**11-877 Advanced Multimodal Machine Learning** **Spring 2022**

# Week 4: Pretraining Paradigm

*Instructors: L.-P. Morency, A. Zadeh, P. Liang* *Synopsis Leads: Karthik Ganesan, David Lin*

*Edited by Paul Liang* *Scribes: Justin Lovelace, Dong Won Lee*

**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/spring2022/

**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. In week 4's discussion session, the class discussed recent trends of large-scale pretrained language and multimodal models. The key themes included large-scale data collection, whether pretrained models are learning or memorizing, the cross-modal interactions learned by these models, and the overall risks and opportunities offered by the pretraining paradigm. The following is a list of provided research probes:

1. Is large-scale pretraining the way forward for building general AI models? What information potentially cannot be captured by pretraining? What are the risks of pretraining?
2. What are the types of cross-modal interactions that are likely to be modeled by current pretrained models? What are the cross-modal interactions that will be harder to model with these large-scale pretraining methods?
3. How can we best integrate multimodality into pretrained language models? What kind of additional data and modeling/optimization decisions do we need?
4. 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.
5. How can we evaluate the type of multimodal information learned in pretrained models? One approach is to look at downstream tasks, but what are other ways to uncover the knowledge stored in pretrained models?

As background, students read the following papers:

1. (Required) Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers [Hendricks et al., 2021]
2. (Required) Multimodal Few-Shot Learning with Frozen Language Models [Tsimpoukelli et al., 2021]
3. (Suggested) Unifying vision-and-language tasks via text generation [Cho et al., 2021]
4. (Suggested) Flava: A foundational language and vision alignment model [Singh et al., 2021]
5. (Suggested) Pretrained transformers as universal computation engines [Lu et al., 2021]
6. (Suggested) On the opportunities and risks of foundation models [Bommasani et al., 2021]
7. (Suggested) Does Vision-and-Language Pretraining Improve Lexical Grounding? [Yun et al., 2021]
8. (Suggested) Behind the scene: Revealing the secrets of pre-trained vision-and-language models [Cao et al., 2020]
9. (Suggested) Integrating multimodal information in large pretrained transformers [Rahman et al., 2020]
10. (Suggested) Zero-shot text-to-image generation [Ramesh et al., 2021]

We summarize several main takeaway messages from group discussions below:

# 1 Billions of Parameters is All You Need?

Recent large pretrained language models have billions of parameters (see Figure 1). Multi-TPU or Multi-GPU setups are required to efficiently pretrain with large-scale data. However, the success of distilled models suggests that we may not need as many parameters as is often used [Sanh et al., 2019]. In addition, the best models are often the models that are undergo task-specific training trained on the data most similar to the downstream task instead of simply using the most data. Instead of worrying about continuously scaling up, we may focus our attention towards developing better and more diverse datasets.

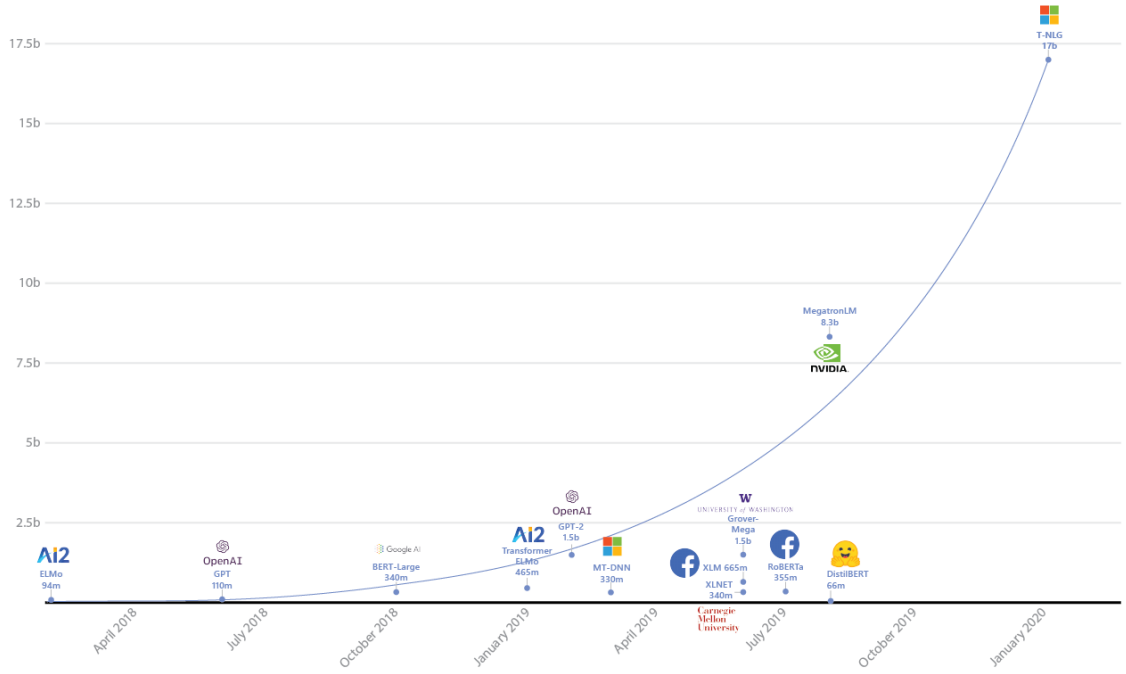


Figure 1: Recent pretrained models have billions of parameters. Source: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft.

