Interpretability in Deep Learning

Sajad Sabzi
MohammadReza AhmadiTeshnizi

December 17, 2023

Introduction

The use of **interpretability** in deep learning has become a focal point in the field of artificial intelligence. This is driven by the need to understand and trust the complex decision-making processes of deep learning models. Interpretability refers to the extent to which a human can comprehend the cause of a decision made by a machine learning model.

Use of Interpretability in Deep Learning

Deep learning models, known for their powerful predictive capabilities, are often seen as ”**black boxes**” due to their complexity. To mitigate this, interpretability methods are employed to make these models more understandable and trustworthy. Two prominent approaches are **Local Interpretable Model-agnostic Explanations (LIME)** and **Shapley Values**.

LIME

- LIME provides a way to interpret black box models by approximating them locally with interpretable models. It focuses on local fidelity, ensuring that the explanation corresponds to how the model behaves in the vicinity of the instance being predicted. LIME is particularly useful for tasks like text and image classification, where it helps in understanding which features of the data lead to certain predictions**.**

Shapley Values

- Shapley Values: These are used for assessing the contribution of each feature in the prediction of a model. This method is model-independent and is generally applied to measure the importance of features in a trained model. However, Shapley Values require significant computational resources, and their effectiveness in explaining the complex interactions in deep neural networks is limited.

Challenges in Interpretability

Deep learning models are complex, with multiple layers and non-linear interactions, making them inherently difficult to interpret. This complexity poses a significant challenge to achieving full interpretability. Additionally, there is a trade-off between fidelity (accuracy of the interpretation relative to the model's operation) and simplicity (ease of understanding the interpretation). Methods like LIME and Shapley Values, while useful, may not always provide a complete or globally faithful explanation of a model's behavior.

Moreover, ensuring the trustworthiness of interpretation algorithms is crucial. These algorithms should accurately reveal the rationale behind a model's decisions. Trustworthiness is especially important when the interpretation is provided by an extrinsic algorithm, which might not be part of the model being interpreted.

Conclusion

Interpretability in deep learning is vital for understanding, trusting, and effectively using AI systems, especially in critical applications. Methods like LIME and Shapley Values have made significant strides in this direction. However, the inherent complexity of deep learning models and the challenge of balancing fidelity with interpretability mean that this is an ongoing area of research. Future advancements in this field are crucial for the development of AI systems that are not only powerful but also transparent and trustworthy.

References

- For a detailed exploration of LIME and its application in interpretability, see *"Why Should I Trust You?" Explaining the Predictions of Any*

Classifier (https://ar5iv.labs.arxiv.org/html/1602.04938)

- For a comprehensive understanding of interpretability and trustworthiness in deep learning, refer to *Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond* (https://ar5iv.labs.arxiv.org/html/2103.10689)
- Additional insights into interpretability challenges and methodologies can be found in *DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems* (https://ar5iv.labs.arxiv.org/html/1906.10263)
- For further exploration of interpretability methods in Natural Language Processing, the article *Model Explainability in Deep Learning Based Natural Language Processing* provides valuable information (https://ar5iv.labs.arxiv.org/html/2106.07410)