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

Week 6: Memory and Long-term Interactions

Instructors: L.-P. Morency, A. Zadeh, P. Liang *Synopsis Leads: Nikhil Yadala, Karthik Ganesan*

Edited by Paul Liang *Scribes: Catherine Cheng, Justin Lovelace*

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 6's discussion session, the class discussed various challenges and approaches in modeling memory and long-term interactions in multimodal tasks. The key themes include: (1) settings where modeling long-range interactions is important, (2) unique challenges in modeling multimodal memory, and (3) modeling techniques and memory-based approaches. The following was a list of provided research probes:

1. 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?
2. 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?
3. 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?
4. 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?
5. 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?

As background, students read the following papers:

1. (Required) Long Range Arena: A Benchmark for Efficient Transformers [Tay et al., 2021]
2. (Required) Large Memory Layers with Product Keys [Lample et al., 2019]
3. (Suggested) Dynamic Memory Networks for Visual and Textual Question Answering [Xiong et al., 2016]
4. (Suggested) Multimodal Memory Modelling for Video Captioning [Wang et al., 2018]
5. (Suggested) Episodic Memory in Lifelong Language Learning [de Masson d'Autume et al., 2019]
6. (Suggested) ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection [Hazarika et al., 2018]
7. (Suggested) Hybrid Computing using a Neural Network with Dynamic External Memory [Graves et al., 2016]
8. (Suggested) History Aware Multimodal Transformer for Vision-and-Language Navigation [Chen et al., 2021]
9. (Suggested) Do Transformers Need Deep Long-Range Memory? [Rae and Razavi, 2020]
10. (Suggested) Neural Turing Machines [Graves et al., 2014]
11. (Suggested) Meta-Learning with Memory-Augmented Neural Networks [Santoro et al., 2016]

We summarize several main takeaway messages from group discussions below:

1 Importance of Long-Range Interactions

Suppose that there are three units of information a, b, and c. The first two units of information, a and b, are close together in some sequence but the third unit of information, c, is very far away. If information from c is critical for contextualizing the information from a and b, models that focus only on local interactions would fail to appropriately contextualize a and b. More concretely, textbook-based question-answering is one example where modeling long-range interactions is critical. Textbooks are very long documents compared to what today's models can typically handle. Furthermore, the content in a textbook often builds very directly on earlier content that may be hundreds of pages in the past, while also requiring the model to associate text with tables and figures. Without the ability to capture long-range multimodal interactions, current models would likely fail in long-range multimodal QA. Similarly, long-range interactions are also necessary for multi-hop QA, long-form document/video/audio understanding, and long-range dialog generation.

2 Challenges in Modeling Long-Range Interactions

In Table 1, we summarize several unique challenges in modeling long-range interactions in multimodal tasks and listed existing solutions attempting to address them:

Table 1: Unique challenges and solutions in modeling long-range interactions in multimodal tasks

Challenges	Definition & Solutions
Data alignment	Using a shared memory for multiple modalities requires strong alignment in data, especially in tasks such as video understanding. Existing attempts to create alignment includes the use of attention mechanism [Zeng et al., 2017].
Cross-modal interactions	Performing early fusion on unimodal memory or having separate memory modules for each modality may not fully capture cross-modal interactions at the early stage. Late fusion technique may be better but it also requires explicit data alignment.
Segmentation	Segmenting sequences into meaningful sub-units can shorten the overall length of the sequence, but determining meaningful segmentation strategies across diverse modalities is challenging. Domain expertise can guide segmentation techniques but requires prior knowledge about the data.
Quadratic complexity of self-attention	Self-attention requires $O(n^2)$ computations for a sequence of length n [Vaswani et al., 2017] which makes scaling self-attention to long sequences challenging. Alternative formulations have been proposed that leverage things like sparsity to improve the scaling behavior [Tay et al., 2021].
Long tail	Important long-range interactions likely occur with lower frequency than local interactions which makes it challenging to learn. Memory-based approaches that explicitly cache information may have an easier time utilizing distant information [Khandelwal et al., 2018].

Self-attention has become one of the dominant techniques for modeling long-range dependencies. However, self-attention requires $O(n^2)$ computations for a sequence of length n [Vaswani et al., 2017]. This makes scaling self-attention to very long sequences challenging. Even beyond the computational complexity of self-attention, the flexibility of the mechanism likely increases the sample complexity of learning interactions between elements. This problem is likely exacerbated for longer-range interactions which may occur with lower frequency. Tay et al. [2021] benchmarked a variety of techniques that attempt to alleviate this problem, but they typically introduce a tradeoff between runtime and performance. Efficiently and effectively modeling long-range interactions is still very much an open problem within the field.

