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

Week 13: Explainability & Interpretability

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

Edited by Paul Liang *Scribes: Martin Q. Ma, Nikhil Yadala*

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 13's discussion session, the class discussed challenges and techniques for interpreting and explaining multimodal models and data, as well as their evaluation. The following was a list of provided research probes:

1. What is a taxonomy of all the multimodal phenomena that we should aim to interpret?
2. In a perfect world, what multimodal information would you expect to be available when interpreting a multimodal model? What multimodal phenomena and characteristics would you want from this "perfect" interpretable model?
3. What aspects of multimodal interpretability extend beyond the unimodal case? What are the dependencies between unimodal and multimodal interpretability? In other words, what needs to be solved on the unimodal side so that we are successful in multimodal interpretability?
4. What approaches and techniques can you imagine being best suited for multimodal interpretation? How should we visualize the results of these multimodal interpretations? Black-box model interpretation vs interpretation by design (white-box)?
5. How can we evaluate that a specific multimodal phenomena (e.g., bimodal interactions) was properly interpreted? How do we measure success in multimodal interpretability?
6. Separate from model interpretation, there is also the topic of dataset interpretation: characterizing and interpreting the multimodal phenomena present in the data itself, independent of a specific model or prediction task. How can we perform multimodal data interpretation, and are there any differences with multimodal model interpretation?
7. What is the best way to visualize relatively understudied modalities beyond language and vision? How can we best analyze and characterize the multimodal interactions present between these other modalities?
8. What are the unique challenges to multimodal explainability, where not only the model is multimodal but also the explanation is potentially multimodal?

As background, students read the following papers:

1. (Required) M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis [Wang et al., 2022]
2. (Required) VL-InterpreT: An Interactive Visualization Tool for Interpreting Vision-Language Transformers [Aflalo et al., 2022]
3. (Suggested) Multimodal Explanations by Predicting Counterfactuality in Videos [Kanehira et al., 2019]
4. (Suggested) "Why Should I Trust You?": Explaining the Predictions of Any Classifier [Ribeiro et al., 2016]
5. (Suggested) The Mythos of Model Interpretability [Lipton, 2018]
6. (Suggested) Interpretable Machine Learning: Moving From Mythos to Diagnostics [Chen et al., 2022]

7. (Suggested) The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective [Krishna et al., 2022]
8. (Suggested) Do Explanations make VQA Models more Predictable to a Human? [Chandrasekaran et al., 2018]
9. (Suggested) Multimodal Neurons in Artificial Neural Networks [Goh et al., 2021]
10. (Suggested) DIME: Fine-grained Interpretations of Multimodal Models via Disentangled Local Explanations [Lyu et al., 2022]
11. (Suggested) Multimodal Routing: Improving Local and Global Interpretability of Multimodal Language Analysis [Tsai et al., 2020]
12. (Suggested) How does this interaction affect me? Interpretable attribution for feature interactions [Tsang et al., 2020]
13. (Suggested) Leveraging Sparse Linear Layers for Debuggable Deep Networks [Wong et al., 2021]
14. (Suggested) Rethinking Explainability as a Dialogue: A Practitioner's Perspective [Lakkaraju et al., 2022]
15. (Suggested) ExplainaBoard: An Explainable Leaderboard for NLP [Liu et al., 2021]

We summarize several main takeaway messages from group discussions below:

1 Taxonomy of Multimodal Interpretability

Multimodal interpretability is different from the unimodal setting because it involves interactions between different inputs. These interactions include dominant, complementary, and conflicting cross-modal interactions [Wang et al., 2022].

Interpretability can come from the dataset side and the model side. Data interpretation attempts to explore the underlying correlation between input data and annotated labels. Model interpretation often considers the specific task, including how data is processed, the kind of task the model performs, and how the model predicts specific results from the data. For example, if we were to focus on multimodal fake-content detection, we would wish to know if there are classes of content, say movie critiques, which are unfairly hit as "fake-content" by our algorithm, as that would indicate our model is learning that the entire class is fake content.

For model interpretation, we can consider two major requirements:

1. The general interactions between modalities in the input (e.g., when the modalities have conflicting information, what does our model trust?)
2. The specifics in how information is paired and correlated (e.g., when given two inputs with similar information, how does our model align them?)

