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

Week 15: Generalization, Low-resource & Robustness

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

Edited by Paul Liang *Scribes: Alissa Ostapenko and Martin Q. Ma*

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 15's session, students discussed challenges in generalization to a large number of modalities and tasks, with a particular focus on low-resource modalities and robustness to noisy and missing modalities.

The following list of research probes was provided:

1. One general claim is that pre-trained models can help with low-resource settings (e.g., few-shot fine-tuning). What are the multimodal problems where the paradigm of pre-training and fine-tuning may not generalize? What are the technical challenges?
2. What are new research paradigms that should be explored to address the challenges of multimodal low-resource problems? Can you propose a taxonomy of the challenges that should be addressed to make progress in this direction, for low-resource modalities?
3. How can we develop new models that generalize across many modalities, going beyond only 2 or 3 modalities? What are the tradeoffs between modality-specific multimodal models and general-purpose multimodal models?
4. What are the commonalities and underlying principles shared across diverse modalities and tasks that can enable good generalization? In other words, what are the pre-requirement for generalization to succeed?
5. What are the limits of generalization? In other words, in which cases is generalization across modalities and tasks not possible due to possibly to data heterogeneity or some other reasons? What are these scenarios where generalization may not be possible?
6. How can we potentially perform generalization of multimodal models in the absence of explicit alignment (e.g., paired data) between modalities? How can we tackle the challenges of learning cross-modal interactions, alignment, reasoning, etc?
7. One other aspect of generalization is with real-world settings where noise is present and modalities may be even missing. How can we robustly handle these noisy situations? How can multimodal help? Can multimodal also make these noisy situations harder?

As background, students read the following papers:

1. (Required) Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language [Zeng et al., 2022]
2. (Required) Robust Contrastive Learning against Noisy Views [Chuang et al., 2022]
3. (Suggested) Unsupervised Vision-and-Language Pre-training Without Parallel Images and Captions [Li et al., 2020]
4. (Suggested) Unsupervised Image Captioning [Feng et al., 2019]
5. (Suggested) LiT: Zero-Shot Transfer with Locked-image Text Tuning [Zhai et al., 2021]
6. (Suggested) Missing Modalities Imputation via Cascaded Residual Autoencoder [Tran et al., 2017]
7. (Suggested) Multimodal Learning with Incomplete Modalities by Knowledge Distillation [Wang et al.,

2020]

8. (Optional) Unsupervised Multimodal Representation Learning Across Medical Images and Reports [Hsu et al., 2018]
9. (Optional) Multimodal Prototypical Networks for Few-shot Learning [Pahde et al., 2021]
10. (Optional) Multibench: Multiscale Benchmarks for Multimodal Representation Learning [Liang et al., 2021]

1 Low-resource learning

1.1 Challenges in low-resource multimodal learning

Alignment is a problem that requires sufficient data to solve since temporal alignment may be rare and inter-modality interaction may not be explicitly annotated. These problems may be exacerbated in a low-resource setup in which the task requires understanding temporal alignment across modalities. Another fundamental challenge in multimodal is the "distance" between modalities (i.e., how different the views are at the underlying phenomena they refer to). Low-resource is particularly difficult when there's high distance (low overlap) between modalities (e.g., EEG and text).

1.2 Potential solutions

Unimodal foundational models can be useful if multimodal labels or interactions are low-resource. If one of the modalities does not have high-resource data, it may also be helpful to use domain expertise to restrict the number of interactions across modalities, or with the low-resource modality. In cases where alignment is the primary concern, data preprocessing using heuristics to get better alignment may help focus the learning process (e.g., filtering out music from tv shows or youtube videos ahead of time). This is particularly difficult on videos, where alignment may be noisy (e.g., the audio of a video being only background music, not the words being spoken).

One way forward could be to generate data to "create" more data for training the low-resource models. For example, structured generation of text modality for some languages: this may not always generate meaningful content, but it help models learn the structure. This may work for languages with certain kinds of syntactic and grammatical encoding (e.g., Swahili).

