11-877 Advanced Multimodal Machine Learning **Spring 2023**

Week 12: Brain and Multimodal

Instructors: Louis-Philippe Morency, Paul Liang *Synopsis Leads: Leena Mathur, Yilin Wang*

Edited by Paul Liang *Scribes: Pratik Joshi, Mehul Agarwal*

Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

Summary: Multimodal machine learning is the study of computer algorithms that learn and improve through the use and experience of multimodal data. The human brain perceives, represents, and reasons about multimodal stimuli, as well, motivating this week's discussion of the brain.

In week 12's discussion session, the class focused on discussing connections among neuroscience, cognitive science, and machine learning. The following was a list of provided research probes:

1. What are the main takeaways from neuroscience regarding unimodal and multimodal processing, integration, translation, and co-learning, that are less known in conventional multimodal ML?
2. How can these insights inform our design of multimodal models, following the challenges we covered previously (connections, interactions, co-learning, pre-training, reasoning etc.)?
3. To what extent should we design AI models with the explicit goal to mirror human perception and reasoning, versus relying on large-scale pre-training methods and general neural networks?
4. How does the human brain represent different modalities (visual, acoustic, touch)? Are these different modalities represented in very heterogeneous ways? How is information linked between modalities?
5. What are several challenges and opportunities in multimodal learning from brain imaging modalities? How do these modalities introduce new challenges not seen in conventional language and vision research?
6. What are some ways in which multimodal learning can help in the future analysis of data collected in neuroscience? What unique challenges arise in this new research direction, beyond classical multimodal learning?

As background, students read the following papers:

1. Multisensory Integration: Methodological Approaches and Emerging Principles in the Human Brain [Calvert and Thesen, 2004]
2. Towards Multimodal Atlases of the Human Brain [Toga et al., 2006]
3. Multimodal Images in the Brain [Kosslyn et al., 2010]
4. Crossmodal Processing in the Human Brain: Insights from Functional Neuroimaging Studies [Calvert, 2001]
5. Modulations of Visual Perception by Sound [Shams et al., 2004]
6. Multi-Modal Perception [Hollier et al., 1999]
7. On the Link Between Conscious Function and General Intelligence in Humans and Machines [Juliani et al., 2022]
8. Multimodal Mental Imagery [Nanay, 2018]
9. Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link(s) in Complex Mental Illness [Calhoun and Sui, 2016]
10. A Multimodal Cell Census and Atlas of the Mammalian Primary Motor Cortex [manuscript editors et al., 2021]

We summarize several main takeaway messages from group discussions below:

1 Multimodal Aspects of the Brain from Neuroscience

Neuroscience research on brain regions has determined that different areas of the brain are responsible for processing different types of information [Toga et al., 2006, Sepulcre et al., 2012]. Information from different modalities is often processed individually before being are passed to regions in the brain designed to process cross-modal interactions and information.

The timing of stimuli across modalities plays a vital component in the brain's response to the stimulus [Bruns and Getzmann, 2008]. For example, temporally-proximate stimuli tend to enhance the brain's response to that joint set of information. However, the temporal structure of data is usually not properly addressed in most ML models, and such temporal information is not available in most ML datasets (this is an avenue for neural and cognitive inspiration to motivate ML research in this area).

2 Neural and Cognitive Inspiration for Machine Learning

Concepts and theories from neuroscience and cognitive science can inform approaches in multimodal machine learning. The bitter lesson [Sutton, 2019] has the potential to emerge while exploring neurally and cognitively-inspired approaches; however, this possibility should not deter scientists from exploring the multimodal human brain as inspiration for multimodal AI systems.

We discussed a computer vision algorithm for **streaming perception** [Li et al., 2020] which performs forecasting during time intervals as a way to integrate **predictive coding** into ML approaches for perception. The approach in this paper was inspired by predictive coding [Huang and Rao, 2011] and is one example of a cognitively-inspired architecture that has been demonstrated as useful for computer vision tasks.

When creating multimodal AI systems, it is worth considering the **temporal aspect** of information processing because neurons do have different responses with respect to the temporal ordering and duration of stimuli across modalities (e.g., seeing something and then hearing something will be represented differently in the cortex vs hearing something and then seeing something) [Mauk and Buonomano, 2004].

Exploration is not enough without **interaction** and **embodiment**. Embodied social interaction (including interaction in pre-linguistic phases of development) is key for human babies to learn language and acquire other linguistic and behavioral skills (supported by decades of psycholinguistic research) [Snow, 1989]. Given these findings, it is worth researching approaches for machines to explore the physical world with curiosity during pretraining in order for their capabilities and understanding to be grounded in the real world [Bisk et al., 2020].

Just as the brain spends **variable compute** on different tasks, we can introduce this into modeling. We discussed **adaptive computation time** as one cognitively-inspired approach for this idea in RNNs and language models [Graves, 2016].

