Rebuttal

March 2024

1 Fidelity

Fidelity score measures faithfulness of attributions. The soft fidelity scores are computed by taking KL-divergence of model's predictive probability for each class and the aggregated attribution scores for each class.

$$\text{Soft Fidelity}(\alpha, y) = \text{KL}(\sum_{i=1}^{m} \alpha_{li} || f(X)_l) \tag{1}$$

where α_{li} is the attribution of feature group $i \in [m]$ and class l where there are a total of m groups, and $f(X)_l$ is the predicted logits for class l.

Table 1 and Table 2 show fidelity of ImageNet using ViT and ResNet respectively. Table 3 shows fidelity of SST using BERT.

We can see that SOP and FRESH by construction are the only ones that achieve 0 or close to 0 on soft fidelity scores. Gradient-based methods have the worst fidelity.

This shows that only built-in methods have perfect faithfulness.

2 Purity Entropy Results on ImageNet-S

Table 4 shows a systematic experiments we do to compare SOP with all the baselines. We compute purity scores for percentage of the mask that belongs to the object vs. the background, and compute entropy for each mask. We then average the entropy scores for a subset of examples in ImageNet-S.

For SOP, we take the top 1 group as our mask. For LIME, we are taking the top 5 features as the default. For SHAP, IntGrad, and GradCAM, as there is no standard way of thresholding the mask, we threshold by the mean score of each image and take the top scored pixels as one mask.

Results in Table 4 shows that SOP is the second best, following Archipelago. This shows that our model is able to select semantically meaningful objects and background without supervision.

Attr. Method	Fidelity Soft (low: better)
LIME	0.29
SHAP	0.059
RISE	6.77
IntGrad	7.30
GradCAM	21.43
Archipelago	21.41
FullGrad	21.42
SOP (ours)	0.0

Table 1: Fidelity for ImageNet on ViT

Attr. Method	Fidelity Soft (low: better)
LIME	1.3334
SHAP	0.1585
RISE	14.4712
IntGrad	8.5138
GradCAM	12.7893
FullGrad	21.9777
SOP (ours)	0.0

Table 2: Fidelity for ImageNet on ResNet

3 MoRF and LeRF

MoRF and LeRF are perturbation methods that evaluate removing from the most important to least important features, or the reverse. Intuitively, if we remove the most important features, the predicted probability for the predicted class should drop the most, thus giving a low AUC. If we remove the least important features, the predicted probability for the predicted class should drop the least, thus giving a high AUC.

Table 5 shows result of ImageNet on ViT.

To do a fair comparison, we remove the same number of pixels (224) each time from ones with the most to least scores. For grouped features, we select the same number of pixels from the top 1 scored group, until that group runs out, and then the next highest group, until the last scored group, and lastly rest of the pixels.

We need to note that although SOP is faithful, it does not necessarily have the best score in perturbation tests. Since SOP selects multiple features into each group, it will unavoidably add some part of the features first and others later. This will result in not the optimal way of deletion if we are deleting on pixel or feature level.

Attr. Method	Fidelity Soft (low: better)
LIME	3.63
SHAP	0.29
RISE	0.04
IntGrad	5.94
Archipelago	2.83
PLS	2.27
FRESH	0.0
SOP (ours)	-2e-10

Table 3: Fidelity for SST - BERT

Attr. Method	Purity (Entropy)
LIME	0.69
SHAP	0.77
RISE	0.77
IntGrad	0.68
GradCAM	0.70
Archipelago	0.64
SOP (ours)	0.67

Table 4: Purity (Entropy) scores for different attribution methods.

Attr. Method	MoRF (low)	LeRF (high)
LIME	0.13	0.68
SHAP	0.20	0.62
RISE	0.35	0.66
IntGrad	0.53	0.51
GradCA,	0.35	0.59
Archipelago	0.32	0.49
FullGrad	0.33	0.62
SOP	0.44	0.53

Table 5: MoRF and LeRF for ImageNet on ViT