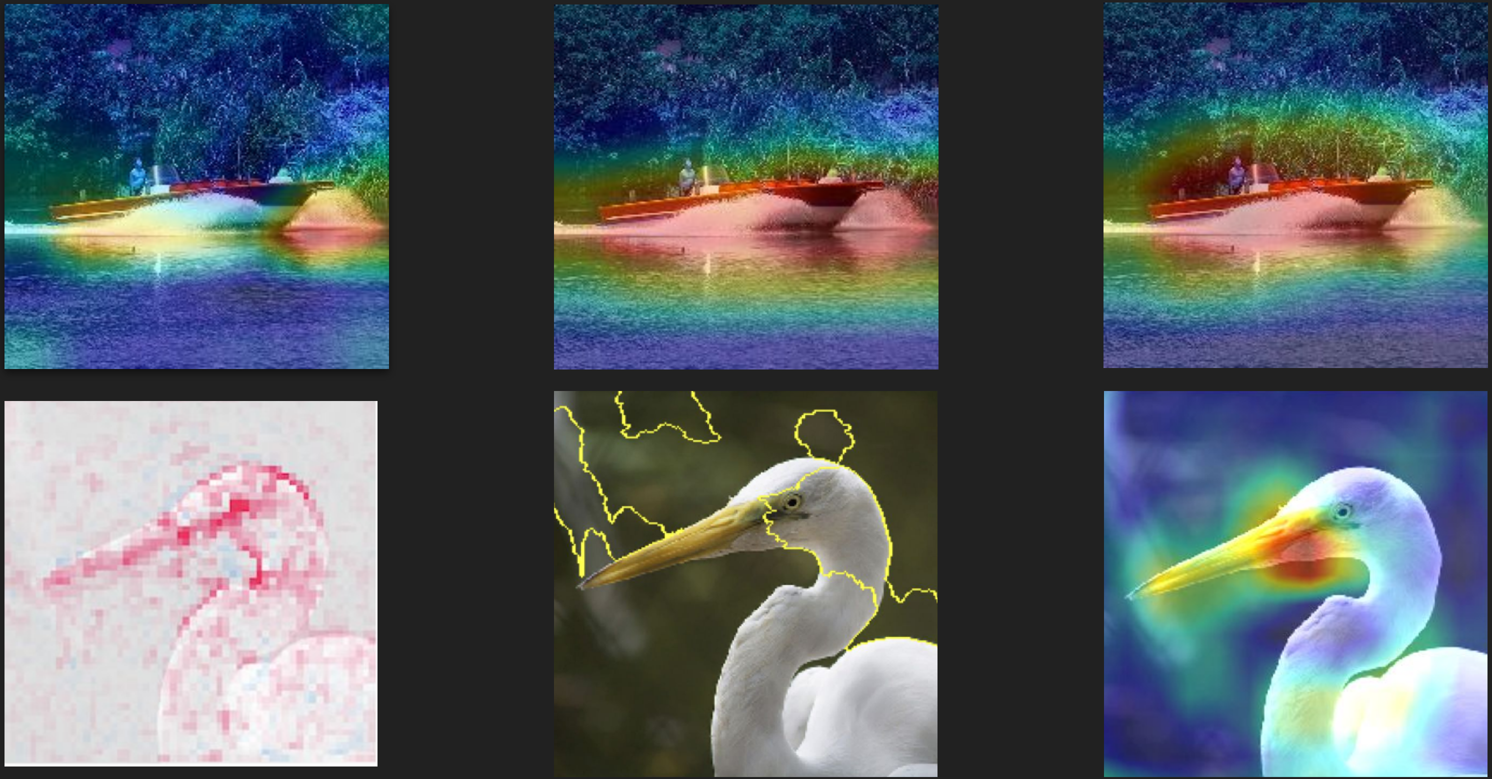
Exploring XAI methods on Image Data

ECE 6960 Explainable ML

Team Name: Raising Hands

Ananya Devarakonda, Heuisu Kim

Motivation: Which explanation is better?



