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

Week 5: Multimodal LLMs 1: Data, pretraining, scaling

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

Edited by Paul Liang *Scribes: wjhan, anwesab*

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

Summary: Pretraining robust, effective multimodal LLMs come with many different challenges. In this week's discussion, we discuss three aspects of the challenge of pretraining multimodal LLMs: 1) Data, 2) Architecture, and 3) Scalability.c

1 Noise in Multimodal Data

- Adversaries in multimodal data often introduce deliberate noise to mislead or confuse models. This noise can manifest as misleading information in one or more modalities, making it challenging for models to make accurate predictions.
- Filtering noise is complex due to the interconnected nature of modalities. In a multimodal setting, noise in one modality (like irrelevant background in images) can significantly impact the overall interpretation.
- **Taxonomy of Noise**:
  1. Adversarial noise: Deliberately manipulated data to fool models.
  2. Label noise: Incorrect labels, which can arise from misinterpretation or misrepresentation of data, including:
     - Sampling bias: A non-representative sample of the population leads to biased predictions.
- The definition of noise is task-dependent. Features considered as noise in one task might be critical in another, highlighting the importance of context in multimodal learning.
- Techniques like Active Learning and Curriculum Learning during pretraining can be employed to mitigate the impact of noise by focusing on the most informative data points and gradually increasing the complexity of the data.
- The complexity of the dataset dictates the model size: a noisy dataset requires a more complex model to accurately capture and distinguish relevant features.
- Conversely, a high-quality dataset with minimal noise allows for simpler models, reducing computational costs without sacrificing performance.
- Cultural and regional differences can lead to varying interpretations of the same data, resulting in different ground truths [Yun and Kim, 2024]. This diversity must be considered when training and evaluating multimodal models.
- The way datasets are labeled (prompt engineering) can significantly impact the quality and utility of the labels, affecting model performance.

2 Scalable, Generalizable, Multimodal Generation Architectures

2.1 Notable Architectures

- Transformer models have shown great success in scaling with data and model size, offering significant improvements in handling multimodal data due to their ability to capture long-range dependencies across different types of inputs.
- Mixture of Experts (MoE) models:
  1. Train experts for different types of datasets, allowing for specialized handling of distinct modalities or data types [Mustafa et al., 2022].

2. Discovering mixture of experts through 2-stage training involves training experts for specific tasks or modalities and then learning to optimally combine their outputs, leveraging their individual strengths.

- Model Merging integrates different specialized models (e.g., culturally specific models, domain-specific query models) to create a more robust and comprehensive understanding of multimodal data.
  - This approach allows for nuanced predictions that consider diverse perspectives and interpretations.

Table 1: Pros and Cons of Treating Data from all Modalities Equally.

Pros	**Cons**
Uniform tokenization across modalities can lead to lower perplexity, indicating better model understanding and generation.	Differentiating modality-specific features becomes challenging, potentially leading to loss of critical information.
Efficient processing across diverse modalities.	Fails to leverage the unique inductive biases inherent in different modalities.
	Increases the complexity and input size, demanding more computational resources.

3 Scaling Laws of Multimodal Models

- With sufficient data variety and volume, multimodal models typically show a reduction in loss, indicating improved learning and generalization.
- Modalities that are conceptually closer (like text and code) tend to have lower perplexity compared to more disparate modalities (like images and audio), suggesting that similar modalities are easier for models to learn and integrate.
- To challenge the current scaling laws, empirical evidence showing deviations from these trends in large-scale multimodal datasets would be necessary.
- The tokenization schema, which converts various modalities into a format understandable by the model, is critical. Effective schemas capture the essence of each modality while enabling integration with others.

References

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts, 2022.

Youngsik Yun and Jihie Kim. Cic: A framework for culturally-aware image captioning, 2024.