END-TO-END AUDIO-ASSISTED LIP-READING USING LSTM

CS 577 Final Project Spring 19’

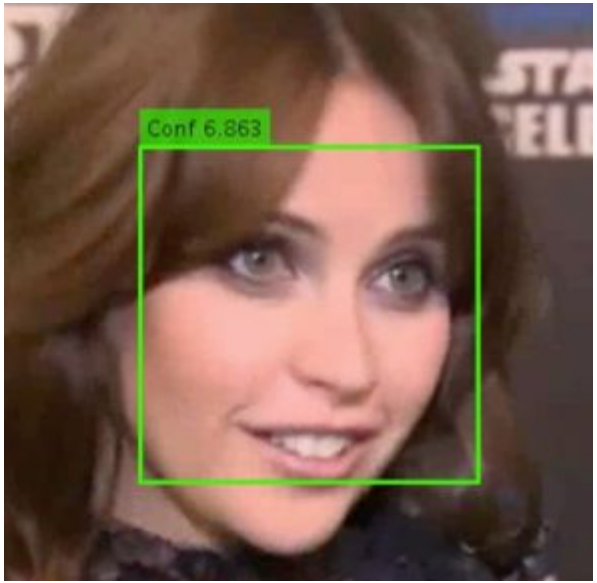
Serg Masis

A20427420

Srirakshith Betageri

A20414667

End to End Lip Reading



- Lip reading is difficult because of homophemes.
- Bear and Pear sound the same.
- Audio during training improves performance.

