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

Week 6: Multimodal LLMs 2: Fine-tuning, aligning, merging

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: Jiya Zhang*

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

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2024/

Summary: Multimodal machine learning (MMML) 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 6's discussion session, the class continued our focus on Multimodal Large Language Models (LLMs) and discussed the taxonomy of AI Alignment, utilizing frozen LLMs into multimodal setting, modality merging with Mixture-of-Expert (MoE), and fine-tuning techniques. The following was a list of provided research probes:

1. Ensuring the effectiveness of multimodal foundation models through high-quality instruction tuning is vital. A primary challenge in this approach is determining which data are most crucial for targeted instruction tuning. How can we accurately identify and select the most impactful data for enhancing instruction tuning in multimodal foundation models? Given the complexity of diverse and multimodal information, what strategies can ensure the effectiveness of instruction tuning data for specific tasks?
2. For model merging, mixture-of-expert-based models enable a new paradigm to utilize multiple expert models for specific tasks. When it comes to multimodal tasks, how might we design a similar system for multimodal tasks that have human-level intelligence? What methodologies could enable the integration of various multimodal models to perform complex tasks such as social interaction effectively?
3. What is the intuition of utilizing frozen large language models as the backbone for multimodal tasks? Which types of encoders would facilitate the integration of diverse information into a format understandable by LLMs? How do these LLMs process and interpret information from different modalities?
4. Considering the various methods available for LLM alignment, is aligning multimodal models perceived to be more challenging or easier? What factors contribute to the difficulty of multimodal alignment, and how might this be related to those previously discussed fundamental parts of multimodal machine learning like interaction and connection?
5. How can we categorize the taxonomy of general AI alignment? Can we classify the AI alignment categories based on the goal of conducting alignment? Assuming the existence of an oracle alignment method, what behaviors would we expect from an aligned AI model? Please list some behaviors that should be exhibited by AI following successful alignment.
6. What is the taxonomy of general AI alignment? Can we classify based on the goal of alignment? Imagine we have an oracle alignment method, what kind of behavior we expect the model to have after alignment? Please list some of the expected behavior that AI should have after alignment.
7. What distinguishes AI alignment from AI personalization? When focusing on AI alignment and personalization, what are the key differences and considerations to keep in mind?Is personalization an easier or harder thing to conduct compared with alignment?

As background, students read the following papers:

1. (Required) MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning [Xu et al., 2023]

2. (Required) Aligning Large Multimodal Models with Factually Augmented RLHF [Sun et al., 2023]
3. (Suggested) Case Law Grounding: Aligning Judgments of Humans and AI on Socially-Constructed Concepts [Chen and Zhang, 2023]
4. (Suggested) Aligning AI With Shared Human Values [Hendrycks et al., 2023]
5. (Suggested) Value Alignment for Advanced Artificial Judicial Intelligence [Winter et al., 2023]
6. (Suggested) Improved Baselines with Visual Instruction Tuning [Liu et al., 2023]
7. (Suggested) Multimodal Few-Shot Learning with Frozen Language Models [Tsimpoukelli et al., 2021]
8. (Suggested) BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models [Li et al., 2023]
9. (Suggested) MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning [Chen et al., 2023b]
10. (Suggested) An Empirical Study of Multimodal Model Merging [Sung et al., 2023]
11. (Suggested) π-Tuning: Transferring Multimodal Foundation Models with Optimal Multi-task Interpolation [Wu et al., 2023]

1 AI Alignment

AI alignment encompasses more than just technical alignment between modalities; it delves into who models should align with, what values and ethics they should adhere to, and the techniques and ultimate goal of alignment, such as promoting helpfulness and mitigating toxicity. This topic is crucial for AI safety, addressing existential risks and long-term threats to humanity.

1.1 Taxonomy Discussion on AI Alignment

Here's a table of AI alignment taxonomy and detailed explanations:

Category	Explanation	Details
WHO	Humanity, Organization, Individuals, Demographics	Alignment targets: universal human being[1]; specific companies or organizations; personalized alignment; considering demographic differences (age, culture, etc.).
WHAT	Moral values, Ethics, Helpfulness, Honesty, Harmlessness (Anthropoic HHH)	Undesirable behavior: biases, spreading misinformation, leaking privacy data, violating personal integrity [Weidinger et al., 2023].
HOW	Time-varying[2], Culture-dependent[3], Context-dependent	Considerations for cultural adaptation, generational differences, and language variation[4].
Techniques	Outer-alignment, Inner-alignment[5], RL-based (RLHF)	Describes approaches to align AI systems with human preferences; considerations for reward functions capturing human values, policy adherence, and interpretability.

Table 1: Taxonomy of AI Alignment

1. Aligning with humans involves two aspects: (1) universal alignment with human ethics and moral values, and (2) personalized alignment. These aspects can conflict when personal intentions diverge from societal norms. This is an interesting topic to look into, involving philosophical, policy, and ethical discussions.
    - Implementing hard-coded constraints and carefully crafted prompts during fine-tuning could mitigate harmful content generation in case of blatant violations. However, defining the severity threshold and enabling the model to grasp nuanced background and context—such as distinguishing

between a joke and a racist remark—requires further research.

