AI internship: enhancing AI assistant performance with early exit and cloud offloading

Dilan Gandhi
Science & Engineering
Manalapan High School
Englishtown, NJ
525dgandhi@frhsd.com

Abstract

The development of AI assistants such as ChatGPT, Bard, and Copilot has become increasingly important as they are integrated into various applications, ranging from customer service to personal productivity. However, many current AI assistants face processing speed, accuracy, and context understanding limitations, especially when handling complex tasks like image processing or real-time decision-making. These limitations often lead to delays, incorrect predictions, and an overall less efficient user experience. Current models, such as traditional machine learning algorithms, struggle to balance efficiency with accuracy, leading to trade-offs that compromise performance, particularly in mobile and real-time environments where computational resources are constrained. This internship, conducted as part of an AI-focused project with the University of Maryland, aims to address these challenges.

In this project, I focused on improving the performance of an AI assistant by developing an optimized model capable of faster, more accurate predictions while maintaining high levels of contextual understanding. By incorporating advanced architectures such as ResNet18 with early exit mechanisms, I aimed to reduce the time taken for processing tasks, especially when handling images uploaded by users. I tested and implemented cloud offloading techniques to ensure that complex computations are processed in the cloud, offloading tasks from local devices. Additionally, I explored different unlearning strategies to enhance the assistant's adaptability to new data and improve its overall efficiency. The result is a more efficient AI assistant that can handle real-time image processing, adapt to new data inputs, and provide quicker responses, improving the user experience while maintaining high accuracy.

Index Terms

AI assistant, early exit, cloud offloading, unlearning, ResNet18, internship