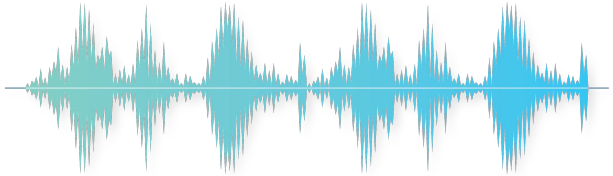
Our Main Paper

DeepMind

Oxford VGG

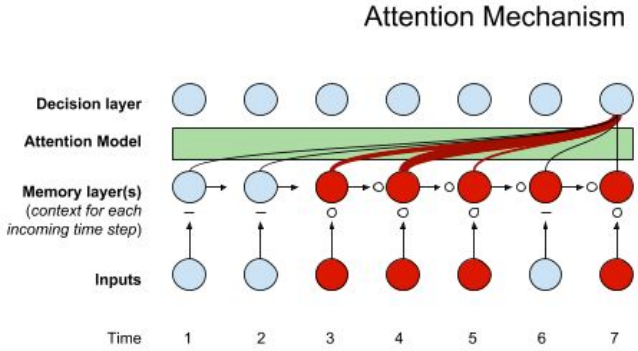
BBC Dataset

Proposed Solution



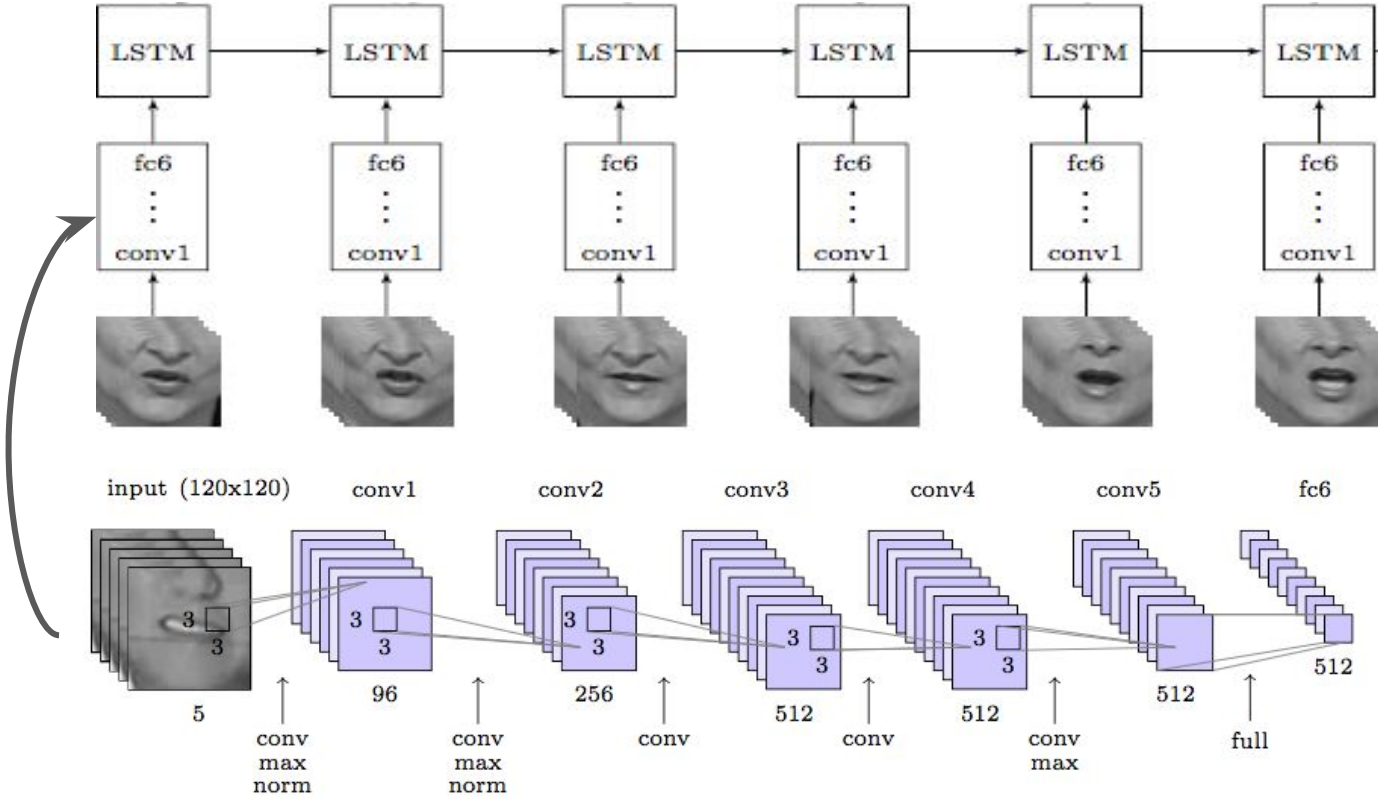
Listen: Audio Encoding

Watch: Video Encoding



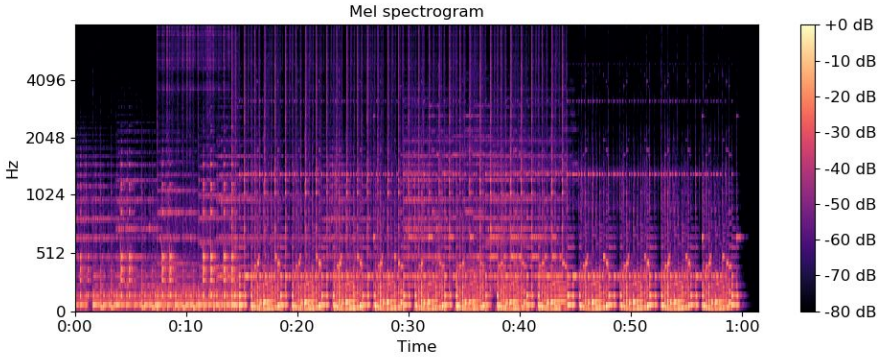
Spell: Audio Video Decoding

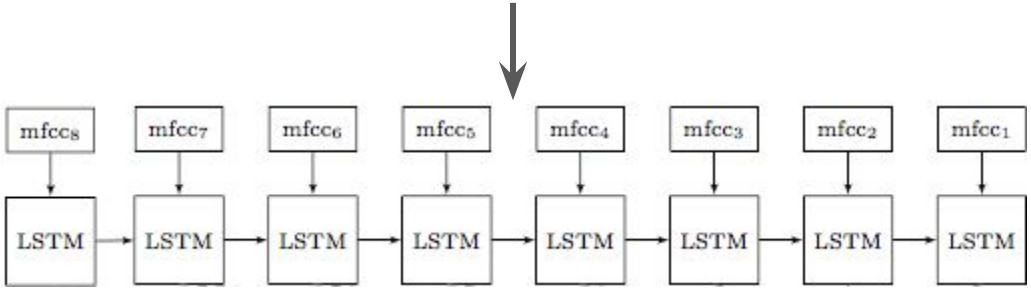
Watch


LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
fc6
conv1
input (120x120)
conv1
conv2
conv3
conv4
conv5
fc6
3
3
5
96
256
512
512
512
512
conv
max
norm
conv
max
norm
conv
conv
conv
max
full

