11-877 Advanced Multimodal Machine Learning **Spring 2024**

Week 13: Efficiency and Privacy

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: Ashwin Pillay*

Edited by Paul Liang *Scribes: Jiya Zhang and Ashwin Pillay*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. It brings unique challenges for both computational and theoretical research given the heterogeneity of various data sources.

During week 13's discussion session, the class started by addressing ways to improve the efficiency of multimodal models. This was followed by identifying some avenues that may compromise the privacy of such models, and the potential ways to resolve them.

The following was a list of provided research probes:

1. Different papers have different definitions of efficiency, including memory, time, space etc. Are there other notions of efficiency that you think current work is missing out on, especially as we build multimodal systems for the real world? How can we make progress on these new notions?
2. How can our study of multimodal connections/interactions help us design more efficient models? How should we balance careful and efficient model design from the start, versus training large models and compressing them as a post-hoc step?
3. How can we scale multimodal models to extremely long sequence lengths, such as over years of human experience? What new capabilities will this enable? How can we start creating benchmarks to make progress toward these capabilities?
4. There has been a lot of work on making language models and vision models more efficient - what ideas here can be translated to other modalities and other multimodal problems? What new domain expertise will we need to build efficient models for these other settings?
5. Most works in improving efficiency fixes the modalities and makes the models more efficiency. Are there potential ideas on changing the modalities themselves so that they can be more efficiently handled (e.g., going from video to images or wireless sensors?)
6. How can we formalize the balance between information, fidelity, efficiency, and privacy of different modalities, and how can we choose which ones to use for a given problem?

As background, students read the following papers:

1. **(Required)** Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference [Zhao et al., 2024]
2. **(Required)** Vid2Doppler: Synthesizing Doppler Radar Data from Videos for Training Privacy-Preserving Activity Recognition [Ahuja et al., 2021]
3. (Suggested) FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness [Dao et al., 2022]
4. (Suggested) FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning [Dao, 2023]
5. (Suggested) SAMoSA: Sensing Activities with Motion and Subsampled Audio [Mollyn et al., 2022]
6. (Suggested) SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models [Xiao et al., 2022]
7. (Suggested) Efficient and accurate prediction of protein structure using RoseTTAFold2
8. (Suggested) Efficiently Scaling Transformer Inference [Pope et al., 2022]

9. (Suggested) Privacy Enhanced Multimodal Neural Representations for Emotion Recognition [Jaiswal2019PrivacyEM]
10. (Suggested) QLORA: Efficient Finetuning of Quantized LLMs [Dettmers et al., 2023]
11. (Suggested) Mamba: Linear-Time Sequence Modeling with Selective State Spaces [Gu and Dao, 2023]
12. (Suggested) Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models [Luo et al., 2023]
13. (Suggested) Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models [Lu et al., 2022]

Scribe notes: We summarize several main takeaway messages from group discussions below:

1 Looking Beyond Conventional Definitions of Efficiency

While discussions on efficiency typically focus on memory, time, and space usages, these factors represent only a subset of what may determine the efficiency of multimodal models. Through our discussions, we identify several alternative parameters critical to assessing model efficiency:

1. **Carbon Footprint (or General Energy Usage)**:
   Multimodal models comprising large neural networks consume substantial energy during training and inference. Previously, [Castaño et al., 2023] explored the carbon footprint of some models hosted on HuggingFace and have subsequently laid guidelines for researchers to report the energy efficiency of their works. However, it is crucial to note that the energy usage for a model could span aspects far beyond its training; [Strubell et al., 2019] performed such analyses on NLP models in factors like Neural Architecture Search (NAS) and hyperparameter optimization. Future efforts should aim to report comprehensive energy usage details across all stages of model deployment, from training to hyperparameter search to inference. This information can even be incorporated into existing scaling laws like those established by [Hoffmann et al., 2022] and [De Vries, 2023] to optimize model designs within energy constraints.
2. **User Accessibility and System Integration**: The accessibility and ease of integration of models like ChatGPT, compared to their open-source counterparts such as LLaMA, Mistral, and Gemma [1], demonstrate the significance of user-friendly interfaces and robust system architecture. Efficient models should not only perform well but also be straightforward to deploy and scale across various platforms, requiring minimal changes for upgrades or feature expansions. This would involve a collaborative effort from ML researchers, software architects, and DevOps experts to create scalable, adaptable systems.

