Advanced NLP Project Proposal

Alec - ameade@g.harvard.edu | Nikhil - nsingh1@mit.edu | Ian - iapalm@mit.edu

Speech driven image generation

Recent work has shown that neural models can generate convincing images given a text caption. We are interested in extending this capacity to speech. In particular, we are interested in building a speech-driven image generation network that learns to synthesize convincing images depicting spoken concepts. To begin with, we plan to use datasets with spoken captions such as *Places Audio Captions* and *SpokenCOCO* to produce domain-specific results. We additionally plan to conduct experiments to interpret the behavior of the model, and review any interesting learned multimodal representations.

Evaluation

One method of evaluating our model is to use an off-the-shelf image captioning model to generate captions for the images produced by our model. We can use a metric like BLEU or ROUGE to evaluate how similar the text caption is to the ASR transcript of the original speech signal. While this is a rough estimate for performance, it should give usas a general idea of how well our model is performing during training and evaluation. We also plan to compare with existing work like S2IGAN on traditional GAN metrics measuring diversity and generated image quality.

If it is difficult to achieve reasonable performance we may pivot our project toward reimplementing Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input and adapt it with modern architectural components such as multi-modal transformers.

Dataset:

- Places Audio Captions
- SpokenCOCO

Available Compute:

Between our team we have access to multiple VMs as well as a compute cluster with several Titans and 1080s which should be sufficient for the proposal above.

Other Ideas:

If we are unable to make significant headway on our project or if the teaching staff has preferences, we also are interested in exploring continuous sign language to text translation. There are numerous available datasets in this space including How2Sign**,** BSL Corpus and SIGNUM**.** In this case we would consider building on the following Joint End-to-End Sign Language Recognition and Translation Paper which uses an encoder-decoder transformer architecture to first convert sign language video to an intermediary gloss representation and then ultimately decode that representation to text translations.