Averaging Improves Saliency Explanations of Image Classifiers

Aadil Ahamed, Kamran Alipour, Michael Pazzani, Sateesh Kumar
(aahamed, kalipour, mpazzani, sakumar)@ucsd.edu

Introduction

- **Problem**: Many common XAI methods do not accurately identify the regions that human experts consider important.
- **Approach**: We propose to use of ensembles of neural networks to increase the accuracy of identifying regions of interest for any XAI algorithm by combining explanations from multiple neural networks.
- **Results**: Regions match more closely to expert annotations, variance across examples is reduced.

Approach

- **Ensemble Explanation**: We generate a diverse ensemble by initializing the classification head of each model with random weights. Once the models are trained, we generate an ensemble explanation by averaging the relevance score for each pixel i in the input image as:

$$score_{ens}(i) = \frac{1}{n}\sum_{j=1}^{n} score_j(i)$$

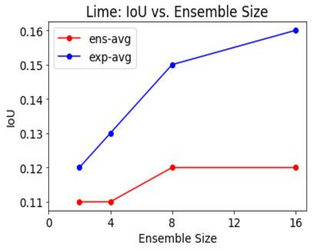
For this work, we combine 11 models for generating ensemble explanation.

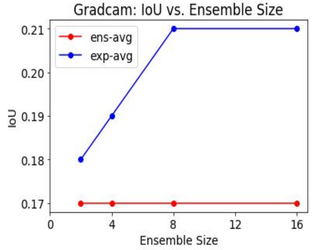
- **Dataset**: We present a new birds dataset called HiRes-Birds containing 14,380 images belonging to 66 bird species. We also experiment with the ISIC2018 melanoma dataset.
- **Metrics**: Distance between centre of mass of generated heatmap and of ground truth ROI and Correlation; IoU for the melanoma dataset.

Quantitative Results

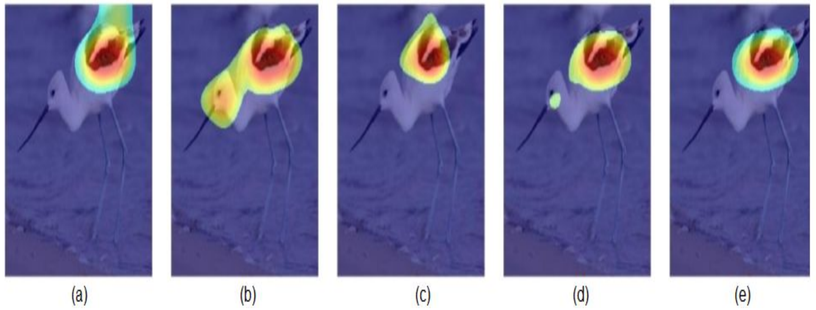
	Method	Ens-Avg	Exp-Avg
HiRes Birds (Beak)	GradCAM	0.395	**0.328**
	Input gradient	0.331	**0.320**
	Gradient Shap	0.346	**0.334**
	Integrated Gradients	**0.**346	**0.334**
	Saliency	*0.328*	***0.322***

	Method	Ens-Avg	Exp-Avg
HiRes Birds (Wings)	GradCAM	0.157	**0.142**
	Input gradient	0.153	**0.145**
	Gradient Shap	0.152	**0.146**
	Integrated Gradients	**0.**156	**0.146**
	Saliency	*0.158*	***0.147***





Qualitative Results



(a-d) are individual classifiers. (e) is average of 11 networks

The red arrowhead points to the CoM of the average while black arrows point to the CoM by individual classifiers

Future Work

- Weighting mechanism for ensembling.
- Reducing cost associated with training multiple models.