Lips Don't Lie
A Fast Low Resource Deep Learning Algorithm to Read Lips

Mitchell Butovsky
Technion
butovsky.mitchell@gmail.com

Tom Bekor
Technion
tom.bekor@gmail.com

Abstract

Lip Reading is the task of understanding speech by visually interpreting the movements of the lips. This task is considered very difficult, even for expert lip readers. Chung et. al [5] claim that an average lip reading human expert can lip-read about 25% of the spoken words. Despite being such a hard task, lip reading has numerous applications in many areas like helping speakers with damaged vocal tracts, speech processing in face-to-face video, biometric person identification [11], etc. With the rise of Deep-Learning algorithms in the last decade, there were few successful tries addressing this problem using such algorithms. The problem however, is that these algorithms are computationally expensive, requiring heavy computational resources. In this work, we designed a new method for solving this problem, using cheaper resources and performing a significantly less amount of computations, being able to both pre-process and train the model under 3 hours.

1 Introduction

Lip reading is a complicated task for humans, as it requires distinguishing the lip motions when spelling different words which may be ambiguous when no context is given. This led to the development of automated machine lipreading methods. Such methods have to extract both spatial and temporal features from a certain video. Initially, such methods were founded on classical computer vision algorithms.

As deep-Learning algorithms have revolutionized the areas of computer vision and natural language processing in the last decade, and due to the development of bigger lip reading datasets (e.g the GRID dataset [2], which is a very synthetic, structured and simple dataset and the LRS dataset [1] which is a much bigger, realistic and unstructured dataset) there have been several tries to leverage deep-learning algorithms to solve the task of lip reading. Although some of these tries have been very successful (LipNet [3] achieves 95.2% accuracy on the GRID dataset. Chung et. al [5] managed to do well on the LRS dataset which is harder in magnitudes than the the GRID dataset) the current methodologies are extremely heavy in terms of computation, requiring very expensive GPUs to run and take a lot of time to train.

2 Related work

In the pre deep-learning era the vast majority of machine lip-reading was done by extracting features from the frames and considering the motion using classical computer vision algorithms (such as optical flow). Zhou et. al [15] provide an extensive review of these algorithms.

Additional approaches make use of classical machine learning methods such as Hidden Markov Models (HMMs) [9] [12].

After the rise of deep-learning algorithms, there were several tries to use deep-learning methods to lip read at the word/phrase level. Such algorithms include a 3D convolution algorithm that was proposed by Zisserman et. al [6]. This algorithm has a good performance, yet it is not scalable for lip reading a whole sequence of words. Another algorithm, an RNN long-short term memory cell (LSTM) based architecture using PCA and HOG features was proposed by Wand et. al [14], but yielded unsatisfactory results.

LipNet [3] was the first architecture to do end-to-end lip-reading in the full sentences scope. It used a combination of spatio-temporal convolutional neural networks (STCNNs) and a multi-layer bidirectional gated recurrent unit (bi-GRU). Jeon et. al [10] showed an extension of this architecture which is the current state of the art on the GRID dataset.

Chung et. al [5] introduced the WLAS (watch, listen, attend and spell) architecture. It applies a convolutional neural network (CNN) on each frame separately to extract visual features, MFCC to extract audio features and combine these features as an input to LSTM encoder-decoder with attention. The audio features are used to enhance the capability of the model to learn. They show that the model can do well even without the audio features.

The major problem with LipNet and WLAS is their training time and the computational resources they require.

Note that the modern techniques for automated machine lip-reading involve some sort of recurrent neural network (RNN) which was commonly used for various natural language processing (NLP) tasks. Recently, the Transformer architecture which was presented in the "attention is all you need" paper [13] has proven to be superior to RNNs in a variety of NLP tasks in particular and in sequence forecasting tasks in general. We therefore make use of the "Transformer" in our architecture.

3 Dataset

In order to be able to train the model, we had to choose an easier dataset than the "Lip-reading sentences in the wild" dataset which is very hard (tons of real life data, a lot of words and sentences spoken, movement of the speakers in a varying backgrounds etc.). We noticed that in the "Lip-reading sentences in the wild" paper, they examined their model on an additional dataset named "GRID dataset". The GRID dataset consists of 34 subjects, each uttering 1000 phrases. The utterances are sequences of the form verb (4) + color (4) + preposition (4) + alphabet (25) + digit (10) + adverb (4). Note that the vocabulary is very small (magnitudes smaller than the LRS dataset), but the number of possible sentences is very high (as well as the number of the sentences that appear in the dataset). In addition, the speakers talk very clearly and are recorded in front of a blue screen.

