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

Week 14: Choose Your Own Topic

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

Edited by Paul Liang *Scribes: Jiya Zhang, Anwesa Bhattacharya*

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 14's discussion session, the class was split into two groups, each introducing a self-selected topic discussing understudied or impactful multimodal machine learning topics not covered in previous discussions. The two topics were (1) beyond text-centric multimodal machine learning, and (2) interpretability of multimodal machine learning. The following is a list of research directions proposed by the students:

1. What key questions/aspects must be addressed to empower non-textual modalities to become powerful language models?
2. How do computational resources differ from textual modalities, and how can we optimize the computational resource problem with non-textual data?
3. Pros/Cons of using non-textual modalities, e.g., vision understanding, instead of text? In what applications are vision understanding models more suited than LLMs?
4. What pre-training objectives are most suitable to train foundational LVMs? Also, what kind of downstream tasks can be addressed by simple fine-tuning of these LVMs?
5. What is the prompt-tuning equivalent in LVMs?
6. How does incorporating compositional reasoning into vision-language models impact their interpretability across modalities?
7. What are the essential criteria for evaluating compositional reasoning benchmarks to enhance interpretability in multimodal models?
8. Besides visualization methods like attention and relevance maps, what alternative strategies exist for enhancing interpretability in multimodal models? Can these approaches effectively capture interactions between different modalities?
9. To what extent does a mixture of experts approach contribute to the development of more interpretable multimodal models?
10. Following the identification of task irrelevant input features through interpretability techniques, what strategies can be employed to refine multimodal models and focus on pertinent features for prediction?
11. What are the feasible methodologies for evaluating the effectiveness and reliability of interpretability approaches in multimodal models?

As background, students picked their own related papers to read and discuss.

1. Here's the list of papers for beyond text-centric multimodal machine learning:
   (a) Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction [Tian et al., 2024]
   (b) Rejuvenating image-GPT as Strong Visual Representation Learners [Ren et al., 2023]
   (c) Sequential Modeling Enables Scalable Learning for Large Vision Models [Bai et al., 2023]
   (d) Video as the New Language for Real-World Decision Making [Yang et al., 2024]
   (e) Finding Visual Task Vectors [Hojel et al., 2024]

   (f) Scalable Pre-training of Large Autoregressive Image Models [El-Nouby et al., 2024]
2. Here's the list of papers for interpretability of multimodal machine learning:
   (a) CREPE: Can Vision-Language Foundation Models Reason Compositionally? [Ma et al., 2023]
   (b) SUGARCREPE: Fixing Hackable Benchmarks for Vision-Language Compositionality [Hsieh et al., 2023]
   (c) DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations [Lyu et al., 2022]
   (d) LVLM-Intrepret: An Interpretability Tool for Large Vision-Language Models [Stan et al., 2024]
   (e) Scaling Vision-Language Models with Sparse Mixture of Experts [Shen et al., 2023]
   (f) MMOE: Mixture of Multimodal Interaction Experts [Yu et al., 2023]

The following is a summary of the presentation and class discussions.

1 Beyond Text-centric Mutlimodal Learning

Text-centric approaches possess dominance in multimodal researches, particularly with the rise of Large Language Models (LLMs). While text-based models excel in reasoning and abstract tasks, they may not always capture the nuances of the physical world effectively. Human cognition often operates beyond textual constraints, relying on various modalities for perception and understanding.

In a multimodal context, emphasizing text as the primary modality within large models can indeed present challenges such as information loss and difficulties in aligning different modalities effectively. This can hinder the overall performance of multimodal systems, especially when dealing with non-textual data.

To address these limitations and explore new avenues for multimodal research, it's essential to investigate recent advancements in non-textual foundational models like Large Vision Models (LVMs), Large Audio Models (LAMs), or equivalents based on physiological or environmental data. These models have the potential to outperform LLMs in multimodal tasks by leveraging the strengths of different modalities and offering more comprehensive representations of real-world phenomena.

By shifting the focus towards non-textual foundational models and exploring their effectiveness in multimodal tasks, researchers can unlock new opportunities for advancing multimodal machine learning and bridging the gap between AI systems and real-world understanding.