In some cases, the problem of "low-resource" is actually the problem of low-resource labels, not low-resource data. In these cases, self supervision signals will be particularly helpful, for example self-supervised contrastive learning, in which samples can be augmented multiple times and networks can be trained using fully self-supervised objectives.

2 Using pretrained models in low resource settings

In this section we describe some ideas on tackling the challenges associated with pretrained models in low resource settings. One idea is to translate a task from one modality to the other with richer information and dataset availability. For example, Zeng et al. [2022] suggests to map the image modality to text embeddings and use generated text tokens for downstream tasks. In the multimodal setting, the same idea translates to mapping multiple modalities into a single latent space that is easier to deal with (for efficient training, testing, and evaluating) the downstream tasks. While we often strive to represent multiple modalities in a single latent space, the biggest unanswered question so far is whether that latent space even exists (theoretically)?

2.1 Modeling a common representation

A promising research direction for the image modality, is to obtain image tokens (like the ones we have for text through word piece and sentence piece token models) that are helpful not just in understanding the interactions across the modalities but also the semantics of the modalities. Zeng et al. [2022] proposes Socratic models guided by LMs. Since language models are easy to train, another advantage is that it's easier to understand the modality (as opposed to subjective evaluation of images).

2.2 Going beyond language

While it is easier to probe language tokens [Wang et al., 2022] leading to better interpretability, visual input can also be interpretable if the training process includes carefully curated intermediate loss functions that reveals the important features of the images for a better interpretation. Therefore, some tasks may be better done in the image modality (for e.g., Interpreting an EEG, or a CT Scan) without any translation. A better approach may be to have an ensemble of models for each task, where each of the model in the ensemble in a unimodal model (i.e. the other modalities of the task are translated to the respective unimodal tokens of the model).

This probes a further fundamental question of whether we should direct research towards training models with only images as the fundamental modality. Towards this suggestion, an argument could be that text datasets collected from the internet has various biases encoded, and hence it's better to train image based models. However, creating large datasets for images is a challenging task, and images drawn from the web have their own biases too.

In the context of pretraining paradigm, one of the suggestions is to train the models end-to-end instead of using several layers of pretrained models as independent blocks. However, the issue with this remains that the parameter space explodes quickly if we were to finetune every pretrained model for the task and only big companies would have enough resources to achieve this.

2.3 Challenges

There are modalities that are hard to describe in language: high frequency modalities like video and audio (e.g., using language to describe the details of your voice tone for identification), and modalities that don't have good coverage in language (e.g., EEG, sensor data from robots). Also, fusion across modalities is hard to describe in language - if it is hard to identify "atomic" elements in a given modality, it will be harder still to refer to multiple across modalities. This relates to the idea of "concreteness" in cognitive science [Buccino et al., 2019]: what can and cannot be expressed in language. There is also the challenge of the language world-state to retrieve things – what happens if there are too many tokens to retrieve? Language may be too coarse a modality to express what is needed in the internals of the network. Either a joint representation space or translation between modalities may be required when an idea can not be effectively covered by language.

3 Generalization in low resource settings

Continuing the discussion from the previous section, it is important to understand which structures among models are important to specific tasks. In low resource generalization, it could be argued that the best approach would be to filter out only the necessary structures from models before using them for few shot/zero shot learning. However, there is no easy way deep learning could allow for such filtering of inherent structures among the models and modalities.

This leads to our next research probe: is deep learning the most optimal way for low resource generalization? Other classes of models, for example, Probabilistic Graphic Models (PGM) [Koller and Friedman, 2009] are more expressive in the terms of their reasoning about the structure being learned. On the other hand, transformers have fewer inductive biases encoded and hence are forced to assume and learn more correlations than what's necessary - potentially creating more issues with generalizability.

3.1 Modality and task specific vs general models

More general models require more resources (data & compute) while more specific models can be more cost efficient. More general models also face the problem of catastrophic forgetting: in finetuning we lose generalizability. One idea that was proposed was to focus on building universal feature extractors which would make the question of task specific vs more general easier.

One question worth considering is how can we incorporate domain expert knowledge in general models? If this is answered, then the gap could be bridged between general and task specific models in an interpretable way. This may require finer grained annotations or collaborating with domain experts. This would be particularly helpful in the medical domain where experts don't trust deep learning models and tend to prefer simpler models that can be partially explainable.

