11-877 Advanced Multimodal Machine Learning **Spring 2023**

Week 6: Pretraining and Scaling

Instructors: Louis-Philippe Morency and Paul Liang *Synopsis Leads: Leena Mathur and Yihan Cao*

Edited by Paul Liang *Scribes: Suzanne Nie and Ryan Liu*

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.*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

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. Recent research directions in multimodal machine learning have focused on **multimodal pretraining** and **scaling**. In Week 6's discussion, the class focused on the following discussion probes:

1. Is large-scale pretraining the way forward for building general AI models? What information potentially cannot be captured by pretraining? What are some potential risks of pretraining and scenarios where we should not use pretrained models?
2. How can we, in an academic environment, do impactful research in multimodal pretraining? What would be your proposed multi-year research agenda in this topic?
3. What are the types of cross-modal interactions that are likely to be modeled by current pretrained models? What cross-modal interactions will be harder to model with these methods? Do you have proposals for different pretraining data, architectures, or objectives that can better capture these interactions?
4. How can we best integrate multimodality into pretrained language models? What kind of additional data and modeling/optimization decisions do we need?
5. What are the different design decisions when integrating multimodal information in pretraining models and objectives? What are the main advantages and drawbacks of these design choices? Consider not just prediction performance, but tradeoffs in time/space complexity, interpretability, and so on.
6. How can we evaluate the type of multimodal information learned in pretrained models? One approach is to look at downstream tasks; what are other ways to uncover knowledge stored in pretrained models?

Students read the following papers (Req = required):

1. (Req) Scaling Laws for Generative Mixed-Modal Language Models [Aghajanyan et al., 2023]
2. (Req) A Generalist Agent [Reed et al., 2022]
3. Scaling Laws for Autoregressive Generative Modeling [Henighan et al., 2020]
4. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models [Li et al., 2023]
5. VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers [Aflalo et al., 2022]
6. Vision-Language Pre-training: Basics, Recent Advances, and Future Trends [Gan et al., 2022]
7. VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks [Sung et al., 2022]
8. HighMMT: Towards Modality and Task Generalization for High-modality Representation Learning [Liang et al., 2022]
9. Training Compute-optimal Large Language Models [Hoffmann et al., 2022]
10. Multimodal Pretraining Unmasked: A Meta-analysis and a Unified Framework of Vision-and-language BERTs [Bugliarello et al., 2021]
11. On the Opportunities and Risks of Foundation Models [Bommasani et al., 2021]

12. Flamingo: A Visual Language Model for Few-shot Learning [Alayrac et al., 2022]

We summarize several main takeaway messages from group discussions below:

1 Is Pretraining the Future for AI?

Pretraining to gain foundational intelligence: Our discussion group agreed that some amount of pretraining will be foundational to the future development of AI systems. An an analogy was drawn between ”pretraining” and the ”basic education” that all humans receive before people choose more specialized (”fine-tuned”) fields to study. Pretraining may be the most effective away to ensure models begin with some priors about the structure and semantics of the world; these priors can then be used for downstream tasks [Choi et al., 2022]. What these priors are, especially across modalities, is an open question and one worth studying by researchers in the coming years (e.g., during visual pretraining, would a model learn visual priors that are useful for multimodal connections, etc).

Can other architectures outperform large pretrained models? Our group discussed a paper on ConvNet approaches for outperforming transformers [Liu et al., 2022] and whether additional architectures or paradigms could be created to outperform pretraining-based approaches.

Counterpoint – is multimodal pretraining necessary? Our group also speculated whether multimodal pretraining is necessary, compared to pretraining unimodal models and then adapting or fusing them to perform a desired task. There is also a separate, but related question, about whether pretraining itself is useful over end-task aware training [Dery et al., 2021].

2 Limitations/Future Directions in Pretraining (Probe 1, 2, 5)

Pretrained models have emerged as powerful tools for solving tasks that require cross-modal knowledge. However, their performance can sometimes falter when applied to modality-specific downstream tasks. As such, the use of pretrained models may not always be the most effective strategy. Our group discussed several limitations of current pretraining/scaling paradigms and future research directions in this area.

Scaling Laws and Bottlenecks: There may be some future bottlenecks in computational resources and instability in the pretraining process, especially when extending pretraining to the multimodal setting. Papers that study scaling laws [Aghajanyan et al., 2023, Henighan et al., 2020], especially these laws in multimodal contexts, will become increasingly important in the years to come.

Knowledge Transfer: Finetuning a pretrained model on another task may have adverse effects on both its robustness and performance [Wortsman et al., 2022]. In particular, finetuning CLIP on ImageNet could reduce the model's robustness to distribution shifts. Not finetuning CLIP and applying it directly to downstream tasks has the potential to yield better performance. These findings suggest that knowledge transfer, especially as it relates to pretraining and robustness, is a challenging task worthy of future research investment.

Architecture Restriction: Generally, multimodal pretrained models use transformers as the model backbone. However, transformers might sometimes not work well on all modalities. For example, ViT [Dosovitskiy et al., 2020], a transformer-based model in vision, requires additional tuning tricks to obtain higher performance. This reveals that transformers might not be suitable for some modalities. It is possible that modality-specific architectures may be more useful in certain contexts; formalizing when and where these contexts occur is an important future research direction.

Domain-Specific and High-Stakes Problems: Pretrained models have demonstrated remarkable performance on general tasks and common research areas such as vision and language. However, they may falter when adapted to downstream tasks requiring highly-specific domain-knowledge. In addition, fields such as healthcare will exhibit low error tolerance for multimodal models deployed in real-world settings. In such scenarios, relying soley on systems built from pretrained models will pose risks, since these models (currently) are black boxes with no formalized approach for controlling their real-time behavior predictably.

