11-877 Advanced Multimodal Machine Learning **Spring 2023**

Week 2: Dimensions of Heterogeneity

Instructors: Louis-Philippe Morency and Paul Liang *Synopsis Leads: Mehul Agarwal, Yiqing Xie*

Edited by Paul Liang *Scribes: Gaoussou Kebe, Santiago Benoit*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

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 brainstormed about the various dimensions of heterogeneity commonly encountered in multimodal ML research. The following was a list of provided research probes:

1. What is a taxonomy of the dimensions in which modalities can be heterogeneous?
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. Modality heterogeneity often implies the design of specialized models capturing the unique properties of each modality. What are some tradeoffs in modality-specific vs modality-general models?
5. What are other modeling considerations that ideally should be informed based on how heterogeneous the input modalities are?
6. What are some risks if we were to ignore modality or task heterogeneity accurately? What if we are unable to estimate modality or task heterogeneity accurately?

As background, students read the following papers:

1. (Required) Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges [Bronstein et al., 2021]
2. (Suggested) A Survey on Heterogeneous Transfer Learning [Day and Khoshgoftaar, 2017]
3. (Suggested) Taskonomy: Disentangling Task Transfer Learning [Zamir et al., 2018]
4. (Suggested) Which Tasks Should Be Learned Together in Multi-task Learning? [Standley et al., 2020]
5. (Suggested) Federated Learning: Challenges, Methods, and Future Directions [Li et al., 2020]
6. (Suggested) Domain Adaptation under Target and Conditional Shift [Zhang et al., 2013]
7. (Suggested) Detecting and Correcting for Label Shift with Black Box Predictors [Lipton et al., 2018]
8. (Suggested) Geometric Dataset Distances via Optimal Transport [Alvarez-Melis and Fusi, 2020]
9. (Suggested) HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Representation Learning [Liang et al., 2022]

We summarize several main takeaway messages from group discussions below:

1 Dimensions of heterogeneity: A taxonomy

1.1 A list of dimensions

Table 1 summarizes the dimensions of heterogeneity discussed in class. We first categorize these dimensions into different granularities: *sensor* studies the nature of devices used to capture these modalities and their subsequent differences, *structure* looks at properties of the individual atoms and their compositionality into global modalities, *feature* studies properties of the model or feature space used to process these modalities, and finally *information* identifies global differences between modalities.

Dimension	Explanation
Sensor	
Source device	Heterogeneity can come in the form of different specialized sensors used to capture raw modalities, such as different sensor equipment, collection environments, sampling rates, time-scales, how raw data is stored and retrieved from files, and data storage formats.
Perceptuality	A modality can be perceptual (e.g., text, image) or non-perceptual (e.g., graph, file system) to humans as a result of different sensors that capture them.
Structure	
Vocabulary/atoms	Heterogeneity in the set of basic atoms (vocabulary) comprising a modality, which can be discrete or continuous and come from different base distributions.
Structure	Heterogeneity in how basic atoms are composed to form global information, which can span spatial, temporal/sequential, hierarchical, graphical, and set-based compositions.
Invariances	When composed at a global level, there lie different invariant transformations that preserve meaning, such as spatial invariance for images and permutation invariance for sets and graphs.
Feature	
Statistics	When representing modalities as features, how do the features differ in their sample space and statistics?
Distribution	Distribution heterogeneity refers to the differences in frequencies and likelihoods of features, such as different frequencies in recorded signals and the density of tokens.
Distance	Different modalities naturally exhibit different similarity metrics to judge similarity between instances.
Information	
Content	Different amounts of information are contained in different modalities, and they can be unique, overlapping, or identical in the context of other modalities.
Density	The frequency of information can be different (e.g., low vs high-frequency sequential data) and potentially over different ranges (e.g., short vs long-term temporal relationships).
Noise	Noise can be introduced at several levels across naturally occurring data and during the data recording process. Noise heterogeneity measures differences in noise distributions across modalities, as well as differences in signal-to-noise ratio.

Table 1: A list of dimensions in which modalities can be heterogeneous.

1.2 Several taxonomic organizations

We can also group these fine-grained dimensions of heterogeneity into different broader categories.

Inter-modality vs intra-modality heterogeneity: An example of inter-modality heterogeneity is the difference between sensors with distinct modalities (e.g., speech vs vision). In comparison, intra-modalities arise from the same sensor, but with varying qualities or settings, such as a smartphone camera versus DSLR for visual inputs, or differing microphone quality levels, etc.

Dataset-level vs instance-level heterogeneity: Furthermore, even within the same modality, different datasets are still heterogeneous. In the visual domain, a dataset of dog photos and dog paintings have different information source. Similarly, instances from the same modality or even the same dataset may also be heterogeneous. For example, "black dog sitting on the grass" and "dog" could be two captions describing the same image, but they still have different information density.

Data vs feature heterogeneity: While some dimensions are dependent only on how the data is naturally

expressed and collected, other dimensions such as feature statistics, feature distance, and information content require one to additionally understand the feature spaces of these modalities.

Task-independent vs task-dependent heterogeneity: Finally, each modality can show different relevance towards certain tasks and contexts - certain modalities may be more useful for certain tasks than others. Heterogeneity should therefore be studied in task and context-dependent ways as well.

2 How can measuring heterogeneity help modeling?

2.1 Modality-specific vs general models

Measuring the heterogeneity between modalities is an important factor when deciding between modality-specific and general models. Modality-specific models may be more useful for particular tasks as they capture the inherent similarities for one modality better, which allows them to make better predictions or decisions. Transformers are often used for NLP while convolutional neural nets (e.g., ConvNext [Liu et al., 2022]) provide advantages when dealing with vision based tasks due to their higher capacity for invariance resulting from convolutional layers capturing visual information better, and could be more suitable for very heterogeneous modalities. However, there are also benefits of general models [highmmt,transferlearning], where it was shown that a model trained on one task can be used as a starting point for a different task for suitably homogeneous modalities. Early fusion or alignment techniques [Barnum et al., 2020] can prove helpful here too. At scale, it is also important to consider which type of model performs better given the data size and network preference; CNNs often provide advantages over Transformers when dealing with small datasets yet have poorer performance on large ones [Liu et al., 2021].

2.2 Noise and robustness

Each modality has a unique noise topology, which determines the distribution of noise and imperfections that it commonly encounters. For example, images are susceptible to blurs and shifts, typed text is susceptible to typos following keyboard positions, and multimodal time-series data is susceptible to correlated imperfections across synchronized time steps. Understanding the heterogeneity in these imperfections can enable more accurate benchmarking in real-world settings and methods that are more robust to noisy or missing modalities.

References

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

George Barnum, Sabera Talukder, and Yisong Yue. On the benefits of early fusion in multimodal representation learning. *arXiv preprint arXiv:2011.07191*, 2020.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Oscar Day and Taghi M Khoshgoftaar. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4: 1–42, 2017.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Ruslan Salakhutdinov. Highmmt: Towards modality and task generalization for high-modality representation learning. *arXiv preprint arXiv:2203.01311*, 2022.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai. Efficient training of visual transformers with small datasets. *Advances in Neural Information Processing Systems*, 34:23818–23830, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. PMLR, 2013.