Data interpretation is more generic, as it could be used on different types of models without considering the tasks. It is important to conduct data interpretation because there is a lot of bias in multimodal datasets, such as datasets accidentally favoring one modality in a task [chandrasekaran2018explanations].

1.1 Considering Stakeholders

For ML engineers and developers who want utmost clarity in our model, we want systems that help us both understand how our model is doing on our target problem, and potential ways we might be able to change the model, like a model-based diff. Here, these stakeholders value having multiple ways to visualize the explanation, as they want to get to the bottom of a problem and hopefully improve their component of the solution.

On the other hand, for the general public, who want to understand local decisions but do not have the background that the above group has, we may prefer systems that focus on making the model's decisions easier to grok, but while conveying that there might be factors that the model is considering which the explanation does not discuss. Given that black-box explanations can fool a user into false conclusions about how a model operates [Lakkaraju and Bastani, 2020], we want to focus on faithfulness and transparency as much as possible. Should a model seem to fail in our visualization, we want ways to explain to the user why they should not trust the model's prediction.

2 Approaches to Interpretability

We now aim to categorize recent approaches to interpreting models and datasets.

2.1 Confirmational vs Exploratory

We can think of techniques as being categorized into **confirmational** or hypothesis-driven, and **exploratory** where we do not have a specific hypothesis to test.

Error analysis and categorization of errors are examples of exploratory interpretation. This exploratory analysis can further lead to hypotheses being generated by researchers about how a particular feature affects the model output (for example, that the model is ignoring visual data and using only language inputs).

From the confirmatory lens, we can perturb images and test whether the prediction changes to test the hypothesis if the model ignores visual input. Confirmatory analysis is easier to define than an exploratory one. We can have hypothesis testing on all types of learning, but we have more explainability methods as datasets become more supervised (gradients, probing, etc.) vs just clustering (for example) on the unsupervised case. Another view of a confirmatory analysis is that we have an expectation of model should do (e.g., count), then set up experiments/tests to check this.

Referring back to the discussion of model and dataset interpretability from Section 1, we can think of model interpretation as more specific, where we focus on how to interpret a hypothesis-driven model (a model with specific inductive bias). We could perform the hypothesis test because we can expect some behaviors or statistics of the variables or certain weights of the model (for example, Generative Addictive Model where all addictive components have specified parametric form). We can set up experiments and tests to check these in a new dataset.

Moreover, we can think of dataset interpretation as exploratory, if we do not have a specific task in mind, and want to analyze association between different features.

2.2 Post-Hoc vs Mechanistic

Another view breaks interpretability into two other categories, **post-hoc** and **mechanistic**.

Post-hoc techniques either assume the model is a black-box or white-box, and afterwards attempt to explain particular decisions a model takes. Models could require post-hoc analysis if they are too big (such as large language models), or too deep, which can make it challenging to get understandable explanations from. This could also be done with techniques like prompt-chaining [Wu et al., 2021], where we perform dialogues with large language models to try to understand their thought process. Post-hoc techniques include VL-InterpreT [Aflalo et al., 2022], which use attention mechanisms to interpret model decisions. While attention mathematically makes sense, it may be difficult to tailor these explanations to stakeholders who are not familiar with the math of attention mechanisms. Similarly, for deep architectures, while gradient-based approaches can lead to understanding, they may generate multiple conflicting explanations. This brings up difficulties of evaluating interpretation techniques, which we discuss further in Section 4.

Mechanistic techniques, on the other hand, derive interpretability from model design. There might be ways to understand the model directly through analyzing the learned algorithm, such as investigating the features and weights of linear models. The model could also be modeled after human reasoning techniques, such as the deep module networks approach [Andreas et al., 2016], making it more interpretable by design.

3 Comparing Interpretability, Explainability, & Controllability

When we discuss interpretability, two other related concepts that commonly come into the discussion are explainability and controllability. Each term is rather overloaded, so we spent some time trying to figure out how best to disentangle what we mean by each of them.

Interpretability typically relies on post-hoc understanding of what a model detects before making a decision. We can see that the model is looking at different aspects of data, but we do not know how the model reasons about those data or makes predictions.

