Explainability for Image Classification with ResNet

Gargeya Sharma
220278025
greyinuk@gmail.com

***Abstract*—Explainable Artificial Intelligence is currently one of the hot topics of research in Applied Artificial Intelligence. There are various domains where it is being applied and Computer Vision is also one of them. Explainable Computer Vision is highly promising since it deals with unstructured data: Images and Videos. Being able to explain and interpret why a model gave certain predictions can highly boost the trust built on those predictions. This document talks about a critical analysis done by experimenting with various attribution algorithms to understand the mapping from model prediction to an input image. Open-source library: Captum has supported the findings of this report.**

I. Introduction

Explainability in Computer Vision is a domain of active research which targets the improvement of model interpretability and output explanations and aims to overcome the black-box nature of machine learning/deep learning models. Multiple techniques have been proposed and are currently used as an additional step in the workflow [1] but a lot needs to be done to develop more self-explanatory models. A few of these approaches are compared in this report but first, we will see why we need explanations in the first place.

II. Problem Definition

Understanding anything is a major pillar to establishing trust in that thing. The same logic applies to the Image Classification task which has taken root in almost every major industry on this planet: Healthcare, Education Entertainment, Retail, Security and more. This calls for added emphasis and sensitivity in understanding how and why these classification models take certain decisions more than others. The report will help us realize which attribution technique is providing a better human understanding of why a certain output is provided and how fast these algorithms operate.

III. Key Works

Most of the techniques are differentiated between *gradient-based* or *feature-based* approaches. Gradients play a major role in any neural network to tune the parametric values with the objective of reducing the loss. They hold the information of what happened during the training which led to the minimization of the objective function. They are further sub-classified into Layer attribution and Neuron attribution. Feature-based methods a.k.a. Primary attributions [2] are intuitively based on the idea of how input features of the image play their individual and collaborative role in influencing an output's behaviour.

IV. Evaluation Criteria

I used a pre-trained ResNet-18 Model on ILSVRC 2012 (ImageNet classification challenge dataset). The testing image of class: Triceratops is taken from this source.

Metric used to evaluate the techniques: *Max Sensitivity*. It uses Monte-Carlo sampling-based approximation to measure the maximum sensitivity of an explanation. Explanation sensitivity in our context measures the degree of explanation change when input is perturbated slightly. The lower the value of this metric, the better the explanation.

V. Discussion

A. Abbreviations and Acronyms

The report includes 8 attribution techniques: Integrated Gradient (IG), which is also selected as the baseline here, SmoothGrad Squared applied on Integrated Gradients (IG-SG), VarGrad applied on Integrated Gradients (IG-VG), Gradient SHAP, Occlusion, Layer-wise Relevance Propagation (LRP), Guided Backpropagation (Guided BP), Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) [3]. Explainable Artificial Intelligence (XAI). Explainable Computer Vision (X-CV)

B. Figures and tables

In this subsection, experiments with multiple techniques, and their implementation details are discussed and presented with the help of Fig 1. Attribution techniques are implemented using an open-source library by 'Meta Open Source': Captum [2]. Captum is a Python implementation to create a unified, open-source model interpretability library for PyTorch. It contains generic implementations of a lot of gradient and perturbation-based attribution techniques. I have used their online GitHub repository and API documentation [3] as a reference for my experimental analysis. Before delving into the techniques, the model is chosen to be a pre-trained ResNet-18 model on the ImageNet dataset because it solved the purpose of my task (classification accuracy on testing image: *99.7%*) and most importantly its' computation complexity with the attribution techniques was lower than other available options.

In Fig 1, the number of outputs from various attributions is divided into 2 sections, each section contains 2 rows of 4 attributions. The first row corresponds to the heat map of visual features which positively influenced the output prediction. The second row shows a blended version of the previously mentioned heat map with the input image to easily visualize the exact part of the input that is being indicated.

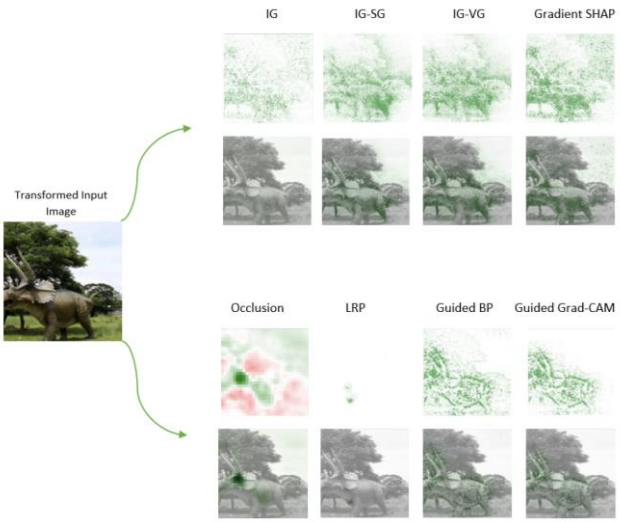

Fig. 1. Heat maps and blended heat maps portraying prediction supporting visual features for different XCV techniques applied on the input image.

