CS 561 Lit Review SDSU

Brendan Barry-Owen
SDSU
bbarryowen5056@sdsu.edu

Eddie Coda
SDSU
ecoda@sdsu.edu

1 General problem/task definition

The core problem is developing models that not only process language quickly, but also maintain a high level of accuracy and relevance in their output. By comparing CrammedBert and UltraFastBert with FACE, we hope to evaluate if the speed gains come at the cost of language generation quality. This is vital for applications where balance between response time and output quality is crucial. This balance is crucial in a wide range of applications, from real-time language translation to efficient content generation for limited-resource devices. This analysis will help in understanding the trade-offs in designing faster language models. Are these models merely fast, or do they also maintain a high level of linguistic coherence and relevance? It's not just about being fast; it's about being effectively fast

BERT's architecture, while groundbreaking, brings with it a set of inherent challenges, particularly in computational efficiency and scalability. The model's reliance on deep transformer networks necessitates substantial computational resources. This is primarily due to the transformer's attention mechanism, which requires calculations over all pairs of input positions in every layer, as noted by Vaswani et al. in their seminal paper introducing transformers. For large inputs, or when processing extensive datasets, this can lead to considerable computational demands, making the model less feasible for use in environments with limited computational resources.

For Abstract: The quintessential challenge in modern NLP is not merely accelerating language processing but achieving this without sacrificing the nuance and fidelity of human-like language understanding. In evaluating CrammedBert and UltraFastBert through the lens of FACE, we aim to discern whether their computational speed is offset by a degradation in language generation quality. This analysis transcends theoretical interests, addressing a pressing need in real-time language applications and content generation for resource-limited environments. Understanding these trade-offs is pivotal, as the pursuit of efficiency in language models should not eclipse the quest for linguistic coherence and relevance. The overarching question isn't just about achieving velocity in computation; it's about redefining what it means to be efficiently articulate.

2 Concise summaries of the articles

BERT paper: BERT's novelty lies in its bidirectional training, which allows it to capture the context of a word based on all surrounding words, unlike previous models that processed text in a unidirectional manner. Key to BERT's success is its pre-training on large datasets, followed by fine-tuning for specific tasks, resulting in unprecedented performance across various NLP benchmarks.

CrammedBert Paper: This paper introduces a model optimized for faster processing times while attempting to maintain the linguistic capabilities of the original BERT model. Its major contribution is in demonstrating how certain optimization techniques can significantly speed up language processing without a substantial loss in performance.

UltraFastBert Paper: Similar in aim to CrammedBert, this paper presents a different approach to achieving high-speed language processing. This paper presents a novel approach to energy accounting in deep learning models, focusing on tensor-based computations. The authors introduce a system for analyzing and optimizing the energy consumption of models like BERT. The paper provides methodologies for breaking down and assessing energy consumption at a granular level, enabling a more targeted approach to energy efficiency.

FACE Paper: This paper introduces a novel set of metrics based on Fourier Analysis of Cross-Entropy for evaluating natural language generation models. The authors propose these metrics to mea-

sure the distance between machine-produced and human language, addressing a critical challenge in NLP. FACE leverages the periodicity of entropy in language and employs Fourier analysis to estimate cross-entropy from different data sources, thereby capturing the similarity between model-generated and human-written languages. The paper demonstrates that FACE can effectively identify the human-model gap, scales with model size, and reflects various outcomes of different sampling methods used in decoding. Importantly, it is shown to correlate well with other evaluation metrics and human judgment scores, making it a robust tool for assessing language model performance.

3 Compare and contrast

Common Goals: All papers share the common goal of advancing NLP - either through developing faster models or through better evaluation techniques.

Methodological Differences: While CrammedBert and UltraFastBert focus on model architecture and optimization, FACE focuses on evaluation metrics like the novel MAUVE did. The first two are about building, the latter about measuring and evaluating how close the industry has come to building an infrastructure we are all proud of. We hold efficiency at a high standard but more importantly, we hold honesty as the leading motivation.

Complementary Nature: FACE can be seen as a tool to assess the trade-offs made in CrammedBert and UltraFastBert. While these models prioritize speed, FACE evaluates the impact of this on language generation quality.

Potential Conflicts: There might be a conflict in objectives - while speed is the primary goal for CrammedBert and UltraFastBert, FACE could reveal that this speed comes at a significant cost to certain aspects of language quality.

4 Future work

Our findings will serve as a benchmark for future language models aiming to achieve a balance between speed and quality. This is crucial for the development of efficient NLP applications in resource-constrained environments. The research will contribute to the broader understanding of efficiency in language models. It's a step towards developing models that are not only computationally efficient but also linguistically competent. Hybrid Model Development: Future work could explore the development of hybrid models that incorporate the strengths of both CrammedBert and UltraFastBert, based on the insights provided by FACE evaluations.

Further Optimization: There's scope for further optimization in these models, perhaps using insights gained from FACE to minimize quality loss while enhancing speed.

Broader Application of FACE: Future explorations might pivot towards hybrid models that synergize the strengths of CrammedBert and UltraFastBert, informed by the meticulous evaluations of FACE. Moreover, extending the application of FACE to a broader spectrum of models could unearth foundational principles of efficient language processing.

Real-World Applications: Real-world testing in scenarios like voice-assisted technology or translation services will be instrumental in translating these theoretical advancements into tangible benefits. This journey is not just about engineering faster models but about sculpting a language technology ecosystem that is as linguistically astute as it is computationally robust.

References

Peter Belcak and Roger Wattenhofer. 2023. Exponentially faster language modelling.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text.

Jonas Geiping and Tom Goldstein. 2022. Cramming: Training a language model on a single gpu in one day.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Li Shen, Yan Sun, Zhiyuan Yu, Liang Ding, Xinmei Tian, and Dacheng Tao. 2023. On efficient training of large-scale deep learning models: A literature review.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding.

Zuhao Yang, Yingfang Yuan, Yang Xu, Shuo Zhan, Huajun Bai, and Kefan Chen. 2023. Face: Evaluating natural language generation with fourier analysis of cross-entropy.

Wangchunshu Zhou and Ke Xu. 2020. Learning to compare for better training and evaluation of open domain natural language generation models.

(Yang et al., 2023)
(Belcak and Wattenhofer, 2023)
(Devlin et al., 2019)
(Dou et al., 2022)
(Geiping and Goldstein, 2022)
(Pillutla et al., 2021)
(Sanh et al., 2020)
(Shen et al., 2023)
(Wang et al., 2019)
(Zhou and Xu, 2020)