Dynamic Activation Function for Efficient Inference of LLMs: Project Proposal Winter 2024-2025

Idan Kashani - idan-kashani@campus.technion.ac.il
Moshe Kimhi - moshekimhi@campus.technion.ac.il

Background

Large Language Models (LLMs), such as GPT, have transformed the field of natural language processing (NLP), enabling machines to perform tasks like translation, summarizing, and question answering with remarkable accuracy. However, these models are computationally intensive and require substantial resources for both training and inference, limiting their usability in real-time applications and on resource-constrained devices.
One of the main challenges in improving LLM efficiency lies in the neural network architecture, particularly in activation functions that introduce non-linearity. Activation functions play a critical role in model performance and efficiency. Most LLMs currently use GELU (Gaussian Error Linear Unit), which, while effective, is computationally complex, making it costly in terms of time and energy consumption.

Project Proposal

This project will focus on the development of a dynamic activation function for efficient inference, providing hands-on experience in optimizing language models. In addition to activation functions, the research will include methods like quantization to enhance model efficiency. The proposed activation function is a dynamic linear combination of ReLU and GELU, evolving during fine-tuning by gradually assigning a greater weight to ReLU over GeLU, with the final convergence to ReLU.

Objectives

- Implement the activation function and modify an existing language model architecture by replacing GeLU with the proposed activation.
- Fine-tune the modified model.

- Evaluate model performance using benchmarks and metrics such as throughput, memory utilization, and environmental impact (e.g., carbon footprint).
- Gain practical experience in running experiments and tracking results using tools like Weights & Biases (wandb).

For additional information, please contact us at the emails provided.