LibiumNet: a Framework for Fast and User-friendly Automatic Lip Reading

Gbetondji J.-S. A. Dovonon
Ashesi University
Berekuso, Ghana
jean.dovonon@ashesi.edu.gh

Mustapha T. Yussif
Ashesi University
Berekuso, Ghana
mustapha.yussif@ashesi.edu.gh

Samuel Atule
Ashesi University
Berekuso, Ghana
samuel.atule@ashesi.edu.gh

Nutifafa Amedior
Ashesi University
Berekuso, Ghana
nutifafa.amedior@ashesi.edu.gh

ABSTRACT

Our aim in this paper is to provide a fast and scalable lip reading framework that is fast enough to be used for mobile inference, light enough to be fine-tuned and reused for other projects. Previous works have only focused on predicting words or alphabets or digits, and reach state of the art (SOTA) performance, which is helped by the availability of large corpus of face videos such as GRID face video corpus or the BBC Lip Reading In-the-wild (LRW) video corpus. In this paper, we develop a new architecture, entirely based on convolutions for fast and scalable lip reading. The proposed system is a combination of spatiotemporal 3D convolutions, dense 2D convolutions, 1D cnvolutions with residual connections and a dense output layer. It achieves a test accuracy of 65% on a subset of the LRW dataset with an inference speed of 0.6 seconds per video.

CCS CONCEPTS

• **IMAGE PROCESSING AND COMPUTER VISION** → *Applications*; • **ARTIFICIAL INTELLIGENCE** → **Applications and Expert Systems**; • **PATTERN RECOGNITION** → Models ,Application,Implementation,Design Methodology; • **COMPUTERS AND SOCIETY** → Social Issues.

KEYWORDS

convolutional neural networks, neural networks, Lip reading, lipreading, speechreading

ACM Reference Format:
Gbetondji J.-S. A. Dovonon, Mustapha T. Yussif, Samuel Atule, and Nutifafa Amedior. 2018. LibiumNet: a Framework for Fast and User-friendly Automatic Lip Reading. In *NORTH RIDGE '18: ACM Compass, July 03–05, 2019, North Ridge, Accra, GH*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/1122445.1122456

North Ridge '19, June 03–05, 2018, North Ridge, Accra, GH

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Lip-reading or speechreading is a technique of understanding speech by visually interpreting the movements of the lips, face, and tongue when normal sound is not available or difficult to capture. It plays an important role in human-human and human-computer interaction for communicating in noisy environments where speech recognition is difficult, and helpful for communicating with the hearing-impaired. Primarily, lipreading is often used by the deaf and hard-of-hearing people. However,there are other used cases where lipreading could be applied. This application could come in handy when a person loses their voice, and may not be able to speak audibly. Integrating this application in software can help the other person understand you, without the stress of forcing to speak.

2 RESEARCH AND RELATED WORKS

- Chung, J. S. et al [2] implements a model that recognizes spoken word given only the video without the audio.They developed an automated pipeline to collect data from TV broadcasting, enabling them to generate relatively large data to predict words rather than part of the words.
- Assael, Y. M., et al [1] solves the problem of lipreading by the observation that features capturing temporal context in an ambiguous communication channel is important. This project implements an architecture for an end-to-end sentence-level lipreading prediction model, using spatio temporal convolutional neural networks (STCNNs), recurrent neural networks (RNNs) and a connectionist temporal classification loss (CTC).The model captures phonologically important regions in the input video and makes few erroneous predictions within visemes.

3 ARCHITECTURE

3.1 General architecture

The general architecture that we use is based on the multiple tower architecture. It consists of an LRCN architecture where all the frames are passed to convolutional neural networks with shared parameters. Our model uses a wide densenet as frame feature extractor. The output is passed to a sequence model, usually a recurrent neural network (RNN) but in our case we used a temporal convolution with residual connections. At the top of the network, we have a single dense layer that outputs the logits.

Table 1: Results with the models tested

Model	inference speed (seconds)	accuracy
LSTM based model	3	65 %
Proposed model	1.3	64 %
Proposed model with RRM	0.6	64 %

We also included a spatiotemporal front-end model that consists of two 3D convolutional layers with kernels of sizes $5 \times 5 \times 5$ and $3 \times 3 \times 3$ respectively. This helps in taking advantage of a part of the mouth movement upfront.

