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

Week 2: Dimensions of Heterogeneity

Instructors: Paul Liang and Daniel Fried *Synopsis Leads: Ashwin Pillay*

Edited by Paul Liang *Scribes: Ashwin Pillay, Anwesa Bhattacharya*

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 2's discussion session, the class aimed to formalize a taxonomy of cross-modal interactions: various ways in which elements from different modalities can relate with each other and the types of new information possibly discovered as a result of these relationships. The following was a list of provided research probes:

1. What is a taxonomy of the dimensions in which modalities can be heterogeneous? What are intuitive definitions of each dimension of heterogeneity?
2. Heterogeneity is also often seen in several other ML subfields (e.g., domain adaptation, domain shift, transfer learning, multitask learning, federated learning, etc). What are some similarities and differences between the notions of heterogeneity between MMML and these fields? Can definitions or methods in each area be adapted to benefit other research areas?
3. How can we formalize these dimensions of heterogeneity, and subsequently estimate these measures to quantify the degree in which modalities are different?
4. Heterogeneity in noise (e.g., due to sensor and system failures) is a relative understudied dimension. How can we reliably understand the unique noise topologies in modalities, to design more robust models?
5. Modality heterogeneity often implies the design of specialized models capturing the unique properties of each modality. What are some trade-offs in modality-specific vs modality-general models?
6. Within each of the 6 multimodal challenges - representation, alignment, reasoning, generation, transference, quantification, how can the study of heterogeneity inform various modeling decisions? What problems could happen in practice if heterogeneity is not properly understood or modeled?

As background, students read the following papers:

1. (Required) Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges [Bronstein et al., 2021]
2. (Suggested) Taskonomy: Disentangling Task Transfer Learning [Zamir et al., 2018]
3. (Suggested) Which Tasks Should Be Learned Together in Multi-task Learning? [Standley et al., 2020]
4. (Suggested) Geometric Dataset Distances via Optimal Transport [Alvarez-Melis and Fusi, 2020]
5. (Suggested) AutoFed: Heterogeneity-Aware Federated Multimodal Learning for Robust Autonomous Driving [Zheng et al., 2023]
6. (Suggested) Natural Image Noise Dataset [Brummer and De Vleeschouwer, 2019]
7. (Suggested) Synthetic and Natural Noise Both Break Neural Machine Translation [Belinkov and Bisk, 2018]

We summarize several main takeaway messages from group discussions below:

1 Taxonomy of Heterogeneity

In Table 1, we summarize some dimensions of heterogeneity discussed during the class. An interesting observation made during this discussion was the existence of heterogeneity even within the modality; In the case of text, heterogeneity could arise from the language used; for example, English has a phonetic writing system while Mandarin is logographic [Monroy]. Hence, it would be beneficial in exploring taxonomies where even intra-modality heterogeneity could be accommodated.

Table 1: Dimensions of Heterogeneity between Modalities

Dimension	Explanation
Source	Heterogeneity can be introduced by how the modality is captured, such as from the difference in sensors being used to capture images (infra-red v/s RGB). In case of language, different language idiosyncrasies add another dimension of heterogeneity and introduce the challenge of incorporating these priors into modelling.
Element Representations	Heterogeneity could arise from the atomic representation of the modality. This can be words and grammatical structure for language, and objects and spatial structure for images.
Structure	Structural/compositional heterogeneity can exist in modalities on properties like their symmetry, such as translational equivariance for images, and temporal invariance for sentiments.

2 Addressing Multimodal Challenges: Representation & Alignment

A significant task in addressing many of the six major multimodal challenges is concerned with identifying and accounting for the heterogeneity in the modalities involved. However, the constituent heterogeneity could also be task-specific (for example, while both speech-to-emotion classification and speech transcription tasks involve the audio and text modalities, the former has temporal-explicit heterogeneity but the latter might not). One possible solution towards addressing the task-specific heterogeneity among the concerned modalities may involve extracting features having similar *structures* [Bronstein et al., 2021] in each of these modalities and aligning them by standard techniques like concatenation (perhaps, with separator-tokens to represent modality boundaries). We discuss relevant techniques as follows:

2.1 Identifying Similar Structures among the Modalities

The first step towards addressing heterogeneity involves identifying similar structures among the available modalities; some of the approaches we discuss in this regard include:

1. Analyzing known-symmetries in other modalities that are most similar to the modality of our interest.
2. Using a combination of modality-specific and modality-agnostic encoders, and a domain-adaptation classifier to determine which encoders must be used for the provided input, as described in [Lu et al., 2022].