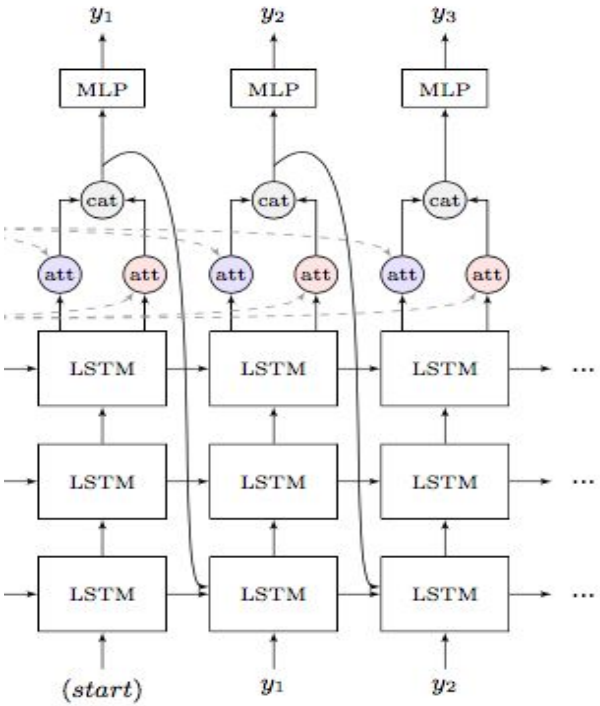
Listen

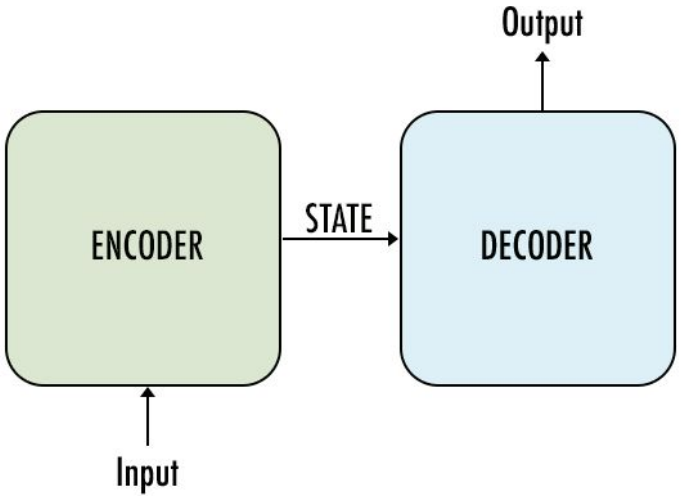
Mel-frequency cepstrum coefficients





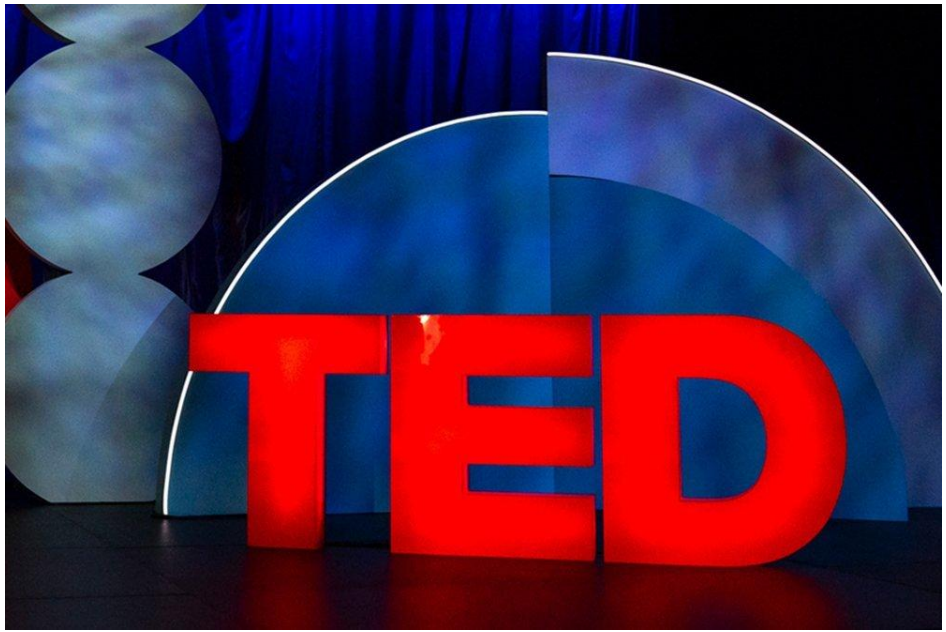
Spell


y_1
y_2
y_3
MLP
MLP
MLP
cat
cat
cat
att
att
att
att
att
att
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
...
...
...
(start)
y_1
y_2


Output
ENCODER
STATE
DECODER
Input

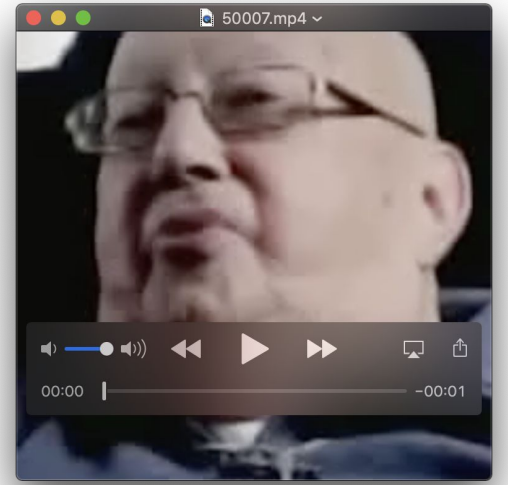
Our Dataset

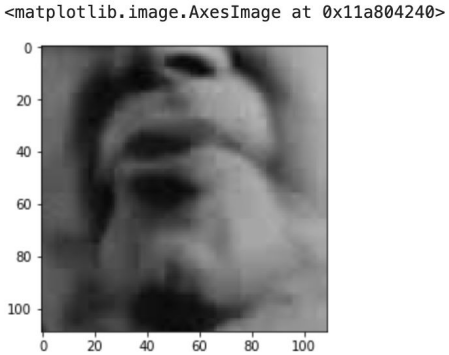
- 500 hours of video
- 4004 mp4 files
- 224x224 resolution at 25fps
