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*

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/spring2024/

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 6's discussion session, the class is focusing on Multimodal Large Langauge Models (LLM) on AI alignment, Fine-tuning with Instruction Tuning, and model merging with Mixture-of-Expert (MoE). 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?

The following is our discussion around the seven probes during the class.

1 Alignment

The Alignment we are referring to is AI alignment, which can be a generic topic instead of only focusing on the technical aspects of modality alignment. We discussed who the model should align to, what aspect should align (e.g. human moral values, ethnics), what's the ultimate alignment target to achieve (e.g. helpfulness, avoiding toxicity), and how to reach alignment with multimodal language based techniques (e.g. RLHF). AI Alignment is an important topic on AI safety, regarding existential risk and long-term threats to humanity as well.

1.1 Taxonomy Discussion on AI Alignment

Here's a table of potential AI alignment taxonomy derived from the discussion below:

Table 1: Taxonomy of AI Alignment

Category	Details
WHO	Humanity, Organization, Individuals, Different demographics (Age group, Cultural, etc.)
HOW	Time-varying, culture-dependent, context-dependent
WHAT	Moral values; Ethnics; Helpfulness, Honesty, Harmlessness (Anthropoic HHH)
Techniques / evalution	Outer-alignment, Inner-alignment, RL-based, etc.

- The DeepMind paper [Weidinger et al., 2023] provides a taxonomy of harm on generative AI systems, which we can apply as an aspect on AI alignment to define whay behavior an aligned system should or shouldn't perform. For example, the expectation on an aligned system is not having social biases, not spreading misinformation, not leaking privacy data or violating personal integrity, etc.
- Human is one of the most important target that AI systems should align with. There are at least two aspects should be included: (1) universal-level alignment on human ethnics and moral values, (2) personalized alignment. These aspects can be contradictory some time, when the personalized intent is against / violating the company's, organization's, or the society's perspectives. It's an interesting topic to look into, which involves philosophical, policy, and ethical issues when this kind situation happens.
  - Hard-coded rules or constraints and well-formulated prompts in fine-tuning might be approaches to avoid harmful content output when obvious violation happens, while how to define the threshold on severity and how to let the model understand the background and context, e.g. a joke v.s. a racist remark, needs further discussion.
- They type of language may also have different output from LLMs, e.g low-resource language can get a less aligned, highest security risk output as LLM's fine-tuning seems to be done on English data. For example, Cross-lingual vulnerability is evaluated in [Yong et al., 2024].
- AI Alignment should be sensitive to culture norms or culture differences. For example, in the Machine Translation (MT) area, it could be a problem that the output is universal while people from all culture backgrounds or demographic groups may not perceive the same meaning. For example, human in different culture treats *argument* differently. One group thinks the word to be peace and pure discussion, while other cultures treat it as conflict. So, besides translation, the model needs to consider the adaptation and localization for the users, known as a concept of cross-cultural competence. Also, the similar situation can extend to images and multimodal settings. However, vision models tend to preserve only a narrow concept of culture, e.g. adding flags in the image to represent culture or nations. So, this is a harder area and there's not much current research on culture adaptation in images.
- Similar to culture differences, generations or time varying has impact on AI alignment as well. Current data amount is significantly larger than 5, 10 years ago and a large portion data are auto-generated, which could have great impact on model training data. The habit on usage of words also changes overtime. There's a interesting analysis tool, Google Ngram viewer, shows how word usage and popularity changes from books.
- On the next level, models can impact human behavior and mindset as well. Good AI alignment might

suggest people to be more objective and inclusive on different cultures and bring new viewpoints for huamn beings. Instead of purely adapting to human habits, values, and context, the model can influence us in certain way, which can be beneficial and risky at the same time as legal problems might involve.

- AI alignment taxonomy introduced in the survey [Shen et al., 2023] includes Outer-alignment (specifying an reward function which captures human preferences, e.g. human values like HHH), Inner-alignment (ensuring that a policy trained on that reward function actually tries to act in accordance with human preferences), and interpretability (being able to reason the process from end-to-end). How do we apply this LLM/general AI alignment into a multimodal setting?

