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

Week 5: Modality Utility, Tradeoffs, and Selection

Instructors: Louis-Philippe Morency, Paul Liang *Synopsis Leads: Durvesh Malpure, Aditya Rathod*

Edited by Paul Liang *Scribes: Yilin Wang, Sean Chang*

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 5's discussion session, the class aimed to formalize a definition of modality utility and various ways to select necessary modalities for a task. The following was a list of provided research probes:

1. What are the different ways in which modalities can be useful for a task? How can we measure the utility of a modality for a task, given only access to the dataset (i.e., before designing and training a model)?
2. What are the criterion for which we should add or select modalities for a task? Is minimizing redundancy the only goal? Are there benefits in maximizing redundancy? Are there other criterion we should consider too?
3. There has been substantial work in feature selection. Is modality selection the same as feature selection? What are the potential differences and new technical challenges in modality selection but not present in conventional feature selection?
4. Given trained models, how can we estimate how important each modality was when making the prediction? How were these modalities used separately and in interaction with other modalities?
5. What are the different ways in which modalities can be harmful for a task? Think about a list of reasons why we would prefer to not use a modality. How can we quantify these potential risks?
6. What are some solutions for tackling these risks and biases in multimodal datasets and models? How can we properly identify, visualize and eventually reduce these risks in multimodal datasets and models?
7. Can we come up with guidelines that compare the tradeoffs between modality benefits and risks? How can we then integrate these insights into multimodal model design? Will integrating these insights help?

As background, students read the following papers:

1. (Required) Greedy modality selection via approximate submodular maximization [Cheng et al., 2022]: An approach for modality selection based on modality importance estimation. Make sure you understand the general idea, algorithm, and results, you can optionally skip the math details.
2. (Required) Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks [Wu et al., 2022]: This paper sets up several hypotheses and experiments to better understand the optimization challenges and shortcut learning in multimodal models.
3. (Suggested) Information Value Theory [Howard, 1966]: This paper proposed the 'value of information' framework enabling the assignment of quantifiable measures that a decision maker would be willing to pay for information prior to making a decision. Can we use this framework to understand the value of modalities and their interactions with existing modalities?
4. (Suggested) A new look at the statistical model identification [Akaike, 1974]: This paper was the first to propose the AIC framework for model comparison, providing a principled way to consider performance relative to model size. Can we extend this to quantify feature or modality selection in a principled way?
5. (Suggested) Efficient Feature Selection via Analysis of Relevance and Redundancy [Yu and Liu, 2004]:

Introduces the min redundancy max relevance approach for feature selection, an important approach influencing many feature selection methods today.

6. (Suggested) Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews [Booth et al., 2021]: This paper studies whether social biases be made worse with multiple modalities.
7. (Suggested) Perceptual Score: What Data Modalities Does Your Model Perceive? [Gat et al., 2021]: Quantifying modality importance in trained models.
8. (Relevant) Shortcut learning in deep neural networks [Geirhos et al., 2020]: General review paper about learning shortcuts due to data and model biases.
9. (Relevant) Multimodal datasets: misogyny, pornography, and malignant stereotypes [Birhane et al., 2021]: This paper highlights some concerns in large-scale multimodal datasets used for modern pretrained models.
10. (Relevant) MultiBench: Multiscale Benchmarks for Multimodal Representation Learning [Liang et al., 2021]: This is a benchmark studying the tradeoffs that modalities can introduce across generalization, efficiency, robustness perspectives.
11. (Relevant) Submodularity issues in value-of-information-based sensor placement [Malings and Pozzi, 2019]: This paper approaches feature selection (the sensor placement problem in their case) with a value of information formalism, and discuss issues regarding its lack of submodularity (which the first required paper assumes in their greedy approach).

We summarize several main takeaway messages from group discussions below:

1 Modality Utility and Selection

In machine learning, selecting the right set of modalities is crucial for the performance and interpretability of the models. Modality selection can be seen as a process of choosing a subset of the available modalities, such as text, images, or audio, to be used for the task at hand. The goal of modality selection is to reduce the dimensionality of the input space and to improve the model's efficiency and effectiveness. It is important to handle noisy or missing data in some modalities and figure out how to deal with them. Bias and overfitting can also be an issue when selecting modalities or features, so it is important to be cautious when interpreting results. Additionally, some tasks may require interactions between modalities, while others may benefit from using only a subset of modalities. Therefore, it is important to consider the specific task and data when making decisions about modality and feature selection.

1.1 Selection Criteria

One common approach in selecting modalities is to use relevance and redundancy metrics to measure the importance of each modality. Relevant modalities contain unique and useful information that can improve the model's accuracy, while redundant modalities provide robustness and help the model learn relationships between different modalities. In some cases, modality selection may not be necessary if the data is already preprocessed and transformed into a feature space.

