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

Week 3: Multimodal Co-Learning

Instructors: L.-P. Morency, A. Zadeh, P. Liang *Synopsis Leads: Arav Agarwal, Alex Kwark*

Edited by Paul Liang *Scribes: Amelia Kuang, Yun Cheng*

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. In week 3's discussion session, the class discussed and compared several ways to achieve multimodal co-learning, the phenomenon of transferring information learned through one (or more) modality to tasks involving another. The following was a list of provided research probes:

1. What are the types of cross-modal interactions involved to enable such co-learning scenarios where multimodal training ends up generalizing to unimodal testing?
2. What are some design decisions (inductive bias) that could be made to promote the transfer of information from one modality to another?
3. How do we ensure that during co-learning, only useful information is transferred, and not some undesirable bias? This may become a bigger issue in low-resource settings.
4. How can we know if co-learning has succeeded or failed? What approaches could we develop to visualize and probe the success of co-learning?
5. How can we formally, empirically, or intuitively measure the additional information provided by auxiliary modality? How can we design controlled experiments to test these hypotheses?
6. What are the advantages and drawbacks of information transfer during co-learning? Consider not just prediction performance, but also tradeoffs with complexity, interpretability, fairness, etc.

As background, students read the following papers:

1. (Required) Multimodal Prototypical Networks for Few-shot Learning [Pahde et al., 2021]
2. (Required) SMIL: Multimodal Learning with Severely Missing Modality [Ma et al., 2021]
3. (Suggested) Multimodal Co-learning: Challenges, Applications with Datasets, Recent Advances and Future Directions [Rahate et al., 2022]
4. (Suggested) Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision [Tan and Bansal, 2020]
5. (Suggested) What Makes Multi-modal Learning Better than Single (Provably) [Huang et al., 2021]
6. (Suggested) Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities [Pham et al., 2019]
7. (Suggested) Zero-Shot Learning Through Cross-Modal Transfer [Socher et al., 2013]
8. (Suggested) 12-in-1: Multi-Task Vision and Language Representation Learning [Lu et al., 2020]
9. (Suggested) A Survey of Reinforcement Learning Informed by Natural Language [Luketina et al., 2019]

We summarize several main takeaway messages from group discussions below:

1 A Taxonomy for Co-learning

Co-Learning aims to transfer information learned through one (or more) modality to tasks involving another. Typically, this involves adding external modalities during the training process, learning a joint representation space, and investigating how the joint model transfers to unimodal tasks during testing. Common examples

include using word embeddings for image classification [Socher et al., 2013], knowledge graphs for image classification [Marino et al., 2017], or video data for text classification [Zadeh et al., 2020].

Co-learning is important since it enables us to improve unimodal systems through incorporating external data, which is particularly useful for low-resource target tasks. Ideas from co-learning towards learning a joint multimodal representation space can also provide independent insights for other multimodal tasks as well. Co-learning methods can be categorized in the following categories based on the type of cross-modal interactions modeled during training, each with its pros and cons:

1.1 Enrichment

Enrichment approaches involve enriching the representation space of unimodal models with additional modalities as input. The representation space combining both modalities can offer more information and structure as compared to prior unimodal representation spaces. Some classic examples in this category include training a joint video-based multimodal model which is then transferred to text-only classification [Zadeh et al., 2020] and integrating knowledge graphs to structure the representation space for image classification [Marino et al., 2017].

Pros: The main benefits include easier tuning and iteration (as it can be easier to make a discriminative pipeline than a generative one), alongside enabling more fine-grained design of the interactions between modalities in their associated latent space. For example, if we know that we have a certain number of examples to train a unimodal predictor, and augment that data with another modality, we can force the training process to prioritize additive interactions rather than multiplicative interactions, which might create easier-to-understand changes in the latent space.

Cons: On the other hand, this type of co-learning is harder to determine and understand. There have been observed cases where training on complete multimodal information still leads to strong unimodal classifiers, despite the model not learning how to handle a dropped modality in training [Zadeh et al., 2020]. It is hard to tell why these cases work, as for some models there is nothing enforcing that the model "knows" that 0 implies a dropped piece of data.

1.2 Translation

Another category involves translating unimodal data into another modality or latent space, essentially learning a joint multimodal space through "hallucination". This approach explicitly forces the latent space to not just handle a single modality, but potentially recreate another with limited training data. Classic examples in this category include Vokenization which maps contextualized text embeddings into images [Tan and Bansal, 2020], projecting image embeddings into semantic word embedding spaces [Socher et al., 2013], and translating language into hallucinated video and audio modalities [Pham et al., 2019].

