Interpretability and Explainability of AI Models

MRP Final Presentation

Shuo Li

307501

Project Summary

- AI models often behave like a “black box,” so people cannot understand why the model makes a certain decision. The motivation of this project is to make AI more transparent and easier to trust. The main idea is to explore both interpretability and explainability methods and show how they work together. As the final result, I built a small demo system that highlights important input features and generates a simple human explanation for the model’s prediction.

What Is Interpretability

- Interpretability means understanding how different parts of the input influence the model’s output. It is usually more technical and low-level. Many methods focus on feature importance, saliency maps, or counterfactual changes. According to the lecture, interpretability can help developers debug models and understand model behavior more clearly.

Interpretability Methods

- Researchers study many interpretability techniques. Some popular ones include Grad-CAM for visual models, LIME and SHAP for tabular or text data, and counterfactual analysis for testing how a prediction changes when we modify a feature. These methods are used to show what part of the input the model pays attention to. This type of work helps reveal hidden logic inside black-box models.

What Is Explainability

- Explainability is more high-level. It focuses on giving people a clear and simple explanation in natural language. Instead of only showing technical signals, explainability tries to tell the “reason” behind the model’s output. In recent years, large language models can generate chain-of-thought explanations that describe the decision step-by-step.

Explainability Methods

- In the research community, explainability includes post-hoc explanations, counterfactual stories, rule-based summaries, and chain-of-thought reasoning. These methods do not change the original model. Instead, they add another layer that makes the model easier to understand. Explainability is important when AI interacts with non-technical users, or when decisions must be justified.

Bias, Fairness, and Ethics

- AI models can be biased toward race, gender, age, or other sensitive attributes. Lecture 9 explains that bias often comes from skewed or unbalanced training data. Fairness means treating different groups equally without discrimination. AI ethics also covers topics such as safety, misuse, environmental impact, and legal responsibility.

AI Regulations

- Many governments now require AI transparency and safety. The EU AI Act controls high-risk AI areas like education, hiring, and law enforcement. In California, several laws discuss transparency, labeling AI-generated content, and developer responsibility. International agreements focus on human rights and safety in AI systems. This shows that interpretability and explainability are not only technical needs, but also legal requirements.

Approach Used

- In my project, I built a small demonstration system inside a Jupyter Notebook. I used a simple logistic regression model for sentiment analysis, trained on short positive and negative sentences with TF-IDF features. The Notebook computes the contribution of each word or bigram to the model’s prediction, which shows interpretability on the feature level. Then it generates a short human-friendly explanation in simple English that summarizes why the model predicts positive or negative. This approach is easy to read, because the code, outputs, and explanations are all visible step by step in one place.

GitHub Link and YouTube Demo

- GitHub Repository:
- https://github.com/ls1991lsok/MSCS-2201

- YouTube Demo:
- https://youtu.be/kSc97VPwUqk

Conclusion

- Through this project, I learned that interpretability works on the low-level input/output relationship, while explainability focuses on giving a clear human explanation. Both concepts support trust and understanding in AI systems. Bias, fairness, ethics, and regulations show that modern AI needs transparent behavior, especially when used in real-world decisions. My demo system shows how these ideas can be combined to make AI more understandable for users.