TABLE I. TABLE TYPE STYLES

Metric	Attribution Methods							
	IG	*IG-SG*	*IG-VG*	*Gradient SHAP*	*Occlusion*	*LRP*	*Guided BP*	*Guided Grad-CAM*
Max Sensitivity	0.400	0.104	0.088	1.040	0.030	3.567	0.152	0.155

[a.] Max Sensitivity values for different attribution performances.

In Table 1, each attribution is given a max sensitivity score. As mentioned above this metric is inversely proportional to the goodness of explanation. IG is a method which approximates the integral of gradients of the model's output with respect to the input image along the path of a baseline (given) /references to inputs [3]. In Fig 1, we can observe that the explanation for classifying 'Triceratops' (green pixels) is noisy and even parts of the background are marked as meaningful features, this results in the max sensitivity score of 0.4 (Table 1). This score can be reduced if pixel density on and around the object is increased by the explanation in their heatmaps.

IG-SG alleviate noise in the process of IG by averaging over the square of the explanations of noisy copies of the input. The noise is derived from a normal distribution. Introduction of such noise and performing averaging increases the emphasis on the object to support the prediction. In Fig 1, features are denser on the parts of the object with respect to what we observed earlier with IG. The same will be achieved from a similar noisy method: IG-VG, which is a variance equivalent to IG-SG [4]. Table 1 supports this analysis with the metric values 0.1 and 0.08 respectively for IG-SG and IG-VG which are significantly lower than IG's 0.4.

One thing to point out here is that IG-SG has a slight increase in sensitivity with respect to IG-VG and the same issue is occurring with the Gradient SHAP algorithm that the noise is more scattered and not heavily clustered around the object hence reducing the meaningfulness of the features in determining the predictions. Gradient SHAP in some sense can be looked at as an approximation of integrated gradients by computing the expectations for different baselines generated by adding white noise to the input sample and then iteratively selecting them at random. This introduced noise increased the max sensitivity value to 1.04 (From Table 1) which is not a great response.

The best candidate attribution in this report based on their performance is Occlusion with a max sensitivity score of 0.03. Such a score is clearly explained by how the most significant region (green patches) is mostly the object in the blended heat map shown in Fig 1, where red patches that are showing the background incur negative dependence on prediction (prediction is not affected even if these areas change). Occlusion attribution is based on the idea that if a major feature is occluded in the image during the inference the prediction will drastically go down. So, a lot of iterative window-based occlusion in a process helps determine which areas are positively helpful in the prediction and vice versa.

The worst candidate in the analysis is LRP, a layer-wise method based on the backpropagation mechanism applied sequentially to each layer in the neural network. The score came to be 3.56 in the case which is terrible sensitivity and as shown in Fig 1, doesn't really explain anything. The next two attribution techniques: Guided BP and Guided Grad-CAM are similar to each other where Guided Grad-CAM uses Guided BP attributions with unsampled (positive) Grad-CAM attributions. . Guided BP computes the gradients of the ground truth with respect to input samples but gradients of ReLU activations are overridden to use only positive gradients at the time of backpropagation. Their similarity can easily be supported by their similar output heat maps and close-by max sensitivity values (0.152 and 0.155 respectively).

My analysis in this report is validating the pattern observed with these attribution algorithms in a similar yet more detailed analysis presented as a paper [5] in the "International Journal of Applied Earth Observation and Geoinformation" in 2021.

VI. CONCLUSION

All these experimentations with various post-hoc attribution techniques point towards how multiple components in a neural network can help figure out what might be the causes of getting a certain output prediction. Some methods like occlusion even when giving the best results are slow to execute and evaluate their metric. The time complexity of XAI methods should be kept as minimum as possible to provide low-latency explanations to the users. These were post-hoc methods, that are applied after the model is trained and ready. Outside the scope of this report, we can further dive into existing self-explanatory X-CV methods and contribute towards developing more of those with higher reliability and robustness in mind to support human decision-making based on these models and intelligently improve and analyse what more could be done by humans on top of the predictions to make fairer, smarter and valuable decisions for the betterment of society.

REFERENCES

[1] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," Proceedings of the IEEE, vol. 109, no. 3, pp. 247–278, Mar. 2021, doi: 10.1109/jproc.2021.3060483.

[2] N. Kokhlikyan et al., "Captum: A unified and generic model interpretability library for PyTorch." Accessed: Dec. 20, 2022. [Online]. Available: https://arxiv.org/pdf/2009.07896.pdf.

[3] Facebook Open Source, "Captum · Model Interpretability for PyTorch," captum.ai. https://captum.ai/api/attribution.html (accessed Dec. 20, 2022).

[4] J. Adebayo, J. Gilmer, I. Goodfellow, and B. Kim, "Workshop track - ICLR 2018 Local Explanation Methods For Deep Neural Networks Lack Sensitivity To Parameter Values," Oct. 2018. [Online]. Available: https://arxiv.org/pdf/1810.03307.pdf.

[5] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," International Journal of Applied Earth Observation and Geoinformation, vol. 103, no. 102520, p. https://www.sciencedirect.com/science/article/pii/S0303243421002270, Dec. 2021, doi: 10.1016/j.jag.2021.102520.