Testing Explanation Algorithms on Transformers

Mingqi Jiang, Saeed Khorram, Li Fuxin
Collaborative Robotics and Intelligent Systems (CoRIS) Institute, Oregon State University

Oregon State University

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Introduction

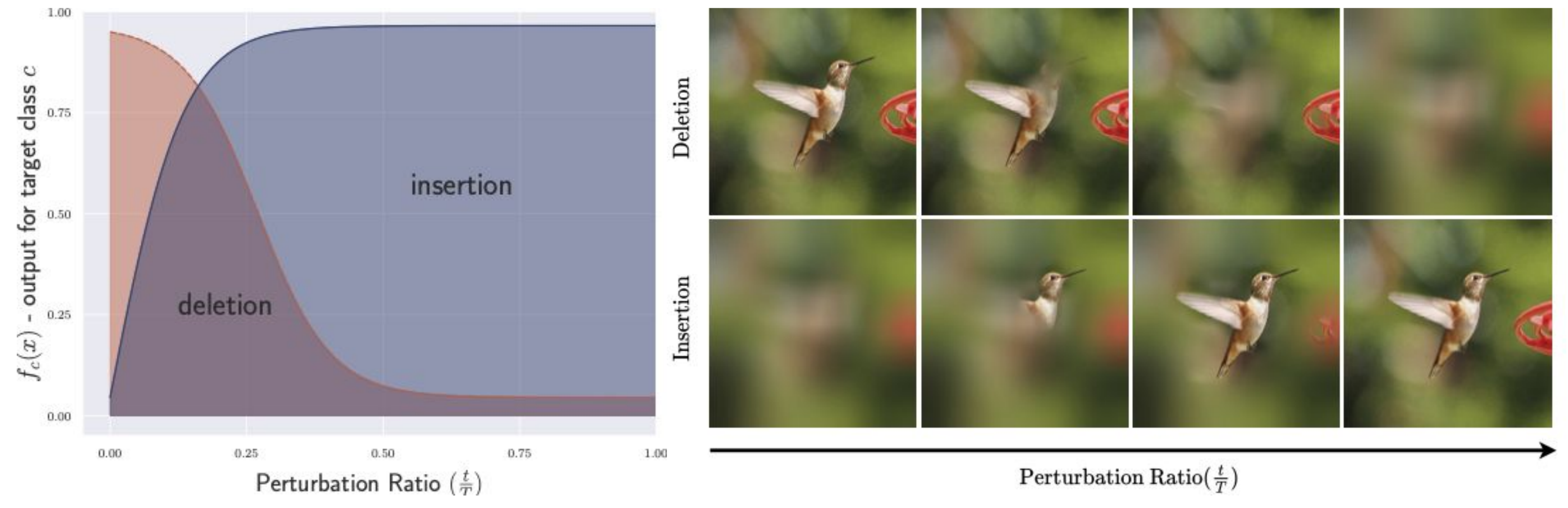
Comparing the underlying decision making process of Transformers and CNNs(ResNet50 and VGG19) by causally perturbing the image regions and observing the change in the output confidence.

Method

- **Cross-Testing**
  - Utilize one specific deep model to generate an attribution map by iGOS++, and use another deep model to assess the insertion/deletion metrics
  - Insertion metric(Deletion is input image I and the baseline image $\tilde{I}$ are swapped):

$$insertion = \frac{1}{T}\left\langle \sum_{t=0}^{T-1} \frac{1}{2}\left(f_c\left(\emptyset^{(t)}(I,\tilde{I},M)\right) + f_c\left(\emptyset^{(t+1)}(I,\tilde{I},M)\right)\right)\right\rangle_{p_{data}}$$

where $\emptyset^{(t)}(I,\tilde{I},M) = I \odot M^{(t)} + \tilde{I} \odot (1 - M^{(t)})$



  - Fair comparison: normalization $\overline{score} = {(score - b)}/{(t - b)}$ where t/b are the top-1/fully-blurred confidences
- **Minimal Sufficient Explanations(MSEs)**
  - Using beam search method of SAG to find diverse set of MSEs:a minimal conjunction /region that achieves a certain high classification confidence P_h

$$f_c(N_i) \geq P_h\, f_c(I), \max_{n_j \subset N_i} f_c(n_i) < P_h\, f_c(I)$$

- **Sub-Explanation Counting**
  - Construct a tree for each MSE by deleting one patch at a time from a parent node to generate child nodes.
  - Stop expansion when the nodes are with a confidence less than 50% compared to the classification confidence on the original image.

