International Journal of Engineering Sciences Paradigms and Researches (IJESPR)
Volume 47, Issue 01, Quarter 01 (January to March 2018)
An Indexed and Referred Journal with Impact Factor: 2.80
ISSN (Online): 2319-6564
www.ijesonline.com

Deep AudioVisual Speech Recognition

Archana Panda[1], Suren Ku. Sahu[2], Jyostnarani Tripathy[3]

[1,3]Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar
[2]Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar

Publishing Date: 3rd March, 2018

Abstract

The goal of this work is to recognize phrases and sentences being spoken by a talking face, with or without the audio. Unlike previous works that have focused on recognizing a limited number of words or phrases, we tackle lip reading as an open- world problem – unconstrained natural language sentences, and in the wild videos. Our key contributions are: (1) we compare two models for lip reading, one using a CTC loss, and the other using a sequence-to- sequence loss. Both models are built on top of the transformer self-attention architecture; (2) we investigate to what extent lip reading is complementary to audio speech recognition, especially when the audio signal is noisy; (3) we introduce and publicly release two new datasets for audio-visual speech recognition: LRS2-BBC, consisting of thousands of natural sentences from British television; and LRS3-TED, consisting of hundreds of hours of TED and TEDx talks obtained from YouTube. The models that we train surpass the performance of all previous work on lip reading benchmark datasets by a significant margin.

Keywords: *Lip reading, audio visual speech recognition, deep learning.*

Introduction

Lip Reading, the ability to recognize what is being said from visual information alone, is an impressive skill, and very challenging for a novice. It is inherently ambiguous at the word level due to homophones different characters that produce exactly the same lip sequence. However, such ambiguities can be resolved to an extent using the context of neighboring words in a sentence, and/or a language model.

A machine that can lip read opens up a host of applications: dictating instructions or messages to a phone in a noisy environment; transcribing and re-dubbing archival silent films; resolving multi-talker simultaneous speech; and, improving the performance of automated speech recognition in general.

That such automation is now possible is due to two developments that are well known across computer vision tasks: the use of deep neural network models; and, the availability of a large scale dataset for training. In this case, the lip reading models are based on recent encoder- decoder architectures that have been developed for speech recognition and machine translation [4, 6].

The objective of this paper is to develop neural transcription architectures for lip reading sentences. We compare two models: one using a Connectionist Temporal Classification (CTC) loss, and the other using a sequence-to-sequence (seq2seq) loss [8]. Both models are based on the transformer self-attention architecture, so that the advantages and disadvantages of the two losses can be compared head-to-head, with as much of the rest of the architecture in common as possible. The datasets developed in this paper to train and evaluate the models, are based on thousands of hours of videos that have talking faces together with subtitles of what is being said.

We also investigate how lip reading can contribute to audio based speech recognition. There is a large literature on this contribution, particularly in noisy environments, as well as the converse where some derived measure of audio can contribute to

lip reading for the deaf or hard of hearing. To investigate this aspect we train a model to recognize characters from both audio and visual input, and then systematically disturb the audio channel.

The first two authors contributed equally to this work. Our models output at the character level. In the case of the CTC, these outputs are independent of each other. In the case of the sequence-to-sequence loss a language model is learnt implicitly, and the architecture incorporates a novel dual attention mechanism that can operate over visual input only, audio input only, or both. The architectures are described in Section 3. Both models are decoded with a beam search, in which we can optionally incorporate an external language model.

We describe the generation and statistics of two new large scale datasets that are used to train and evaluate the models: LRS2-BBC, Lip Reading Sentences based on BBC broadcasts; and LRS3-TED, Lip Reading Sentences based on TED videos. Both contain talking faces together with subtitles of what is said. The videos contain faces in the wild with a significant variety of pose, expressions, lighting, backgrounds and ethnic origin. Section 5 describes the network training, where we report a form of curriculum learning that is used to accelerate training. Finally, Section 6, evaluates the performance of the models, including for visual (lips) input only, for audio and visual inputs, and for synchronization errors between the audio and visual streams. On the content: This submission is based on the conference paper [11]. We replace the WLAS model in the original paper with two variants of a Transformer-based model [48]. One variant was published in [2], and the second variant (using the CTC loss) is an original contribution in this paper. We also update the visual front-end with a ResNet-based one proposed by [44]. The new front-end and back-end architectures contribute to over 20% absolute improvements in Word Error Rate (WER) over the model proposed in [11]. Finally, we publicly release two new datasets, LRS2-BBC and LRS3-TED, that supersede the original LRS dataset in [11] which could not be made public due to license restrictions.