2 Improving Multimodal Efficiency

As multimodal models evolve, ensuring their efficient scaling is paramount, particularly when performance enhancements typically stem from increased model size. Strategies to enhance efficiency include:

1. **Training New Multimodal Models**:
   - `Temporal Segmentation for Time-based Modalities`: Inspired by techniques in [Carreira and Zisserman, 2017], segmenting and aligning temporal data like video and audio can significantly reduce computational demands, especially when paired with architecture optimizations specific to each modality.
   - `Direct Multimodal Integration`: Moving beyond LLMs that rely on text as a ”core modality”, exploring models with the ability for direct planning and decision-making using visual, audio, or environment state data can mitigate the inefficiencies of modality conversion and leverage the inherent strengths of each data type.
2. **Fine-tuning and Adapting Existing Models**:
   - `Parameter-Efficient Fine-Tuning (PEFT) Techniques`: Exploring advanced, non-intrusive methods such as AdaLink [Wang et al., 2023] and Representation Finetuning (ReFT) [Wu et al., 2024] can enhance model adaptability while maintaining architectural integrity, offering substantial

[1] We are referring to the locally-run versions of these models using services like `https://ollama.com/`

improvements over traditional fine-tuning approaches.

3 Privacy in Multimodal Models

The integration of multiple modalities in models not only enhances capabilities but also broadens potential security and privacy-related vulnerabilities. Recent incidents like the XZ backdoor [2], underscore the risks associated with systems even if they are open-sourced and widely adopted. Additionally, the black-box nature of DL models exacerbates this issue; for example, [Liu et al., 2024] highlighted how LoRA models can be used to discreetly inject backdoors into public LLMs. Potential solutions to these problems include:

1. **Verifiable Builds**: Leveraging platforms like CodaLab [3] for building models from source can enhance security by allowing community verification and publicly-accessible build processes.
2. **Personalized Models**: Developing models trained exclusively on user-provided data can tailor experiences and enhance security by limiting data exposure.
3. **Privacy-preserving Modalities**: Employing non-invasive modalities such as millimeter wave (mmWave) sensors over traditional visual systems for tasks like activity recognition [Ahuja et al., 2021] can maintain functionality without compromising privacy.

By addressing these expanded definitions of efficiency and integrating advanced security measures, the development of multimodal systems can proceed in a manner that is not only technologically advanced but also sustainable and secure.

References

Karan Ahuja, Yuemin Jiang, Mayank Goel, and Chris Harrison. Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. URL https://api.semanticscholar.org/CorpusID:233987106.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. URL https://api.semanticscholar.org/CorpusID:206596127.

Joel Castaño, Silverio Mart'inez-Fern'andez, Xavier Franch, and Justus Bogner. Exploring the carbon footprint of hugging face's ml models: A repository mining study. *2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–12, 2023. URL https://api.semanticscholar.org/CorpusID:258762605.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *ArXiv*, abs/2307.08691, 2023. URL https://api.semanticscholar.org/CorpusID:259936734.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher R'e. Flashattention: Fast and memory-efficient exact attention with io-awareness. *ArXiv*, abs/2205.14135, 2022. URL https://api.semanticscholar.org/CorpusID:249151871.

Harm De Vries. Go smol or go home, 2023. URL https://www.harmdevries.com/post/model-size-vs-compute-overhead/.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314, 2023. URL https://api.semanticscholar.org/CorpusID:258841328.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv*, abs/2312.00752, 2023. URL https://api.semanticscholar.org/CorpusID:265551773.

[2] https://en.wikipedia.org/wiki/XZ_Utils_backdoor

[3] https://codalab.org/

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022. URL https://api.semanticscholar.org/CorpusID:247778764.

Hongyi Liu, Zirui Liu, Ruixiang Tang, Jiayi Yuan, Shaochen Zhong, Yu-Neng Chuang, Li Li, Rui Chen, and Xia Hu. Lora-as-an-attack! piercing llm safety under the share-and-play scenario. *ArXiv*, abs/2403.00108, 2024. URL https://api.semanticscholar.org/CorpusID:268201486.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *ArXiv*, abs/2211.01095, 2022. URL https://api.semanticscholar.org/CorpusID:253254916.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shen Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *ArXiv*, abs/2305.15023, 2023. URL https://api.semanticscholar.org/CorpusID:258865326.

Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. Samosa: Sensing activities with motion and subsampled audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(3), sep 2022. doi: 10.1145/3550284. URL https://doi.org/10.1145/3550284.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *ArXiv*, abs/2211.05102, 2022. URL https://api.semanticscholar.org/CorpusID:253420623.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *ArXiv*, abs/1906.02243, 2019. URL https://api.semanticscholar.org/CorpusID:174802812.

Yaqing Wang, Jialin Wu, Tanmaya Shekhar Dabral, Jiageng Zhang, Geoff Brown, Chun-Ta Lu, Frederick Liu, Yi Liang, Bo Pang, Michael Bendersky, and Radu Soricut. Non-intrusive adaptation: Input-centric parameter-efficient fine-tuning for versatile multimodal modeling. *ArXiv*, abs/2310.12100, 2023. URL https://api.semanticscholar.org/CorpusID:264289072.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Daniel Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models. 2024. URL https://api.semanticscholar.org/CorpusID:268889731.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. *ArXiv*, abs/2211.10438, 2022. URL https://api.semanticscholar.org/CorpusID:253708271.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. *ArXiv*, abs/2403.14520, 2024. URL https://api.semanticscholar.org/CorpusID:268553791.