XAI: Cluster Variational Inference - draft

Jack Li

October 2023

1 Introduction

In the realm of machine learning and artificial intelligence, two critical areas of research have emerged: adversarial attacks and Explainable Artificial Intelligence (XAI). Adversarial attacks involve the deliberate manipulation of input data to deceive machine learning models, thereby exposing vulnerabilities in their predictive capabilities[1][2]. These attacks can significantly undermine the reliability and robustness of AI systems, particularly in high-stakes applications such as autonomous driving and healthcare. It is noteworthy more than any other defense technique, as it is an essential component of AI systems[3][4].

Conversely, Explainable Artificial Intelligence (XAI) aims to enhance the transparency and interpretability of AI models[5]. XAI techniques provide insights into the decision-making processes of complex models, enabling users to understand, trust, and effectively manage AI systems. While adversarial attacks exploit the opacity of AI models to generate misleading outputs, XAI seeks to demystify these models, offering explanations that can be scrutinized and validated by human experts.

The interplay between adversarial attacks and XAI is multifaceted. On one hand, XAI can be leveraged to develop more robust models that are resistant to adversarial perturbations. On the other hand, adversarial examples can serve as tools for generating local explanations, thereby contributing to the broader objectives of XAI. Understanding the distinctions and interactions between these two domains is crucial for advancing the field of AI and ensuring the development of secure, transparent, and trustworthy AI systems.

However, the adversarial attacks method is aimed to deceive the model, meaning it still have different perspectives from XAI method. In this paper, we will provide variational inference to measure the critical features in computer vision models. Meanwhile, it will also be more meaningful to evaluate models in terms of the robustness and sensitivityness.

References

[1] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, March 2015.

[2] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, February 2017.

[3] Andrew Slavin Ross and Finale Doshi-Velez. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients, November 2017.

[4] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks, March 2017.

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619, May 2002.