Figure 1: The structure of the sentences in the GRID dataset.

4 Methods

4.1 Face and Landmarks

Following the preprocess in [5], we focused on the speaker lips in order to improve our learning process. To do so, we cropped a bounding box of the lips for each frame of the video. This was made using face recognition (Facial HOG Pyramid [7, 8]) and facial-landmarks predictor (BlazeFace[4]) pretrained models. In our first idea, we used the landmarks we found to get a bounding box of the lips. The BlazeFace model returns 68 (x, y) facial landmarks, where landmarks 48-68 indicates the lips, and in order to get the lips' bounding box we took a fixed size rectangle, which centered over the centroid of these 20 landmarks. Later we changed our attitude, and instead of using the lips' bounding box, we fed our model with the lips' facial landmarks themselves.

4.2 Architecture

4.2.1 Initial Model

As our first model, we used a combination of a pretrained VGGNet, as backbone, and a classical Transformer model. This model worked with the first version of the preprocessing as follows: after the preprocessing was done, we got the lips' bounding box for each video frame. These bounding boxes is in gray scale, they where up scaled with 1×1 conv and loaded into the VGG. We took from the VGG only its backbone, which was trained to capture more local and general features in the input image, and we replaced its linear layers by new ones, so they will precept the unique global features in our inputs. The VGG created from the bounding boxes feature vectors that creates our vector sequence, which was loaded into the transformer. The transformer outputs were loaded into a fully connected neural network, and created our final output sequence, that is, the transcribed sentence. See figure 2.

4.2.2 Landmarks Model

After we thought about how can we improve the model, we came to the idea of using only the landmarks, without the need for the lips' bounding box. Logically, the landmarks are exactly the features we would want to extract from the bounding box, so instead of using the pretrained VGG, we fed the landmarks into a 3-

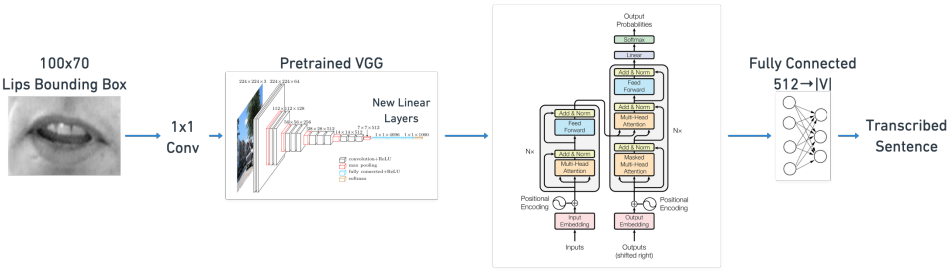

Figure 2: First architecture - mouth bounding box into 1x1 convolution to increase the number of channels, then into pretrained VGG, where the linear layers are replaced with new ones, each sequence fed into the transformer, and using a fully-connected neural network the final predictions are created.

layer MLP , with ReLU activations and dropout of 0.3, to create more complex relations between them. The output of the MLP has the same shape as the transformer inputs shape.

4.2.3 Enhancing the learning process with Temporal information

After some thought about the model we designed, we noticed that the absolute location of the landmarks is not important, but how these landmarks change during time is really the interesting feature in our case. Moreover, if one speaker moves from one point to another in different videos this would make the learning harder. Aiming to solve this issue we applied a very simple trick which yielded a great improvement in the results! instead of considering the raw landmarks of a video we considered the element-wise shift of the landmarks. Namely, instead of feeding a raw sequence of landmarks: $(\bar{l}_1, ..., \bar{l}_n)$ we fed to the model the sequence: $(\bar{d}_1, ..., \bar{d}_{n-1})$ (where $\bar{d}_i = \bar{l}_{i+1} - \bar{l}_i$). Afterwards, we took this idea even further; intuitively, if we would take a glance on few frames before the current one, they would add an important information - "information from the past" which helps to understand the current frame. Following this intuition, we concatenated to every frame landmarks (or, to be more precise, landmark shifts) the landmarks shifts from the previous $w-1$ frames. Thus, the final input to our model was of the form: $(\bar{t}_1, ..., \bar{t}_{n-w})$ where $\bar{t}_i = concat(\bar{d}_i, \bar{d}_{i+1}, ..., \bar{d}_{i+w-1})$.

