CIS 519: Project Proposal: Automatic Captioning

Shiva Suri, Kristin Li, and Rohith Venkataraman

Objective: Our ultimate goal for this project is to successfully be able to caption an English video of a single person speaking face-front to a camera. We envision implementing the following phases (in order and/or parallelized across group members) through the course of the project:

1. Image Sequence → Subtitles (Lip reading)
2. Sound → Subtitles (Audio analysis)
3. Using both (1) and (2) as features for final model.

We have also compiled a list of possible enhancements we could explore, given that we achieve this baseline:

1. Bucket word identifications by strength, and relay to user (interface to user by dimming less confident words, inspired by YouTube's auto-captioning)
2. (a) Handle multiple individuals facing camera instead of just a single person, and be able to identify who is speaking.
   (b) Auto-caption any video using variable-length features (i.e. if lips can be read, read them, and if there is audio, capture it).

These latter goals are certainly hyper-stretch ones, but we are ambitious, and will achieve as much as we can, maintaining the baseline as priority.

Today's approach: To no one's surprise, Google has chipped into this industry and essentially set the standard. YouTube auto-captioning works by combining Google's Automatic Speech Recognition (ASR) with YouTube's captioning system. [1] Google Cloud's Speech-to-Text API [2] uses a neural net in its conversion scheme. The API has separate optimizations for videos and phone calls, as well as a generic model. Overall, this auto-captioning focuses on the audio-to-speech portion of auto-captioning. Google does not reveal the details behind its neural nets, and the API is a paid subscription.

Novelty behind our approach: We are approaching the speech-to-text from primarily a visual approach as opposed to strictly an audio-based approach like a lot of models out there (only the very best are mentioned in the previous section), and this skews the real-world applications of our final model to the ones described in the next section. Unlike the Cornell University deep learners, we will also incorporate audio as features in our model to hopefully result in even better performance. Thus, the emphasis on visual features in our model distinguishes us from Google a bit, and the incorporation of audio distinguishes us from the Cornell University researchers. We realize that there is no 'beating Google' at their own game, but hopefully our model can serve a narrower population, and help us gain acumen in computer vision, deep learning, and audio analysis as a result of this project. [3]

Target audience and Impact: This project is particularly interesting because we aim to utilize both image sequences and audio. Depending on our success, this could represent an avenue for improving communication for people with impaired hearing. Additionally, this incorporation of the actual images from the video could present other applications in analyzing security footage or understanding participants in video chatting when the call drops a bit (cases where the images of the video remain, but perhaps not the audio).

Risks and Payoffs: We also recognize that this proposal is ambitious because we are tackling the problem from two angles - video images and audio - which will require an additional layer of understanding and parsing our instances in preparation for training. This being said, we believe that the incorporation of both types of features will lend more accuracy to our model.

Sources of Real-world Data:

The first collection is a Kaggle dataset with 3000 instances of lip movements. The second is a collection of 3 different datasets from 3 British publications in the lip-reading space. It will be particularly interesting to examine the accent on the performance and challenges we encounter while training the models. The third dataset we plan on using is a celebs saying words from Vox interviews. This will provide the diversity of different pronunciations for common words as well as variety of speech. Finally the last dataset is several hundred hours of aligned wikipedia being read. Finally we will employ the industry standard for text performance measure: Word Error Rate. $WER = \frac{S+I+D}{N}$ where the S,I,D stand for substitutions, insertions, deletions and N is the number of words.

Source Links:

1. `https://www.kaggle.com/apoorvwatsky/miraclvc1?fbclid=IwAR2qe6rc`
   `BQYjn5RDbWLBztGVZm75tCVok7FR_m6oA5gCdPA0yc8bp-B88wA`

2. `http://www.robots.ox.ac.uk/~vgg/data/lip_reading/?fbclid=IwAR1XKkvkXpaeO3a0NTvxlvPAV`
   `_ZHCrdBJ8sWtZssK2JKP6Q550E6qnGeNiQ`

3. `http://www.robots.ox.ac.uk/~vgg/software/yousaidthat/`

4. `https://nats.gitlab.io/swc/`

Notes

[1] *Automatic captions in YouTube* (2009)

[2] Google Cloud's Speech-to-Text API

[3] However, in June 2018, Deep Learning researchers at Cornell University published a paper*Deep Lip Reading: a comparison of models and an online application* (2018) on visual speech recognition (i.e. lip reading), exploring ”a recurrent model using LSTMs; (ii) a fully convolutional model; and (iii) the recently proposed transformer mode” *(Triantafyllos Afouras,Joon Son Chung,Andrew Zisserman).*