2.2 Extracting Similar Structures among the Modalities

The task of extracting features having common structures from each modality could be largely classified as applying the corresponding inductive bias during the feature extraction process; we discuss several ways in which this could be achieved:

1. **Theoretically-motivated**:
   - The Geometric Deep Learning paper [Bronstein et al., 2021] applies techniques from the Erlangen program towards identifying the levels of symmetry that is preserved by commonly used neural architectures. CNNs, for example, generate features with translation-equivariance.

- The concepts discussed in this work provide a strong theoretical basis towards selecting (and possibly developing new) neural architectures that can address the heterogeneity in multi-modal data.

2. **Empirically-motivated**:
   - Many state-of-the-art multimodal systems employ transformer-based architectures that are known to have a high-degree of generalizability. Their success, combined with the complexities involved in the aforementioned theoretical approaches empowers an argument that it might be more practical to empirically extract the features that provide the best results for the multimodal task at hand (for example, the basis of selecting sinusoidal positional encoders employed in [Vaswani et al., 2017] was based on empirical results than theoretical validations).
   - Another supporting point here is that such empirically-derived methods could already have theoretical groundings that are not formally studied during their conception (for example, the greater success of diffusion models [Ho et al., 2020] in vision experiments than for text could be ascertained to the higher robustness of vision data to noise). In this case, we expect a capable model to learn such structures directly from the data without needing any inductive biases.
   - We also discuss models pre-trained on other tasks and modalities, showing positive results when transferred to the target multimodal task; for example, [Papadimitriou and Jurafsky, 2020] observe that training on MIDI data or Java code improved test performance on NLP tasks. This could be seen as an empirical form of applying inductive bias where a general model has been configured to focus on certain structures in the target data as a result of pre-training.

3 Measuring Heterogeneity

1. One of the approaches for quantifying heterogeneity could be to mathematically prove which layers focus on learning from what modalities. This has foundations in the idea of Geometric Deep Learning, which shows that certain kinds of architectures are better suited to exploit symmetries in certain modalities.
2. The idea of optimal transport (OT) can help in estimating how similar one modality is to another. VoLTA [Pramanick et al., 2023] uses graph optimal transport for patch token alignment. This gives rise to the idea of using OT between transformer layers on the projections as an attempt to quantify heterogeneity. At the same time, it's necessary to ensure that these comparisons are done at network layers that share symmetries (for example, avoiding comparison in layers where one modality is sequential while the other is explicit).
3. Another possible solution would be to attempt alignment of modalities. Since negative examples cause the representations to spread out in the space, the CLIP training objective [Radford et al., 2021] could be used without negative examples. Subsequently, OT could be used to quantify how closely the various modality embeddings map to each other.

4 Adapting Strategies from other ML Subfields

In the field of federated learning, heterogeneity refers to differences in devices, particularly in terms of parameters like computation power, network settings etc. One of the approaches [Eichner et al., 2019] adaptively chooses between global and device specific models to handle cyclic patterns in data samples. Parallels could be drawn between this approach and the idea of adaptively using modality general and modality specific encoders in MMML.

References

David Alvarez-Melis and Nicolò Fusi. Geometric dataset distances via optimal transport, 2020.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation, 2018.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.

Benoit Brummer and Christophe De Vleeschouwer. Natural image noise dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Hubert Eichner, Tomer Koren, H. Brendan McMahan, Nathan Srebro, and Kunal Talwar. Semi-cyclic stochastic gradient descent, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks, 2022.

Marco Monroy. Chinese alphabet: Why it doesn't exist | A useful language guide. URL https://www.berlitz.com/blog/chinese-alphabet.

Isabel Papadimitriou and Dan Jurafsky. Learning music helps you read: Using transfer to study linguistic structure in language models, 2020.

Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik Shah, Yann LeCun, and Rama Chellappa. Volta: Vision-language transformer with weakly-supervised local-feature alignment, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning?, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Amir Zamir, Alexander Sax, William Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning, 2018.

Tianyue Zheng, Ang Li, Zhe Chen, Hongbo Wang, and Jun Luo. Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving, 2023.