et al., "Towards better understanding of gradient-based attribution methods for deep neural networks" (2018)

Sensitivity-n (cont.)



Ancona et al., "Towards better understanding of gradient-based attribution methods for deep neural networks" (2018)

Variable cardinality version

- Calculate the same correlation, but with subsets of different cardinalities
  - Require a distribution $\mathrm{p}(S)$ over all cardinalities
  - Uniform over all $S \subseteq \{1, \ldots, d\}$, or uniform over cardinalities

$$\mathrm{Corr}\left(f_y(x_S), \sum_{i \in S} a_i\right)$$

Related metrics

- Insertion/deletion
  - Samek et al., "Evaluating the visualization of what a deep neural network learned" (2015)
  - Lundberg et al., "From local explanations to global understanding with explainable AI for trees" (2020)
- Sensitivity-n
  - Alvarez-Melis & Jaakkola, "Towards robust interpretability with self-explaining neural networks" (2018)
  - Bhatt et al., "Evaluating and aggregating feature-based model explanations" (2020)

Feature removal choice

- Ablation metrics mirror removal-based explanations
- Same question of how to remove features
  - Likely no good default value
  - If we retrain, we're not analyzing the original model
  - Replacing with random values is an option, but which distribution do we use?
    - Marginalizing with conditional gives best-effort predictions with partial input, but difficult to implement

Feature removal choice (cont.)

- A metric's feature removal choice favors similar explanations
  - E.g., when using insertion/deletion with zeros masking, SHAP with zeros beats SHAP with marginal distribution
  - See illustrative experiment in Covert et al.

Covert et al., "Explaining by removing: a unified framework for model explanation" (2021)

Remarks

- **Pros:**
  - Ablation metrics test an explanation's correctness for the model, rather than what's important to humans
  - No extra data annotation required
- **Cons:**
  - Difficult choice of how to remove features
  - In some cases, not focused on the original model (ROAR)

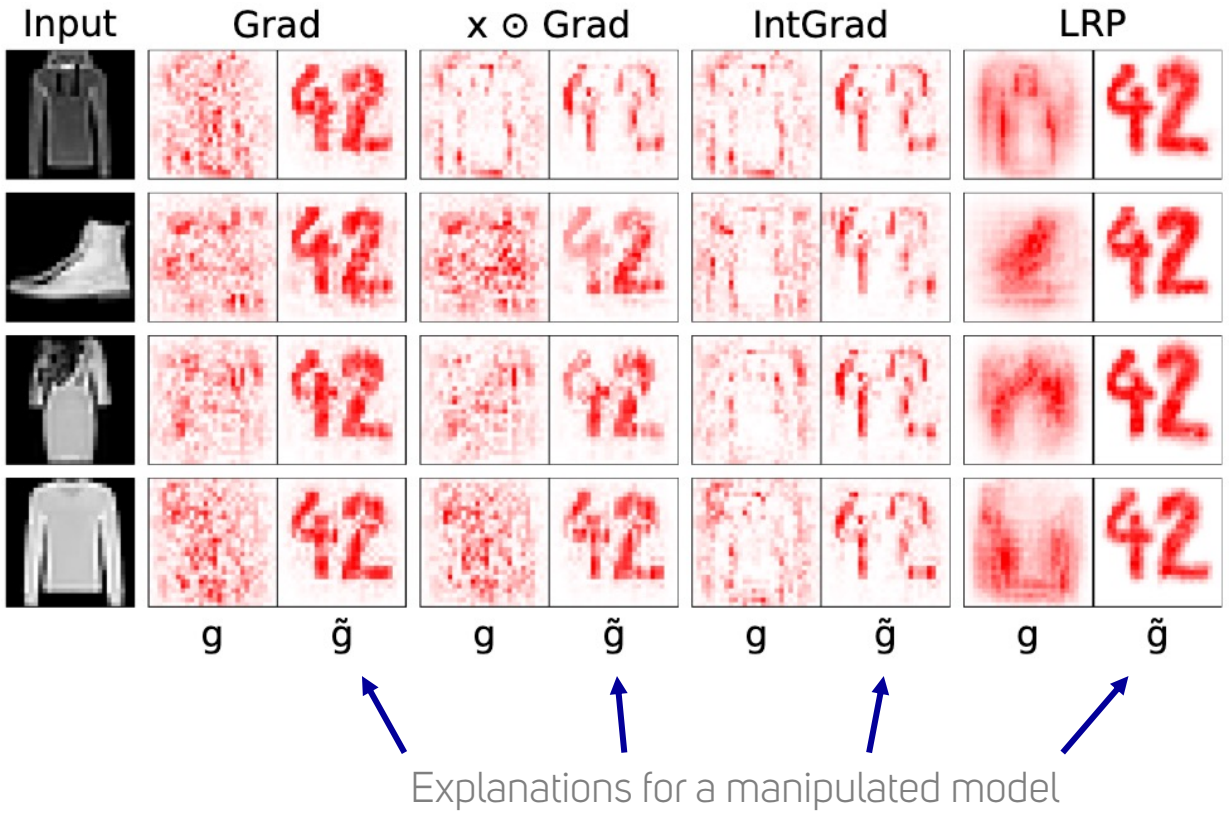
Today

- Section 1
  - Sanity checks
  - Ground truth comparisons
- Section 2
  - Ablation metrics
  - Other criteria 

Robustness

- Adversarial examples: imperceptible changes that affect the prediction
  - Szegedy et al., "Intriguing properties of neural networks" (2013)
  - Similar ideas have been explored in XAI
- Are explanations robust to small changes in the **data**?
  - Ghorbani et al., "Interpretation of neural networks is fragile" (2018)
- Are explanations robust to small changes in the **model**?
  - Anders et al., "Fairwashing explanations with off-manifold detergent" (2020)
  - Slack et al., "Fooling LIME and SHAP: Adversarial attacks on post-hoc explanation methods" (2019)

Robustness (cont.)



Anders et al., "Fairwashing explanations with off-manifold detergent" (2020)

Hyperparameter sensitivity

- Many methods have hyperparameter choices
  - Number of samples (LIME)
  - Baseline/removal approach (IntGrad)
  - Superpixel size (occlusion)

- Problematic when a parameter…
  1. Has large impact on results
  2. Doesn't have a clear "right" choice

Bansal et al., "SAM: The sensitivity of attribution methods to hyperparameters" (2020)

Human utility

- How **useful** is an explanation?

- Must specify the use-case
  - Human-AI team setting
    - E.g., calibrating confidence in model decisions
    - We'll discuss this in a later lecture
  - Scientific setting
    - E.g., identifying biological hypotheses that are later verified
    - Difficult to test at scale

Conclusions

Summary

- Sanity checks
  - Failing these is not okay, but many methods will pass
- Ground truth comparisons
  - Extra annotations can be laborious
  - Tests both model and explanation, which may or may not reflect intended usage
    - E.g., identify regions to direct doctor focus
- Ablations
  - Best option to test explanation's correctness for the model
  - Several good metrics: insertion/deletion, sensitivity-n
  - Tricky choice: how to hold out features

When to use these metrics?

- Mainly when developing a new method
  - Prove that it works
  - Show benefits over prior methods

- Additionally, when deciding what to use with a new model/dataset
  - Verify implementation choices
  - Bhatt et al., "Explainable machine learning in deployment" (2020)

Perspective

- No method is *wrong*, but some are misaligned with user questions
  - Metrics effectively formalize user questions
  - Can design metrics for other user objectives as needed