Background

For the most part, end-to-end deep learning approaches for sequence prediction can be divided into two types.

The first type uses a neural network as an emission model which outputs the likelihood of each output symbol (e.g., phonemes) given the input sequence (e.g., audio). These methods generally employ a second phase of decoding using a Hidden Markov Model [24]. One such version of this variant is the Con-nectionist Temporal Classification (CTC) [21], where the model predicts frame- wise labels and then looks for the optimal align-ment between the frame-wise predictions and the output sequence. The main weakness of CTC is that the output labels are not conditioned on each other (it assumes each unit is independent), and hence a language model is employed as a post-processing step. Note however that some alternatives to jointly train the two step process has been proposed [20]. Another limitation of this approach is that it assumes a monotonic ordering between input and output sequences. This assumption is suitable for ASR and transcription for example, but not for machine translation.

The second type is sequence-to-sequence models [8, 45] (seq2seq) that first read all of the input sequence before predicting the output sentence. A number of papers have adopted this approach for speech recognition [9, 10]: for example, Chan et al. [6] proposes an elegant sequence-to-sequence method to tran- scribe audio signal to characters. Sequence-to-sequence decodes an output symbol at time t (e.g. character or word) conditioned on previous outputs 1, . . . , t − 1. Thus, unlike CTC-based models, the model implicitly learns a language model over output symbols, and no further processing is required. However, it has been shown [6, 25] that it is beneficial to incorporate an external language model in the decoding of sequence-to-sequence models as well. This way it is possible to leverage larger text-only corpora that contain much richer natural language information than the

limited aligned data used for training the acoustic model.

Regarding architectures, while CTC-based or seq2seq approaches traditionally relied on recurrent networks; recently there has been a shift towards purely convolutional models [5]. For example, fully convolutional networks have been used for ASR with CTC [50, 54] or a simplified variant [15, 31, 53].

Related works

There is a large body of work on lip reading using non deep learning methods. These methods are thoroughly reviewed in [55], and we will not repeat this here. A number of papers have used Convolutional Neural Networks (CNNs) to predict phonemes [36] or visemes [28] from still images, as opposed to recognising to full words or sentences. A phoneme is the smallest distinguishable unit of sound that collectively makes up a spoken word; a viseme is its visual equivalent.

Independent performance on the constrained grammar and 51 word vocabulary of the GRID dataset [16]. A deeper architecture than LipNet [3] is used by [44], who propose a residual network with 3D convolutions to extract more powerful representations. The network is trained with a cross-entropy loss to recognize words from the LRW dataset. Here, the standard ResNet architecture [23] is modified to process 3D image sequences by changing the first convolutional and pooling blocks from 2D to 3D.

In our earlier work (Chung et al. [11]), we proposed a WLAS sequence-to-sequence model based on the LAS ASR model of [6] (the acronyms are Watch, Listen, Attend and Spell (WLAS), and Listen, Attend and Spell (LAS)). The WLAS model had a dual attention mechanism – one for the visual (lip) stream, and the other for the audio (speech) stream. It transcribed spoken sentences to characters, and could handle an input of vision only, audio only, or both.