Data Description

Dataset: **Imagenet50**[1]

- Dataset of similar images to original ImageNet dataset but randomly collected through Google image search
- Pros:
  - Extremely diverse image dataset
  - Focus on explaining the models trained on ImageNet data while avoiding leakage
  - low computation overhead (compared to ImageNet test set 100,000 images)
- Cons
  - Small Image sizes
  - Ground truth labels are uncertain
- Preprocessed dataset to run in backbone, XAI methods

Example images

1. https://shap.readthedocs.io/en/latest/generated/shap.datasets.imagenet50.htm (Compiled by the authors of SHAP)

Exploration

Dataset

Imagenet50 (*n=50*)

Backbone Models

1. VGG 16
2. Resnet50
3. Densenet 121

**Used pretrained model weights (on Imagenet) for generating predictions on the dataset*

Explainable methods

1. SHAP (**gradient explainer**)
2. LIME
3. GradCAM

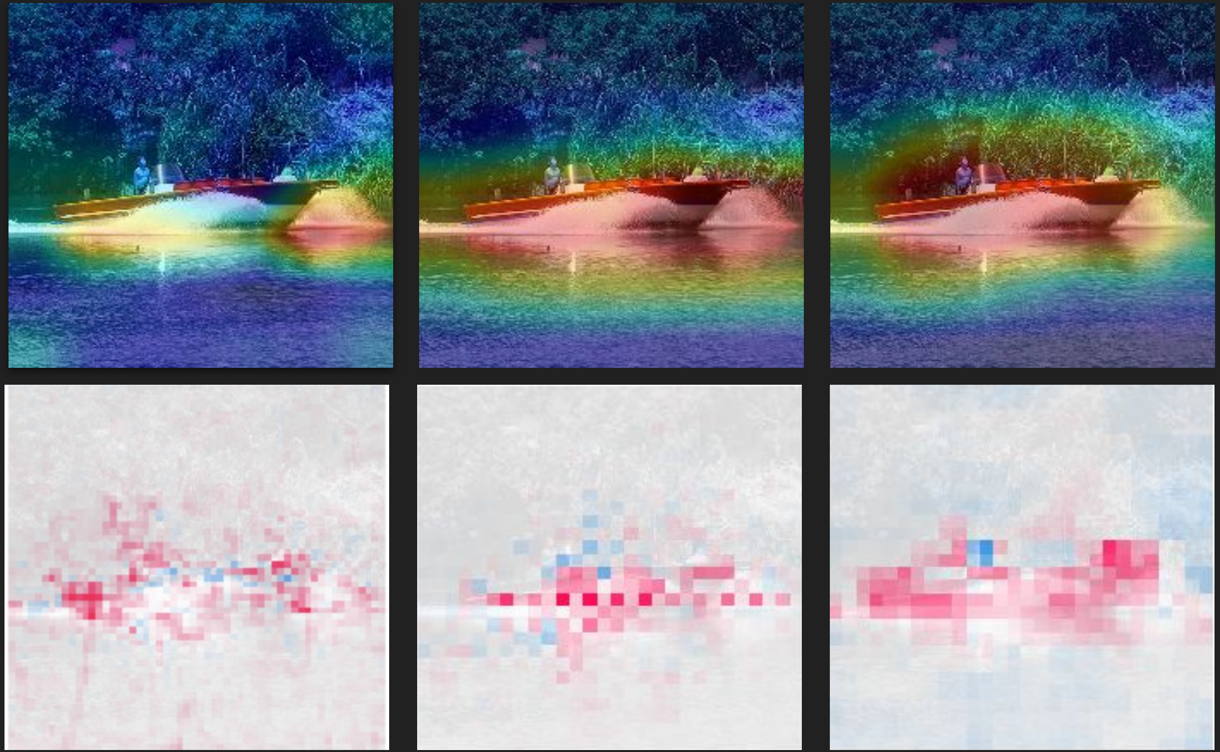
Qualitative Results

Original Image
“speedboat”

VGG16 Resnet50 Densenet121

GradCAM

SHAP

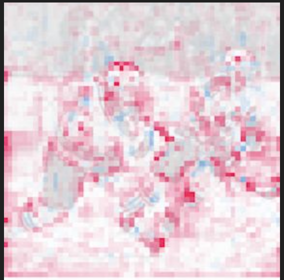


Qualitative Evaluation

SHAP

- Generally captured more helpful regions to understand the ground truth
- Sometimes the highly **relevant regions were scattered** & still identified even when the ground truth label was questionable

E.g. “puck”



GradCAM

- Smaller heatmaps were reshaped to overlap image, appearing blurry and distorted
- Most activated regions of the heatmap highlighted areas that were meaningful and helpful to understanding the ground truth