- 16 bit Single Channel
- 16 kHz format
- 11 GB of data

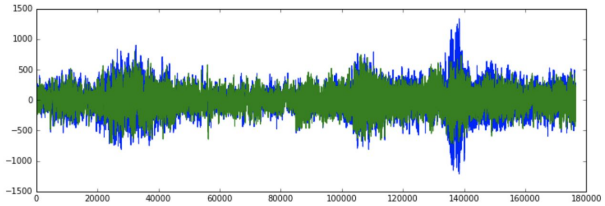


Data Preparation

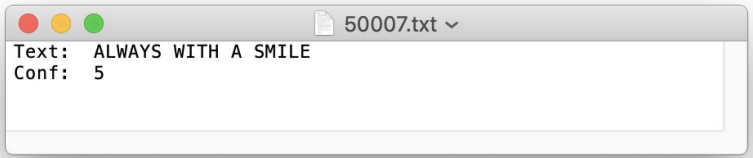
x_i







y_i



Implementation and Training

- Weights for CNN were initialised using weights from SyncNet
- Watch and Listen LSTMs have cell size of 256 each
- Spell LSTM has a cell size of 512
- Attention Network has a hidden size of 512 with a max length of 800
- Initial Learning rate of 0.1, with reduction of 10% when training error did not improve over 2000 iterations.