Robustness in terms of Insertion and Deletion

Model	Deletion ↓	Insertion ↑
Swin-T	0.1291	0.9433
ResNet50	0.1109	0.9437
VGG19	0.1254	0.9255

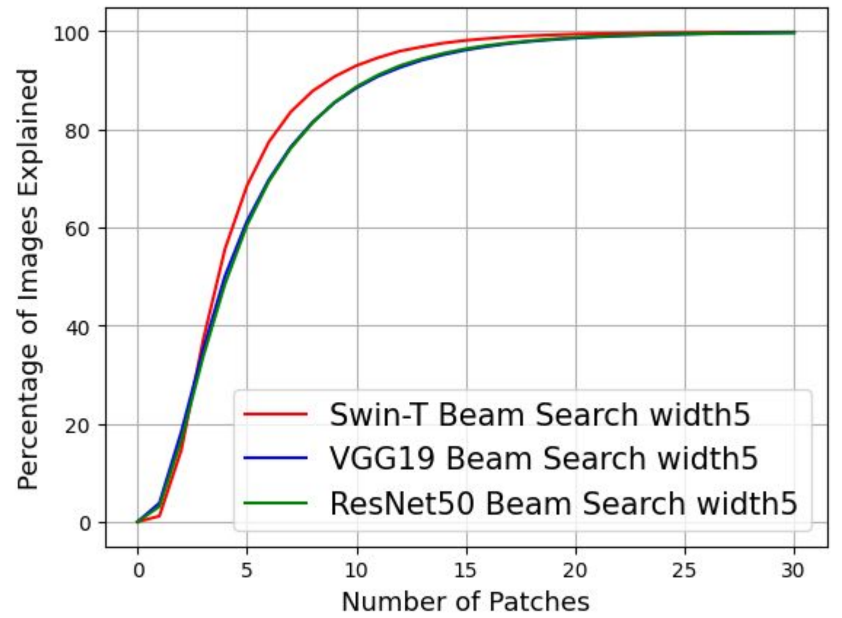
Model Pair	Deletion ↓	Insertion ↑
Swin-T → ResNet50	0.1771 (+0.0480)	0.7670 (-0.1763)
ResNet50 → Swin-T	0.2322 (+0.1068)	0.8179 (-0.1076)
Swin-T → VGG19	0.1731 (+0.0440)	0.7151 (-0.2282)
VGG19 → Swin-T	0.2449 (+0.1340)	0.8236 (-0.1019)

Quantitative results of deletion/insertion metrics using themselves and their counterpart heat maps