2. Generational shifts and temporal changes significantly influence AI alignment. The present abundance of data far surpasses that of five or ten years ago, with a considerable portion being auto-generated, altering model training data. Additionally, linguistic habits evolve over time, impacting word usage. The Google Ngram viewer provides a fascinating tool for analyzing changes in word usage and popularity gleaned from books.
3. AI alignment must accommodate cultural norms and differences. In machine translation, individuals from diverse cultural backgrounds interpret the same words differently in terms of meaning or sentiment. Models require adaptation and localization for users, termed cross-cultural competence. Similarly, this challenge extends to images and multimodal settings. However, vision models often possess limited cultural perspectives, resorting to simplistic representations like flags or landmarks to denote culture or nationality. Achieving visual culture adaptation remains a challenging area with limited research conducted thus far.
4. The language used can influence the output of Large Language Models (LLMs); for instance, low-resource languages may yield less aligned outputs, posing higher security risks, as LLM fine-tuning is more extensive on English data. For instance, cross-lingual vulnerability is examined in [Yong et al., 2024].
5. The survey [Shen et al., 2023] elaborates on three concepts: Outer-alignment (defining a reward function that reflects human preferences), Inner-alignment (ensuring that a policy trained on the reward function aligns with human intent), and interpretability (the ability to reason the process from end-to-end).

Conversely, models can also influence human behavior and mindset. Well-aligned AI can encourage users to adopt more objective and inclusive attitudes towards different cultures, offering fresh perspectives. Rather than solely conforming to human habits, values, and context, models have the potential to shape our behavior, which presents both benefits and risks, potentially leading to legal ramifications.

2 Frozen LLMs as Backbone

Large language models are inherently auto-regressive, posing challenges for integrating them with image patches in vision tasks. Unlike large pre-trained CNNs, which were commonly used for image/video tasks, leveraging pre-trained LLMs in a multimodal context is challenging. What techniques enable the effective utilization of frozen pre-trained LLMs in a multimodal settings?

2.1 Autoregressive Models

The Diffusion model shows promise for processing vision and audio modalities in conjunction with LLMs. Techniques for integrating them include

- Converting visual images into sequences of discrete tokens, such as with the ViT-VQGAN model [Yu et al., 2022], and feeding them into auto-regressive transformers. This approach is advantageous for large-scale data and models, as frozen LLMs might have already learned representations of images before generating the first token.
- Leveraging agentic LLMs, like the LangChain Paradigm, where LLMs act as multiple agents or experts making decisions based on encoded modalities. However, scalability may be limited.
- Combining multiple models, where interface design is crucial. For example, an LLM produces a text-based vector output for a diffusion image generator, but obtaining stable diffusion representations for each image in the dataset can be challenging, particularly at large scale.
- Considering Text Diffusion models, e.g. [Chen et al., 2023a], with minor adjustments, given the discrete nature of text. However, estimating the length of the diffusion model output can be challenging. This technique is beneficial for control or fast generation tasks.

2.2 Non-autoregressive Models

High-performing non-auto-regressive LLMs offer an alternative approach. One example is the Fill-in-the-Middle (FIM) or Casual Masking model [Bavarian et al., 2022], which operates in a sliding-window fashion. Within each window, the model receives input and signals on masked or missing text. After traversing the

entire document, the model predicts and fills in the missing parts. The ordering, number, and size of the windows are randomly selected. FIM is effective for training Codepilot and GPT 3.5 models.

3 Mixture-of-Expert-Based (MoE) Models

To integrate various multimodal models or perform complex tasks like social interactions, techniques such as Mixture-of-Experts (MoE) can be employed. For instance, fine-tuning multimodal models to detect specific interaction characteristics, such as pose or social cues, and combining them as experts in a mixture-of-experts fashion. Utilizing a general dataset, rather than task-specific data, helps mitigate limitations and enables a weakly-supervised approach where model experts are combined on a prediction level using late fusion. This flexible approach allows for differences in each model's structure and facilitates addressing biases or limited data. Additionally, to mitigate bias, a model trained on biased features can serve as a biased expert, with subsequent training aimed at discarding or penalizing its biased predictions to debias the overall model. MoE proves effective when sufficient domain knowledge of tasks, data, or modality properties is available.

4 Fine-tuning / Instruction Tuning

To ensure the effectiveness of instruction tuning in handling the complexity of diverse multimodal information, different approaches are explored in recent research. For instance, Flan-T5 [Longpre et al., 2023] experimented with zero-shot, few-shot, and chain-of-thought (CoT) settings across over 1,000 tasks, while MultiInstruct [Xu et al., 2023] focused on 62 diverse downstream tasks with zero-shot and very few instructions for each task. The variation in both the number of tasks and instructions reflects the limited availability of multimodal data compared to pure text. While MultiInstruct demonstrates that even a few high-quality instructions can yield good performance, there is still room for exploring larger sets of instructions.

Another approach involves generating instructions using pre-trained LLMs and filtering them with Reinforcement Learning from Human Feedback (RLHF). This method can produce more instructions and undergo human review with hallucination control. Alternatively, a potential variation is to replace the reward model with Proximal Policy Optimization (PPO) during model fine-tuning on traditional datasets, thus circumventing human review as a bottleneck.

References

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle, 2022.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters, 2023a.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023b.

Quan Ze Chen and Amy X. Zhang. Case law grounding: Aligning judgments of humans and ai on socially-constructed concepts, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2023.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning, 2023.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. Large language model alignment: A survey, 2023.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023.

Yi-Lin Sung, Linjie Li, Kevin Lin, Zhe Gan, Mohit Bansal, and Lijuan Wang. An empirical study of multimodal model merging, 2023.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models, 2021.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative ai systems, 2023.

Christoph Winter, Nicholas Hollman, and David Manheim. Value alignment for advanced artificial judicial intelligence. *Am. Philos. Q.*, 60(2):187–203, April 2023.

Chengyue Wu, Teng Wang, Yixiao Ge, Zeyu Lu, Ruisong Zhou, Ying Shan, and Ping Luo. π-tuning: Transferring multimodal foundation models with optimal multi-task interpolation, 2023.

Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning, 2023.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak gpt-4, 2024.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022.