**11-877 Advanced Multimodal Machine Learning** **Spring 2022**

# Week 10: Beyond Language and Vision

*Instructors: L.-P. Morency, A. Zadeh, P. Liang* *Synopsis Leads: Alissa Ostapenko, Alex Wilf*

*Edited by Paul Liang* *Scribes: Alex Wilf and Arav Agarwal*

**Disclaimer**: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.* Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/

**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 10's session, students discussed learning problems involving many heterogeneous modalities, including information coming from sensor data. They discussed challenges in fusion, scalability, and model evaluation. They considered which existing techniques that work well for 2-3 modalities, especially ones pertaining to language or visual input, may fail or succeed when dealing with heterogeneous data from many (10+) modalities. The following list of research probes was provided:

1. What are the modalities beyond language and vision that are important for real-world applications? What unique structure do they contain, and what are the main challenges in performing multimodal learning with them?
2. When reflecting on the heterogeneous aspect of multimodal learning, how are the other modalities different from language, speech, and vision? What dimensions of heterogeneity are important for these other modalities?
3. What are the cross-modal interactions that you expect in these other modalities? Could you see ways to model cross-modal interactions with these other modalities and with language and vision?
4. How do the core research problems of unimodal and multimodal processing, integration, alignment, translation, and co-learning generalize to modalities beyond language and vision? What core insights from these 'common' modalities have yet to be explored in understudied modalities?
5. What is the best way to visualize these relatively understudied modalities? How can we best analyze and characterize the multimodal interactions present between these other modalities?
6. How to learn models for many modalities (10+ modalities)? What are the chances to create multimodal learning algorithms that work for all modalities? What are the tradeoffs between modality-specific multimodal models and general-purpose multimodal models?
7. If two modalities are very far from each other (strong heterogeneity and/or encoding very different information), how can we address the problem of multimodal learning?

As background, students read the following papers:

1. (Required) A comprehensive survey on multimodal medical signals fusion for smart healthcare systems [Muhammad et al., 2021]
2. (Required) Multimodal Sensor Fusion with Differentiable Filters [Lee et al., 2020a]
3. (Suggested) Integration of EEG/MEG with MRI and fMRI [Liu et al., 2006]
4. (Suggested) A Multi-Sensor Fusion System for Moving Object Detection and Tracking in Urban Driving Environments [Cho et al., 2014]
5. (Suggested) Detect, Reject, Correct: Crossmodal Compensation of Corrupted Sensors [Lee et al., 2021]
6. (Suggested) Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks [Lee et al., 2020b]
7. (Suggested) Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges

[Gravina et al., 2017]

8. (Suggested) Multi-Sensor Fusion: A Perspective [Hackett and Shah, 1990]
9. (Suggested) MultiBench: Multiscale Benchmarks for Multimodal Representation Learning [Liang et al., 2021]
10. (Suggested) HighMMT: Towards Modality and Task Generalization for High-Modality Representation Learning [Liang et al., 2022]
11. (Suggested) Sensor Fusion for Mobile Robot Navigation [Kam et al., 1997]
12. (Suggested) Multi-source information fusion based on rough set theory: A review [Zhang et al., 2021]
13. (Suggested) Combining EEG and fMRI: A Multimodal Tool for Epilepsy Research [Gotman et al., 2006]
14. (Suggested) Kalman Filter, Sensor Fusion, and Constrained Regression: Equivalences and Insights [Jahja et al., 2019]

We summarize several main takeaway messages from group discussions below:

# 1 Challenges in learning from heterogeneous data

## 1.1 Heterogeneity and representation

Some modalities, such as brain scans, are analogous to language and vision. For example, brain scans are still images so we can use the same techniques (such as Convolution Neural Networks) [Bernal et al., 2019]. In addition, Lee et al. [2020b] use WaveNet [van den Oord et al., 2016] for processing data from forces and torque, even though WaveNet is originally used for audio data. This presents the possibility of using the same processing technique despite the different data input (force/torque vs audio) so that we can adopt prior techniques for new data types.