5 Experiments and Results

5.1 Training Process

We trained the model for 20 epochs with a batch size of 64, where each sample in the batch is a video. We used CE Loss, Stochastic Gradient Descent (with $lr = 1 \times 10^{-1}$) and Adam optimizer with the default parameters. We used the "reduce learning rate on plateau" scheduler on the loss with $patience = 2$ and $factor = 0.1$ and updated it every 50 iterations (namely, we reduced the learning rate by 10% each time the loss didn't decrease for $2 * 50 = 100$ iterations). The lips landmarks delta were concatenated through WinSize=5 frames, and were sent to 2-layered MLP with hidden dim of 256 and output dim of 512, that was trained with dropout of 0.3. The trained transformer was composed of 512 input features in the encoder/decoder, 4 attention heads, 2 layers in the encoder/decoder, 2048 sized feed-forward model, and 0.1 dropout.

5.2 Evaluation

Note that the task of lip reading is a sequence-to-sequence task. In general, evaluating the performance on such tasks is difficult because of the fact that the recognized word sequence can have a different length from the true word sequence. One simple and common approach for evaluating such tasks is using the WER (word error rate) metric. This metric is calculated as follows:

$$WER = \frac{S + D + I}{N}$$

where S, D, I are the numbers of word substitutions, insertions, and deletions made in the generated sequence with respect to the ground truth sequence. It is also common to define the $W_{acc} := 1 - WER$. We noticed that in our case, all of our purposed models always generated sentences which match the structure of the spoken sentences. This is due to the fact that all the dataset sentences are fixed structured, as mentioned in section 1.2. Note that

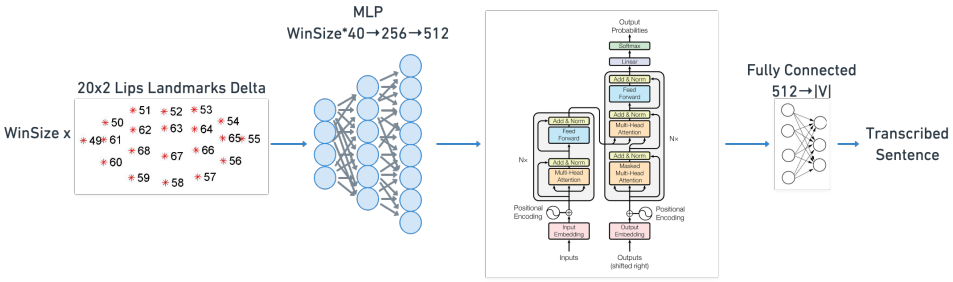

Figure 3: Final Model - lips landmarks delta are concatenated through WinSize number of frames, fed to MLP, which creates sequences that are fed into the transformer, and using a fully-connected neural network the final predictions are created.

in this case, the W_{acc} metric exactly equals to the per-word accuracy.

5.3 Results & Conclusions

We present here the plots for the loss and accuracy 4, 5. Our trained model achieves word accuracy of 80.4% on the test set, with a training time of only 31 minutes (note that the landmark creation pre-processing using BlazeFace [4] inference took about 2 hours and is the bottleneck of our methodology). See table 1 for comparison with some of the aforementioned methods. We therefore conclude that our architecture is indeed superior with its efficiency to the existing methods (to the best of our knowledge). This is due to the relatively fast preprocessing proposed and the high parallelism achieved by replacing the RNN with Transformer. Moreover, we were able to significantly reduce the amount of model parameters, allowing it to run the on low resource devices.

Figure 4: Training loss as a function of epoch number.

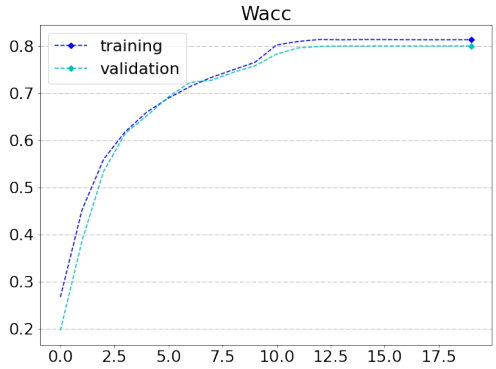

Figure 5: Word accuracy as a function of epoch number.

