Causality for Inherently Explainable Transformers: CAT-XPLAIN

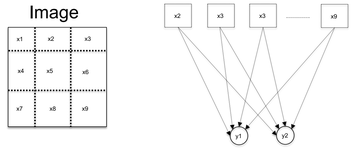
Subash Khanal[1,2*], Benjamin Brodie[1], Xin Xing[1,2], Ai-Ling Lin[2], Nathan Jacobs[1]
[1]Department of Computer Science, University of Kentucky, Lexington, KY
[2]Department of Radiology, University of Missouri, Columbia, MO

Overview:

- Decisions of a black-box neural network are usually explained through a post-hoc explainer.
- Explainers trained in the post-hoc scheme might be learning different feature representations than the black-box model. This raises concerns regarding the faithfulness of post-hoc explanations.
- As an alternative to post-hoc explainers, we propose to build an inherently explainable model that identifies the input regions having the most causal impact on the model's decision.

Causality for Explaining CV model:

- Any black-box model that consumes a set of input features to produce the task-related outputs can be represented as a structural causal model (SCM).
- There exists a causal relationship between the input space and the output space.



Black-box as a Structural Causal Model [1]

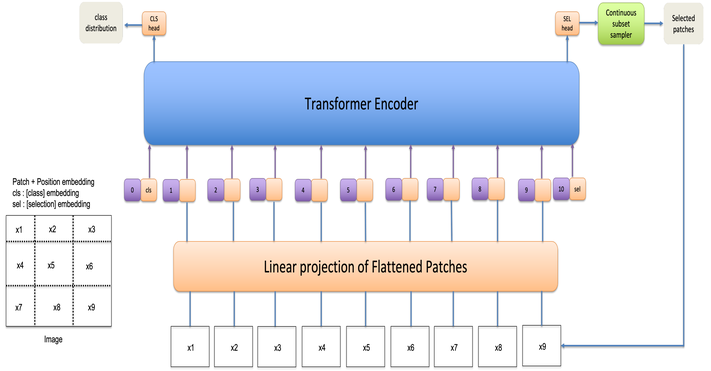
- By simulating the breakage of this causality, the top-k input regions contributing to the black-box's decision can be identified.
- Panda et al.[1] propose to explain a black-box model (B) through a selector network (E) trained to produce a categorical distribution from which a fixed set of features (*s*) is sampled.
- For a given input image (X) of class y, a set of patches (s) selected by E, and the black-box model B, the loss function to train E is:

$$L_{\text{sel}} = \sum_{y=1}^{C} P(y|X) log(P(y|X_{\bar{s}}))$$

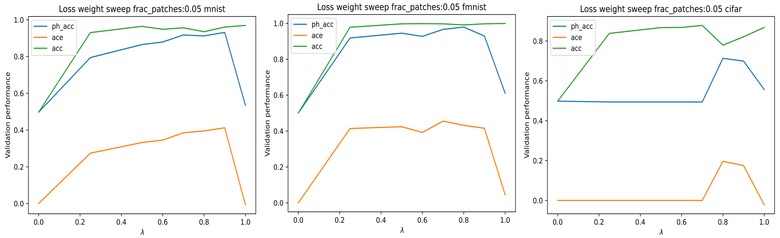
where, $X_{\bar{s}}$ is the input image where the patches selected by *E* are zeroed out. $P(y|X)$ and $P(y|X_{\bar{s}})$ are distributions across *C* classes, obtained from the *B* for inputs X and $X_{\bar{s}}$, respectively.

Inherently Explainable Vision Transformer:

- We bring the causal feature selection into the training of a vision transformer (ViT), hence making it an inherently explainable transformer (expViT).
- The attention mechanism of expViT leverages the relationship between the image patches.
- The continuous subset sampler of expViT uses an additional learnable token (sel) to sample the top *k* patches having a strong causal relationship with the model's output.



$$Loss = \lambda * \text{CE} + (1 - \lambda) * L_{\text{sel}}$$



Evaluation metrics:

1. Post-hoc accuracy (PA):

$$\text{PA} = \frac{1}{|X_{\text{T}}|} \sum_{x \in X_{\text{T}}} \mathbb{1}\big(\underset{y}{\text{argmax}}(P(y|x) = \underset{y}{\text{argmax}}(P(y|x_s))\big)$$

2. Average Causal Effect (ACE):

$$\text{ACE} = \frac{1}{|X_T|} \sum_{x \in X_T} (P(y|x_s) - P(y|x_{random}))$$

Results:

Post-hoc ViT vs. expViT on MNIST [3,8]

Methods	post-hoc		expViT		
frac	PA	ACE	PA	ACE	ACC
0.05	0.744	0.215	**0.936**	**0.422**	0.968
0.1	0.891	0.332	**0.953**	**0.430**	0.983
0.25	0.952	0.302	**0.969**	**0.415**	0.982
0.5	**0.969**	0.196	0.968	**0.318**	0.986

*Black-box ACC=0.993

Post-hoc ViT vs. expViT on FMNIST [t-shirt, shoe]

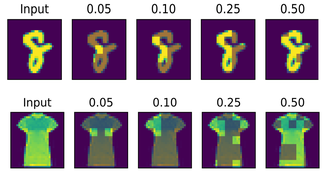
Methods	post-hoc		expViT		
frac	PA	ACE	PA	ACE	ACC
0.05	0.871	0.308	**0.970**	**0.462**	0.997
0.1	**0.992**	0.389	0.991	**0.472**	0.997
0.25	**0.995**	0.195	0.992	**0.449**	0.994
0.5	0.986	0.033	**0.987**	**0.316**	0.992

*Black-box ACC=0.999

Post-hoc ViT vs. expViT on CIFAR [bird, truck]

Methods	post-hoc		expViT		
frac	PA	ACE	PA	ACE	ACC
0.05	**0.702**	0.095	0.700	**0.171**	0.825
0.1	0.779	0.122	**0.808**	**0.280**	0.832
0.25	0.774	0.134	**0.820**	**0.275**	0.830
0.5	0.778	0.130	**0.832**	**0.269**	0.847

*Black-box ACC=0.895



Conclusion

- Our inherently explainable model mostly performs better than the post-hoc explainer on two explainability metrics.
- However, there is a trade-off between explainability and the true accuracy of the model with full input (ACC)
- https://github.com/mvrl/CAT-XPLAIN

References

[1] Panda Pranoy et al. Instance-wise causal feature selection for model interpretation. In Causality in Vision CVPR 2021.