Vision transformer medical imaging explanations using Relevance Propagation

Piotr Komorowski PK406736@STUDENTS.MIMUW.EDU.PL
Kajetan Husiatyński KH406160@STUDENTS.MIMUW.EDU.PL
Szymon Antoniak SA394197@STUDENTS.MIMUW.EDU.PL
University of Warsaw, Poland

Abstract

In this study, we demonstrate the application of the Layer-wise Relevance Propagation (LRP) method to explain the predictions of Vision Transformers (ViT) trained on radiographs of lungs during the COVID-19 pandemic. The aim is to classify the patients' health condition based on their radiographs. The evaluations of the generated explanations were performed using the Quantus Python library. Furthermore, the resulting LRP explanations will be compared with those generated by the Local Interpretable Model-Agnostic Explanations (LIME) method.

1. Introduction

The objective of our research was to compare the LRP method (Montavon et al., 2019) with the LIME method (Ribeiro et al., 2016). LRP is a technique for propagating prediction backwards in a neural network by using a set of designated propagation rules. Until recently, it was not possible to use LRP on Transformers. In our study, we tested a novel method that implements LRP for ViT (Chefer et al., 2021). We trained a ViT model to classify radiograph images of lungs.
To evaluate our results, we utilized the Quantus Python library. Quantus is a toolkit that offers over 30 metrics in 6 categories of XAI evaluation. Although it is still in the early stages of implementation, we were able to improve the original package with our own patch to use it with our model. More information on this can be found in the appendix.

2. Dataset

For this study, we used the COVID-QU-Ex dataset (Tahir et al., 2021), which consists of 33K chest Xray (CXR) images divided evenly to COVID-19, non-COVID infections (Viral or Bacterial Pneumonia) and healthy chest X-ray images. For sample images, see Figure 1.

3. Background

3.1 Vision Transformer

Vision Transformers (Dosovitskiy et al., 2020) have recently been attracting attention in the field of medical imaging due to their ability to effectively capture the spatial information

present in medical images. Compared to traditional deep learning models such as Convolutional Neural Networks (CNNs), Vision Transformers differ in their use of self-attention mechanisms to process information. This enables them to effectively capture long-range dependencies between different regions of the input data, which is particularly useful in medical imaging where features such as anatomical structures may span large distances in the image.

3.2 Layer-wise relevance propagation

Layer-wise Relevance Propagation (LRP) is an explanation technique that provides a way to interpret decisions made by deep neural networks. It works by assigning a relevance score to each unit in the network, which represents its contribution to the final prediction. In particular, the 'relevance' received by a given layer/component/neuron from the layer above must be fully redistributed to the layer below. The layer-wise approach of LRP allows for a hierarchical and interpretable view of the model's decision-making process. For a visualisation of the LRP mechanism, see Figure 2.

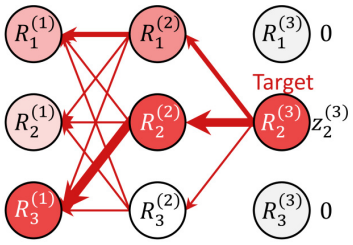

Figure 2: The relevance scores are propagated back through the intermediate layers, and ultimately to the input layer. The relevance scores are used to determine the contribution of each unit in the network to the final prediction.

In (Chefer et al., 2021), the authors show that one of the most common explanation algorithms from the LRP family cannot be assumed to work effectively on Transformer models in the same manner as they do on standard deep neural networks. Specifically, within the framework of Layer-wise Relevance Propagation, it is demonstrated that Generalized Inversion (GI) fails to implement the conservation of relevance. Furthermore, the analysis pinpoints that attention heads and LayerNorms need to be specifically addressed. In this

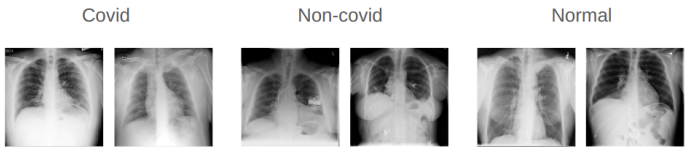

Figure 1: Samples from every class in the dataset. It's challenging for individuals without specialized knowledge to accurately categorize these images.

work, we base our implementation of the modified LRP on the official github repository released by the authors.