Computational Requirements: Large pretrained models have computing restrictions due to their size and complexity, which can make training and inference prohibitively-expensive. This has prompted researchers to explore efficient model architectures and optimization techniques, such as distillation, pruning, and quantization, to reduce the computational cost of large pretrained models [Khalili et al., 2022, Sanh et al., 2019, Zhang et al., 2021].

Robustness: Despite their impressive performance on many tasks, large pretrained models can be vulnerable to adversarial attacks [Elsayed et al., 2018]. This vulnerability has important implications for the deployment of large pretrained models in high-stakes applications.

3 Academic Agendas for Multimodal Pretraining (Probe 2)

Merging pretrained models: Even if academics cannot train large models and conduct pretraining from scratch with purely academic resources, there are a wealth of research questions and directions to pursue in relation to how to **merge** the capabilities of unimodal and multimodal pretrained models. Several papers were mentioned in our discussion towards this direction [Matena and Raffel, 2021, Raffel, 2023, Khanuja et al., 2021].

Questions during continued pretraining: Academics can study new approaches for masking during continued pretraining (given a pretrained model, what additional pretraining tasks can be designed to enable the model to perform well on a desired downstream task) [Gururangan et al., 2020]. Academics can also study when when it is beneficial to **stop** pretraining in this setting (when will continued pretraining no longer be effective?). Can continued pretraining reverse any effects of the base model's pretraining? Can there be iterative pretraining approaches to save compute?

Emergence in models: A model can sometimes perform well on a task for which it was not trained [Wei et al., 2022]. How does this emergent ability develop? Academics can research this phenomena and try to develop mathematical formalisms for emergence.

Model safety, bias, toxicity: Since the overall goals of industry are not the same as academic, academics can be well-equipped to explore questions related to multimodal toxicity and bias, and the relationship between these phenomena and continued pretraining [Ousidhoum et al., 2021].

4 Cross-Modal Interactions and Multimodality (Probes 3, 4)

Multimodal pretraining aims to capture connections and interactions between data across modalities. One perspective on the role of **transformers** in this process is that they homogenize the format in which information is represented at a low level (converts every modality into a sequence of tokens).

Parameters in large models will force a constraint on the types of interactions and amount of cross-modal information that can be captured by the model. The goal is for models to learn multimodal alignment across modalities during multimodal pretraining. However, inducing alignment is also possible through approaches that operate on top of frozen models, as in [Koh et al., 2023].

Multimodal pretraining tasks can influence a model to learn cross-modal relationships. For example, a **next-modality text-audio prediction task** (given input text T and possible audio tokens A_x and A_y, which audio token comes next? In this way, sequences of text would be aligned or related to audio. Another example give in the discussion was that of word-region alignment [Zhao et al., 2021], which has the model learn semantic correlation between textual and visual modalities.

There may be **subnetworks** that represent the individual modalities in a multimodal model [Lee et al., 2020]. Our group identified this as a potential line of cross-modal research to further explore.

5 Language-Oriented Multimodal Pretraining (Probe 4)

In our discussion, we proposed the idea that natural language could be the best modality to guide multimodal pretraining architectures, for the following reasons:

- **Sequential Nature:** Language, as a modality, is inherently sequential. This characteristic is useful for guiding the integration of other modalities. One example of language guiding the combination and use of other modalities is Socratic Models [Zeng et al., 2022], which use language to guide the zero-shot composition of foundation models to perform multimodal tasks. Our group speculated that language could, similarly, be used in the pretraining stage align and fuse information across modalities.
- **Direct interaction with humans:** Natural language is the intermediate modality that humans use to describe objects and communicate with others. Using the text modality as a decoder for large pretrained models can help humans interpret and probe what the model is learning during multimodal pretraining.
- **Large Language Models (LLMs) are strong learners:** Recently, LLMs have been demonstrated to be strong learners in various diverse downstream task domains [Liu et al., 2023, Zhou et al., 2022, Tsai et al., 2019], with many of these downstream tasks being multimodal.

References

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21406–21415, 2022.

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. *arXiv preprint arXiv:2301.03728*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. *Transactions of the Association for Computational Linguistics*, 9:978–994, 2021.

Kristy Choi, Chris Cundy, Sanjari Srivastava, and Stefano Ermon. Lmpriors: Pre-trained language models as task-specific priors. *arXiv preprint arXiv:2210.12530*, 2022.

Lucio M Dery, Paul Michel, Ameet Talwalkar, and Graham Neubig. Should we be pre-training? an argument for end-task aware training as an alternative. *arXiv preprint arXiv:2109.07437*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Leila Khalili, Yao You, and John Bohannon. Babybear: Cheap inference triage for expensive language models. *arXiv preprint arXiv:2205.11747*, 2022.

Simran Khanuja, Melvin Johnson, and Partha Talukdar. Mergedistill: Merging language models using pre-trained distillation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2874–2887, 2021.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023.

Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. *arXiv preprint arXiv:2012.04124*, 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Shengtong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. Highmmt: Towards modality and task generalization for high-modality representation learning. *arXiv preprint arXiv:2203.01311*, 2022.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. *arXiv preprint arXiv:2111.09832*, 2021.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. Probing toxic content in large 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 4262–4274, 2021.

Colin Raffel. Building machine learning models like open source software. *Communications of the ACM*, 66 (2):38–40, 2023.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, 2019.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, et al. Cpm-2: Large-scale cost-effective pre-trained language models. *AI Open*, 2: 216–224, 2021.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:244–259, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022.