LIME

- Initially explored, but omitted
- Superpixels were often not captured for most images
- Experimented thresholds, but ultimately very low threshold resulted in **superpixel segmentation not particularly helpful/ accurate**

Quantitative Evaluation

1. % Increase in Confidence
2. % Drop in Confidence
3. % Increase in Confidence with ROAD
4. % Drop in Confidence with ROAD
5. IOU

$$\left(\sum_{i=1}^{N} \frac{\mathbb{1}_{Y_i^c < O_i^c}}{N}\right)100$$

% Increase in Confidence

$$\left(\sum_{i=1}^{N} \frac{max(0, Y_i^c - O_i^c)}{Y_i^c}\right)100$$

% Drop in Confidence

Quantitative Evaluation: Definitions

Original Image

Class: “speedboat”

Occluded “b”

(background is occluded)

Occluded “s”

(subject is occluded)

Quantitative Evaluation: GradCAM

Model \ Metric	**% ↑ in Confidence (Occlusion “b”)[a]**	**% ↓ in Confidence (Occlusion “b”)[b]**	**% ↑ in Confidence (Occlusion “s”)[b]**	**% ↓ in Confidence (Occlusion “s”)[a]**
VGG-16	16.00	36.08	8.00	33.11
ResNet-50	**26.00**	15.64	**2.00**	**56.72**
DenseNet-121	14.00	**15.13**	**2.00**	50.73

a. *Higher the better*
b. *Lower the better*

Quantitative Evaluation: SHAP

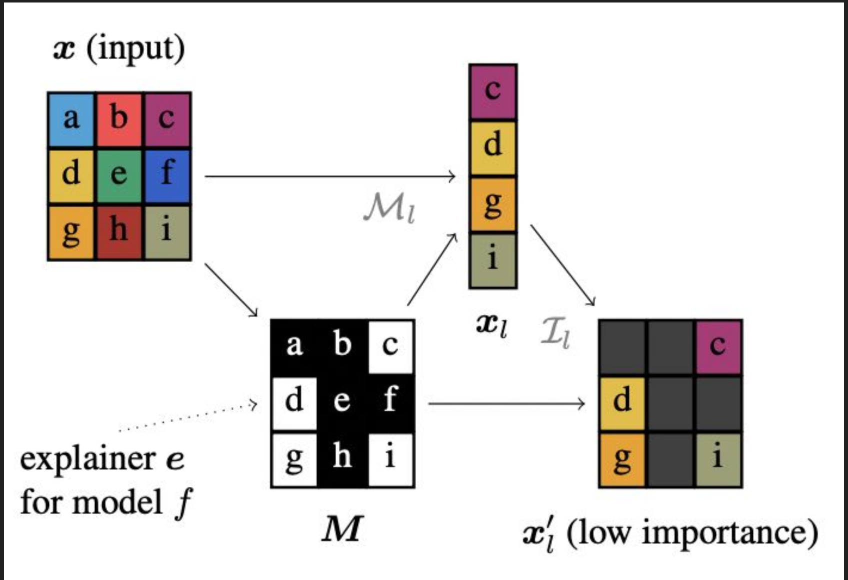
Model \ Metric	**% ↑ in Confidence (Occlusion “b”)[a]**	**% ↓ in Confidence (Occlusion “b”)[b]**	**% ↑ in Confidence (Occlusion “s”)[b]**	**% ↓ in Confidence (Occlusion “s”)[a]**
VGG-16	4.00	66.04	0.0	**86.62**
ResNet-50	**4.00**	64.66	**2.00**	85.47
DenseNet-121	2.00	**66.73**	**0.0**	75.98

a. *Higher the better*
b. *Lower the better*

Side track on evaluation methods: ROAD

- X = input image x (9 pixels a–i)
- M = mask produced by explanation method where important pixels are indicated in black
- Ml = masking operator that extracts remaining, less important pixel values xl and transforms to an imputed variant of the input x 0 l

Goal is to **separate the information** contained in the binary mask M



*Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "A consistent and efficient evaluation strategy for attribution methods."

Intuition: Remove and Debias (ROAD)

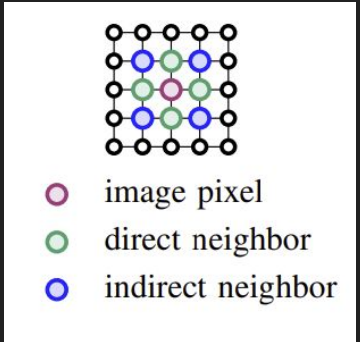
*“…classifier that **infers the class just from the location of the masked out pixels** and obtain high accuracy.”*

*Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, “A consistent and efficient evaluation strategy for attribution methods.”