4. Experiments

4.1 Model and training

We fine-tuned a pretrained 12−layer Vision Transformer of width 768 for 10 epochs with batch size of 32 and learning rate of 10^{-4}. Early stopping was applied, finding that the best performing model is attained after the 3rd epoch of training.

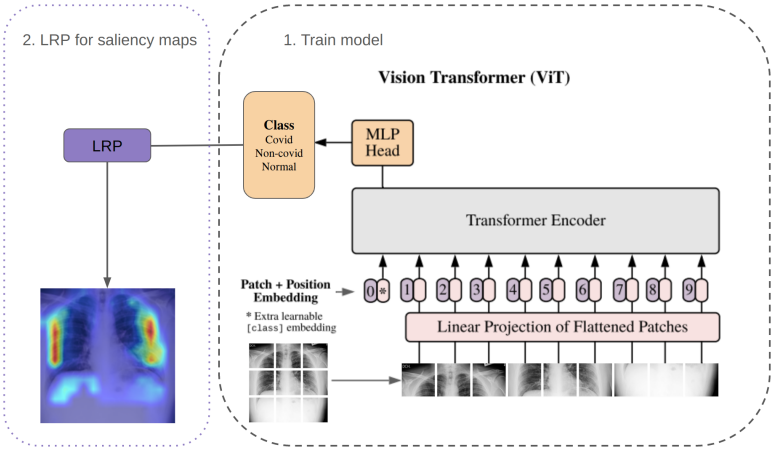

Figure 3: A schema of our explanation pipeline.

4.2 Explanation evaluation metric

Faithfullness estimate, as implemented in the Quantus package, is based on (Alvarez-Melis and Jaakkola, 2018) is. This metric evaluates the quality of explanation by computing the correlations of probability drops and the relevance scores when randomly perturbing the base image.

4.3 Results

The comparison between LRP and LIME methods shows that LRP outperforms LIME according to the Faithfulness Estimation metric (as seen in Table 1). The example of visual results, shown in Figure 4, reveal that LRP provides more accurate explanations for covid and non-covid samples. However, for healthy sample, LRP focuses on irrelevant artifacts while LIME focuses on the lungs. Expert consultation is suggested for a comprehensive understanding of these explanations.
Additionally, Figure 5 presents the values of the Faithfulness Estimation for the LRP estimator with respect to specific classes. The boxplot indicates that the metric is lower for non-covid explanations, necessitating future exploration.

Explanation type	Faithfulness Estimation↑
baseline (LIME)	0.08
TransformerLRP	**0.27**

Table 1: Quantitaive comparison.

5. Conclusion

The study applies the Layer-wise Relevance Propagation (LRP) method to explain the predictions of a Vision Transformer (ViT) model trained on radiographs of lungs during the COVID-19 pandemic to classify patients' health condition. The model was trained on the COVID-QU-Ex dataset and the resulting LRP explanations were evaluated using the Quantus Python library. This work also compared the LRP explanations with those generated by the Local Interpretable Model-Agnostic Explanations (LIME) method. The results showed the effectiveness of the LRP method for explaining the predictions of ViT models in medical imaging. However, it is important to note that the validity of the results is dependent on the choice of evaluation metrics and their implementation, and there may be limitations to the generalizability of the findings to other medical imaging tasks or datasets.

References

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks, 2018. URL `https://arxiv.org/abs/1806.07538`.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791, June 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. URL `https://arxiv.org/abs/2010.11929`.

Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. URL `http://jmlr.org/papers/v24/22-0142.html`.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: An overview. In *Explainable AI*, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

Anas M. Tahir, Muhammad E.H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, 2021. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2021.105002. URL `https://www.sciencedirect.com/science/article/pii/S0010482521007964`.

