*Lip Reading Recognition Using ANNs*

**Christopher Papanagiotou**   
Department of Computer Science  
Ryerson UniversityToronto, Canada  
cpapanagiotou@ryerson.ca

***Abstract***

***This paper explores the hypothesis whether a lip reading recognizer can function with the use of a single standard ANN. The procedure followed to solve this problem involved the construction and processing of a dataset, the analysis of facial points and feeding of the dataset into the neural network. At the end, the results showed that further development was needed before this problem could be solved.***

1. **Introduction**

This paper describes the procedure of implementing an algorithm that will be able to recognize a spoken word through the movement of the lips. The task will be achieved with the use of a standard deep artificial neural network. The reason behind this is to inquire whether can lip recognition be achieved without the use of a CNN or RNN. As input to this procedure a dataset of sort video clips will be used.

1. **Method**
   1. **The dataset**

The data originally consisted of only 500 mp4 video clips of a time length in the range of 1-3 seconds. The videos represent one of the 100 words chosen for this task. Therefore, in total we have 5 instances per word.

* + 1. **Data augmentation**

Due to the fact that 400 videos are insufficient to train a neural network for such a task and recording videos is time-consuming, a few data augmentation techniques had to me implemented. The first technique applied to the original dataset was a steady decremental resolution reduction. The dataset was reduced in resolution down to 30% of it’s original size, any reduction further than that would not function properly for this task (see figure 1). This procedure led to the creation of an additional 5,200 videos. The second technique applied was the rotation of the whole dataset by 5 **°** (see figure 2). Any further rotation would result to faulty or no detections. The result of this procedure was the addition of 5,500 more videos. In total, the whole dataset consists of 11,100 videos.

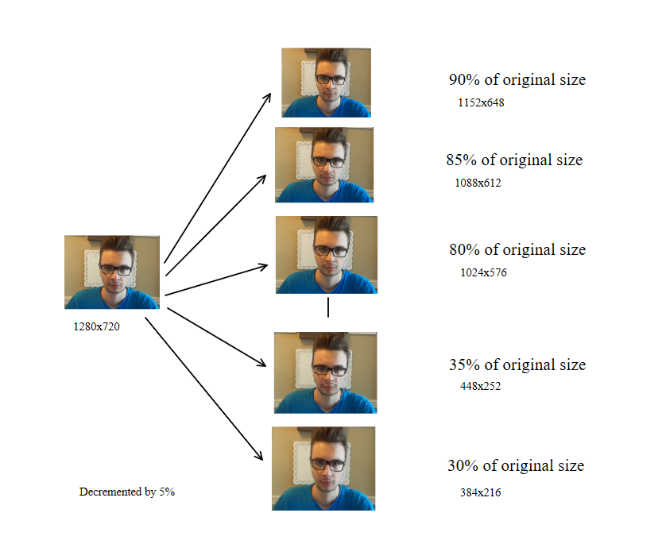


Figure 1. Resolution reduction.

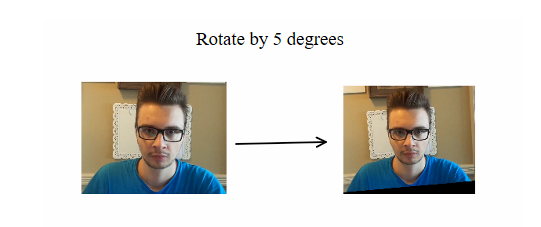


Figure 2. Video rotation.

* 1. **Face detection**

A haar cascade[1] face detector was applied to the frames of every video. The reason behind this is to minimize the area on which to try and apply the 68 point facial landmark detector.

* 1. **Facial landmarks**

After the detection of the faces a 68 point facial landmark mask[2] was applied on the face. From those 68 points only the 20 (see figure 3) around the mouth were extracted.[3] 

Figure 3. Mouth landmarks.

