Future Computing Technologies Lab: Creative Inquiry

VISUAL LIP READING

December 10, 2019

Bradley Selee
Clemson University
Department of Electrical and Computer Engineering
bselee@clemson.edu

1 Introduction

The project I chose to explore was a Time series visual lip reading data set. The data set contains ten different words and ten different phrases, each spoken by fifteen people, ten females and ten males. Each word and phrase is uttered ten times as an iteration of several color and depth images. The data set was found on Kaggle but its origin can be found in the Jupyter Notebook.

The goal of this data set was to train a convolutional neural network, which classifies a word or phrase being spoken, given a set of images. I chose this project because I found image analysis interesting and wanted to explore different methods of analyzing images. Also, this is my first semester learning about machine learning and true data analysis, so I also wanted to choose a project that was doable given the limited time in a semester.

2 Materials and Methods

My project was based almost completely on the convolutional neural notebook we covered in class. Similarly, on Kaggle, there were others who also explored this data set and uploaded their notebook as a reference. There are many other components that could have been explored greatly improve my results; however, due to time, many of these were not completed and will be discussed in the results and future work.

2.1 Understanding the Data

The very first step was to understand the data I was working with. This is a necessity in order to load in the images and group these by classification.

ID	Words		ID	Phrases
1	*Begin*		1	*Stop navigation.*
2	*Choose*		2	*Excuse me.*
3	*Connection*		3	*I am sorry.*
4	*Navigation*		4	*Thank you.*
5	*Next*		5	*Good bye.*
6	*Previous*		6	*I love this game.*
7	*Start*		7	*Nice to meet you.*
8	*Stop*		8	*You are welcome.*
9	*Hello*		9	*How are you?*
10	*Web*		10	*Have a good time.*

Figure 2.1: All classifications in data set [1]

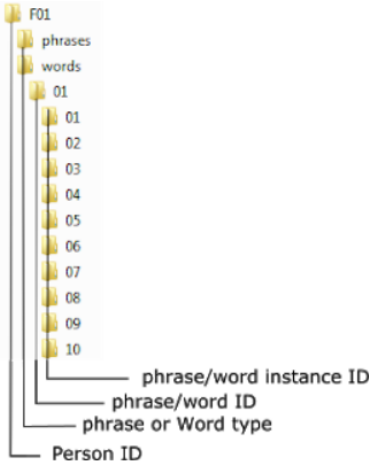

Figure 2.2: Data set folder structure [1]

Figure 2.1 shows all of the possible labels that can be classified by the images. In this project, I only used the images of people uttering words, not phrases. Figure 2.2 shows the structure of the data set. This is necessary in order to load the data into python.

2.2 Loading the Data

Once the data was understood, the data was loaded into Palmetto and Jupyter. To me, this was the hardest part. The previous notebooks had simple functions to acquire the data, but this data had to be manually loaded. In my previous python experience, I have not experienced file I/O in this structure, so this took some time.

One major problem I was experiencing was GPU resource exhaustion. Initially, I thought the node I was using was too small. However, after requesting the largest node Jupyter Hub can handle, I was still receiving this error. After a while of debugging, I realized the error was with dimensions of the pictures. For the amount of data within the data set, the original dimensions of the picture, 640x480, was way too large. Therefore, while loading in the images, I resized each picture to 96x96. This resulted in a lower quality analysis.

2.3 Displaying the Data

Next, once the data was loaded, the images were randomly shuffled and displayed with their labels for verification:

Figure 2.3: Image data after resizing

Surprisingly, the images were not too distorted, however, the number of pixels around the mouth is significantly reduced and will result in lower quality results.

2.4 Create, Train, and Evaluate the CNN

Once the data set was properly loaded and all the prepossessing was finished, the convolutional neural network (CNN) was configured using the ReLU activation function. Next, the data was normalized, One-hot labels were computed, and the network began to train. Once completed, the results were analyzed and discussed in the Results section.

