Two-Stage Selective Explanation Approach: Efficient Feature Attribution for High-Dimensional AI Models

DISSERTATION SYNOPSIS

SUBMITTED TO
BABU BANARASI DAS UNIVERSITY
LUCKNOW

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

MASTER of COMPUTER APPLICATIONS
Data Science & Artificial Intelligence)

SUBMITTED BY
Alaka Singh
1240259030
MCADS2

UNDER THE SUPERVISION OF
Dr Anshu Gupta
Associate Professor

SCHOOL OF COMPUTER APPLICATIONS
BABU BANARASI DAS UNIVERSITY
BBD CITY, AYODHYA ROAD, LUCKNOW (U.P.) - 226028, INDIA

1. Introduction

The rapid adoption of Artificial Intelligence (AI) in critical decision-making domains such as healthcare, finance, and security has made **explainability** a crucial requirement. Traditional Explainable AI (XAI) methods like SHAP, LIME, and Layer-wise Relevance Propagation (LRP) provide valuable insights but face major **computational challenges** when applied to high-dimensional data such as images, text embeddings, or genomic datasets.

When thousands of features are present, computing explanations for every feature becomes infeasible in terms of time and resources. This creates a need for an **efficient XAI approach** that can identify the most influential features without exhaustively analyzing all inputs.

This dissertation proposes a **Two-Stage Selective Explanation Approach** that combines a fast, lightweight filtering stage with a precise, high-fidelity explanation stage to achieve both scalability and accuracy in feature attribution.

2. Motivation

- Traditional XAI techniques scale poorly with feature dimensionality, leading to **excessive runtime** and **high computational cost**.
- In practical scenarios, decision-makers are often interested only in the **top few important features** rather than a full explanation.
- A **selective explanation strategy** can drastically reduce computation while still providing reliable insights.
- By combining **gradient-based estimations (fast)** and **precise methods (accurate)**, we can design an XAI pipeline that is **both efficient and interpretable**.

3. Brief Literature Survey

- **LIME (Local Interpretable Model-Agnostic Explanations)** – Perturbs input features and builds a surrogate model to approximate feature importance. Computationally expensive for high dimensions.
- **SHAP (SHapley Additive exPlanations)** – Based on cooperative game theory, ensures fairness but is **NP-hard** and requires many samples.

- **Integrated Gradients (IG)** – A gradient-based method that attributes importance by integrating gradients along an interpolation path. More scalable but still expensive for full features.
- **Recent work on dimensionality reduction in XAI** – Suggests grouping or sampling features but does not formalize a **two-stage pipeline** for selective precision.

Gap Identified: No systematic framework exists that combines **fast feature filtering** with **precise explanations** to provide **top-1% feature attribution** efficiently.

4. Problem Formulation

Research Problem:

"How can we design a scalable XAI technique that efficiently identifies and explains only the most important features in high-dimensional AI models without computing full feature attributions?"

5. Objectives

1. To design a **two-stage selective explanation pipeline** for AI models.
2. To implement **Stage-1: Fast feature filtering** using gradient-based or statistical methods.
3. To implement **Stage-2: Precise feature attribution** (e.g., SHAP, IG, LIME, LRP) restricted to selected features.
4. To evaluate the proposed method on **image and tabular datasets** with thousands of features.
5. To compare efficiency (runtime, memory) and fidelity (accuracy of top features) against baseline XAI methods.

6. Methodology / Planning of Work

1. **Literature Review**
    - Study existing XAI methods (SHAP, LIME, IG, LRP).
    - Identify their scalability limitations.
2. **Stage-1: Fast Filtering (10%)**
    - Implement gradient-based saliency maps.
    - Explore perturbation-based sensitivity analysis.
    - Rank features and retain only top candidates (e.g., top 10%).
3. **Stage-2: Precise Attribution (1% of 10%)**
    - Apply IG/SHAP/LIME only to filtered features.

    - Derive accurate explanations for top-1% of features.
4. **Evaluation**
    - Datasets: MNIST, CIFAR-10, and high-dimensional synthetic tabular dataset.
    - Metrics: Precision, Deletion/Insertion AUC, runtime, stability.
    - Compare against full SHAP/LIME/IG.
5. **Implementation Tools**
    - Programming: Python
    - Frameworks: PyTorch/TensorFlow, SHAP library, Captum (for IG)
    - Hardware: GPU-enabled environment for scalability.

7. Expected Outcomes

- A novel **Two-Stage Selective Explanation Framework** for XAI.
- Significant **reduction in computation cost** (time and memory) compared to existing methods.
- Explanations focused on **top-1% most important features**, which are often sufficient for human interpretation.
- Demonstrated applicability on both **image and tabular datasets**.
- A comparative study showing that selective explanations preserve fidelity while improving scalability.

8. References

1. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *ACM SIGKDD*.
2. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.
3. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *ICML*.
4. Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*.
5. Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*.