However, these 20 points in 2D space do not offer a lot of information in regards to the word spoken unless the actual relation between them is analyzed. Therefore, I calculated the distance between certain points of interest (see figure 4 & 5).



Figure 4. Distances between points.

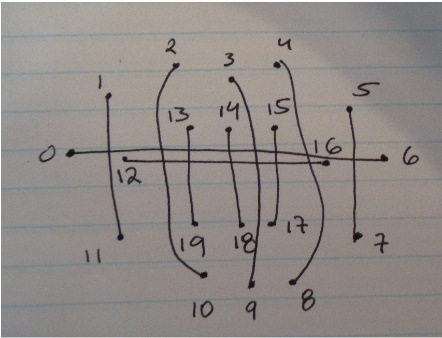


Figure 5. Point association.

* 1. **Dimensionality reduction**

The return of the above procedure will result in a 2D matrix with every row representing the 10 distances for each frame. However, the neural network is use for this project can only receive a 10 element 1D array, therefore a dimensionality reduction must take place. As a result, I took the standard deviation for each feature/distance across all the frames of the video. This method made the necessary array to pass as input to the neural network.

1. **Neural network**

A series of network configurations were attempted however the results were similar with no obvious improvement. The neural network has 10 input neurons and 100 output neurons. In regards to the number of hidden neurons, the number of layers and the learning rate, a few test runs were performed (see Table A).

Table A.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Learning rate | Hidden Layers | Hidden Units | Colour | Figure |
| 0.1 | 1 | 10 | r | a |
| 100 | b |
| 1000 | g |
| 0.1 | 2 | 10x10 | r | b |
| 100x100 | b |
| 1000x1000 | g |
| 0.1 | 3 | 10x10x10 | r | c |
| 100x100x100 | b |
| 1000x1000x1000 | g |
| 0.1 | 2 | 10x10 | r | d |
| 100x10 | b |
| 1000x10 | g |
| 0.1 | 2 | 10x100 | r | e |
| 100x100 | b |
| 1000x100 | g |
| 0.1 | 2 | 10x1000 | r | f |
| 100x1000 | b |
| 1000x1000 | g |
| 0.1 | 4 | 10x10x10x10 | r | g |
| 100x100x100x100 | b |
| 1000x1000x1000x1000 | g |
| 0.01 | 1 | 10 | r | h |
| 100 | b |
| 1000 | g |
| 0.001 | 1 | 10 | r | i |
| 0.01 | 10 | b |
| 0.1 | 10 | g |
| 0.001 | 2 | 10x10 | r | j |
| 0.01 | 10x10 | b |
| 0.1 | 10x10 | g |

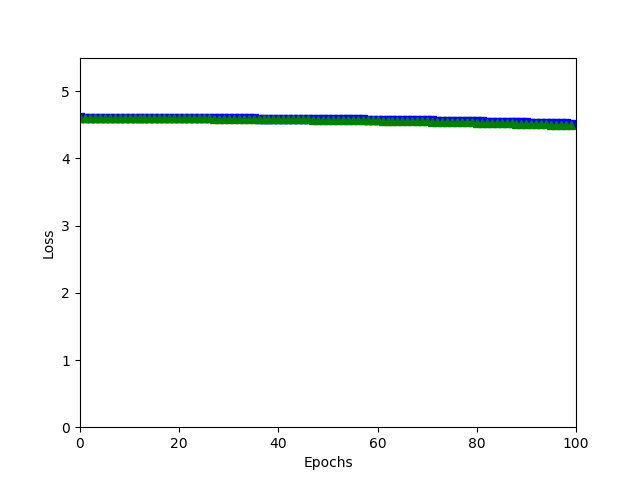


Figure 6a.

