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

Week 7: Multimodal LLMs 3: Generative Models

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: wjhan*

Edited by Paul Liang *Scribes: wjhan, skhanuja*

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

Summary: In this week's discussion we talk about generative models. More specifically, we talk about two main points: Advancements in recent generative models, and ethical considerations about the AI generated content.

1 Advancements in Generative Modeling

1.1 Use of Frozen Encoders

It was discussed that utilizing frozen encoders in multimodal models raises concerns about the learning of intricate interactions between modalities. Adapters can be employed in such scenarios to enable the capturing of necessary interactions by updating only a few layers, thus offering a balance between model stability and adaptability.

1.2 Latent Interactions and Generative Planning

Capturing latent interactions in generative models is challenging. A model should not solely focus on generation but also incorporate elements of planning. This integration can enhance the model's ability to understand and predict complex multimodal interactions.

1.3 Extending Masked Modeling

The extension of masked modeling to generative tasks was a point of interest. If masked modeling leads to a better understanding of data, exploring ways to transfer this knowledge to generative tasks could be beneficial. This approach might involve developing techniques that leverage the strengths of masked modeling in generative contexts. In the context of the V-JEPA model [Bardes et al., 2024], utilizing masked representations could lead to improved motion understanding and ensure spatial and temporal consistency within video frames. This notion of learning grounded representations via unsupervised learning was initially introduced with images by Assran et al. [2023]. This approach could enhance the model's ability to predict and generate realistic sequences.

1.4 Computational Resources vs. Architectural Innovations

A critical discussion point was whether to apply more computational resources or to focus on developing better loss functions and architectures. A consensus seemed to be that a mix of both computational power and architectural innovation is essential for advancing multimodal learning models.

2 Ethical Considerations and Safety

2.1 AI Generated Content Detection

The detection of AI-generated content is crucial for maintaining authenticity and trust. Strategies such as watermarking (with an emphasis on minimizing false positives) and ethical considerations (e.g., the need to label AI-generated content) are vital. The nuances of what constitutes AI-generated content were also discussed, especially in the context of text and images.

2.2 Misinformation and Bias

Issues such as misinformation, hallucinations, and copyright infringements were identified as significant challenges. Different modalities face varying difficulties in watermarking. The potential for AI to generate misleading content, such as manipulated speeches, was also noted.

2.3 Bias and Representation

The discussion acknowledged the presence of gender and racial biases in generative models, often stemming from the data used for training. The challenge of overfitting and mode collapse in generative models was highlighted. The group emphasized that as long as humans are involved in data labeling, unconscious biases will inevitably influence the training data. It was also noted that digital representations of humans do not always align with real-world diversity.

References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint*, 2024.