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, USA
$^2$Department of Radiology, University of Missouri, Columbia, MO, USA
subash.khanal.cs@gmail.com

Abstract

There have been several post-hoc explanation approaches developed to explain pre-trained black-box neural networks. However, there is still a gap in research efforts toward designing neural networks that are inherently explainable. In this paper, we utilize a recently proposed instance-wise post-hoc causal explanation method to make an existing transformer architecture inherently explainable. Once trained, our model provides an explanation in the form of top-$k$ regions in the input space of the given instance contributing to its decision. We evaluate our method on binary classification tasks using three image datasets: MNIST, FMNIST, and CIFAR. Our results demonstrate that compared to the causality-based post-hoc explainer model, our inherently explainable model achieves better explainability results while eliminating the need of training a separate explainer model. Our code is available at https://github.com/mvrl/CAT-XPLAIN.

1. Introduction

Explainable AI (XAI) aims at designing methods or explainers that provide reasoning for the decisions made by a model trained for a specific task. XAI approaches can be broadly grouped under two categories: post-hoc explainers and explanation through inherently explainable models. Post-hoc approaches use backpropagation-based techniques [4], perturbation-based methods [8], or train a post-hoc explainer model [5] to highlight regions of the input instance considered important for the decision of a pre-trained black box model.

Post-hoc explanation methods are often model-agnostic and do not affect the black-box model’s performance during the explanation. However, there can be differences in inductive bias between the black-box and the post-hoc explainer. Moreover, the post-hoc explainers are trained in isolation while being guided by the output of a pre-trained black box. These explainers can be learning different feature representations and focus on different input regions than the black box. Therefore, there have been raising concerns regarding the faithfulness of post-hoc explanations [7]. Accordingly, there has been a push towards developing inherently explainable models for high-stakes decisions such as AI-based medical diagnosis, biomarker discovery, etc. In this line of thinking, we propose a small modification to an existing vision transformer architecture [2] and its training to build an inherently explainable model which identifies the most causally significant input regions that contribute to its decision.

2. Causality based post-hoc explanation

Causality-based interpretation methods draw motivation from cognitive psychology of human reasoning [6] and are accepted as the unifying approach for interpretability [1]. In a recent work by Panda et al. [3], instance-wise causal feature selection is proposed. Their approach explains a given 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. To back-propagate through this network, the sampling operation should be differentiable. This is achieved by the continuous subset sampler built using the Gumbel-Softmax trick. $E$ is trained with the objective function that maximizes the likelihood of the patches.
selected by it. 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 \(E\) as introduced in [3] is as follows:

\[
L_{sel} = \sum_{y=1}^{C} P(y|X) \log(P(y|X_{s}))
\]

where, \(X_{s}\) is the input image with the set of patches selected by \(E\) zeroed out. \(P(y|X)\) and \(P(y|X_{s})\) are distribution across \(C\) classes, obtained from the black-box model for inputs \(X\) and \(X_{s}\) respectively.

3. Inherently explainable transformer

We hypothesize that instead of building a separate explainer model, we can build an inherently explainable model. For this, we bring the causal feature selection approach developed in [3] into the training of a transformer. The approach in [3] assumes that there is a causal relationship between the input space and output space. Therefore, by simulating the breakage of this causality, we can identify the top-\(k\) input regions contributing to the black-box model’s decision. However, this assumption does not take into account relationships within the input space. Transformers, which are built from a series of self-attention layers leverage such relationships, hence should be able to better identify causally important regions. This motivated us to choose vision transformer (ViT) [2] as our base model.

3.1. Explainable ViT

ViT consumes an image as a sequence of flattened patches through a linear projection layer. The linearly projected patches are concatenated with their respective positional embeddings and passed through the transformer encoder, along with a learnable class embedding \((cls)\). Into the existing ViT architecture, we include an additional learnable embedding token \(sel\) which is passed through the transformer encoder. The output embedding corresponding to \(sel\) is used by the selection head to produce probability distribution across the total number of patches which is then used to sample the \(k\) patches most important for the model’s decision. This inherently explainable ViT (expViT) is trained by a loss function which is a weighted \((\lambda)\) sum of the standard cross-entropy (CE) loss for the classification and the selection loss as defined in Eq. (1):

\[
Loss = \lambda \cdot CE + (1 - \lambda) \cdot L_{sel}
\]

3.2. Evaluation metrics

Two causality based interpretation metrics used in [3] are: post-hoc accuracy (PA), defined as:

\[
PA = \frac{1}{|X_T|} \sum_{x \in X_T} \mathbb{1}(\text{argmax}_y P(y|x) = \text{argmax}_y P(y|x_s))
\]

and, Average Causal Effect (ACE), defined as:

\[
ACE = \frac{1}{|X_T|} \sum_{x \in X_T} (P(y|x_s) - P(y|x_{rand}))
\]

Here, \(X_T\) refers to the test set, and \(x_s\) refers to image instances where the top \(k\) patches from image \(x\) are sampled from the learned categorical distribution, whereas for \(x_{rand}\), \(k\) patches are sampled from a uniform random distribution. Eq. (3) computes how often the prediction of the model matches when it is passed the full input image as compared to the input image with only the top selected \(k\) patches retained while zeroing out the rest. Eq. (4) evaluates the causal strength of the image regions selected by the explainer compared to those selected at random.

3.3. Experiments

Similar to [3], for binary classification, we use a subset of the classes \((3, 8)\), (t-shirt and shoe), and (bird and
4. Results

| Method  | post-hoc | expViT  |
|---------|----------|---------|
| frac    | PA ACE   | λ PA ACE ACC |
| 0.05    | 0.744 0.215 | 0.9 0.936 0.422 0.968 |
| 0.1     | 0.891 0.332 | 0.7 0.953 0.430 0.983 |
| 0.25    | 0.952 0.302 | 0.6 0.969 0.415 0.982 |
| 0.5     | **0.969** 0.196 | 0.7 0.968 **0.318** 0.986 |

Table 1. Post-hoc explainer ViT vs. expViT trained on MNIST subset. Black-box model had ACC of 0.993

| Method  | post-hoc | expViT  |
|---------|----------|---------|
| frac    | PA ACE   | λ PA ACE ACC |
| 0.05    | 0.871 0.308 | 0.7 0.970 0.462 0.997 |
| 0.1     | **0.992** 0.389 | 0.9 0.991 0.472 0.997 |
| 0.25    | **0.995** 0.195 | 0.5 0.992 0.449 0.994 |
| 0.5     | 0.986 0.033 | 0.6 **0.987** 0.316 0.992 |

Table 2. Post-hoc explainer ViT vs. expViT trained on FMNIST subset. Black-box model had ACC of 0.999

5. Conclusion

As evident from the results, 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). This trade-off for the proposed expViT can be tuned based on two parameters: λ in Eq. (2) and frac of the total patches desired to be selected.

References

[1] Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

[2] Alexey D. et al. An image is worth 16x16 words: Transforming convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[3] Panda Pranoy et al. Instance-wise causal feature selection for model interpretation. In *Advances in neural information processing systems*, pages 10220–10230, 2019.

[4] Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Advances in neural information processing systems*, 2017.

[5] Schwab et al. Explain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems*, 2019.

[6] Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3):232–257, 2007.

[7] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

[8] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

This paper was accepted for spotlight presentation at the Explainable Artificial Intelligence for Computer Vision Workshop at CVPR 2022.