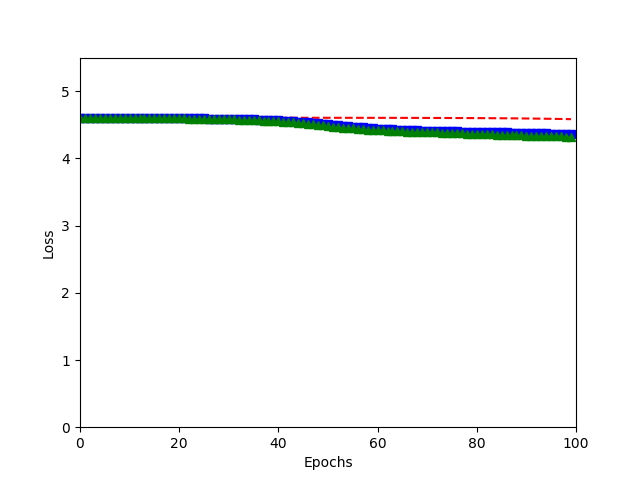


Figure 6b.

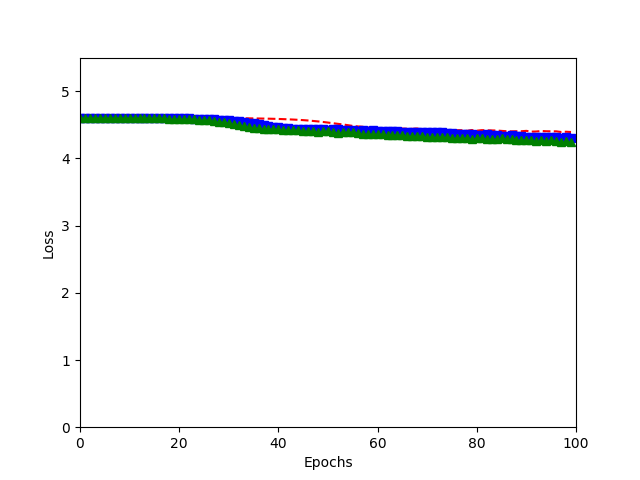


Figure 6c.

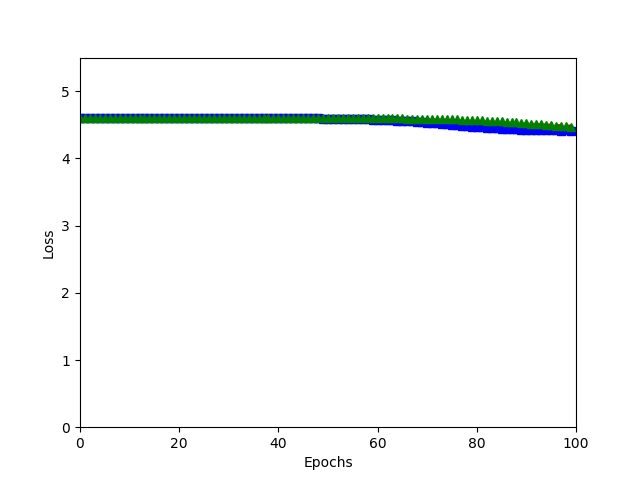


Figure 6d.

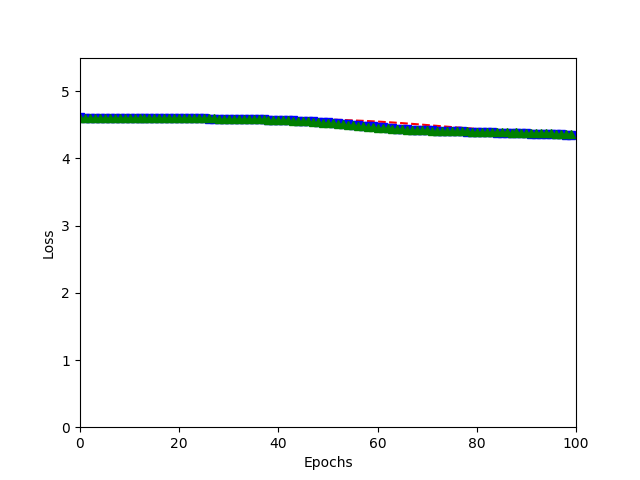


Figure 6e.

