Research Proposal - Model Multiplicity in Image Processing Through the Lens of Robustness and Explainability

Aryan Kakadia

January 2024

1 Research Objective

The main goal of this project is to investigate model multiplicity in image processing, also known as the Rashomon Effect. For a given dataset, there may exist many models with equally good performance but different solution strategies. In this research project, we would evaluate different machine learning models for image processing, including neural network architectures as well as adversarially-robust architectures and critically compare them using different explainability techniques [1][2][3].

2 Overview

The recent advancements in computational hardware and development of efficient machine learning algorithms have resulted in machine learning being used in a vast range of applications. Some of these applications are very critical and sensitive. Machine learning algorithms generally work as a black box. It receives an input and the model generates an output. The user has no information about the reasoning behind the decisions made by the model.

Explainability in the context of machine learning refers to the degree to which the predictions made by a model can be understood by humans. Different features in an input result in different activations in the model. Explainability in machine learning explores how much these features contribute in generating a particular output [4]. Robustness on the other hand is about reliability of model prediction under perturbations of their inputs, by an adversary.

Applying posthoc explainability techniques to these black box models can help us better understand the inner workings of them. This is crucial in fields where the model is implemented to make critical decisions. Explainability techniques can help the professionals in making better informed decisions. It becomes critical to ensure that systems are not only working most of the times,

but also make sure that they are robust and reliable [5].

Explainability can also help in cultivating a better understanding of the model, which makes it easier to improve current models. Lastly, with increasing popularity of machine learning and its implementation in different aspects of life, it is also becoming necessary to explain the inner workings of these black boxes for ethical reasons and to comply with new legislation governing the use of machine learning.

As mentioned earlier, there may exist many models with equally good performance but different solution strategies (Rashomon effect). Applying post hoc explainability techniques to these regular and adversarially-robust models and critically comparing and evaluating the results can reveal interesting facts about feature importance. This could potentially help in identifying features that contribute to robustness. It can also shed light on why adversarially robust models might be more resilient to attacks.

3 Student Involvement

Research project conducted by: Aryan Kakadia
Supervisor: Sanghamitra Dutta

4 Methodology

This research project will involve a significant amount of coding, data analysis and data gathering. Coding will primarily be done in python. Additionally, I will be extensively using pytorch library for creating and training regular and adversarially-robust machine learning image processing models.

In order to evaluate the explainability of the models, I will be using Grad-CAM and deep SHAP (SHapley Additive exPlanations) framework. Grad-CAM belongs to the group of Activation Based Methods. The main idea behind this is to combine features that are considered important by the network for an output. This highlights the areas that have a greater impact on the output decision. The explanation is presented as a coarse heat map which is relatively easy to interpret [6]. SHAP is an occlusion based method. It is a unified measure of feature importance that assigns each feature an importance value for a particular prediction. In this technique occluded versions of the input image are computed and passed through the model to compute an output value. Compared to other methods, this approach leads to a significant computation overhead. However, it creates more understandable results which are independent of model type and architecture [6].

In this project, I will also attempt to generate my own dataset and train

models using those datasets. I would mainly rely on the internet and generative AI to gather the image data for classification. I feel like a mix of both will help in generating diverse datasets. Once I have collected the images, I would convert them into a format suitable for image processing. This would involve resizing the images, normalizing the pixel values, and other transformations. Preprocessing helps in ensuring consistency in the dataset. Next is the labeling step. Each image would be labeled according to its corresponding category. This is a crucial step. The accuracy of the labels impacts the performance of the model. Lastly, the preprocessed and labeled images will be compiled into a dataset. I will be splitting the dataset into training, validation, and test sets. The quantity of data generated may be limited as this is being generated over the span of a week and by a single person.

One interesting observation that Zico Kolter and Aleksander Madry make in their "Adversarial Robustness: Theory and Practice" (slide 45) presentation is that the gradient for adversarially robust model resembles closely to the input image [7]. Perhaps, during this project, I can explore it in detail and gain insight into why that is the case.

5 Project Learning Schedule

Week 1 and 2: Train several image classification models on MNIST Dataset
Week 3 and 4: Apply Posthoc Explainability Techniques
Week 5 and 6: Train Adversarially-robust models and evaluate their explainability
Week 7: **Mid Semester Report.**
Week 8 and 9: Research on gathering data and generate my own dataset
Week 10 and 11: Use my datasets to train previously used models.
Week 12: Apply Posthoc Explainability Techniques on these newly trained models
Week 13 and 14: Compiling all results and making a **Final Report.**

6 Student Learning Outcomes

1. Understanding Machine Learning Concepts:
   (a) Gain a deeper understanding of machine learning models for image processing and learn about their strengths, weaknesses, and appropriate use cases.
2. Practical Experience:
   (a) Learn to work with cutting edge machine learning libraries and implement different machine learning models.
   (b) Understand how datasets are generated and learn to generate datasets that match required specifications to train machine learning models.

3. Evaluation Skills and Ethical Considerations:

   (a) Learn to critically evaluate different machine learning models.

   (b) Learn how adversarial perturbation works and techniques to make robust machine learning models.

   (c) Understand the importance of transparency, accountability and explainability of machine learning models.

7 Advanced laboratory requirement

This research project is intended to satisfy the advanced laboratory requirements. As discussed in the proposal, this project involves a significant amount of data analysis and data gathering, which are intended to meet the requirements.

References

[1] S. Müller, V. Toborek, K. Beckh, M. Jakobs, C. Bauckhage, and P. Welke, "An empirical evaluation of the rashomon effect in explainable machine learning," 2023.

[2] C. T. Marx, F. du Pin Calmon, and B. Ustun, "Predictive multiplicity in classification," 2020.

[3] F. Hamman, E. Noorani, S. Mishra, D. Magazzeni, and S. Dutta, "Robust counterfactual explanations for neural networks with probabilistic guarantees," 2023.

[4] A. Bueff, I. Papantonis, A. Simkute, and V. Belle, "Explainability in machine learning: a pedagogical perspective," 2022.

[5] Z. Kolter and A. Madry, "Chapter 1 - introduction to adversarial robustness," 2023. Accessed: 2024-01-20.

[6] E. A. P. B. D. D. e. a. Pierre Dardouillet, Alexandre Benoit, "Explainability of image semantic segmentation through shap values," Aug 2022.

[7] Z. Kolter and A. Madry, "Adversarial robustness - theory and practice: Tutorial slides, parts 1 and 4," 2018. Accessed: 2024-01-20.