LIPS: Let's think of a better title...

John C. Merfeld
Department of Computer Science
Boston University
Boston, MA 02215
jcmerf@bu.edu

Allison Mann
Department of Computer Science
Boston University
Boston, MA 02215
ainezm@bu.edu

Abstract

The abstract paragraph should be indented ½ inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Introduction

2 Related work

3 Method

Transforming video footage into a lip-reading prediction requires a significant processing pipeline. In this section, we discuss the training dataset, how it was prepared for modeling, and the overall architecture of our model.

3.1 Data

The Oxford-BBC Lip Reading Sentences 2 (LRS2) Dataset was selecting for training. Details of its creation can be found in [1]. A brief summary of the data will follow.

The atomic units of raw data are video files, each around ten seconds long. The video clips come from BBC news and talk shows, but each frame's viewing window is shrunk to track the face of an individual speaker. A single speaker may span multiple clips, but LRS2 was designed specifically to include as many different faces as possible. Accompanying each video is a like-named metadata text file; this file contains an ordered list of words spoken in the clip along with timestamps indicating when the speaker starts and stops saying each word. These times are precise to the hundredth of a second.

We did not use the entirety of LRS2; we limited our study to the "pre-train" segment because the "main" segment's text files do not include word timestamps. Still, the model's potential training examples encompassed over 2 million utterances of 41,427 unique words; 47GB of data in all.

3.2 Feature Processing

Consider an example pipeline whose output is a model that, given a video clip, predicts whether the speaker in that clip is saying the word "yes" or the word "no." So as to act within the realm of computational feasibility, we assign a parameter N_{files} as a stand-in for the size of this model's training set.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

First, a dictionary is prepared, assigning unique integer values to the strings 'YES' and 'NO'. Next, N_{files} text files are parsed word by word. If a word in the file is 'YES' or 'NO,' it is recorded, along with its beginning and ending timestamps.

Next, the videos corresponding to those N_{files} text files are loaded and converted to greyscale. If a clip contains a word belonging to $\{$'YES', 'NO$\}$, every frame between that word's two timestamps is recorded into its own 2D array of integers between 0 and 255. Thankfully, the videos in LRS2 are already of a uniform height and width, so no resizing is necessary. Together, this list of 2D arrays composes a 3D object we call a **word clip**. Note that at this stage, audio information is no longer present in the data representing the video.

Of course, each word instance may be composed of a different number of frames, depending on how long the speaker took to say the word. The model, however, will expect every input to be of the same dimension. Thus, we find the median length in frames of every word clip already created. If a word clip's frame count is higher than the median, it is evenly downsampled. Word clips that are too short have some of their frames repeated until they reach the median frame count. At the end of this preliminary process, then, we have a list of data objects pairing a word with a 3D word clip object of uniform size.

Talk about face detection with dlib...

Talk about model...

3.3 Model

4 Experimental Results

5 Conclusion

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Remember that you can use more than eight pages as long as the additional pages contain *only* cited references.**

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, A. Zisserman Deep Audio-Visual Speech Recognition arXiv:1809.02108