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

Week 9: Multimodal Co-learning

Instructors: Louis-Philippe Morency and Paul Liang *Synopsis Leads: Ryan Liu, Aditya Rathod*

Edited by Paul Liang *Scribes: Jingyi Zhang, Santiago Benoit*

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

Summary: Multimodal reasoning is a very relevant research direction in today's world, looking at how models can reason about the world in an accurate and interpretable manner.

In week 9's discussion session, the class aimed to understand **multimodal co-learning** and recent works around it. The following was a list of provided research probes:

1. We define co-learning broadly as multimodal data and training helping performance on unimodal tasks. Under what scenarios will co-learning occur? Why is that research has demonstrated both positive and negative results? What assumptions do we have to make on the heterogeneity of data sources and the nature of connections and interactions between modalities for co-learning to be successful?
2. How can we formally, empirically, or intuitively measure the additional information provided by auxiliary modalities for co-learning? How can we design controlled experiments to test these hypotheses?
3. What are some design decisions (modeling, training, objective functions) that could be made to promote co-learning from one modality to another? What is a taxonomy of approaches and their pros and cons?
4. Text is usually the modality used for additional supervision. Why is text such a popular choice? Can other modalities also be used for additional supervision, and how would co-learning methods work differently?
5. How do we measure what information is transferred during co-learning? How do we ensure that only useful information is transferred, and not some undesirable bias or shortcuts?
6. 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, beyond target task performance?
7. What are the advantages and drawbacks of information transfer during co-learning? Consider not just prediction performance, but also tradeoffs with increased complexity, interpretability, biases, etc. Can we come up with a guideline for when we should use co-learning, when the benefits outweigh the additional costs?

As background, students read the following papers (REQ = required, SUG = suggested, REL = relevant):

1. (REQ) Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision [Jia et al., 2021]
2. (REQ) Does Vision-and-Language Pretraining Improve Lexical Grounding? [Yun et al., 2021]
3. (SUG) Multimodal Co-learning: Challenges, Applications with Datasets, Recent Advances and Future Directions [Rahate et al., 2022]
4. (SUG) Cross-Modal Data Programming Enables Rapid Medical Machine Learning [Dunnmon et al., 2019]
5. (SUG) Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision [Tan and Bansal, 2020]
6. (SUG) Grounding 'Grounding' in NLP [Chandu et al., 2021]
7. (REL) A Survey of Reinforcement Learning Informed by Natural Language [Luketina et al., 2019]

8. (REL) Analyzing the Effectiveness and Applicability of Co-training [Nigam and Ghani, 2000]
9. (REL) Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering [Lu et al., 2022]
10. (REL) An Information Theoretic Framework for Multi-view Learning [Sridharan and Kakade, 2008]
11. (REL) Experience Grounds Language [Bisk et al., 2020]

We summarize several main takeaway messages from group discussions below:

1 Definition and motivation of co-learning

Co-learning is an innovative technique that involves leveraging the power of multiple modalities to improve the performance of a model on a given task. The central idea behind this approach is to use cross-modal information to enhance the performance of a model beyond its unimodal baseline. Co-learning can be viewed as a form of multi-task learning, where a model is trained to perform several related tasks simultaneously. This approach allows knowledge from one domain to be transferred to another domain, leading to improved performance across all tasks [Liang et al., 2022, Rahate et al., 2022].

One of the key motivations for co-learning is to increase the amount of data available for training. By using multi-modal data, it is possible to obtain weakly supervised labels, which can help to reduce the amount of annotation required for training. This, in turn, can make it easier and more cost-effective to train models on large-scale datasets.

Another advantage of co-learning is that it can be used to enrich the representations of the modalities being used. By leveraging the unique information present in each modality, it is possible to create a richer and more informative representation of the data. This can lead to better performance on the task at hand, as the model is better able to capture the underlying patterns in the data.

