Review On:   
LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING

LIPNET is the first of its kind that has been able to do lip reading end to end. That is do a complete lip reading of a sentence. In earlier days we have seen word reading at best and the accuracy was around 86% where, LIPNET can generalize with an accuracy of 88.6% through unknown speakers in the GRID corpus, LIPNET's performance has also been compared with that of people with hearing impairment who can lip read on the GRID corpus task.

The basic idea:   
One phenome’s audio can be dubbed on top of a video of a different phenome and a new third phenome can be created.

Dataset:   
GRID corpus (2) which has audio and video recording of 34 speakers with 1000 sentences each for a total of 28 hours across 34000 sentences. The accuracy on sentence level achieved was 95.2% here.

Building Blocks:  
LIPNET is a neural network architecture for lip-reading that maps video sequences into text sequences and has been trained end to end.

1. Spatiotemporal Convolutions:   
   Which is mainly used for video analysis and study their effects on action recognition (3). STCNN can process video data by convolving across time and other spatial dimensions simultaneously.
2. Gated Recurrent Unit:  
   GRU(4) is a type of recurrent neural network (RNN) that can improvise earlier RNNs by adding more cells and gates to propagate information over more time-steps. The formula has been standardized and simplified.
3. Connectionist Temporal Classification:  
   CTC (5) loss is used in speech recognition as it eliminates the need for training data and alignment and addresses variable length sequences.

The LIPNET architecture starts with a 3x (STCNN, channel-wise dropout, spatial max-pooling). The features extracted are followed by two Bi-GRUs, then a final linear transformation is applied in each time-step followed by a softmax over the augmented CTC blank and then CTC loss.

Baseline:   
Evaluations have been made in 3 criteria:

1. Hearing impaired people:

They were introduced to the GRID corpus and they were asked to annotate the videos.

1. Baseline-LSTM:  
   Sentence level training set up was used to replicate state-of-the-art of GRID corpus.
2. Baseline-2D (6) and Baseline-NoLM show 14% and 31% poorer performance of their STCNNs compared to the 2D architectures in the dataset.

Over all we can say that LIPNET exhibits 2.3x higher performance in the overlapped compared to the unseen speaker split (1).

We can conclude that LIPNET eliminates the need to segment videos into words before predicting a sentence (1). It does not require a hand-engineered model or a separately trained sequence model.

References:

1. LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING, arXiv:1611.01599v2 [cs.LG] 16 Dec 2016
2. M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006.
3. A Closer Look at Spatiotemporal Convolutions for Action Recognition, by Du Tran1, Heng Wang1, Lorenzo Torresani1,2, Jamie Ray1, Yann LeCun1, Manohar Paluri1
4. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.
5. A. Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In ICML, pp. 369–376, 2006.
6. J. S. Chung and A. Zisserman. Lip reading in the wild. In Asian Conference on Computer Vision, 2016a.