Conversely, explainability refers to a model which is laid out closer to a simple program. Every step of the computation is laid out and easily readable, and a user is able to simulate model decisions, at least in part. When we look at large models like large language models, it may be difficult to see how the model actually performs the task due to the sheer number of computations and the lack of model modularity. By aligning what the model does with human thinking, we might be able to make the decisions of models more useful to humans. For example, when users are told the weakness of a model, this will help humans predict when the model will fail, providing a good security check for the model [Wong et al., 2021].

Lastly, controllability refers more to counterfactual reasoning, where we wish to understand how a model's inputs can be changed to produce a desired output. Statistically, controllability can be evaluated through interventions, which involves causal analysis. For example, should we know that some language causes a censorship filter to filter out a message, we can change the language accordingly to potentially bypass the filter. There have been efforts to perform these sorts of comparisons on language generation through the use of linearizing generated outputs [Geva et al., 2022]. The major problem with this category, however, is that it becomes hard to disentangle confounding affects of models given the sheer amount of play we have in continuous space.

Overall, comparing the three, explainability is a higher standard than interpretability, and true controllability is of an even higher standard, in terms of reliability and trust. When you can perform local interpretation of a model's outputs, you can generally locally understand why a decision was made. Conversely, when a model is explainable, you should be able to understand why all decisions are made, as explainability further includes global interpretability. Lastly, when a model is controllable, you are not only able to understand all global decisions, but are able to change a decision by changing some variable accordingly, making it the highest form of interpretability of the three.

4 Evaluating Interpretability

Evaluating interpretability is a task that came up time and time again during our discussions. While most people who study interpretability consider human-experiments as the highest standard for determining if some explanatory tool is better than another, such approaches are both costly and can provide less information than desired.

Part of the reason that interpretability models are hard to evaluate is that a majority of the most general models (e.g., LIME [Ribeiro et al., 2016]) work at the instance or local level in explaining individual predictions rather than explaining predictions globally. It may be hard to tell how much local understanding translates to global understanding.

One way to evaluate interpretability is to ask people to simulate the model predictions based on interpreted evidences. While these techniques can work for smaller models, they do not scale for large numbers of parameters easily. For example, while VL-InterpreT faithfully visualizes attention maps in multimodal transformers, it may be hard for a user to look through all attention maps to simulate model decision-making. It can also be difficult to tell how to extend such strategies to new model architectures, and how to ensure that those explanations are faithful regardless.

Alternatively, we can try to establish causal or correlation-driven relationships from our models. By deriving rules from models, and then seeing if those rules correlate to the actual model accuracy, we can perhaps think of explanation as a form of model compression, where instead of having to compress to discrete logical operations we just need to compress the model into more comprehensible forms.

Furthermore, by dividing our current model into different steps of reasoning, we can use different interpretability techniques at each stage of reasoning, which might let us better taxonomize and rank different explainability techniques. First, we can identify atomic objects or steps (not task specific), how they are used in cross-modal interactions, as well as the hierarchical reasoning process. Each of these are different levels of multimodal interpretability.

Lastly, what further complicates evaluation is modalities that are less understood or have fewer resources. While language and vision are easily visualized and their models relatively more understood, it is difficult to see how to best extend these explanations to modalities like audio, sensors, and time-series where it may be difficult to visualize.

References

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. *arXiv preprint arXiv:2203.17247*, 2022.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016.

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make vqa models more predictable to a human? In *EMNLP*, 2018.

Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Interpretable machine learning: Moving from mythos to diagnostics. *Queue*, 19(6):28–56, 2022.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. Multimodal explanations by predicting counterfactuality in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8594–8602, 2019.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective, 2022. URL https://arxiv.org/abs/2202.01602.

Himabindu Lakkaraju and Osbert Bastani. "How do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875*, 2022.

Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. Explainaboard: An explainable leaderboard for nlp. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, 2021.

Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Yao-Hung Hubert Tsai, Martin Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1823–1833, 2020.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. *Advances in neural information processing systems*, 33:6147–6159, 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.

Eric Wong, Shibani Santurkar, and Aleksander Madry. Leveraging sparse linear layers for debuggable deep networks. In *International Conference on Machine Learning*, pages 11205–11216. PMLR, 2021.

Tongshuang Wu, Michael Terry, and Carrie J. Cai. AI chains: Transparent and controllable human-ai interaction by chaining large language model prompts. *CoRR*, abs/2110.01691, 2021. URL https://arxiv.org/abs/2110.01691.