Intuition: ROAD (Continued)

Propose a *Noisy Linear Interpolation strategy* – approximate each pixel by a weighted mean of its neighboring pixels.

(perturbations are more difficult to detect) – still not ideal, but is an improvement



*Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, “A consistent and efficient evaluation strategy for attribution methods.”

Quantitative Evaluation: ROAD-Definitions

Original Image

Class: “speedboat”

ROAD “b”

(background is hidden)

ROAD “s”

(subject is hidden)

Quantitative Evaluation: ROAD-GradCAM

Model \ Metric	**% ↑ in Confidence (ROAD "b")**[a]	**% ↑ in Confidence (ROAD "s")**[b]	**Mean ↑ in Confidence (Combined)**[a]
VGG-16	-13.33	-17.78	2.22
ResNet-50	-3.00	-26.10	11.55
DenseNet-121	**-2.97**	**-41.48**	**19.44**

a. *Higher the better*
b. *Lower the better*

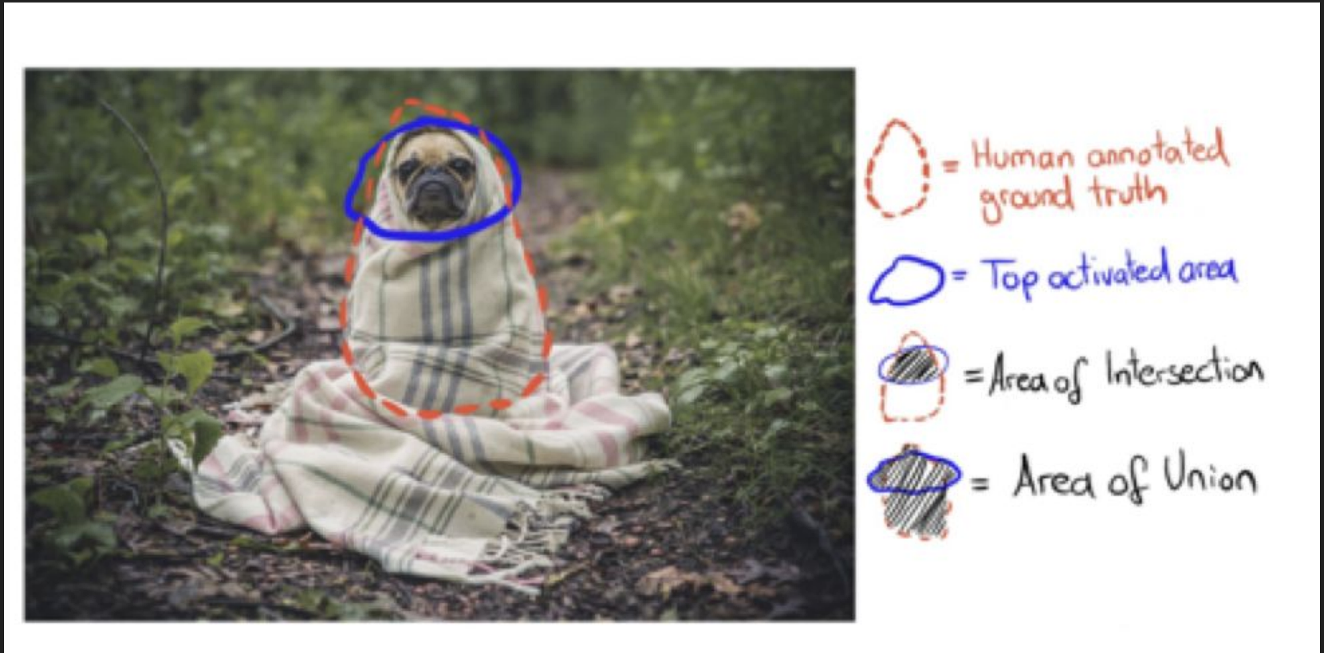
Quantitative Evaluation: ROAD-SHAP

Model \ Metric	**% ↑ in Confidence (ROAD "b")[a]**	**% ↑ in Confidence (ROAD "s")[b]**	**Mean ↑ in Confidence (Combined)[a]**
VGG-16	**-20.54**	-19.84	**-0.355**
ResNet-50	-32.39	-31.12	-0.636
DenseNet-121	-48.26	**-35.40**	-0.643

