Informatics Institute of Technology

In Collaboration With

University of Westminster, UK

University of Westminster, Coat of Arms

ANNOE: Pushing the Boundaries of Neural Network Optimization through Adaptive Technique Selection and Application within Computational Constraints

A dissertation by

Mr. Kariyawasam Bovithanthri Thavindu Nimendra Kariyawasam

W1761259 / 2019264

Supervised by

Mr. Sharmilan Somasundaram

May 2023

This Project Proposal is submitted in partial fulfillment of the requirements for the BEng (Hons) Software Engineering degree at the University of Westminster.

Abstract

As the demand for deploying deep learning models on high-end mobile devices and IoT devices increases, the need for efficient machine learning model optimization becomes critical to execute on-device AI tasks. One of the primary challenges in this context is the retraining process of these models on IoT and mobile devices with new data, which is constrained due to limited processing power. This limitation hinders the generation of accurate domain-specific inference results, as models may not be as well-trained with real-world.

To address this problem, a novel prediction system was developed that predicts the most suitable optimization technique using historical data. This system incorporates a neural network specifically designed for making predictions under available computational resources. By devising an adaptive neural network training and optimization approach, the system can interact with new data effectively and accurately, enhancing the overall performance of deep learning models on mobile and IoT devices.

The proposed system was tested under various resource utilization scenarios, with a focus on data science metrics to evaluate its performance. Compared to the random selection of optimization techniques, the novel prediction system demonstrated significant improvements in the effectiveness of the overall system, increasing it by 15-25%. These results highlight the potential of the adaptive neural network training and optimization approach in enhancing the performance of deep learning models deployed on resource-constrained devices, paving the way for more accurate domain-specific inference outcomes.

Keywords: Prediction Systems, Neural Network Optimizations Techniques, Machine Learning, IoT, Resource-constrained Environments.

Subject Descriptions:

- Computing methodologies → Artificial intelligence → Computer vision → Computer vision problems → Object recognition
- Computing methodologies → Machine learning → Machine learning approaches → Neural networks
- Hardware → Sensor applications and deployments