3 Results

Images are complex and difficult to analyze. While a human may be able to analyze and recognize a picture immediately, a computer requires a much more precise process in order to recognize an image. Therefore, specific methods of approaching this time series data will lead to better results.

This data set contained images from both males and females, each speaking against a different background. However, the most important pictures are found around the speakers mouth. To gain better results from the CNN, cropping the person's mouth and performing an edge detection on the new image would help define the region that needs to be targeted.

Due to time constraints and the difficulty of learning certain image processing, the images were not cropped nor was the edge detection used. Therefore, the accuracy was substantially decreased. While training the network, the resources of the GPU were being exhausted, even with the largest node jupyter hub can handle. Once, I discovered the problem was with the image dimensions, 640x480 was too large, each image was resized to 96x96 pixels. A resize this large also caused a great loss in the CNN's accuracy. With many of these considerations accounted for, the CNN was very poor and the accuracy, loss, and confusion matrix are shown below.

Figure 3.1: Training and Test accuracy per epoch

Figure 3.2: Training and Test loss per epoch

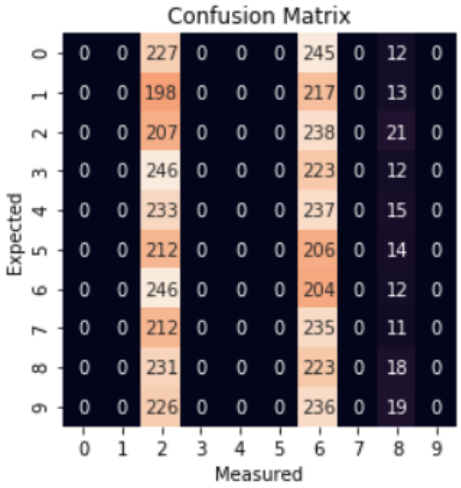

Figure 3.3: Confusion Matrix for Testing

The overall accuracy was about 9%. From viewing the confusion matrix, this percentage seems to be accurate.

4 Experience

Participating in this CI was a fantastic experience. The knowledge we were provided with was very powerful and high level. Material such as machine learning is relatively new so this is not something simple google search would be able to teach us. I enjoyed having meetings in the FCTL laboratory and being exposed to graduate students working on their projects. The jupyter notebooks were also another great tool to help us learn and most of them were very clear and not too complex; allowing us to "teach ourselves". Prior to the CI, my Python knowledge was very mediocre, as I have only worked on a

handful of small projects that included Python. But, while applying the knowledge we learned in our semester project, my python skills improved drastically. I joined this CI to gain new experience in a sub-field within my major, in an attempt to find which path I would like to pursue in my future. I have always planned on pursuing a master's degree and this CI furthered my confidence in pursuing one. After hearing about Ben's research, bioinformatics appears to be a field of computer engineering that I would like to explore in the near future. I wish Junior year was not so hard so I could put more effort into improving my CI project. This would allow me to apply theoretical concepts learned in school into real world applications.

5 Conclusions and Future Work

After understanding the amount of work that would be required to achieve accurate results, I knew from the beginning that my results would not be extremely accurate. However, as this is my first semester in this CI, I believe the main idea was to be able demonstrate that I learned something from this CI. If I had more time I would have liked to greatly improve the accuracy by performing pre-processing the images and find a better solution than shrinking the dimensions of the image.

Like many engineering students, I would like my future work to be meaningful to society. Hopefully, I will be able to begin research with Dr. Feltus and explore the realm of merging biology with computer science.

References

[1] A. Rekik, A. Ben-Hamadou, and W. Mahdi. A new visual speech recognition approach for RGB-D cameras. In *Image Analysis and Recognition - 11th International Conference, ICIAR 2014, Vilamoura, Portugal, October 22-24, 2014, Proceedings, Part II*, pages 21–28, 2014.