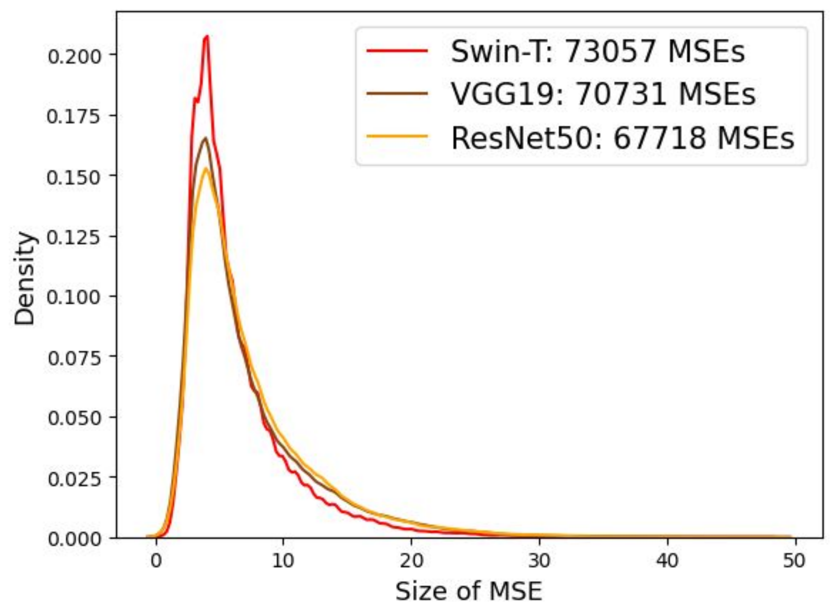
VGG19	ResNet50	**Swin-T**	VGG19	ResNet50	**Swin-T**	VGG19	ResNet50	**Swin-T**	VGG19	ResNet50	**Swin-T**	VGG19	**ResNet50**	Swin-T
0.0494	0.0731	**0.8593**	0.0775	0.0794	**0.8345**	0.1646	0.0914	**0.9573**	0.1609	0.1841	**0.7739**	0.4301	**0.3113**	0.8467

Qualitative Cross-Testing Results

Minimal Sufficient Explanations



Percentage of images explained by different number of patches



Distribution of MSEs over 5000 images

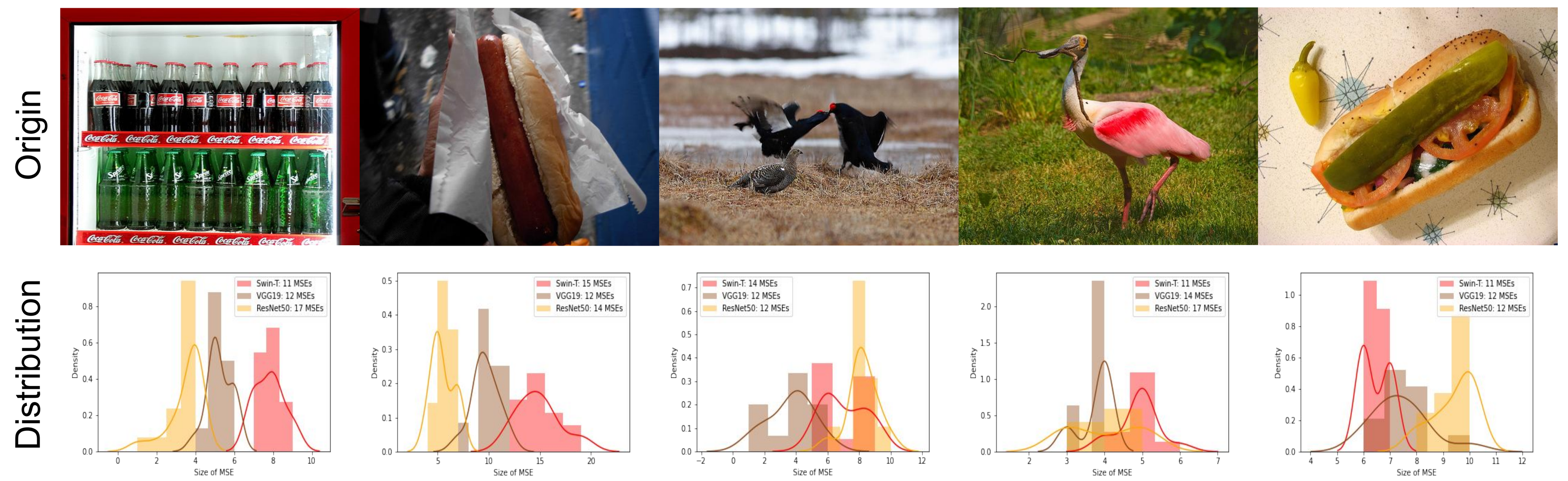
Model	MSEs Mean	MSEs Std	MSEs Median	Overlap=0 Mean	Overlap=0 Std	Overlap=0 Median	Overlap=1 Mean	Overlap=1 Std	Overlap=1 Median
Swin-T	**14.66**	**5.12**	**13.00**	1.26	0.67	1.00	**2.12**	2.04	1.00
ResNet50	13.52	4.42	12.00	1.25	0.66	1.00	1.98	1.96	1.00
VGG19	14.10	4.97	12.00	**1.27**	**0.69**	1.00	2.08	**2.12**	1.00

