

BACHELOR THESIS PROPOSAL

THE DISAGREEMENT PROBLEM IN XAI ON IMAGE DATA VIA SALIENCY MAPS

*Vấn đề bất đồng trong giải thích của XAI khi ứng dụng trên dữ liệu
ảnh bằng saliency map*

1 GENERAL INFORMATION

Thesis Supervisor:

– Prof. Lê Hoài Bắc (Faculty of Information Technology)

Students:

1. Mai Duy Nam (ID: 19120298)
2. Nguyễn Hữu Bình (Id: 19120460)

Project Type: Research

Timeline: From *March 2023* to *July 2023*

2 CONTENTS

2.1 Introduction

In recent years, there has been a high rise in the use of AI models due to their high accuracy and usefulness. However, these models are often black boxes with very deep neural network structures and have an enormous number of parameters, meaning their internals are obscure and beyond human comprehension. In many real-world scenarios, AI systems are used to make critical decisions that affect human lives, such as medical diagnoses, autonomous vehicles, security, and surveillance. Interpreting these models' output is crucial in determining the appropriate response to these threats. Explainable AI (XAI) has emerged as a solution to prevent blind faith in model outputs and improve the interpretability of AI models. However, many AI practitioners who utilize visual AI models pay less attention to or are not aware of the disagreement problem between the XAI methods—the problem in which different XAI methods generate inconsistent explanations when evaluation on the same black box. Additionally, there is a lack of research examining the extent of disagreement between XAI methods in general [1], and methods that target images or speech input in particular.

Therefore, this work aims to investigate the extent of such disagreements and find a proper way to quantify their differences by conducting experiments with multiple visual XAI methods that explain a black-box model trained on different popular pre-trained image datasets. A variety of metrics will be used to quantify the differences. Moreover, we would propose solutions in case discrepancies arise by observing factors that made the explanations disagree.

2.2 Target

2.2.1 Rationale

As mentioned in 2.1, AI models are increasingly being used in various fields, but they are often black boxes that are difficult for humans to understand the decision-making process behind them.

Computer Vision (CV) is the most widely used among the different AI fields. Visual AI systems are now used to make critical decisions in various real-world scenarios that affect human lives, such as medical diagnoses, autonomous vehicles, security, and surveillance. Interpreting these models' output is crucial in determining the appropriate response to these threats. If these decisions are made solely by an AI system without any explanation, it can be difficult for humans to understand how they were made and trust the system's output. This can lead to legal, ethical, and social issues.

Numerous studies have been conducted to address this problem to improve the interpretability of these models. Many methods were created to explain black boxes using saliency map explainers, such as CAM [2], Grad-CAM [3], layer-wise relevance propagation [4], deep Taylor decomposition [5], etc.). While they all aim to explain the black-box models' decision-making processes, they use different methods to identify which features in the input image are important for the model's output. As a result, they may produce different saliency maps that highlight different regions of the image as having higher importance, leading to discrepancies in their explanations.

According to the paper [1] by Krishna et al., machine learning practitioners use multiple explanation models to gain insights into how a black-box works. Krishna et al. presented some experiments on the disagreement problems between various XAI methods, such as LIME, SHAP, and Gradient-based models. After that, they compared these methods using some metrics to quantify the extent

of disagreement. However, when applied to images, the aforementioned methods measure the disagreement using feature importance by regarding each pixel as one feature, which is not semantically meaningful. In this work, we explore the methods whose outputs are in the form of saliency maps by conducting experiments on many other suitable metrics to find whether the disagreement problem persists when applied to image data, what makes these disagreements arise, and how to quantify and resolve them.

2.2.2 Significance

This comparative study has several potential contributions to the field of XAI.

Firstly, it can help improve the transparency and interpretability of ResNet family models by investigating how their behaviors vary when pre-trained on different datasets. This is especially important as the quality of the data used to train a model can significantly impact its performance and reliability. By comparing the interpretability of a pre-trained model, we can identify which datasets are most suitable for achieving a particular level of transparency and interpretability. This can help guide the development of more transparent and trustworthy AI systems.

Secondly, identifying the reasons for saliency map disagreement or inconsistency can help researchers develop new XAI methods and metrics that are more effective in explaining black-box models. This can contribute to the advancement of the field of XAI and improve the overall performance of visual AI systems.

Finally, conducting a comparative study of different visual XAI methods across different datasets can provide valuable insights into best practices and guide AI practitioners in selecting the most reliable and robust method for a given task.

2.3 Scope

By utilizing multiple popular XAI methods, this study conducts experiments by explaining a pre-trained ResNet model on some common image datasets. The resulting explanations are then evaluated using some types of metrics to draw

practical conclusions about all the factors joining in the train model process and the disagreement issue between visual explanations.

2.4 Approach

In this study, we focus on factors that may affect the accuracy of saliency map explanations, such as datasets, the architecture of ResNet, and the mechanism of XAI methods.

With a pre-trained model on any given dataset, we will first use various XAI methods on ResNet to evaluate the interpretability level of this model family. Because different XAI methods may focus on different aspects of the data, leading to different explanations. So if a disagreement arises, we can evaluate the performance in terms of the accuracy of each algorithm to get insight into which types of methods fit with each dataset. Then, we will quantify the disagreement between the XAI methods by applying several metrics to quantify the disagreement between explanations to find a suitable metric for this task.