5.4 Future Work

Despite the substantial improvement in the training time and the number of model parameters, our approach has sub-optimal word accuracy. Thus a future work would be to try restore the state-of-the-art results (or at least improve the results) without compromising on the presented efficiency and memory. In addition, note that in this research, we managed to lip-read the GRID dataset's speakers' lips using only the visual features. Thus, an additional future expansion of our work can be the utilization of both visual and auditory features during the training process. Inspired by [5], such an approach can be done by using all of the features at the start of the training and gradually applying random noise to the auditory ones. That way, the model will rely more on the visual ones. On inference time, the model will be able to transcribe a video using only the visual features, while a random noise will represent the audio of the video. The

Method	Model size	Training time	Training + preprocess time	Word accuracy
LipNet	45.7M	72.8 hours	$>$ 72.8 hours	95.2%
Jeon et. al	34.5M	70.6 hours	$>$ 70.6 hours	**99**%
LipsDontLie (ours)	**15M**	**31 minutes**	**2 hr 31 min**	80.4%

Table 1: Mentioned methods' performances.

advantage of such training can result in faster convergence time and possibly a better word accuracy on test time.

References

[1] Triantafyllos Afouras et al. "Deep Audio-Visual Speech Recognition". In: *CoRR* abs/1809.02108 (2018). arXiv: 1809.02108. URL: http://arxiv.org/abs/1809.02108.

[2] Najwa Alghamdi et al. "A corpus of audio-visual Lombard speech with frontal and profile views". In: *Acoustical Society of America Journal* 143.6 (June 2018), EL523–EL529. DOI: 10.1121/1.5042758.

[3] Yannis M. Assael et al. "LipNet: Sentence-level Lipreading". In: *CoRR* abs/1611.01599 (2016). arXiv: 1611.01599. URL: http://arxiv.org/abs/1611.01599.

[4] Valentin Bazarevsky et al. "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs". In: *CoRR* abs/1907.05047 (2019). arXiv: 1907.05047. URL: http://arxiv.org/abs/1907.05047.

[5] J. Chung et al. "Lip Reading Sentences in the Wild". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, July 2017, pp. 3444–3453. DOI: 10.1109/CVPR.2017.367. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.367.

[6] Joon Son Chung and Andrew Zisserman. "Lip Reading in the Wild". In: Mar. 2017, pp. 87–103. ISBN: 978-3-319-54183-9. DOI: 10.1007/978-3-319-54184-6_6.

[7] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

[8] Pedro F. Felzenszwalb et al. "Object Detection with Discriminatively Trained Part-Based Models". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (2010), pp. 1627–1645. DOI: 10.1109/TPAMI.2009.167.

[9] Alan J Goldschen, Oscar N Garcia, and Eric D Petajan. "Continuous automatic speech recognition by lipreading". In: *Motion-Based Recognition*. Springer, 1997, pp. 321–343.

[10] Sanghun Jeon, Ahmed Elsharkawy, and Mun Sang Kim. "Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition". en. In: *Sensors (Basel)* 22.1 (Dec. 2021).

[11] J. Luettin, N.A. Thacker, and S.W. Beet. "Speaker identification by lipreading". In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*. Vol. 1. 1996, 62–65 vol.1. DOI: 10.1109/ICSLP.1996.607030.

[12] Chalapathy Neti et al. *Audio visual speech recognition*. Tech. rep. IDIAP, 2000.

[13] Ashish Vaswani et al. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[14] Michael Wand, Jan Koutnık, and Jürgen Schmidhuber. "Lipreading with Long Short-Term Memory". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai, China: IEEE Press, 2016, pp. 6115–6119. DOI: 10.1109/ICASSP.

2016.7472852. URL: https://doi.org/10.1109/ICASSP.2016.7472852.

[15] Ziheng Zhou et al. "A review of recent advances in visual speech decoding". In: *Image and Vision Computing* 32.9 (2014), pp. 590–605. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2014.06.004. URL: https://www.sciencedirect.com/science/article/pii/S0262885614001036.