In independent and concurrent work, Shillingford et al. [42], design a lip reading pipeline that uses a network which outputs phoneme probabilities and is trained with CTC loss. At inference time, they use a decoder based on finite state transducers to convert the phoneme distributions into word sequences. The network is trained on a very large scale lip reading dataset constructed from YouTube videos and achieves a remarkable 40.9% word error rate. Audio-visual speech recognition. The problems of audio-visual speech recognition (AVSR) and lip reading are closely linked. Mroueh et al. [35] employs feed-forward Deep Neural Networks (DNNs) to perform phoneme classification using a large non-public audio-visual dataset. The use of HMMs together with hand-crafted or pre-trained visual features have proved popular – [47] encodes input images using DBF; [19] used DCT; and [37] uses a CNN pre-trained to classify phonemes; all three combine these features with HMMs to classify spoken digits or isolated words. As with lip reading, there has been little attempt to develop AVSR systems that generalise to real-world settings.

Petridis et al. [39] use an extended version of the architecture of [44] to learn representations from raw pixels and waveforms which they then concatenate and feed to a bidirectional recurrent network that jointly models the audio and video sequences and outputs word labels.

Discussion

Overall the TM-seq2seq model performs significantly better for lip-reading in terms of WER, when no audio is supplied. For audio-only or audio-visual tasks, the two methods perform sim- ilarly. However the CTC models appear to handle background noise better; in the presence of loud babble noise, both the audio-only and audiovisual TM-seq2seq models perform significantly worse that their TM-CTC counterparts.

Training time: The TM-seq2seq models have a more complex architecture and are harder to train, with the full audiovisual model taking approximately 8 days to complete the full curriculum for both datasets, on a single GeForce

Titan X GPU with 12GB memory. In contrast, the audiovisual TM-CTC model trains faster i.e. in approximately 5 days on the same hardware. It should be noted however that since both architectures contain no recurrent modules and no batch normalization, their implementation can be heavily parallelized into multiple GPUs.

Inference time: Decoding of the TM-CTC model does not require auto-regression and therefore the CTC probabilities need only be

7. Conclusion

In this paper, we introduced two large-scale, unconstrained audio-visual datasets, LRS2-BBC and LRS3-TED, formed by collecting and preprocessing thousands of videos from the British television and YouTube respectively.

We considered two models that can transcribe audio and video sequences of speech into characters and showed that the same architectures can also be used when only one of the modalities is present. Our best visual-only model surpasses the performance of the previous state-of-the-art on the LRS2- BBC lip reading benchmark by a large margin and sets a strong baseline for LRS3-TED. We finally demonstrate that visual information helps improve speech recognition performance even when the clean audio signal is available. Especially in the presence of noise in the audio, combining the two modalities leads to a significant improvement.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016. 6

[2] T. Afouras, J. S. Chung, and A. Zisserman. Deep lip reading: A comparison of models and an online application. In INTERSPEECH, 2018. 1

[3] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. Lipnet: Sentence-level lipreading. arXiv:1611.01599, 2016. 2, 4

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine trans-lation by jointly learning to align and translate. Proceedings of the International Conference on Learning Representa-tions, 2015. 1

[5] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271, 2018. 2

[6] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals. Listen, attend and spell. arXiv preprint arXiv:1508.01211, 2015. 1, 2, 4

[7] C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani. State-of-the-art speech recognition with sequence-to-sequence mod-els. CoRR, abs/1712.01769, 2017. 3

[8] K. Cho, B. Van Merrienboer,¨ C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical ma-chine translation. In EMNLP, 2014. 1, 2

[9] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio. End-to-end continuous speech recognition using attention-based recurrent NN: first results. In NIPS 2014 Workshop on Deep Learning, 2014. 2

[10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In Advances in Neural Information Processing Systems, pages 577–585, 2015. 2

[11] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1, 2, 5, 6

[12] J. S. Chung and A. Zisserman. Lip reading in the wild. In Proceedings of the Asian Conference on Computer Vision, 2016. 2, 4, 5.

[13] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip- reading, ACCV, 2016. 4

[14] J. S. Chung and A. Zisserman. Lip reading in profile. In Proceedings of the British Machine Vision Conference, 2017. 4, 5

[15] R. Collobert, C. Puhrsch and G. Synnaeve. Wav2letter: An end-to-end convnet-based speech recognition system. CoRR, abs/1609.03193, 2016.