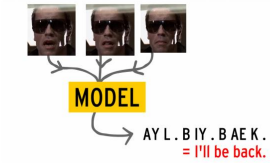
Read My Lips

Cary K. Huang (carykh@stanford.edu) and James Y. WoMa (jaywoma@stanford.edu)

Predicting

When watching noisy videos, it's often hard to figure out what somebody's saying, even if you see their lips.



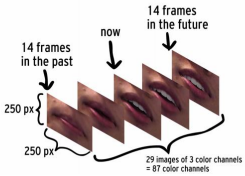
We built a neural network to convert image sequences of people speaking into the phonemes they said.

Our working model was a convolutional neural network that takes in 29 consecutive video frames centered on the speaker's lips, and outputs the corresponding phoneme spoken. (1 out of 41)

Features

Our features include the 29 input frames of the "frame neighborhood". Each frame is 250x250 px (enough to contain the mouth) and has 3 colors, so the size of the features of each element was 250x250x87.

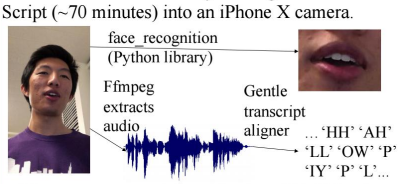
250 x 250 x 87 features



Since our task is image classification, we do not use other features besides the sequences of images.

Data

Our dataset includes Cary reading the Bee Movie Script (~70 minutes) into an iPhone X camera.

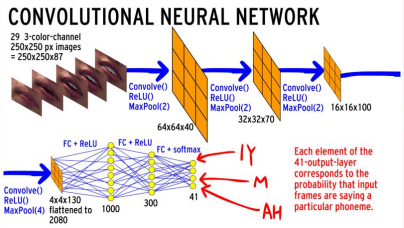


Then, we separated the video into 131,000 color images, cropped them around his mouth using a face recognizer [1], and then grouped the images into "frame neighborhoods" (see Features section).

We also converted the audio into a list of phonemes using the Gentle Transcript Aligner [3], giving us a desired output for each "frame neighborhood": a phoneme out of 41 options.

Models

Our main model was a convolutional neural network based on Chablani's auto-encoder [2]. Its structure is depicted below:

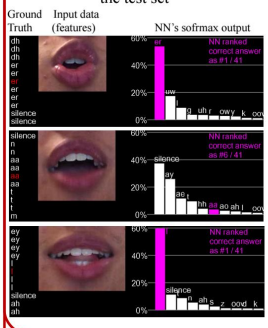


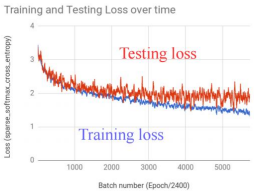
$$\text{ReLU}(x) = \max(0,x)$$

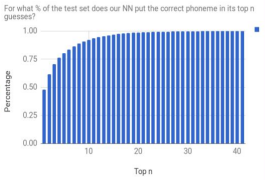
$$\text{softmax}(x) = e^x / \text{sum}(e^x)$$

Results

Analyzing the NN's performance on the test set







	0.01 epochs	3.00 epochs
Training Loss (first 120,000 frames, ~67 min)	3.4416	1.3337
Testing Loss (last 10,000 frames, ~6 min)	3.3763	1.8987

According to the graph above, our final NN guessed the correct phoneme as its top choice with 48% accuracy, and within its top 3 choices with 70.5% accuracy.

Discussion

Although an accuracy rate of 48% does not seem effective, this accuracy is high considering the NN chose from 41 different phonemes. Additionally, many phonemes look indistinguishable ('f' & 'v'). Our accuracy is also high compared to human level lip-reading accuracy, which have ranged from 12.4% to 52.3%.

In addition, we were surprised by our NN's performance in detecting silence. The NN even detected silence while Cary was inhaling with his mouth open. The NN also could distinguish between silence and 'm' in many cases.

We believe using a 29 frame neighborhood (from 14 frames before to 14 after) allowed the NN to be so effective. The mouth distorts differently depending on the sound produced, and the NN was likely able to use that information in its predictions.

Future

As of now, we only trained on videos of Cary's mouth. As an extension, we could train on other mouths so the NN generalizes.

Another extension includes turning the NN's outputted phonemes into actual sound, so it can generate audio instead of a string of phonemes.

References

[1] Geitgey, Adam, *Face Recognition*. GitHub, 2018.
[2] M. Chablani, *Autoencoders Introduction and Implementation in TF*. Medium, 2017.
[3] lowerquality, *Gentle*. GitHub, 2017.
[4] Rouzic, Michael, *The ARSS*. SourceForge, 2008.