3 Should AI Models Mirror Human Cognitive/Neural Processes?

We have come to an (empirical) consensus that when working on a newly-established task, neuroscience findings can be helpful in building an initial model. In such cases, it may be intuitive to ask the model to mimic human behavior. Also, for tasks that require a long chain of reasoning (e.g., mathematics), it might be preferable to build more human-like models, as logic (especially deductive logic) is hard to be captured using probabilistic modeling approaches.

3.1 Can AI be better than humans?

AI could be better than humans in more specific tasks, especially those tasks that require a lot of memorization and bottom-up reasoning. This is due to the fundamental differences in the learning paradigm for humans and AI. Human learning resembles reinforcement learning: we receive rewards based on our actions, and we learn from feedback. The current-state AI (e.g., LLMs) are trained using primarily unsupervised learning methods (and then finetuned using supervised learning to align with human instruction). Such a learning paradigm

makes them excel at capturing patterns in the data, yet less so in top-down reasoning and generalization.

More forward-looking, the group believed that AI can be better than humans. The group believed that AI research is searching for potential forms of general intelligence in a broader space, where the structure of humans may only be a local optimum.

3.2 Calibration

When facing uncertainties in reasoning, humans are able to reflect on their cognitive process and calibrate any logical fallacies while responding to the stimulus. However, most generative AI produces the output through only one forward pass, which does not enable the model to reflect on its output. This problem is especially prominent for LLMs, as they often tend to hallucinate the answer when encountering uncertainties.

3.3 What stops us from developing human-like models?

The major obstacle in creating human-like agents is the limited knowledge we have about the brain. Moreover, modern ML models usually concentrate on a single modality or task, which does not provide enough motivation for building a human-like agent.

References

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. *arXiv preprint arXiv:2004.10151*, 2020.

Patrick Bruns and Stephan Getzmann. Audiovisual influences on the perception of visual apparent motion: exploring the effect of a single sound. *Acta psychologica*, 129(2):273–283, 2008.

Vince D Calhoun and Jing Sui. Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(3):230–244, 2016.

Gemma A Calvert. Crossmodal processing in the human brain: insights from functional neuroimaging studies. *Cerebral cortex*, 11(12):1110–1123, 2001.

Gemma A Calvert and Thomas Thesen. Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology-Paris*, 98(1-3):191–205, 2004.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Mike P Hollier, Andrew N Rimell, David S Hands, and Rupert M Voelcker. Multi-modal perception. *BT Technology Journal*, 17(1):35–46, 1999.

Yanping Huang and Rajesh PN Rao. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593, 2011.

Arthur Juliani, Kai Arulkumaran, Shuntaro Sasai, and Ryota Kanai. On the link between conscious function and general intelligence in humans and machines. *arXiv preprint arXiv:2204.05133*, 2022.

Stephen M Kosslyn, Giorgio Ganis, and William L Thompson. Multimodal images in the brain. *The neurophysiological foundations of mental and motor imagery*, pages 3–16, 2010.

Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards streaming perception. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 473–488. Springer, 2020.

Principal manuscript editors, Analysis coordination, Integrated data analysis Armand Ethan 42 Yao Zizhen 5, ATAC seq data generation, processing Fang Rongxin 45 Hou Xiaomeng 10 Lucero Jacinta D. 18 Osteen Julia K. 18 Pinto-Duarte Antonio 18 Poirion Olivier 10 Preissl Sebastian 10 Wang Xinxin 10 97, Epi retro-seq data generation, processing Dominguez Bertha 53 Ito-Cole Tony 1 Jacobs Matthew 1 Jin Xin 54 99 100 Lee Cheng-Ta 53 Lee Kuo-Fen 53 Miyazaki Paula Assakura 1 Pang Yan 1 Rashid Mohammad 1 Smith Jared B. 54 Vu Minh 1 Williams Elora 54, et al. A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature*, 598(7879):86–102, 2021.

Michael D Mauk and Dean V Buonomano. The neural basis of temporal processing. *Annu. Rev. Neurosci.*, 27:307–340, 2004.

Bence Nanay. Multimodal mental imagery. *Cortex*, 105:125–134, 2018.

Jorge Sepulcre, Mert R Sabuncu, Thomas B Yeo, Hesheng Liu, and Keith A Johnson. Stepwise connectivity of the modal cortex reveals the multimodal organization of the human brain. *Journal of Neuroscience*, 32 (31):10649–10661, 2012.

Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. Modulations of visual perception by sound. 2004.

Catherine E Snow. Understanding social interaction and language acquisition; sentences are not enough. 1989.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1), 2019.

Arthur W Toga, Paul M Thompson, Susumu Mori, Katrin Amunts, and Karl Zilles. Towards multimodal atlases of the human brain. *Nature Reviews Neuroscience*, 7(12):952–966, 2006.