3 Modeling Long-Term Cross-Modal Interactions

Many of the structures designed for unimodal memory can be adapted for multiple modalities. For example, a compositional memory that fuses linguistic information and visual features [Jiang et al., 2015] was useful for VQA when there is a strong alignment in the input modalities. As pointed out by Lample et al. [2019], memory tends to work better on abstract representations than low-level features. A natural question is: do we also need to model low-level features or should multimodal memory and interactions be mainly on high-level abstractions?

Another question is how to combine design decisions in fusion and memory. One could perform early fusion

and store multimodal evidences, or store unimodal evidences and perform late fusion on extracted memory. Furthermore, the concept of "long-range" may represent different temporal scopes for different modalities, which makes alignment particularly challenging in capture long-range interactions [Zeng et al., 2017].

4 Memory-Based Approaches

Memory-based approaches are another interesting line of work that can help model long-range interactions [Chen et al., 2021], enable better reasoning capabilities [Xiong et al., 2016], and efficiently increase the parameterization of models [Lample et al., 2019]. Many approaches along this path involve storing explicit memory states which are updated as relevant information is seen.

However, memory-based approaches come with similar challenges to some of the techniques proposed to model long-range interactions. Information will likely be compressed as it is stored in the memory states and valuable information may be lost. This means that some critical long-range interactions may not be captured. There is also the challenge of designing retrieval approaches to access the correct memory states given some new information. Multimodal memory further brings unique challenges in requiring alignment of unimodal evidences into cross-modal interactions to be effectively stored in memory.

5 Connections to Neuroscience

Insights from human memory can also provide several inspirations for designing better memory modules in multimodal tasks. For humans, the hippocampus regulates both long-term and short-term memory. In particular, we have episodic memory that focuses on specific information (what has happened, where, and when). Meanwhile, abstract memory stores abstract concepts of experience and links pieces of information to create new knowledge. Therefore, the separation of short-term and long-term memory in human memory may suggest separate modules for modeling short-range and long-range interactions of modalities.

Another inspiration from neuroscience comes from memory updates. We observe in human brains that long-term memory is slowly updated as we gather more information while short-term memory is updated at a much faster pace. This idea is similar to "fast weights" that model the recent past in Ba et al. [2016]. Both storing and querying of memory in human brains can be multimodal in nature, which suggests a similar pattern in model learning. For example, an olfactory signal can trigger the memory of a vision and vice versa. Finally, content in short-term memory after being reinforced multiple times can become a long-term memory, which indicates that the connections between short-term and long-term memory are not static.

6 Future Directions

Khandelwal et al. [2018] observed different sensitivities of models to tokens with short-range and long-range dependencies. Specifically, changes in long-range interactions are more ambiguous and harder to detect. Therefore, for small changes such as a negation that affect long-range interpretation, how should it be stored in the long-term memory module? Another important question is about storing and updating memory. Can memory modules be updated during test time? Existing memory modules like the differentiable neural dictionary (DND) [Pritzel et al., 2017] are only learned in training.

Although many alternative formulations of self-attention that scale better to long sequences have been proposed, the comprehensive benchmark of such approaches by Tay et al. [2021] demonstrated that there is still a meaningful trade-off between scalability and performance. Designing improved attention mechanisms or alternative architectures altogether that minimize these trade-offs is still an open research direction.

References

Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. *Advances in Neural Information Processing Systems*, 29, 2016.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer

for vision-and-language navigation. In *NeurIPS*, 2021.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic memory in lifelong language learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/f8d2e80c1458ea2501f98a2cafadb397-Paper.pdf.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014. URL http://arxiv.org/abs/1410.5401.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, Oct 2016. ISSN 1476-4687. doi: 10.1038/nature20101. URL https://doi.org/10.1038/nature20101.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1280. URL https://aclanthology.org/D18-1280.

Aiwen Jiang, Fang Wang, Fatih Porikli, and Yi Li. Compositional memory for visual question answering. *arXiv preprint arXiv:1511.05676*, 2015.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, 2018.

Guillaume Lample, Alexandre Sablayrolles, Marc' Aurelio Ranzato, Ludovic Denoyer, and Herve Jegou. Large memory layers with product keys. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/9d8df73a3cfbf3c5b47bc9b50f214aff-Paper.pdf.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control, 2017.

Jack Rae and Ali Razavi. Do transformers need deep long-range memory? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7524–7529, 2020.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1842–1850. JMLR.org, 2016.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Junbo Wang, Wei Wang, Yan Huang, Liang Wang, and Tieniu Tan. M3: Multimodal memory modelling for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7512–7520, 2018.

Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406. PMLR, 2016.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.