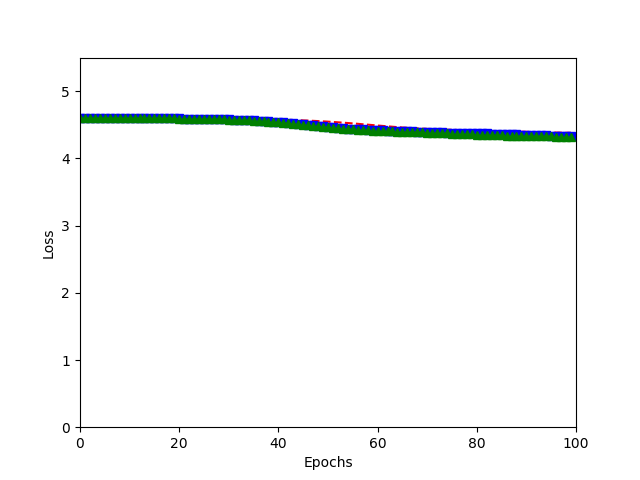


Figure 6f.

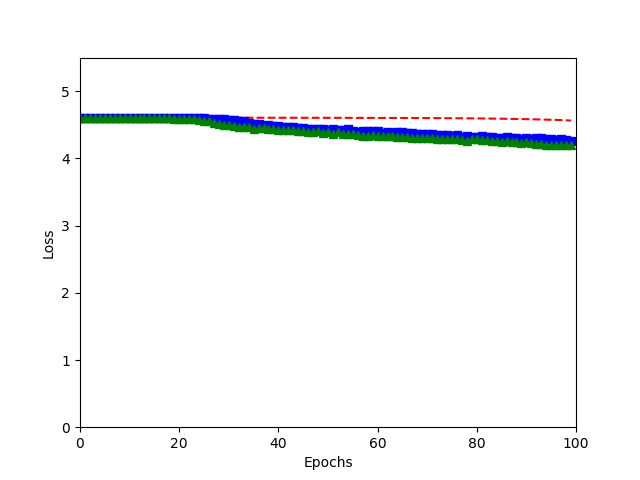


Figure 6g.

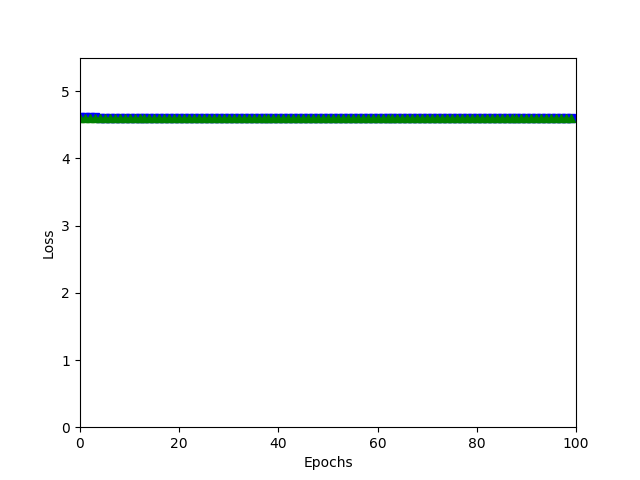


Figure 6h.

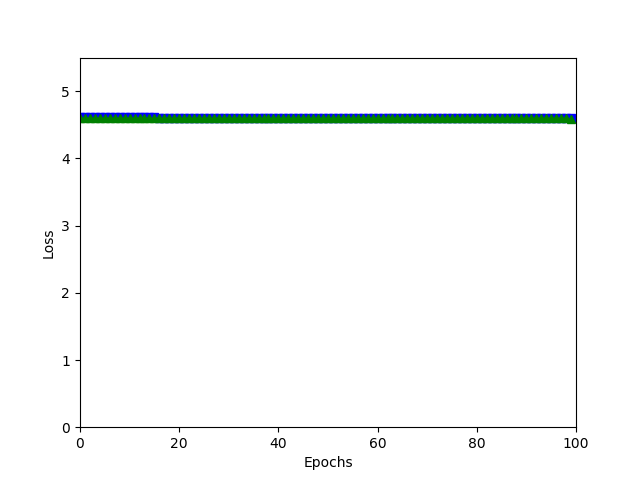


Figure 6i.

