BAI Ph.D. Proposal

Faculty: R. Venkatesh Babu (CDS) and S.P. Arun (CNS)

Neuroscience-Inspired Methods for Improving the Robustness of Deep Networks

Introduction:

Deep learning has been one of the most promising fields of machine learning that has achieved remarkable success in various domains such as image and speech processing, natural language processing, and recommendation systems. However, deep models are highly vulnerable to adversarial attacks, where an attacker can manipulate the input data to mislead the model without significantly changing the original data's perception. Deep Networks are also brittle against distribution shifts between the training and inference data. These vulnerabilities can have severe consequences, especially in safety-critical applications such as autonomous driving, medical diagnosis, and financial fraud detection.

Objective:

The objective of this research proposal is to investigate the feasibility of developing robust deep models against adversarial attacks and distribution shifts based on biologically inspired or neuroscience-inspired ways. Specifically, we aim to explore the potential of incorporating the principles of the human visual system such as attention mechanisms, visual saliency, and object recognition to enhance the robustness of deep models.

Methodology:

In this research, we will first conduct a comprehensive review of the literature on deep learning, adversarial attacks and defenses, vulnerability of deep models to distribution shifts, and existing methods of improving their robustness. We will also investigate the recent advances in the understanding of the human visual system and its potential implications for deep learning.

Based on the literature review, we will develop biologically inspired or brain-inspired defense mechanisms against adversarial attacks and distribution shifts. We will explore various strategies, such as incorporating attention mechanisms to prioritize the important features of the input data, using visual saliency to filter out irrelevant information, and leveraging object recognition to detect and classify the objects in the input data.

To evaluate the effectiveness of the proposed defense mechanisms, we will conduct extensive experiments on benchmark datasets such as MNIST, CIFAR-10, and ImageNet, and the popular domain generalization datasets on the DomianBed benchmark, and compare the performance of our approach with the state-of-the-art methods.

Expected Outcomes:

We expect that this research will contribute to the development of robust deep models against adversarial attacks and distribution shifts, which is particularly important for their safe

and reliable deployment. By incorporating the principles of the human visual system, we anticipate that our approach can enhance the interpretability and robustness of deep models, leading to more reliable and trustworthy AI systems.

Conclusion:

This research proposal aims to investigate the feasibility of developing robust deep models against adversarial attacks and distribution shifts based on biologically inspired or neuroscience-inspired ways. We believe that this research can pave the way for future studies on incorporating the principles of the human visual system into deep learning, opening up new opportunities for enhancing the interpretability, robustness, and reliability of AI systems.