Results

CER = 34.6%

WER = 43.1%

Can deep learning help solve lip reading?

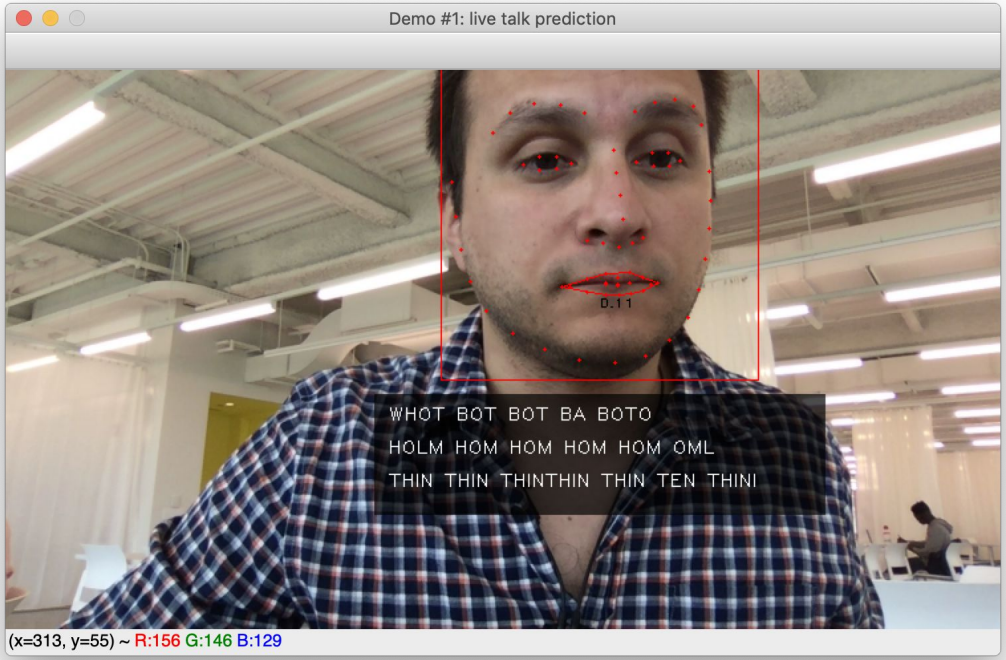
New research paper shows AI easily beating humans, but there's still lots of work to be done

By James Vincent | Nov 7, 2016, 12:50pm EST

Researchers from Google's AI division DeepMind and the University of Oxford have used artificial intelligence to create the most accurate lip-reading software ever. Using thousands of hours of TV footage from the BBC, scientists trained a neural network to annotate video footage with 46.8 percent accuracy. That might not seem that impressive at first — especially compared to AI accuracy rates when transcribing audio — but tested on the same footage, a professional human lip-reader was only able to get the right word 12.4 percent of the time.

The research follows similar work published by a separate group at the University of Oxford earlier this month. Using related techniques, these scientist were able to create a lip-reading program called LipNet that achieved 93.4 percent accuracy in tests, compared to 52.3 percent human accuracy. However, LipNet was only tested on specially-recorded footage

Demo



Bad predictions due to model that predicts probabilities on a character-basis made worse by poor frame rate and worse sound quality

Stack

GTX 1070

References

- http://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs3.html
- https://towardsdatascience.com/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d
- https://pytorch.org/
- https://stackoverflow.com/
- https://skymind.ai/wiki/attention-mechanism-memory-network

- Chung, J.S., Senior, A.W., Vinyals, O., & Zisserman, A. (2017). *Lip Reading Sentences in the Wild*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3444-3453. (ISBN: 978-1-5386-0457-1).

- Afouras, T., Chung, J.S., & Zisserman, A. (2018). *LRS3-TED: a large-scale dataset for visual speech recognition*. CoRR, abs/1809.00496.

THANK YOU!


This is RALPH