Next, we will use the most stable XAI algorithm to explain ResNet’s outputs on multiple datasets and observe the vary of interpretability through multiple datasets. Computer Vision datasets used to train computer vision models are vast in size, with a significant portion readily accessible online and might be combined from various sources. These sources may be unreliable and biased due to human error, which can lead to the models themselves inheriting such unfairness.

Finally, from these above results, plus the insights gained from analyzing the ResNet architecture & pre-trained datasets, we can draw conclusions about the disagreement between saliency maps.

2.5 Expectation

Expected outcomes of this research project include:

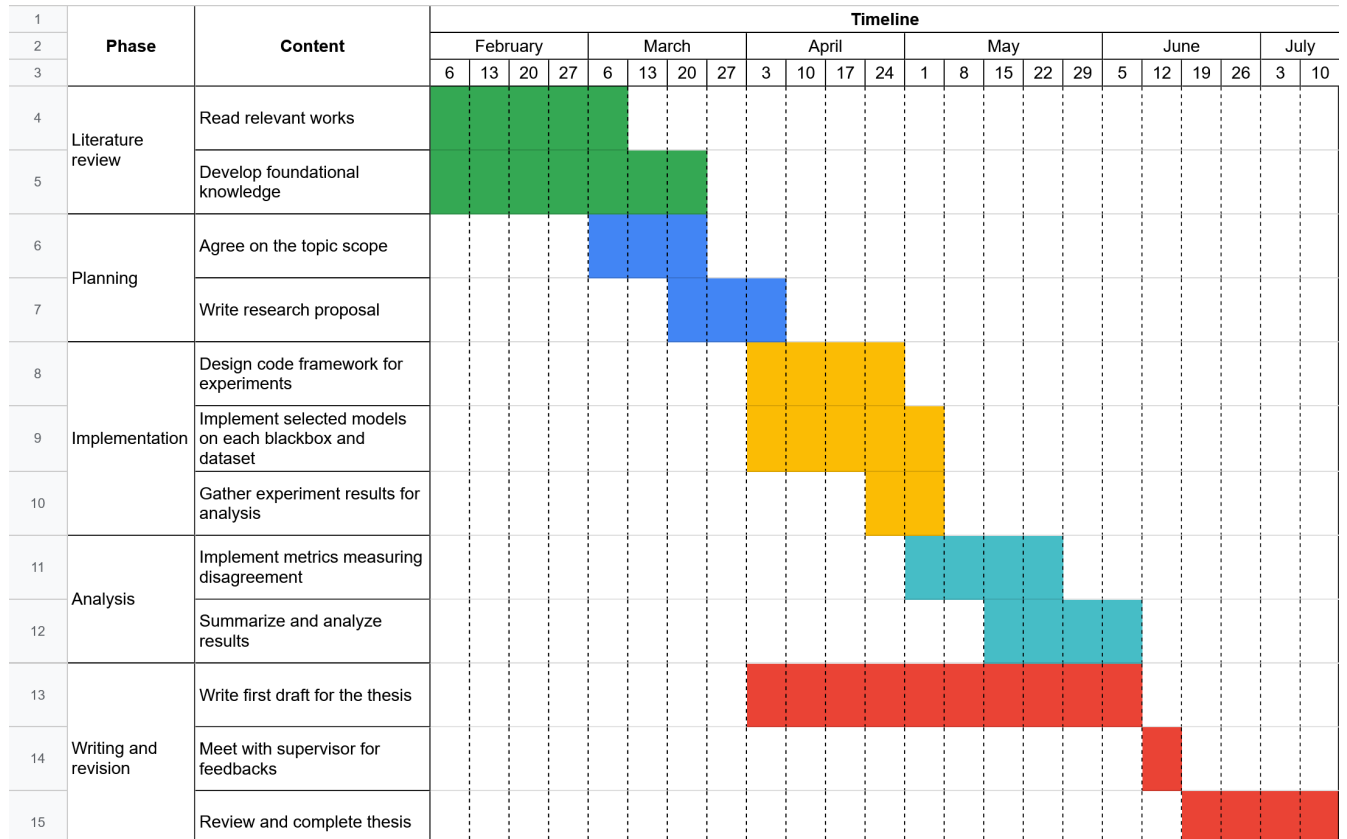
- A comparative study will show practical accuracy & performances of popular

visual XAI methods on ResNet.

- Identification of the reasons for inconsistency or disagreement between the models.
- Improvement in the transparency and interpretability of the ResNet model and understanding of its behaviors when pre-trained on different datasets.
- Practical way to choose which XAI methods to explain ResNet & metrics to measure disagreement degree.

2.6 Project Plan

The project plan is summarized in figure 1.



Hình 1: The project plan summarized in a Gantt chart, which includes the research phases, the activities in each phase, and the timeline for each activity.


References

- [1] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, “The disagreement problem in explainable machine learning: A practitioner’s perspective,” *CoRR*, vol. abs/2202.01602, 2022.
- [2] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” *CoRR*, vol. abs/1512.04150, 2015.
- [3] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, “Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization,” *CoRR*, vol. abs/1610.02391, 2016.
- [4] A. Binder, G. Montavon, S. Bach, K. Müller, and W. Samek, “Layer-wise relevance propagation for neural networks with local renormalization layers,” *CoRR*, vol. abs/1604.00825, 2016.
- [5] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep taylor decomposition,” *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

Ho Chi Minh City, April 2nd 2023


SUPERVISOR

(Signature)


Lê Hào Bửu

STUDENTS

(Signature)


Mai Duy Nam

Nguyễn Hữu Bình