Multi-modal deep learning with NLP focus

My two takeaways from this lecture are 1) multi-modal foundation models are coming and 2) framework for how multi-modal fusion is implemented.

I had no idea how furiously researchers are working multi-modal foundation models such as FLAVA that generalizes across datasets. As a working professional it's difficult to see the trends in Artificial Intelligence (AI) research because it is progressing so quickly. It was surprised to hear that *'we're running out of text data'* and that means that multimodal is the next logical direction. While I knew about the high-profile multi-modal models like DALL-E and Stable Diffusion, I was surprised to hear about the progress in audio (Whisper) and video (MERLOT). I was disappointed to hear that excitement has waned in multi-modal learning in simulated environments because I work in the robotics industry and was more optimistic.

I liked Douwe's breakdown of the five ways to implement multi-modal fusion: similarity, linear/sum, attention, multiplicative, and bilinear. I liked that he pointed out that *when* to do fusion is important: early fusion (mix inputs), middle fusion (concatenate features) or late fusion (combine final scores). It's a neat framework. I learnt that images are high bandwidth information while language is low bandwidth. This is a fact that I hadn't appreciated. Other surprising nuggets of knowledge I learnt include: multimodal pretraining doesn't work, there exists open source multi-modal data sources like Laion, and that Winoground showed that SOTA multi-modal models are still very limited.

Finally, it was extremely encouraging to hear that the state of the art is entirely based on the transformer architecture we are learning in CS224n!