Number of diverse MSEs obtained and by allowing for different degrees of overlap

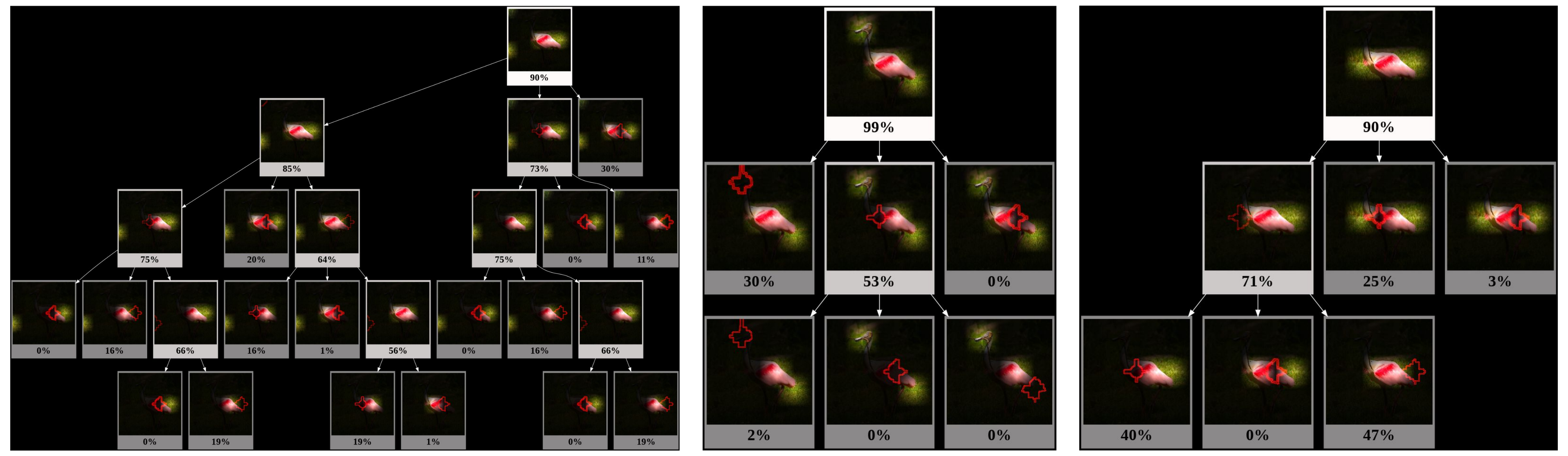
Sub-Explanation Counting

Model	Confidence ≥ 90 %	≥ 80%	≥ 70%	≥ 60%	≥ 50%
Swin-T	3.16	**21.01**	**85.19**	**255.47**	**601.72**
ResNet50	**3.21**	11.17	40.70	145.19	373.76
VGG19	3.19	13.86	57.59	158.42	439.52

Quantitative results of Sub-Explanation Counting



A few example distributions of MSE sizes for different algorithms on random images



(a) Swin-T (b) VGG19 (c) ResNet50

Sub-explanations of different models on an image of the ***Spoonbill*** class

Conclusion

- ❖ Swin-T is more robust to occlusion than CNNs using informative regions generated by iGOS++ and beam search method of SAG
- ❖ Swin-T can handle CNNs regions but CNNs can not handle Swin-T regions

This research is supported by National Science Foundation (1751402)