Appendix A. Quantus

Quantus (Hedström et al., 2023) package has issues with its implementation. Quantus needs additional libraries to work properly. There are two possible alternatives: ‘captum‘ and ‘zennit‘. The first package raises error while being imported so it’s necessary to use ‘zennit‘ instead. Our visual transformer is registering backward hooks, which requires gradient flow. Gradient is disabled by default and there is no interface to change it. In order to fix this issue we’ve forked this repository and turn the gradients on.

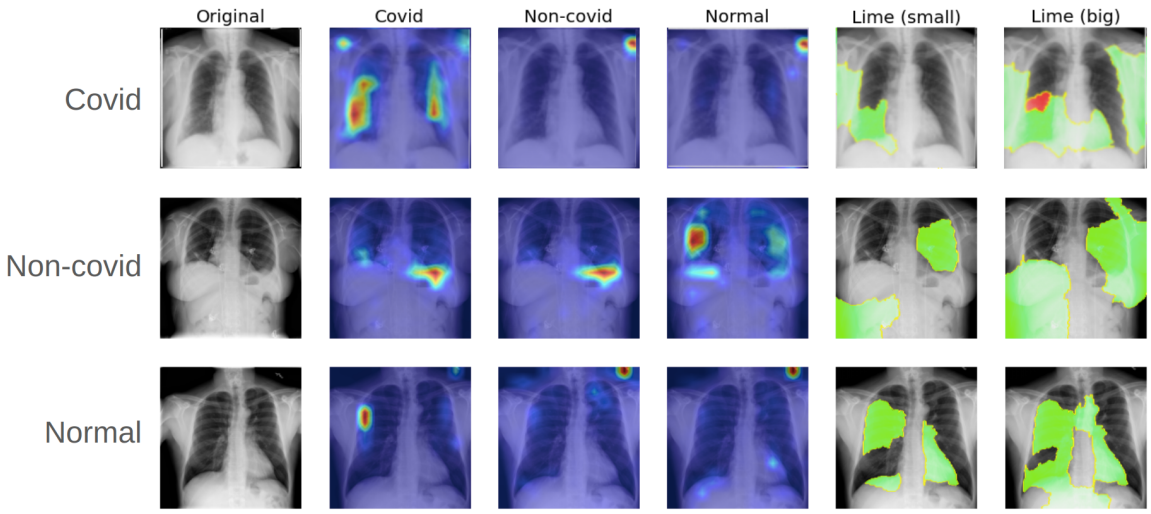

Figure 4: Illustration of the visual outcomes. Columns 2 to 4 show the explanations generated by the LRP analysis performed on ViT with respect to specific classes, while columns 5 and 6 present the explanations generated by LIME with 3 and 8 visible features respectively.

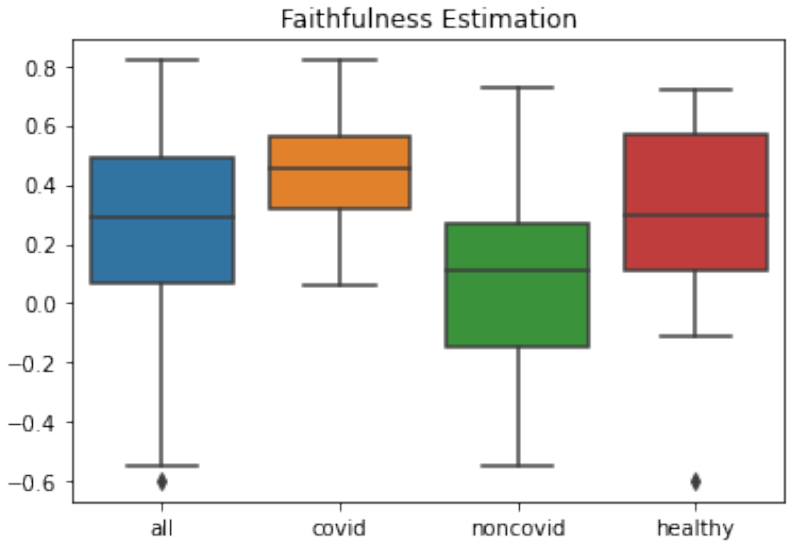

Figure 5: Boxplot aggregated over 87 random samples with respect to the specific classes.