3.2 On foundation models

The current trend is to depend heavily on certain large "foundation models" for further fine-tuning and downstream tasks, especially in tasks with low resources. However, understanding and interpreting these models is crucial for guiding better generalization. Industry is better accompanied with resources and infrastructure to deploy these large models in various products, through heavy distillation [Aguilar et al., 2019] and custom hardware & software specialized [dee] for low latency inferences on CPUs. This allows for big tech companies to gain sufficient feedback from the model in real time, and update or iterate with continued learning for better generalization.

In order to efficiently use these foundation models in the pretraining paradigm, it's probably the best to define a curriculum for foundation models. This curriculum could be custom designed for specific modality and tasks. However, how to define the best curriculum for the pretrained foundation models remains an answered question.

Moreover, how do we quantify the amount of prior knowledge required in such custom curriculum? Human brains are pretrained for billions of years through the process of evolution. However, any deep learning model that's randomly initialized doesn't have necessary inductive biases to learn efficiently. This reinforces the idea that, ideally, we would require multiple foundation models. A single model may not be capable of adapting quickly to low resource settings.

4 Dimensions of robustness

Before describing the challenges in designing more robust models, it is worth considering what robustness means and how we evaluate it. It may be difficult to evaluate robustness when models are trained on large datasets, because any test we devise may be covered by some sample in the dataset, causing the model to memorize it. What is needed, then, is more rigorous benchmarking for robustness. It is also worth mentioning that there is often a tradeoff between robustness and task-specific accuracy. When the task is well known, optimizing for robustness may hinder performance.

A few years ago, it was very common to tailor models to specific tasks using domain expertise. It is an open question whether this will still be a possible research direction for academia in the age of large pretrained models. One line of thinking is that this opportunity exists but comes in the form of fine annotations. Very fine annotations could encode domain expertise, and allow for efficient adaptation of pretrained models to specific tasks.

4.1 Adversarial training and robustness

Can we borrow ideas from adversarial training to design more robust models? The argument is as follows: if robustness is defined as a model performing well on a "similar" or increasingly dissimilar space of inputs and labels to the ones it was trained on, then couldn't the problem of robustness be conceived as an adversarial training problem? The objective then would be for the adversary to attempt to explore this space of inputs during training so the model is prepared at test time. In theory, this may be a sufficient condition, but in practice it will be very difficult to implement. For an adversary to be able to perform this role effectively, it would need to know which samples break the pattern we see as humans. For example, if we are testing our model to be robust to changes in lighting in an image, our adversary could alter an image to be more dimly lit. But if it makes it too dimly lit, a human would not be able to recognize it. Then the question is not: how can our adversary explore *towards* the regions we want it to find for the sake of robustness, but how can we limit the exploration *away* from only the regions that break the label function we are attempting to approximate? Current perturbation based techniques are limited in how they preserve semantics while changing form.

Perhaps prompting large pretrained models could be a way forward to enhance this diversity.

References

URL https://www.microsoft.com/en-us/research/blog/deepspeed-accelerating-large-scale-model-inference-a #toc-heading-9.

Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. Knowledge distillation from internal representations, 2019. URL https://arxiv.org/abs/1910.03723.

Giovanni Buccino, Ivan Colagè, Francesco Silipo, and Paolo D'Ambrosio. The concreteness of abstract language: an ancient issue and a new perspective. *Brain Structure and Function*, 224(4):1385–1401, 2019.

Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. *arXiv preprint arXiv:2201.04309*, 2022.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134, 2019.

Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive Computation and Machine Learning series. MIT Press, 2009. ISBN 9780262013192. URL https://books.google.com/books?id=7dzpHCHzNQ4C.

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. *arXiv preprint arXiv:2010.12831*, 2020.

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.

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.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1828–1838, 2020.

Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812, jan 2022. doi: 10.1109/tvcg.2021.3114794. URL https://doi.org/10.1109%2Ftvcg.2021.3114794.

Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *arXiv preprint arXiv:2111.07991*, 2021.