ProTran: Profiling the Energy of Transformers on Embedded Platforms

Shikhar Tuli, *Student Member, IEEE,* and Niraj K. Jha, *Fellow, IEEE*

***Abstract*—The abstract goes here.**

***Index Terms*—Machine Learning, Transformers, Embedded Platforms.**

I. Introduction

THIS is the paper skeleton for the ProTran project for the course ECE 464.

- Why transformers have gained so much attention by industry and academia.
- How transformer architectures have been increasing in model size. Challenges of running these architectures on embedded devices for edge-AI.
- Prior work on pruning transformers and running hardware-aware NAS. What inspiration can be obtained from these works and what are the challenges left. Limitations of previous works and how they can be countered.
- Briefly introduce proposed approach for creating an inclusive surrogate benchmark of transformer performance measures on diverse embedded platforms. Bullet points for contributions.
- A paragraph on the organization of the paper.

II. Background and Related Work

- A section on background and related work in both hardware-aware NAS and pruning methods. Every section highlights the drawbacks of that work and how our paper suggests improvements.

A. Transformer Architectures

- Mention popular transformer architectures proposed in the literature.
- Hint towards a design space of transformer architectures. Show previous works on design space generation.
- Show how transformers are getting larger-and-larger. Resulting problems in memory footprint and energy consumption on edge devices. Taking motivation from HAT, show how activations can be reduced and network can be 8-bit quantized, etc.

This work was supported by NSF Grant No. —. S. Tuli and N. K. Jha are with the Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, 08544, USA (e-mail: {stuli, jha}@princeton.edu).
Manuscript received —; revised —.

B. Hardware-Aware Neural Architecture Search

- Show traditional NAS works in CNNs. How these works have shown model reduction and improvements in energy/latency measures. Cite TCAD.
- Show transformer-specific NAS works. Cite FlexiBERT, HAT. Previous works do not target simultaneous latency/energy/power consumption for optimization.
- Need for a benchmarking platform that gives model accuracy/latency/energy/power on diverse edge platforms. Show how FlexiBERT can be leveraged.

C. Energy Profiling of ML Models

- Show previous works for CNN design spaces. Cite ChamNet. Show co-design approaches like MCUNet that use a range of off-the-shelf micro-controllers.
- Show how FTRANS profiles energy for FPGA platforms, but HAT, FlexiBERT, etc. do not. Need for surrogate models (trained using active learning) on diverse embedded platforms for a design space of transformer architectures. Will aid co-design of transformer architectures in edge applications.

III. Motivation

A. Energy Reduction on Mobile Platforms

- Show reduction in energy consumption of running inference of an off-the-shelf transformer architecture on a mobile device (*e.g.*, an iPhone), compared to CPU and GPU energy consumption.

B. Challenges and Proposed Solutions

- Show challenges in trying to fit large models. Where are the memory bottlenecks? Show latency increase due to gradient accumulation. Show advantages of quantization and reduced activation models.

IV. The ProTran framework

V. Experimental Setup

VI. Results

VII. Conclusion

References