ProTran: Profiling the Energy of Transformers on Embedded Platforms

Shikhar Tuli

February 2022

Project Proposal

In recent years, self-attention-based transformer models [1, 2] have achieved state-of-the-art results on tasks that span the natural language processing (NLP) domain and, more recently, even the computer vision domain [3]. This burgeoning success has largely been driven by large-scale pre-training datasets, increasing computational power, and robust training techniques [4]. However, a challenge that remains is to get pruned models that can *efficiently* be run on embedded devices.

Many works have tried to tackle this problem by running neural architecture search (NAS) on a large design space of transformer models, thanks to recent advancements in enhancing the flexibility of transformer architectures [5–7]. Others aim at pruning off-the-shelf pre-trained models, in order to fit these models on edge devices. Such works often target reduction in the floating-point operations per second (FLOPs) or even the latency on certain devices [8–10]. However, to the best of our knowledge, there is no such work that profiles the energy and power of running these models on such devices [11]. This is important as many edge applications have strict constraints on energy (being consumed from a battery) or the power consumption (from an intermittent power supply).

Hence, in this work, we propose to not only profile the latency, but also the energy and power for a diverse range of transformer models. Based on this setup, we leverage active learning to train a predictor function that models the latency, energy and power for each transformer model in the design space,

for a given embedded platform. The target embedded devices (along with baseline server platforms) are:

- Apple iPhone (already available).
- Google Pixel phone.
- Raspberry Pi 4 (quad-core ARM A72 SoC; approx. cost: $75).
- NVIDIA Jetson Nano (quad-core ARM A57 + 128-core NVIDIA Maxwell embedded GPU SoC; approx. cost: $100).
- NVIDIA Jetson TX2 (quad-core ARM A57 + dual-core NVIDIA Denver CPU + 256-core NVIDIA Pascal embedded GPU SoC; approx. cost: $980).
- Intel Movidius Neural Compute Stick 2 (Intel Myriad X Vision Processing Unit; approx. cost: $80).
- 128-core 2.6 GHz AMD EPYC Rome CPU (already available).
- NVIDIA A100 GPU (already available).

The design space of models would be selected as a subset of FlexiBERT, based on pruned models from BERT-Base that show higher performance on the GLUE benchmark, while also having fewer number of parameters. Evidently, optimal model sizes for each platform would be different [5]. The models run on these platforms need not be trained yet, since we do not target a joint modeling of model performance with energy consumption (although weights would be transferred from the nearest trained neighbors in the FlexiBERT design space). Both training and inference metrics would be modeled.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 4171–4186, 2019.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[5] Hanrui Wang, Zhanghao Wu, Zhijian Liu, Han Cai, Ligeng Zhu, Chuang Gan, and Song Han. HAT: Hardware-aware transformers for efficient natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7675–7688, July 2020.

[6] Ashish Khetan and Zohar Karnin. schuBERT: Optimizing elements of BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2807–2818, July 2020.

[7] Anonymous. FlexiBERT: Expanding the flexibility of transformer architectures and exploring a heterogeneous design space. *International Conference on Machine Learning*, 2022. In review.

[8] Jin Xu, Xu Tan, Renqian Luo, Kaitao Song, Jian Li, Tao Qin, and Tie-Yan Liu. NAS-BERT: Task-agnostic and adaptive-size BERT compression with neural architecture search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, page 1933–1943, 2021.

[9] Yichun Yin, Cheng Chen, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. AutoTinyBERT: Automatic hyper-parameter optimization for

efficient pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5146–5157, Online, August 2021. Association for Computational Linguistics.

[10] Forrest Iandola, Albert Shaw, Ravi Krishna, and Kurt Keutzer. SqueezeBERT: What can computer vision teach NLP about efficient neural networks? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 124–135, Online, November 2020. Association for Computational Linguistics.

[11] Bingbing Li, Santosh Pandey, Haowen Fang, Yanjun Lyv, Ji Li, Jieyang Chen, Mimi Xie, Lipeng Wan, Hang Liu, and Caiwen Ding. FTRANS: Energy-efficient acceleration of transformers using fpga. In *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design*, ISLPED '20, page 175–180, New York, NY, USA, 2020. Association for Computing Machinery.