Co-learning can be achieved in multiple ways: First, a pipeline could focus on the shared mutual information between modalities, and increase the accuracy or confidence of existing predictions. Another approach is to change the interpretation of the initial modality, perhaps making it easier to understand through a filter. Furthermore, there are also possibilities of emergent information that is in neither individual modality (e.g. bit-wise XOR). Our discussions from previous weeks on interactions between modalities are relevant here, including special cases such as temporal and negative interaction.

One example for better performance is in image captioning: co-learning can be used to improve the performance of a model by incorporating information from both images and text [Jia et al., 2021]. By using cross-modal information, the model can learn to generate more accurate and informative captions that capture the key elements of the image. In speech recognition, co-learning can be used to improve the accuracy of the model by incorporating information from both audio and text. By using cross-modal information, the model can learn to recognize speech more accurately, even in noisy environments.

1.1 Relation to multimodal learning

When humans learn, we learn by combining all modalities, and use it to perform both unimodal and multimodal tasks. Thus, from our own perspectives, multimodal learning, where we learn to solve multimodal tasks, and co-learning, where we learn to solve unimodal tasks, are not necessarily different. Furthermore, the boundaries between modalities themselves are vague: GPT-4 has a unimodal output of text, but this can also include computer code, and depending on if we treat this as the same output modality, we get co-learning or multimodal learning. Motivated by this, we attempt to demarcate co-learning from multimodal learning. One example that we agreed upon was that late fusion is not co-learning, as a post-processing combination step does not suffice as co-learning. Instead, modalities must be intuitively "learned together".

Another perspective that was brought up is that we should treat co-learning as a property of the training and models rather than a type of task. This way, co-learning is a feature of the training process that can be achieved alongside multimodal learning.

2 What is needed for co-learning?

In order for co-learning to be effective, the different modalities should complement each other and provide synergistic information and interactions. This synergy can be quantified by measuring the mutual information between the two modalities. Mutual information is a measure of the amount of information that is shared between two variables, and it only arises when both variables are present. Thus, when two modalities are used together in co-learning, the amount of mutual information can indicate how much they are working together to improve the model's performance.

In addition to mutual information, co-learning can also benefit from redundant information, such as grounding and alignment. Grounding refers to the process of associating words or concepts with specific visual or sensory experiences. For example, when learning about the concept of a cat, a co-learning model might use images of cats as well as text descriptions to improve its understanding of what a cat looks like and what characteristics it has. Alignment refers to the process of aligning information from different modalities so that they can be compared and combined. For example, in a co-learning model that uses both text and audio data, the model might align text transcripts of spoken words with the corresponding audio waveforms. Redundancy is especially important in scenarios where information from one modality is noisy or limited, as the model will benefit from the supplemental understanding provided by the other modality.

Another important factor in co-learning is ensuring that one modality does not dominate the other. In other words, the model should be able to extract useful information from both modalities, rather than relying too heavily on one or the other. This can be achieved through careful design of the co-learning task and the architecture of the model. Multi-modal data can be particularly useful for building smaller models that can still achieve high performance, as the different modalities provide two completely different types of information that can be used to improve the model's overall understanding of the task at hand.

There are some particular nuances at hand when considering the effectiveness of co-learning. Firstly, having data from the pretraining phase is not enough to solve tasks (e.g. visual patches unable to solve commonsense reasoning). Next, consider the contrastive learning experiments in Jia et al. [2021]. Here, contrastive learning on noisy data is inefficient, as the proposed method only does as well as CLIP, which has 400 million data pairs. This shows that when we do co-learning, we also need to take into account the curation of datasets, including potentially focusing on tackling specific domains. Co-learning is also task-dependent: whether adding data from a modality works or not varies on a very fine-grained degree [Yun et al., 2021].

One interesting side note is that co-learning can theoretically be used to test our theory of learning. For example, one could test if listening to a radio is enough to learn a language, or whether we would need additionally modalities of information (e.g. embodiment). However, a limitation of this method would be that co-learning requires a unimodal downstream task for evaluation.