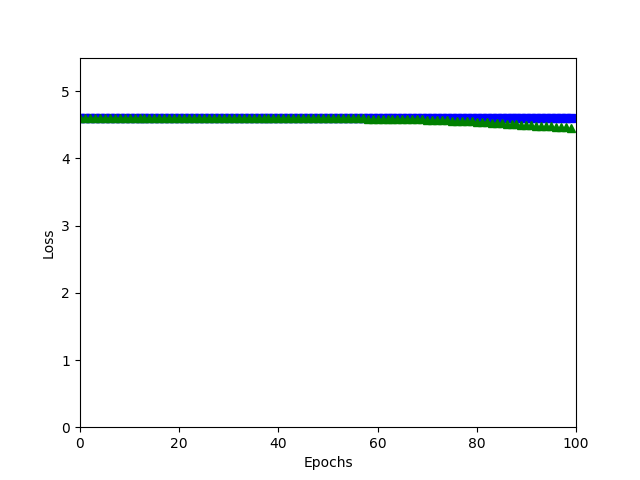


Figure 6j.

In regards to activation function used in the hidden layers the following test was performed(see Table B).

Table B.

|  |  |  |
| --- | --- | --- |
|  | Activation Function | |
|  | Sigmoid | Relu |
| Learning Rate | 0.1 | |
| Hidden Layers | 2 | |
| Hidden Units | 2 | |
| Colour | r | b |

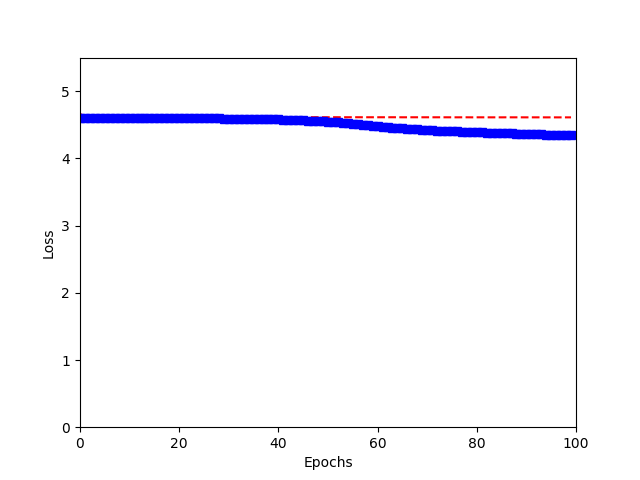


Figure 7.

From all of the tests done above I decided to go with a 4 hidden layer deep neural network with 1000 neurons per layer, at 0.1 learning rate and the use of the Relu function for the hidden layers and softmax for the output layer. Even with this configuration the accuracy of the network was at best 11%.

1. **Conclusion**

The conclusion that I arrived at was that this network needs further development and a bigger dataset to achieve any significant results. However, it seems plausible that you can construct a lip reading recognizer with just an ANN.

**References**

[1] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*. Vol. 1. IEEE, 2001.

[2] C. Sagonas, E. Antonakos, G, Tzimiropoulos, S. Zafeiriou, M. Pantic. [**300 faces In-the-wild challenge: Database and results**](https://ibug.doc.ic.ac.uk/media/uploads/documents/sagonas_2016_imavis.pdf). Image and Vision Computing (IMAVIS), Special Issue on Facial Landmark Localisation "In-The-Wild". 2016.

[3] Kazemi, Vahid, and Josephine Sullivan. "One millisecond face alignment with an ensemble of regression trees." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014.