Other data structures may require additional finetuning before they can be processed. For example, Electronic Health Records (EHRs) have language information, so we can use similar techniques as for other language processing tasks. However, we would want to fine-tune them to process domain-specific vocabulary and numerical inputs [Li et al., 2021]. Other data sources like point cloud representations [Qi et al., 2017] have both real and complex domains, which could pose challenges.

## 1.2 Fusion

Fusion poses additional challenges because different data sources have different structural properties: for example, EHR data is typically in a tabular format which is difficult to fuse with other data structures such as images. To address this, one could use a heterogeneous graph that encodes the information sources into a common structures [Hu et al., 2020]. Different data sources can also have different scaling techniques and may not standardize in the same way, posing alignment challenges. Moreover, with new modality types, there may be new types of cross-modal interactions beyond the typical additive [Hessel and Lee, 2020] and multiplicative [Jayakumar et al., 2020] interactions we expect between vision and language. Thus, there may not be any general way to fuse modalities. Until alignment is solved, the latent spaces we perform fusion on may not contain sufficient cross-modal representations for a general fusion algorithm to be effective.

Moreover, we can fuse modalities at different points in the network, opening another research problem. Early fusion may create large, unstructured data representations. Late fusion allows the model to learn some high-level structures of the modalities which could be lost through early fusion. This may increase computational cost, revealing a trade-off between complexity and location of modality fusion.

## 1.3 Modality Bias

Aside from data structure differences, there is a risk that representations will be biased towards a handful of the modalities. For example, if we have a better model for one modality (e.g. a large pretrained transformer for text [Devlin et al., 2018] and low level statistical signals for acoustic [Degottex et al., 2014]) it may bias learning to rely much more on the better pretrained text transformer. This problem may also be pronounced when low-resource modalities are present, as their representation spaces will be less expressive than their large pretrained counterparts. Thus, we may have to integrate some modalities into the model at different points during training or weight modalities differently to prevent one or a few modalities from dominating. We could draw from information theory to determine how to weigh modalities. For example, we can measure information flow from one modality to another and up-weight the modality that carries more information than another one. On the decision level, we also have to use different techniques to address different confidence scores from each modality, such as in the Bayesian Filtering paper [Lee et al., 2020a]. We have to figure out which modality is most useful or trustworthy in a given situation.

### 1.4 Pre-training

Pre-training for heterogeneous data is difficult when there is limited data. In healthcare, for example, data is limited and hard to access, so it's difficult to build the same large, pre-trained models we use for language and vision. Also, data is collected for very specific tasks (i.e., classifying one type of disease or finding one particular robotic target), so it may be hard to build a pre-trained model that can generalize to multiple tasks. Other pairwise techniques, like contrastive learning [Le-Khac et al., 2020], may not scale to many modalities. Contrastive loss is typically pairwise, and would result in many error terms, risking the model ignoring some or many of the terms, and therefore, the model may not learn cross-modal interactions between certain modalities. Co-learning may also be difficult because the number of modalities can vary; some modalities may be missing and the model must recover from that. A recent model, HighMMT, may address some issues we discussed, including recovering from missing modalities and integrating different data structures into a single model in the context of multi-task learning and partially observable modalities [Liang et al., 2022].

### 1.5 Low-Resource Settings

There are also some unique challenges to consider when learning from low-resource modalities. There may not be enough data for effective model learning; thus, we could incorporate inductive biases for our models to better fit domain specific data. This "top down" approach relies on human expertise instead of creating algorithms that can learn patterns from the data alone, but it can fail in cases where inductive biases cannot be readily defined. However, some approaches also infuse inductive biases in a "gated" way that maintains bottom-up data learning as well [Silvestri et al., 2021]. Model architectures could also be designed to adapt knowledge from similar (but higher resource) tasks to the low-resource task. In this direction, domain adaptation of models may be helpful [Bousmalis et al., 2016, Ajakan et al., 2014, Gideon et al., 2019].

