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

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

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: William Jongwon Han*

Edited by Paul Liang *Scribes: William Jongwon Han, Anwesa Bhattacharya*

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. The following was a list of provided research probes:

1. What types of multimodal data noise are typically present in multimodal datasets, and how can they negatively impact the performance of a model during training? Can you provide examples of multimodal data points that might be considered noisy? Furthermore, how might we develop estimators capable of distinguishing between noisy and noise-free multimodal data pairs? If you have unlimited fundings to use for data filtering and data cleaning, what would be the ideal way to clean the multimodal dataset?
2. Given the demonstrated effectiveness of high-quality pretraining, as evidenced by projects like Mistral, imagine you have access to a large-scale, high-quality multimodal dataset for pre-training purposes. What types of generalization or additional capabilities might this enable the model to acquire compared to those trained on lower-quality data? Why do models trained with high-quality data obtain such abilities?
3. Considering the diversity of model architectures available for multimodal generation, which architecture would be most suitable for scaling general multimodal generation tasks? Moreover, which model architecture is best equipped to learn complex multimodal interactions effectively?
4. What are some pros and cons of treating data from all modalities equally (throwing them into a single large generative Transformer, after tokenizing the data)?
5. If you were leading a multimodal foundation model project equipped with extensive resources, including a skilled team and significant GPU capabilities, what multimodal architecture and types of multimodal data would you prioritize for an initial pilot study?
6. In exploring the scaling laws of multimodal models, different papers have different definitions for scaling law formula. Which factors should be incorporated into the scaling law formula, and which among these do you believe is the most critical to consider?

As a background for the mentioned content, students were required and suggested to read the following papers:

1. **(Required)** ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision [Kim et al., 2021]
2. **(Required)** Scaling Laws for Generative Mixed-Modal Language Models [Aghajanyan et al., 2023]
3. (Suggested) Scaling Multimodal Pre-Training via Cross-Modality Gradient Harmonization [Wu et al., 2022]
4. (Suggested) Scaling Laws for Autoregressive Generative Modeling [Henighan et al., 2020]
5. (Suggested) Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens [Liu et al., 2024]
6. (Suggested) CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation [Tang et al., 2023]
7. (Suggested) Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP [Nguyen et al., 2023b]

8. (Suggested) Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved with Text [Zhu et al., 2023]
9. (Suggested) Improving Multimodal Datasets with Image Captioning [Nguyen et al., 2023a]

1 Noise in Multimodal Data

Noise in multimodal data can be particularly challenging due to the interconnected nature of modalities. Adversaries may introduce deliberate noise to mislead or confuse models, which can manifest as misleading information in one or more modalities, making it difficult for models to make accurate predictions. The complexity of the dataset dictates the model size, with a noisy dataset requiring a more complex model to accurately capture and distinguish relevant features. On the other hand, a high-quality dataset with minimal noise allows for simpler models, reducing computational costs without sacrificing performance.

In a multimodal setting, noise in one modality, such as irrelevant background in images, can significantly impact the overall interpretation. The definition of noise is task-dependent, and 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.

Moreover, cultural and regional differences can lead to varying interpretations of the same data, resulting in different ground truths. This diversity must be considered when training and evaluating multimodal models. Additionally, the way datasets are labeled (prompt engineering) can significantly impact the quality and utility of the labels, affecting model performance.

To address these challenges, it is essential to have a clear understanding of the characteristics of the data and the task at hand. By recognizing the different types of noise present in the data, such as adversarial noise, label noise, and sampling bias, we can develop strategies to mitigate their impact. For instance, techniques like data augmentation, transfer learning, and ensembling can be used to improve the robustness of models to noise.

In summary, noise in multimodal data presents unique challenges that must be addressed through a combination of techniques that take into account the interconnected nature of modalities, cultural and regional differences, and the task-dependent definition of noise. By developing a deeper understanding of these factors, we can create more accurate and robust multimodal models that can effectively handle noisy data.

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

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models, 2023.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling, 2020.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens, 2024.

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

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning, 2023a.

Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. Quality not quantity: On the interaction between dataset design and robustness of clip, 2023b.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context, interleaved, and interactive any-to-any generation, 2023.

Junru Wu, Yi Liang, Feng Han, Hassan Akbari, Zhangyang Wang, and Cong Yu. Scaling multimodal pre-training via cross-modality gradient harmonization, 2022.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text, 2023.