Piseth Ky

Project Proposal:

Create an automatic image captioning with an audio readout system. This type of system will be useful primarily in IoT products (including self-driving vehicles) which need to understand the images it is seeing and to communicate with people. In order for these products to feel more natural and seamless it is vital for AI systems to be able to communicate in natural language what they are seeing.

Data Required:

For the first part annotated captioned images will be required, the COCO and flickr30k datasets should be sufficient for this. For the audio portion, google provides a high quality audio set.

Methodology:

Microsoft Research recently published a unified Vision-Language Pre-training (VLP) model. My model architecture will most likely use this model. The audio readout model will essentially be an additional text to speech system which processes the output from the VLP model. Both submodels will utilize deep learning -- primarily transformer architectures. Transfer learning will be tested, and may be used to reduce training time/improve quality of output.

Deliverables:

I will deliver an API and web page for this. Primarily I envision users will upload a photo and it will display the caption as well as read it out. Additionally I hope users will rate how well it was captioned or provide what they think is the best caption. This can add to potential future training data. For extra credit, I may allow uploading of a second image and do a neural style transfer for fun. I am also interested in seeing if we can transfer a music style to an image.

Resources Required:

Resources required is TBD, as with most vision tasks a GPU is all but required. Early testing/prototyping will help determine resources.