1.1 Interesting Recent Work Highlights

Here're are two of the six interesting works presented from the students:

1. Sequential Modeling Enables Scalable Learning for Large Vision Models [Bai et al., 2023]
   - This paper introduces a novel approach to training a Large Vision Model (LVM) by defining visual sentences to represent individual or sequences of images and visual annotations (without using text). The model is trained by minimizing cross-entropy for next token prediction, similar to Large Language Models (LLMs). Additionally, this work presents the Unified Vision Dataset (UVD), a purely visual dataset comprising single images, image sequences, images with annotations, and image sequences with annotations, which can be used to train LVMs.
   - Furthermore, this work demonstrates that LVMs can generalize and combine multiple tasks effectively. The approach outperforms Visual Prompting in nearly all tasks and exhibits the ability to grasp abstract visual patterns and extrapolate from them.
2. Video as the New Language for Real-World Decision Making [Yang et al., 2024]
   - The paper proposes an intriguing idea that video generation is to the physical world as language modeling is to the digital world. To substantiate this standpoint, the paper compares the characteristics of video generation with three key components of language generation. Below is the table showing the comparison details:

Language	Video
A Unified Representation (i.e. Text)	Unified Representation of Information, e.g. Visual and Spatial Information, Physics and Dynamics, Behavior and Action Information.
A Unified Interface (i.e. Text Generation)	Unified Task Interface, e.g. CV Tasks as Video Generation, Video as Answers, Video Reasoning.
Interaction with External Environments	Video Generation as Simulation

1.2 Works Beyond Visual

Besides textual modalities, recent works have been primarily focusing on vision (images, videos), such as the emergence of Large Vision Models (LVMs) as a prominent direction. However, there are other modalities, such as audio or sensor data, that hold potential for leveraging large models.

An interesting example is the Contrastive Audio-Visual Masked Autoencoder (CAV-MAE) proposed by Gong et al. [Gong et al., 2023], which extends the Masked Auto-Encoder (MAE) model from a single modality to audio-visual multimodalities. This model combines contrastive learning and masked data modeling approaches to learn meaningful representations from unlabeled audio and visual data, without relying on text.

Applying large foundation models to all modalities can pose challenges. Primarily, the lack of large, annotated multimodal datasets, especially for sensor data, presents a significant hurdle for developing robust multimodal models. Additionally, efficiently handling the high dimensionality and computational demands of sensor data, particularly for real-time applications like robotics, is another challenge in developing efficient large foundation models.

1.3 Combine LLMs with LVMs vs Directly Use VLMs

If both Large Language Models (LLMs) and Large Vision Models (LVMs) are powerful, the decision to use a combination of LLMs and LVMs versus a Vision-Language model (VLM) depends on several factors, including the specific task, input and output modalities, and data distribution.

VLMs are suitable for a broader range of multimodal tasks involving both vision and language, such as visual question answering, image captioning, and grounded language understanding. Their architecture may facilitate better fusion and alignment between vision and language modalities, thereby enhancing modality interactions. On the other hand, combining LLMs with LVMs might be preferable when one modality dominates, and optimal performance of a single-modality model is desired.

An intriguing question arises regarding the design of a proper combined model architecture. This architecture should enable the model to seamlessly process various modalities of data and produce diverse modalities according to the downstream tasks.

2 Interpretability of Multimodal Learning

Interpretability in multimodal models is pivotal for understanding how different modalities such as text, image, and audio contribute to predictions, especially in complex tasks like image captioning or video understanding.

One significant challenge lies in deciphering how these modalities interact and influence each other during the prediction process. Unlike unimodal models, where interpretability may focus solely on one modality, multimodal connections and interactions add layers of complexity. For instance, in a task like image-text matching, understanding how visual features align with textual semantics demands capturing intricate cross-modal relationships. This interplay introduces challenges in interpreting which modalities or features are crucial for decision-making, necessitating novel interpretability techniques capable of unraveling these intricate connections. Additionally, the dynamic nature of multimodal interactions requires interpretable models to not only provide insights into individual modalities but also elucidate how these modalities synergize to form coherent predictions. Thus, achieving interpretability in multimodal models necessitates a nuanced understanding of the intermodal dynamics inherent to the task at hand.

It also prompts exploration into designing architectures that emulate human-like interpretability in reasoning processes. Architectures like Socratic models, leveraging compositional reasoning, offer insights into how linguistic and visual elements combine to make predictions, enhancing interpretability by mimicking human-like reasoning processes. Similarly, Mixture of Experts models provide interpretability by decomposing complex tasks into simpler components, allowing for a clearer understanding of each modality's contribution to the final decision.