a. *Higher the better*
b. *Lower the better*

Quantitative Evaluation: IoU

Intersection Over Union (IoU) or Jaccard Index



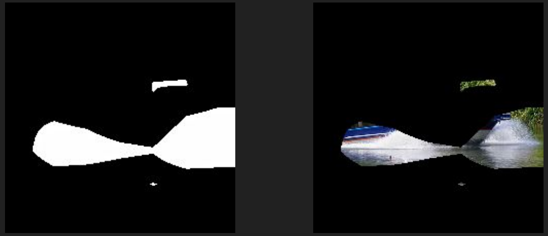
*Lecture Slides, C. Molnar

Quantitative Evaluation: IoU-GradCAM

Intersection Over Union (IoU) or Jaccard Index; Masks:

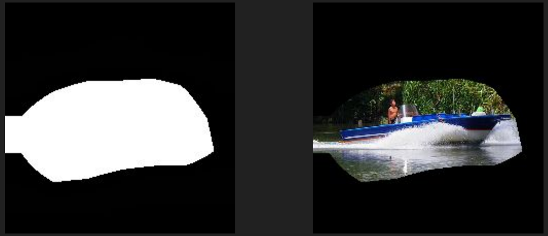
Original Image

Ground truth



VGG-16

ResNet-50



DenseNet-121

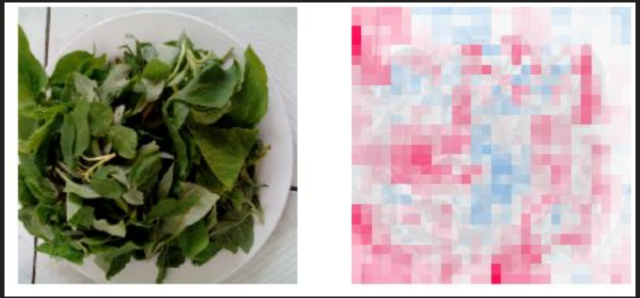
Quantitative Evaluation: IoU-Overview

XAI Method	Metric	VGG-16	ResNet-50	DenseNet-121
GradCAM	mIoU (%)	12.28	31.78	**38.75**
SHAP	mIoU (%)	19.48	26.24	27.65

Challenges

- Many images without meaningful masks (especially LIME)
- Consistency when comparing XAI methods
  - Gradient Explainer –results of course drastically change depending on the identified layer and architecture of backbone model
- Evaluation methods
  - How do we quantify “good” explanations?
  - Do quantified metrics fairly capture the explanation accuracy? Does it make sense to us, humans?
  - Technical challenges of just getting the explanation masks

“Picket fence”
LIME using VGG16



“Pot”
SHAP from Densenet121 (layer 7)

Practical Challenges

- Make segmentations by hand

“snow”

“flower”

“coast”

Conclusion

1. Both perturbation methods (occlusion, ROAD) and object localization metrics (IoU) have advantages and disadvantages, a good metric could be a combination of both
2. There is inherent subjectivity to explanations
3. There is scope to analyze more XAI methods and other evaluation metrics

References

1. M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?” explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135–1144, 2016.
2. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in neural information processing systems, vol. 30, 2017.
3. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017.
4. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929, 2016.
5. A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in 2018 IEEE winter conference on applications of computer vision (WACV), pp. 839–847, IEEE, 2018.
6. S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, “A benchmark for interpretability methods in deep neural networks,” Advances in neural information processing systems, vol. 32, 2019.
7. Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, “A consistent and efficient evaluation strategy for attribution methods,” arXiv preprint arXiv:2202.00449, 2022.
8. D. Li, H. Ling, S. W. Kim, K. Kreis, A. Barriuso, S. Fidler, and A. Torralba, “Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations,” 2022.
9. G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Mu ̈ller, “Explaining nonlinear classification decisions with deep taylor decomposition,” Pattern recognition, vol. 65, pp. 211–222, 2017.
10. S. Gadgil, M. Endo, E. Wen, A. Y. Ng, and P. Rajpurkar, “Chexseg: Combining expert annotations with dnn-generated saliency maps for x-ray segmentation,” arXiv preprint arXiv:2102.10484, 2021.
11. Imagenet50: https://shap.readthedocs.io/en/latest/generated/shap.datasets.imagenet50.htm

Thank you