11-877 Advanced Topics in Multimodal Machine Learning

Week 10: Beyond language and vision

Due date: 11PM EST, Wednesday, March 23 2022

Submission: https://forms.gle/g2mNfcDPmFWqyme77

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 10 discussion probes:

- What are the modalities beyond language and vision that are important for real-world applications? What unique structure do they contain, and what are the main challenges in performing multimodal learning with them?
- When reflecting on the heterogeneous aspect of multimodal learning, how are the other modalities different from language, speech, and vision? What dimensions of heterogeneity are important for these other modalities?
- What are the cross-modal interactions that you expect in these other modalities? Could you see ways to model cross-modal interactions with these other modalities and with language and vision?
- How do the core research problems of unimodal and multimodal processing, integration, alignment, translation, and co-learning generalize to modalities beyond language and vision? What core insights from these 'common' modalities have yet to be explored in understudied modalities?
- What is the best way to visualize these relatively understudied modalities? How can we best analyze and characterize the multimodal interactions present between these other modalities?
- How to learn models for many modalities (10+ modalities)? What are the chances to create multimodal learning algorithms that work for all modalities? What are the tradeoffs between modality-specific multimodal models and general-purpose multimodal models?
- If two modalities are very far from each other (strong heterogeneity and/or encoding very different information), how can we address the problem of multimodal learning?

Required papers (you should read these papers in detail)

- Multi-sensor for healthcare (focus on sections 3.3 and 4): https://www.sciencedirect.com/science/article/pii/S1566253521001330
- Multi-sensor in robotics: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9341579

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

- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1657787
- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6907100
- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9561847
- https://arxiv.org/abs/1907.13098
- Multi-sensor survey: https://www.sciencedirect.com/science/article/pii/S156625351630077X

Other relevant papers:

- Old survey: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=126184
- https://arxiv.org/abs/2107.07502
- https://arxiv.org/abs/2203.01311
- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=554212
- https://www.sciencedirect.com/science/article/pii/S1566253520304085
- https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/jmri.20577
- https://arxiv.org/abs/1905.11436