3 Dominance in co-learning

In a co-learning setup, it is important to consider the dominance of language as a modality. This is because language has the unique ability to express almost all modalities in some way, and there is a vast amount of language data available for use. Consequently, language models can contain a wealth of information and can serve as a strong baseline for many tasks.

However, it is also important to recognize the potential limitations of relying solely on language as the main modality in a co-learning setup. In some cases, other modalities such as images, video, or audio may provide information that is difficult or impossible to express in language. Thus, incorporating multiple modalities can help to enrich the representation of the task at hand and improve performance.

Moreover, using multiple modalities in a co-learning setup can help to reduce the risk of harmful bias. For instance, if language is the only modality used, it may inadvertently perpetuate certain stereotypes or biases that are present in the language data. However, by incorporating other modalities, the model can learn

to recognize patterns that are not solely dependent on language, thus reducing the risk of bias. Overall, co-learning with multiple modalities can lead to more robust and accurate models that are better suited to a wider range of tasks.

4 What are the risks of co-learning?

In machine learning, bias refers to the systematic error that occurs when the model consistently makes incorrect predictions due to flawed assumptions in the data. Harmful bias, in particular, can arise when the data contains artifacts or prejudices that are embedded in the features or labels, leading to unfair or discriminatory outcomes.

In a co-learning setting, the risk of harmful bias can increase when multiple modalities are used. For instance, in a visual recognition task, if the training data consists of mostly male doctors and female nurses, the model may learn to associate the profession with gender, leading to harmful biases. While it is possible to remove such biases from text, removing biases from images or other modalities can be more challenging.

Another challenge is that when multiple modalities are used, the model may learn shortcuts that are specific to the artifacts in the data, rather than generalizing to new instances. This can lead to overfitting and poor performance on real-world data.

Therefore, it is important to carefully curate and preprocess the data to mitigate the risk of harmful bias. Additionally, regularizing the model or incorporating fairness constraints can help prevent the model from learning shortcuts and improve generalization performance.

5 Evaluating co-learning

Once the model using co-learning is trained, we also need to evaluate whether co-learning was indeed successful. Here, one challenge is the imbalance of data: as we add modalities, we are also adding net data that is being fed into the model. Thus, it might be worth considering whether the same amount of data in the original modality would also provide such an improvement compared to the co-learning case. Another method is to randomly remove a portion of the data in the original modality and replace it with data in the new co-learning modality, and compare performance.

However, this is not the only perspective that we should be evaluating upon. The above methods treat data of the same modality and new modalities as equal, but this is not always the case. Often, collecting more data in the same modality is harder than using an existing dataset from another modality. When we aggregate multiple instances of this together, though adding a modality for a single task requires additional data, sharing across multiple co-learned tasks and datasets could result in savings in labeled data overall. Thus, it might make sense to have an effort-based metric, where effort is lower when data has already been collected, and effort also depends on the amount of dedication required to gather new annotations. However, this metric may be subjective and not necessarily well-defined.

Some other methods of evaluating if co-learning was successful are measuring a shift of representation space, or, as mentioned before, an improvement in the confidence of prediction. There may also be additional benefits such as results becoming more interpretable. This, and standard metric-based evaluations, can be completed using ablation studies.

References

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, 2020.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. Grounding 'grounding'in nlp. In *Findings of*

the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4283–4305, 2021.

Jared Dunnmon, Alexander Ratner, Nishith Khandwala, Khaled Saab, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew Lungren, Daniel Rubin, and Christopher Ré. Cross-modal data programming enables rapid medical machine learning, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

J Luketina, N Nardelli, G Farquhar, J Foerster, J Andreas, E Grefenstette, S Whiteson, and T Rocktäschel. A survey of reinforcement learning informed by natural language. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, August 10-16 2019, Macao, China.*, volume 57, pages 6309–6317. AAAI Press (Association for the Advancement of Artificial Intelligence), 2019.

Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 86–93, 2000.

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, may 2022.

Karthik Sridharan and Sham M Kakade. An information theoretic framework for multi-view learning. 2008.

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.

Tian Yun, Chen Sun, and Ellie Pavlick. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, 2021.