Instead of relying on model transfer, one could also map the style of low-resource data to the style of high-resource tasks, then use the models trained on the high-resource task to finetune on the low-resource data. A fundamental challenge in this approach would be how to transfer the style of the data without problems such as mode collapse [Creswell et al., 2018, Lin et al., 2018] and other issues that undermine the effectiveness of style-transfer, particularly in non-interpretable latent spaces. Another solution could be to use large pretrained models to help gather more data within the low-resource domain, thus expanding the dataset while minimizing the loss in authenticity from the original domain space.

## 2 Developing a pipeline for scaling to many modalities

All of the above challenges considered, it is useful to develop a systematic way for approaching problems using heterogeneous data from diverse and understudied modalities. One example workflow could be:

1. Start with unimodal data and filter out anomalies.
2. Scale data and transform it (for example, standardizing sampling rate differences) to enable better alignments.
3. Create a simple baseline by concatenating inputs from all modalities.
4. Look for unimodal or multimodal pre-trained models and decide whether to use a pre-trained model.
5. Analyze interactions between modalities and filter ones as necessary.
6. Determine evaluation (discussed more in Section 3).

For a baseline model, we could simply concatenate different modality representations to see the performance of the model without cross-modal interactions, then add more complex interactions afterwards, depending on the baseline performance. The memory and parameter complexity would scale linearly with the number of modalities, but concatenation may miss out on important feature combinations that a method which scales exponentially might capture (e.g. Tensor Fusion [Zadeh et al., 2017]).

Going beyond the baseline (step 3), we could filter out modalities using different criteria, reducing the complexity of our input space. Aggregators at different stages [Yang et al., 2020] or clustering techniques [Müllner, 2011] could combine modalities and thus reduce loss terms in objectives like the contrastive loss. Another option is to discard modalities that are not useful for the task. For high-dimensional data, we can use techniques like manifold learning and Principal Component Analysis (PCA) to reduce the dimensionality of data. Moreover, we could run experiments to determine where in the network to perform early or late fusion, and train models to automatically find an optimal set of modalities out of the original input set.

The quality of fused representations heavily depends on the quality of unimodal feature extractors, which are mostly trained in a self-supervised fashion. We could analyze the self-supervised losses each unimodal feature extractor is trained on and understand how those signals influence each other. For example, perhaps a word-masking approach [Devlin et al., 2018] is the best way to train language models, but by masking frames in a higher frequency signal such as audio [Hsu et al., 2021], the pretrained model we create may not generate as cohesive a latent space when combined with the text model. For this reason, it is important to look into end-to-end multimodal self-supervised learning and downstream signals to learn coherent representations.

## 3 Measuring the quality of heterogeneous multimodal spaces

One interesting open question is how to determine the quality of multimodal latent spaces. This is a difficult problem to solve with real world data, as the properties of alignment and co-learning are both difficult to define and costly to annotate. Designing synthetic datasets and tasks will be helpful to clarify what is meant by co-learning and alignment, test these hypotheses in an empirical way, and transfer understanding to real-world data. One core challenge is to understand how modalities can be similar or different, and how that affects co-learning and alignment.

In considering how to evaluate multimodal spaces, we could follow different schemes, including evaluating on the downstream task, performing multitask evaluation, and evaluating representations through retrieval or reconstruction tasks. For cases in which there are multiple "correct" outputs, we could use contrastive learning to teach the model the similarities and differences between plausible and implausible outputs. A clear taxonomy of the kinds of evaluation metrics and approaches we define for joint multimodal spaces will be important to unlocking insights about the next generation of multimodal models beyond language and vision.

## References

Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

José Bernal, Kaisar Kushibar, Daniel S. Asfaw, Sergi Valverde, Arnau Oliver, Robert Martí, and Xavier Lladó. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. *Artificial intelligence in medicine*, 95:64–81, 2019.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.