3.2 Mobile inference

The fully convolutional model we developed presents several advantages over models that use RNNs for sequence modelling. One of them is to use recurrent residual modules (RRM) [3] to speed up inference. This technique takes advantage of the difference between two consecutive frames to avoid computing the convolutions for every single frame.

- fewer weights and smaller model size making it suitable for mobile inference
- possibility to improve on the inference speed using recurrent residual modules
- faster training time

3.3 Implementation details

The implementation makes use of PyTorch 1.0 and the networks are trained on a single GPU Tesla P100 with 13 GB of RAM. We train for 30 epochs with SGD with a momentum of 0.9 with different learning rates for each block: 10^{-3} for the dense output layer, 10^{-5} for the sequence model, 10^{-4} for the shared frame feature extractor, 10^{-4} for the spatiotemporal 3D convolutional frontend.

3.4 EXPERIMENTS AND RESULTS

We tested several architectures and techniques to increase our accuracy, reduce the model size and increase training and inference speed. The test were run on just two classes (LEADER and HOSPITAL) with a very limited amount of data (100 videos per class for training and 20 videos per class for testing). The LSTM based model achieved a test accuracy of 65% but the fully convolutional model we propose got an accuracy of 64% with less than half its training and inference time. This can be further improved when using RRM.

4 CONCLUSION AND FUTURE WORKS

We proposed a convolutional neural network based architecture that is fast and lightweight to be used in smaller devices. We plan on testing it further with the rest of the LRW dataset. Also, we would investigate further the impact of the spatio-temporal frontend and the residual convolutions used.

5 FIGURES

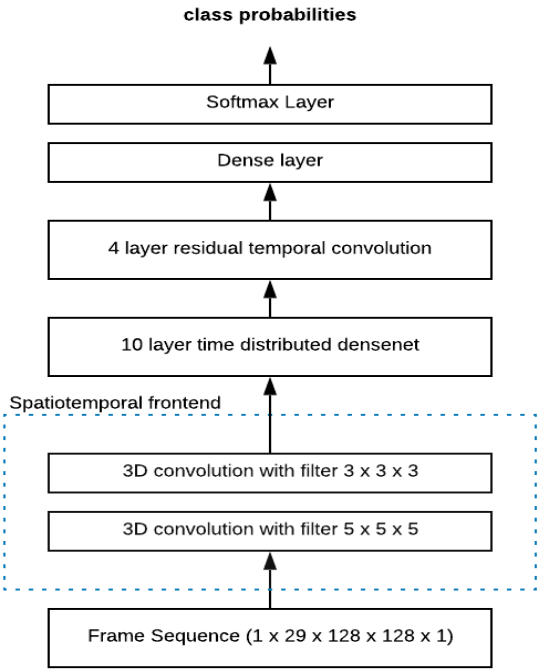

Figure 1: Block diagram of the proposed architecture

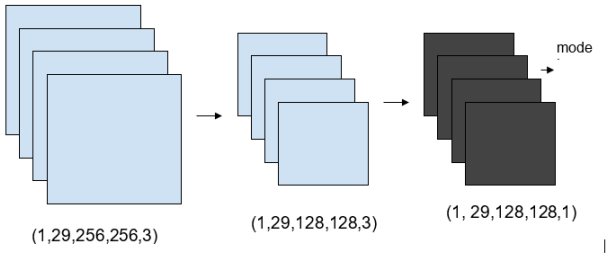

Figure 2: 29 frames of (256, 256) images from a video is first reshaped into (1, 29, 128, 128, 3); it is then converted to gray scale, (1, 29, 128, 128, 1), and then feed into the model for prediction.

ACKNOWLEDGMENTS

To Oxford university, for providing us the datasets.

REFERENCES

[1] Shillingford B. Whiteson S. de Freitas N. Assael, Y. M. 2017. *LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING.*
[2] Senior A. Vinyals O. Zisserman A Chung, J. S. 2016. Lip Reading Sentences in the Wild. *Cornell university* (2016). http://arxiv.org/abs/1611.05358
[3] Lin W. Fang X. Huang C. Zhou B. Lu C Pan, B. 2018. Recurrent Residual Module for Fast Inference in Videos. *Cornell university* (2018). http://arxiv.org/abs/1802.09723