11-877 Advanced Topics in Multimodal Machine Learning

Week 13: Explainability and Interpretability

Due date: 11PM EST, Wednesday, April 13 2022

Submission: https://forms.gle/RTtyvFHYoCqesVVc8

We designed the reading assignments to help you prepare for the live discussions. Discussion probes were drafted related to this week's topic. These were written to help conceptualize the problem and guide your thought process. Take the time to read them first. The goal is not to answer each of these questions and probes individually, but they are meant to be taken as a whole. We also selected research papers relevant to this topic. Required papers should be read completely. Suggested papers should at least be skimmed. The purpose of the reading assignment is to start your critical thinking process, so your responses should demonstrate constructive thoughts, with a good understanding of the current research in this area, and express your own insights.

Your response to this reading assignment should be submitted in the online Google Form (see link above). Your response should consist of four main components:

(1) **Scouting:** As you start thinking about the discussion probes, it is always good to also scout papers, blog posts, and other resources related to the topic. We ask that you search for related resources and share with us 2 extra links to these new resources. For each extra link, include 1-2 sentences explaining the value and relevance of this extra resource.
(2) **Reading notes:** As you read the required papers, suggested papers, and the extra resources you scouted, please write down at least 4-6 notes related to the discussion probes. Each note should be 1-3 sentences long. These can be empirical results you observed, ideas or theories expressed by other researchers, or any interesting fact that is worth noting when summarizing your reading.
(3) **Your thoughts:** Separate from your reading notes, we ask that you reflect more holistically about the discussion probes. Please write 3 discussion points you would like to share on this topic. Each discussion point should be one paragraph (3-5 sentences). These discussion points should go beyond the reading papers, and try to address as many aspects of the discussion probes as you can. We do not expect that you answer all discussion probes. For example, it would be ok to focus on only 1 or 2 probes if these bring the most ideas and thoughts for you.
(4) **Clarification requests [OPTIONAL]:** Please take a moment to suggest parts of the papers where clarifications would be useful. Try to be as specific as possible in your clarification requests. These requests will be shared with the Reading Leads in charge of creating a short presentation for the beginning of Friday's course and answering other requests directly on Piazza.

Week 13 discussion probes:

- *What is a taxonomy of all the multimodal phenomena that we should aim to interpret?*
- 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?
- 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?
- 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)?
- How can we evaluate that a specific multimodal phenomena (e.g., bimodal interactions) was properly interpreted? How do we measure success in multimodal interpretability?
- 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?
- 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?
- What are the unique challenges to multimodal explainability, where not only the model is multimodal but also the explanation is potentially multimodal?

Required papers (you should read these papers in detail)

- https://arxiv.org/abs/2107.08264
- https://arxiv.org/abs/2203.17247

Suggested papers (you should skim through these papers, at the minimum)

- https://arxiv.org/abs/1812.01263
- https://arxiv.org/abs/1602.04938
- https://arxiv.org/abs/1606.03490
- https://arxiv.org/abs/2103.06254
- https://arxiv.org/abs/2202.01602
- https://arxiv.org/abs/1810.12366

Other relevant papers:

- https://distill.pub/2021/multimodal-neurons/
- https://arxiv.org/abs/2203.02013
- https://arxiv.org/abs/2004.14198
- https://arxiv.org/abs/2006.10965
- https://arxiv.org/abs/2105.04857
- https://arxiv.org/abs/2202.01875
- https://arxiv.org/abs/2104.06387