2 Frozen LLMs as Backbone

Large language models are auto-regressive in nature and some are in seq-to-seq structure, which may not work well for image patches on vision. What are the techniques to make better performance with frozen pre-trained LLMs in multimodal setting?

2.1 Diffusion Models

Diffusion model is very powerful for vision modalities (image, video) and even audio (text-to-audio), which can be a potential model to consider on integrating with LLMs. For tasks like text to other modalities generation, there are two ways we can do today:

- We can tokenize input modality as discrete tokens and feed into auto-regressive transformers. This might be better when the data and model size is at large scale, as the frozen LLMs might already have an representation of who the image would look like before a first token of image is generated. We can also leverage agentic LLMs, e.g. LangChain Paradigm, which LLMs are serving as multiple agents or experts to make decisions based on a holistic view after text, image, or any other modality is encoded with its own encoder, instead of projecting modalities into a frozen Language space. Though, this approach has limitations in terms of scalability.
- Another way is to combine multiple models together, in which the interface between those models matters a lot in terms of scalability. For example, a LLM produces a vector-like output for a text and feed into a diffusion image generator. However, this approach requires to get stable diffusion representation of every image in the data set and properly cache, which is hard for large scale. In this case, the previous approach is easier, which is a trade-off between inductive biases and scalability.
- Text Diffusion models could also be considered with small changes since text is discrete. The disadvantages in a rough estimation on the length of diffusion model output is needed. This technique is good for control generation or fast generation.

2.2 Non-autoregressive Models

Needs further research on if people use non-regressive frozen LLMs as backbone on multimodal setting, but well-performed non-auto-regressive LLMs could also be considered as an alternative. One example is called Fill-in-the-Middle (FIM) or Casual Masking [Bavarian et al., 2022], which uses a sliding-window fashion: within each window the model can get input and signals on masked/missing text; after sliding through the entire document, the model starts to predict and fill-in the missing parts. The ordering, number, and size of the window is randomly chosen. The FIM is objective for training Codepilot and GPT 3.5 models.

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

What are the techniques or architectures that can integrate various multimodal models, or even perform complex tasks like social interactions? Mixture-of-Expert is one of the approaches. For example, we can consider fine-tuning on multimodal models to detect particular characteristics of interactions, e.g. pose or social interaction tasks. Then, combine those models as experts for each of those characteristics in a mixture-of-experts fashion. We may consider using a general dataset, not limited to dataset for specific downstream tasks, to avoid limitations on a limited dataset. This can be a weak-supervised approach and just combine model experts on a prediction level, allowing the use of late fusion and difference between each model's structure. This flexibility and generalized approach enables MoE to facilitate issues when there's biased or limited data. In order to avoid bias, it's also a possible approach that a specific model is trained on biased features as a biased expert. Then, train the model to throw away the biased expert or penalizing on the expert's choice, thus de-bias the entire model. The MoE will be powerful is one have enough domain knowledge on the tasks, data, or modality properties.

4 Fine-tuning / Instruction Tuning

Given the complexity of diverse multimodal information, how can we ensure the effectiveness of instruction tuning? The Flan-T5 [Longpre et al., 2023] tried zero-shot, few-shot, and chain-of-thought (CoT) as mixed prompt setting with more than 1k tasks, while the MultiInstruct [Xu et al., 2023] uses only 62 diverse downstream tasks with zero-shot and very few instructions for each task. There's a big difference in terms of both the number of tasks and the number of instructions between the two papers. This is because multimodal data are much more limited compare to pure text. Even though few high-quality instructions could result in good performance in multimodal tasks from MultiInstruct, it's still a further direction to have larger number of instructions.

Another approach to generate instructions, without too much limitation on high-quality human experts, is from pre-training LLMs and then filter with RLHF. This approach can generate more instructions and also reviewed by human with hallucination control. A potential variation is to replace reward model with Proximal Policy Optimization (PPO) when fine-tuning the model in a traditional data set, to avoid human review as the 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.

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.

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.

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.