Pros: These approaches are typically more flexible and easy to design since all the user needs to define is a reconstruction loss on the modality's input space. Furthermore, it is easier to visualize what a latent space "knows", as we can visualize the reconstructed modality or measure reconstruction losses to determine the amount of information contained in the latent space.

Cons: From the literature on GANs and other generative models, it is known that such models suffer from sample inefficiency since it is often challenging to reconstruct high-dimensional data. This may result in the hallucination of false correspondences between the limited data so it may be challenging to tell if reconstruction is a good idea.

2 Assessing Strength of Co-Learning Methods

Despite the recent success of co-learning, analyzing the strength of co-learning and determining when they exist remains a challenge. There are several possible ways to better understand how these interactions unfold and how they yield different outcomes for their respective unimodal problems:

2.1 Data Curation

One way to better understand the strength of these approaches is through better data curation to study how much multimodal data is needed to generate a better unimodal model compared to a similar amount of unimodal data.

To do so, one could imagine creating datasets where there are pre-defined "trends" or false associations that could be learned in the data, that a model would have to correct by using data from another modality that is not present in test-time. For example, in some language-classification problems, one might try putting in questions that compare the relative sizes of objects, and add false noise making the sizes of objects indistinguishable from each other. By augmenting the dataset with corresponding image data, one could investigate modeling approaches to enable co-learning from both image and text.

Besides creating new datasets, further work in co-learning will require a better understanding of the degree of data parallelism in the co-learning process. For example, in Vokenization [Tan and Bansal, 2020], data parallelism enables models to learn fine-grained one-to-one relationships between contextualized words and images. Other approaches, such as Pham et al. [2019], only provide aligned data at a more global level. It is worth investigating the impact of the resolution of aligned data on co-learning performance.

2.2 Model-based Assessment

Another way to analyze co-learning could be through comparing differences between the latent spaces of models before and after co-learning. For example, if one could see if co-learning representation spaces become more separable with respect to latent concept classes.

One central challenge lies in measuring the variations in a latent space, and how to reduce the dimensionality of the latent in a way that preserves these variations while not hurting interpretability. Given that current techniques like t-SNE are frequently mis-read [Wattenberg et al., 2016], such an analysis would require careful analysis of other high-dimensionality reduction techniques, like UMAP [McInnes et al., 2020] to see if they are suitable for this particular type of comparative analysis. Alternatively, one could try to use a clustering technique or a gradient-based prompting approach [Gao et al., 2021] to better ground the latent spaces together, but in general, the question becomes that of measuring the differences between two latent spaces of different models.

3 Unsupervised Co-Learning

Besides looking at co-learning as a supervised learning problem, it was discussed how we could try to use co-learning to create better unsupervised learning methods. From efforts like Data2Vec [Baevski et al., 2022] and other multimodal pre-training paradigms, learning a joint representation space is more powerful than unimodal representation spaces. Future work can also study self-supervised co-learning where pretraining is performed in large-scale diverse multimodal tasks [Liang et al., 2021] before transferring knowledge to a subset of multimodal or unimodal tasks.

4 Neuroscience and Modality Strength

The class also discussed understanding the importance of multimodal co-learning through the lens of neuroscience. From studies looking at aphantasia, a disorder where people who are unable to visualize what is going on in their head do worse on memory tasks [Zeman et al., 2015, Jacobs et al., 2018], we can draw insights that the ability to take a latent space and convert it between modalities is an important aspect of human cognition and memory. Conversely, it appears that the more one is able to convert and visualize their memory, the easier it is to memorize, as evidenced by memory-improving techniques like the memory palace.

References

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *URL https://ai. facebook. com/research/data2veca-general-framework-for-self-supervi sed-learning-in-speech-vision-and-la nguage/. Accessed*, pages 01–27, 2022.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners, 2021.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34, 2021.

Christianne Jacobs, Dietrich S Schwarzkopf, and Juha Silvanto. Visual working memory performance in aphantasia. *Cortex*, 105:61–73, 2018.

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.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A survey of reinforcement learning informed by natural language. In *IJCAI*, 2019.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2302–2310, 2021.

Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20–28. IEEE, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2644–2653, 2021.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899, 2019.

Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.

Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, 2020.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016. doi: 10.23915/distill.00002. URL http://distill.pub/2016/misread-tsne.

Amir Zadeh, Paul Pu Liang, and Louis-Philippe Morency. Foundations of multimodal co-learning. *Information Fusion*, 64:188–193, 2020.

Adam Z Zeman, Michaela Dewar, and Sergio Della Sala. Lives without imagery-congenital aphantasia. 2015.