However, the 'maximize relevance and minimize redundancy' principle may not be the most appropriate sometimes. For example, when the data is noisy, then having redundant modalities may help the model enhance its judgement. Furthermore, in unsupervised learning like the translation between thermal and RGB images, we would need redundant information for temporal alignment, which is essential to the task.

Modality selection can also affect the synergy between different modalities. Synergy refers to the interactions between modalities, which can improve the model's performance. On the other hand, some tasks may benefit from using only a subset of modalities, as the model can learn simpler and more interpretable representations.

One of the primary ways we can determine the modality importance is by human simulation and finding out which modality is more used by humans for the task. Humans also prioritize certain modalities based on the

task at hand, such as focusing on text when taking a test. In some cases, only a small subset of modalities is needed, while in others, modality selection is based on the task, data, and model.

1.2 Modality selection vs feature selection

There has been substantial work in feature selection in machine learning [Yu and Liu, 2004]. Is modality selection the same as feature selection? What are the potential differences and new technical challenges in modality selection but not present in conventional feature selection? We believe that one major difference between feature selection and modality selection is that the feature selection is constrained to only one task at hand. This may not apply so well to modality selection because we want our multimodal model to learn more than just the current task. Furthermore, each modality has its own structure, which should receive separate treatment respectively, whereas typical feature selection operates on numerical features that are often of the same structure. Finally, a possible answer to distinguishing between two modalities is that there exists some ambiguity when connecting two modalities. If there is ambiguity in translation between two kinds of data, then they should be called modalities. Handwritten text and computer generated text are different modalities as they have different structures and some ambiguity when translating between each other.

1.3 Risks of a modality

From a statistical/information-theoretic standpoint, including more modalities generally does not hurt given that the training data is enormous and our objective is general information extraction. However, having too many modalities can introduce information irrelevant to the task, and may cause the model to pick up spurious correlations and introduce bias to the model.

To mitigate the bias of multimodal models, one approach is to explicitly state the bias and train the model in a generative-adversarial manner to make sure that the model does not pick up the bias in data. Other ways to account for the bias would include (1) creating fair and unbiased training datasets and (2) using de-biasing methods such as re-weighting and fair learned representations.

Many bias/fairness papers mentioned that one de-biasing method would be to remove features that contains biased information. However, it is also mentioned in another paper [Acosta et al., 2021] that there seems to be a tradeoff between having low bias and high performance. Therefore, we concluded that it is acceptable for the model (pre-trained encoders) to learn features that contains biased information, but when it is used in application (e.g., adding classification head or decoders on encoders), we must ensure that the model would be unbiased. Yet, such regularization should be added in the inference/application step, not during (pre)-training.

References

Halim Acosta, Nathan Henderson, Jonathan Rowe, Wookhee Min, James Minogue, and James Lester. What's fair is fair: Detecting and mitigating encoded bias in multimodal models of museum visitor attention. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, ICMI '21, page 258–267, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384810. doi: 10.1145/3462244.3479943. URL https://doi.org/10.1145/3462244.3479943.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *CoRR*, abs/2110.01963, 2021. URL https://arxiv.org/abs/2110.01963.

Brandon M. Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K. D'Mello. Bias and fairness in multimodal machine learning: A case study of automated video interviews. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021.

Runxiang Cheng, Gargi Balasubramaniam, Yifei He, Yao-Hung Hubert Tsai, and Han Zhao. Greedy modality selection via approximate submodular maximization. In James Cussens and Kun Zhang, editors, *Proceedings*

of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, volume 180 of *Proceedings of Machine Learning Research*, pages 389–399. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/cheng22a.html.

Itai Gat, Idan Schwartz, and Alexander G. Schwing. Perceptual score: What data modalities does your model perceive? *CoRR*, abs/2110.14375, 2021. URL https://arxiv.org/abs/2110.14375.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020. URL https://arxiv.org/abs/2004.07780.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1): 22–26, 1966. doi: 10.1109/TSSC.1966.300074.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multibench: Multiscale benchmarks for multimodal representation learning. *CoRR*, abs/2107.07502, 2021. URL https://arxiv.org/abs/2107.07502.

C. Malings and M. Pozzi. Submodularity issues in value-of-information-based sensor placement. *Reliability Engineering System Safety*, 183:93–103, 2019. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2018.11.010. URL https://www.sciencedirect.com/science/article/pii/S0951832017306270.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24043–24055. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wu22d.html.

Lei Yu and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.*, 5:1205–1224, dec 2004. ISSN 1532-4435.