11-877 Advanced Topics in Multimodal Machine Learning

Week 6: Memory and Long-term Interactions

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

Submission: https://forms.gle/dTsbUPjSR1PM6rQR8

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 6 discussion probes:

- What are the scenarios in which memory for long-term interactions is required in multimodal tasks, where data comes from heterogeneous sources? What could be a taxonomy of long-range cross-modal interactions that may need to be stored in memory?
- What are certain methods of parametrizing memory in unimodal models that may be applied for multimodal settings, and the various strengths/weaknesses of each approach?
- How should we model long-term cross-modal interactions? How can we design models (perhaps with memory mechanisms) to ensure that these long-term cross-modal interactions are captured?
- What are the main advantages of explicitly building memory-based modules into our architectures, as compared to the large-scale pre-training methods/Transformer models discussed in week 4? Do Transformer models already capture memory and long-term interactions implicitly?
- To what extent do we need external knowledge when performing reasoning, specifically multimodal reasoning? What type of external knowledge is likely to be needed to succeed in multimodal reasoning?
- A related topic is multimodal summarization: how to summarize the main events from a long multimodal sequence. How can we summarize long sequences while keeping cross-modal interactions? What is unique about multimodal summarization?

Required papers (you should read these papers in detail)

- Baselines + benchmarks: https://arxiv.org/abs/2011.04006
- https://arxiv.org/abs/1907.05242

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

- Memory: https://arxiv.org/abs/1603.01417
- Memory: https://arxiv.org/abs/1611.05592
- Memory: https://arxiv.org/abs/1906.01076
- Memory and interactions: https://aclanthology.org/D18-1280.pdf

Other relevant papers:

- https://www.nature.com/articles/nature20101
- Memory and interactions: https://arxiv.org/abs/2110.13309
- Memory in Transformers: https://arxiv.org/pdf/2007.03356.pdf
- Transformer XL: https://arxiv.org/abs/1901.02860
- https://arxiv.org/abs/1410.5401
- https://proceedings.mlr.press/v48/santoro16.pdf