# 2 Collecting Large-scale Datasets

There is the notion of strong pairs vs weak pairs for multimodal datasets. Strong pairs are typically represent exact alignments between modalities and are often more expensive to obtain (e.g., human-annotated image captioning datasets where all regions of the image are exactly described in text). Weak pairs represent imperfect alignments between modalities but can be obtained more cheaply from large-scale web data (e.g., Instagram captions of images are easy to obtain but may not exactly describe all details in the image). As a result, large-scale multimodal datasets used for pretraining typically leverage noisy weak pairs [Ramesh et al., 2021] scraped from the web. We may obtain strong pairs by asking humans, such as Mechanical Turk workers, to explicitly describe the images. However, due to the involvement of humans, we may need to also consider potential in-group or regional biases of the workers. We may also use strong pretrained models to label more aligned data [Schuhmann et al., 2021], structured knowledge databases like WikiHow that provide quality aligned data [Koupaee and Wang, 2018], or large-scale multimodal benchmarks that contain diverse sets of modalities and tasks [Liang et al., 2021].

## 3 Learning or Memorizing?

Given large datasets, are pretrained models meaningfully learning or are they simply "nearest neighbor machines that can memorize big hash tables"? We may try adding perturbations to examine whether the model is learning the spirit of the tasks, such as adding a "not" in the front of VQA questions [Gokhale et al., 2020], blocking out the answer part of the image, asking the model to answer sub-questions, or other measures of robustness. Other ideas include generating a heatmap of the attention scores to visualize whether the model is focusing on relevant information [Selvaraju et al., 2017] or analyzing the multimodal embedding space with unsupervised techniques like clustering to gain further insight.

## 4 Beyond "Language" Models

Masked-language modeling seems to be very helpful while the visual analog (masked-region modeling) does not seem as helpful [Hendricks et al., 2021]. How can we represent images to work in conjunction with language models? Object detection allows abstractions for the visual domain, but it seems like it doesn't always help in practice. Can we develop universal models learned via multimodal pretraining? In Table 1, we list several recent attempts in this direction:

Table 1: Several recent attempts at universal models learned via multimodal pretraining.

| Paper | Modalities | Tasks Tested |
|---|---|---|
| UniVL [Luo et al., 2020] | video, text | text-based video retrieval, video captioning, action segmentation and step localization, sentiment analysis |
| VL-T5/VL-BART [Cho et al., 2021] | image, text | visual question answering, referring expression comprehension, visual commonsense reasoning |
| FLAVA [Singh et al., 2021] | image, text | vision tasks, NLP tasks (GLUE), VQAv2, SNLI-VE, Hateful Memes, Flickr30K, COCO retrieval |
| Perceiver [Jaegle et al., 2021] | text, images, sets | NLP, optical flow, image classification, StarCraft II |
| PolyViT [Likhosherstov et al., 2021] | video, image, audio | standard video- and audio-classification datasets |
| VATT [Akbari et al., 2021] | video, audio, text | video action recognition, audio event classification, image classification, text-to-video retrieval |
| UFO [Wang et al., 2021] | image, text | VQA, COCO image captioning and nocaps, image-text retrieval |
| data2vec [Baevski et al., 2022] | image, text, speech | speech recognition, image classification, natural language understanding |
| BLIP [Li et al., 2022] | image, text | image-text retrieval, image captioning, VQA |
| OFA [Wang et al., 2022] | image, text | image captioning, text-to-image generation, VQA, SNLI-VE, referring expression comprehension |

## 5 Cross-modal Interactions

In this section, we aim to further investigate the cross-modal interactions captured during pretraining. One way to analyze this is to take a pretrained Transformer model and visualize the type of interactions they capture [Li et al., 2020]. Another method that was proposed separately trained unimodal transformers before training a cross-modal transformer to learn cross-modal interactions while retaining unimodal information [Singh et al., 2021]. For future work, can we derive interpretable cross-modal interactions and compare them with mathematical, formulaic expressions (additive, multiplicative)? Can we design self-supervised tasks that reflect the specific cross-modal interactions we want the model to learn?

A related area of discussion is the role of co-learning on multimodal pretraining. Existing pretraining paradigms require a large bank of language data to learn representations. How much additional information from the image modality can we use to supplement or replace the information from text corpora? Future work can explore vision and language models that capture alignment between two modalities [Tsimpoukelli et al., 2021] and compare it with language-only pretraining paradigms on language-only tasks.

## 6 Future Directions

First, a promising direction to evaluate whether a pretrained model is really learning instead of memorizing is to start with building probing datasets for specific domains [Zadeh et al., 2019]. Second, as we shift towards a paradigm of developing large-scale pretrained models, more and more computational resources are required to train these models. In such a scenario, how can academia and industry work together?

## References

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 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.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer, 2020.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer, 2020.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does BERT with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.469. URL https://aclanthology.org/2020.acl-main.469.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv preprint arXiv:2111.12993*, 2021.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 2021.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access, 2020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

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.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. *arXiv preprint arXiv:2112.04482*, 2021.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34, 2021.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. Ufo: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022.

Tian Yun, Chen Sun, and Ellie Pavlick. Does vision-and-language pretraining improve lexical grounding? *arXiv preprint arXiv:2109.10246*, 2021.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.