2.1 Interpretability Techniques

Interpretability techniques in multimodal models encompass a spectrum of approaches, ranging from visualization methods to adaptations of unimodal interpretability techniques like LIME. One notable work, the DIME paper [Lyu et al., 2022], focuses on disentangling multimodal models into their unimodal contributions and multimodal interactions. Similarly, human-interpretable explanations are generated using LIME, with each modality analyzed individually while holding the other constant, resulting in four fine-grained explanations delineating unimodal contributions and their influence on multimodal interaction. These techniques are evaluated across diverse tasks including CLEVR and VQA 2.0 using models such as LXMERT, MLP, and MDETR. Another noteworthy approach, LVLM-Interpret [Stan et al., 2024], facilitates interactive analysis for VLMs like LLaVA through layer attention, relevancy maps, and causal interpretation of attention. Notably, when the generated outputs exhibit a greater relevance to one modality, changes in the other modality do not significantly affect model accuracy.

2.2 Compositional Reasoning

Large vision-language models have showcased remarkable performance across various tasks like captioning and visual question answering (VQA); however, they encounter challenges with compositional reasoning. Benchmarks such as CREPE [Ma et al., 2023] assess this ability by evaluating systematicity and productivity. Sytematicity is evaluated by generalization to both seen and unseen compounds. To evaluate productivity, CREPE introduces nine complexities, each with three types of hard negatives. This assessment involves performing random walks on the scene graphs of an evaluation dataset to generate subgraphs of varying complexities. However, these benchmarks tend to exhibit significant biases, making them susceptible to manipulation, with blind models even outperforming state-of-the-art vision-language models. To address this vulnerability, SUGARCREPE [Hsieh et al., 2023] ntroduces a novel benchmark that leverages large language models to generate fluent hard negatives and employs an adversarial refinement mechanism to minimize biases. Unlike previous benchmarks relying on procedurally-generated hard negatives, SUGARCREPE aims to ensure logical and grammatically correct evaluations.

2.3 Mixture of Experts

Mixture of Experts (MoE) models inherently possess interpretability due to the specialized training of each expert for specific tasks. Additionally, MoE models offer scalability for large Vision-Language Models (VLMs), as demonstrated in the paper on Scaling Vision-Language Models with Sparse Mixture of Experts [El-Nouby et al., 2024], where token routing decisions on COCO dataset are qualitatively analyzed. Vision tokens exhibit clear specialization, routed to specific experts such as food and vegetable experts or OCR experts, while language tokens demonstrate syntax specialization, with experts processing different linguistic components. Moreover, MoE models excel in capturing various multimodal interactions crucial for interpretability, as highlighted in the MMOE paper [Yu et al., 2023], which trains expert models for uniqueness, redundancy, and synergy interactions.

2.4 Future Directions

- Future research may focus on utilizing advanced Language and Vision Models (LLMs) to generate higher quality hard negative captions or images, as identified by limitations in benchmarks like CREPE and SUGARCREPE.

- Studies such as Interpretability in the Wild [Wang et al., 2022] have revealed distinct attention heads within models like GPT-2 for specific tasks, indicating specialized functionalities such as identifying indirect objects. Future inquiries could extend this exploration to Vision-and-Language Models (VLMs) to understand if attention heads can discern attributes like objects or colors.
- Recent work like MMoE designs experts to capture various multimodal interactions, offering insights into routing mechanisms. However, these approaches need label information to know of multimodal interactions. Future directions might involve designing experts that capture multimodal alignment based on data connections without relying on label information.
- Evaluation approaches for interpretability methods could include saliency-based methods, where humans analyze whether models identifies the right input features for predictions, offering insights into model interpretability.

References

Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models, 2023.

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models, 2024.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder, 2023.

Alberto Hojel, Yutong Bai, Trevor Darrell, Amir Globerson, and Amir Bar. Finding visual task vectors, 2024.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023.

Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations, 2022.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally?, 2023.

Sucheng Ren, Zeyu Wang, Hongru Zhu, Junfei Xiao, Alan Yuille, and Cihang Xie. Rejuvenating image-gpt as strong visual representation learners, 2023.

Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts, 2023.

Gabriela Ben Melech Stan, Raanan Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. Lvlm-intrepret: An interpretability tool for large vision-language models, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022.

Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making, 2024.

Haofei Yu, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Mmoe: Mixture of multimodal interaction experts, 2023.