Hyunggi Cho, Young-Woo Seo, BVK Vijaya Kumar, and Ragunathan Raj Rajkumar. A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1836–1843. IEEE, 2014.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

John Gideon, Melvin G McInnis, and Emily Mower Provost. Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog). *IEEE Transactions on Affective Computing*, 12(4):1055–1068, 2019.

Jean Gotman, Eliane Kobayashi, Andrew P Bagshaw, Christian-G Bénar, and François Dubeau. Combining eeg and fmri: a multimodal tool for epilepsy research. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 23(6):906–920, 2006.

Raffaele Gravina, Parastoo Alinia, Hassan Ghasemzadeh, and Giancarlo Fortino. Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges. *Information Fusion*, 35:68–80, 2017.

Jay K Hackett and Mubarak Shah. Multi-sensor fusion: a perspective. In *Proceedings., IEEE International Conference on Robotics and Automation*, pages 1324–1330. IEEE, 1990.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.62. URL https://aclanthology.org/2020.emnlp-main.62.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. *Proceedings of The Web Conference 2020*, 2020.

Maria Jahja, David Farrow, Roni Rosenfeld, and Ryan J Tibshirani. Kalman filter, sensor fusion, and constrained regression: equivalences and insights. *Advances in Neural Information Processing Systems*, 32, 2019.

Siddhant M. Jayakumar, Jacob Menick, Wojciech M. Czarnecki, Jonathan Schwarz, Jack W. Rae, Simon Osindero, Yee Whye Teh, Tim Harley, and Razvan Pascanu. Multiplicative interactions and where to find them. In *ICLR*, 2020.

Moshe Kam, Xiaoxun Zhu, and Paul Kalata. Sensor fusion for mobile robot navigation. *Proceedings of the IEEE*, 85(1):108–119, 1997.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

Michelle A Lee, Brent Yi, Roberto Martín-Martín, Silvio Savarese, and Jeannette Bohg. Multimodal sensor fusion with differentiable filters. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10444–10451. IEEE, 2020a.

Michelle A Lee, Yuke Zhu, Peter Zachares, Matthew Tan, Krishnan Srinivasan, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. *IEEE Transactions on Robotics*, 36(3):582–596, 2020b.

Michelle A Lee, Matthew Tan, Yuke Zhu, and Jeannette Bohg. Detect, reject, correct: Crossmodal compensation of corrupted sensors. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 909–916. IEEE, 2021.

Irene Li, Jessica Pan, Jeremy Goldwasser, Neha Verma, Wai Pan Wong, Muhammed Yavuz Nuzumlali, Benjamin Rosand, Yixin Li, Matthew Zhang, David Chang, Richard Andrew Taylor, Harlan M. Krumholz, and Dragomir Radev. Neural natural language processing for unstructured data in electronic health records: a review. *ArXiv*, abs/2107.02975, 2021.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

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

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. *Advances in neural information processing systems*, 31, 2018.

Zhongming Liu, Lei Ding, and Bin He. Integration of eeg/meg with mri and fmri. *IEEE engineering in medicine and biology magazine*, 25(4):46–53, 2006.

Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, Mansour Alsulaiman, and Tiago H. Falk. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, 76:355–375, 2021. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2021.06.007. URL https://www.sciencedirect.com/science/article/pii/S1566253521001330.

Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

C. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017.

Gianluigi Silvestri, Emily Fertig, Dave Moore, and Luca Ambrogioni. Embedded-model flows: Combining the inductive biases of model-free deep learning and explicit probabilistic modeling. *arXiv preprint arXiv:2110.06021*, 2021.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *ArXiv*, abs/1609.03499, 2016.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. Mtgat: Multimodal temporal graph attention networks for unaligned human multimodal language sequences. *ArXiv*, abs/2010.11985, 2020.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017.

Pengfei Zhang, Tianrui Li, Guoqiang Wang, Chuan Luo, Hongmei Chen, Junbo Zhang, Dexian Wang, and Zeng Yu. Multi-source information fusion based on rough set theory: A review. *Information Fusion*, 68:85–117, 2021.