CS230

Multi-modal Model for Sentence Level Captioning

Hikaru Hotta
Department of Computer Science
Stanford University
hhotta@stanford.edu

Tom Pritsky
Department of Biology
Stanford University
tom5@stanford.edu

Abstract

Captioning refers to the transcription of speech into time-coded blocks. Traditional approaches to captioning have relied on transcribing audio data using End-to-End Automatic Speech Recognition models (ASR) (DeepSpeech citation). However the performance of ASR models in real world settings is limited due to factors such as high noise and speaker distance. Other approaches have relied on transcribing visual data using lipreading models (LipNet citation). However, the performance of lipreading models have been variable with limited spatio-temporal robustness. To build on the existing state of the art, we developed a multi-modal model for captioning that integrates both auditory and visual cues. Our model splits video data into visual and audio data and makes use of a DeepSpeech ASR Model and LipNet, an end-to-end sentence-level lipreading model, which are fused together with fully connected softmax and Connectionist Temporal Classification (CTC) layers.

1 Introduction

Around 48 million Americans have a hearing loss which adversely impacts speech understanding. This reduced comprehension is particularly pertinent in academic settings where precise discernment of speech is critical. Real time captioning is a highly pertinent accommodation; however, accuracy in real world settings is limited due to factors such as noise and distance to speaker. Inspired by an approach called lip-reading, a technique commonly used by the hearing impaired to better understand speech, we developed a multi-modal model for captioning that integrates both auditory and visual cues for higher captioning accuracy. Our model is particularly useful as a noise reduction technique, allowing a single speaker to be isolated from a crowd's background noise.

The input to our model is a video. We use a parser to split our video into audio and visual data (frames of images). We then feed our audio data through independent layers of our audio modality module comprised of three 1D Convolutions, three Bi-GRUs, and a dense layer. Simultaneously, visual data is passed through the independent layers of our visual modality comprised of three STCNNs, each followed by spatial pooling, two Bi-GRUs and a dense Layer. The dense layers of both modules are processed by a softmax layer and trained with CTC loss. Our multi-modal model outputs a predicted text captioning of the input video.

2 Related work

In this section we outline various approaches to ASR, automated lip-reading, and multi-modal neural networks.

Automatic Speech Recognition: Traditional ASR models have used multiple algorithms and processing stages (Hannun et al., 2014). The simplicity of our modules and their integration with each

,

other were critical design considerations in developing our multi-modal model. Therefore, we decided to use a pretrained end-to-end speech recognition model, Deep Speech, which was developed by Baidu Research (Silicon Valley AI Lab). Combined with an N-gram language model, Deep Speech achieved a word error rate of 16.0% on the full Switchboard Hub5'00 test set and 19.1% on a noisy speech recognition data set (Hannun et al., 2014). It achieves this by applying end-to-end deep learning using RNNs (Hannun et al., 2014). Other ASR models such as Google's Listen, Attend, and Spell (LAS) system were also considered for our audio module (Chiu et al. 2018). The model exhibits superior performance with a word error rate of on a diction task (Chiu et al. 2018). However, the complexity of the model compared to Deep Speech, which is delineated by its Decoder, Attention, Encoder module architecture, made Deepspeech a better candidate for our multi-modal system (Chiu et al. 2018).

Automated Lip-reading: As both a computer vision and NLP task, a variety of models have been developed for the task of lip-reading. One such approach is a Fully Convolutional (FC) deep neural network architecture trained on a CTC loss on the Lip Reading in the Wild and Lip Reading Sentences 2 data sets (Afouras et al., 2018). This FC approach achieves a word error rate of 55% and trains faster than other types of models because it has a smaller number of parameters (Afouras et al., 2018). The same 2018 paper outlines the use of three stacked Bidirectional LSTM (BLSTM) recurrent networks. The first BLSTM layer ingests the vision feature vectors and the final outputs the character probability for each output frame (Afouras et al., 2018). The approach achieves a word error rate of 62.2% which is significantly higher than the FC approach, despite the fact that the recurrent model has context on every decoding time-step as compared to the FC model (Afouras et al., 2018). To take advantage of the fast training and high performance of the FC approach and the context retention characteristic of the recurrent model, we investigated Lip Net, a lip reading model that is characterized by by its spatiotemporal convolutional and Bi-GRU layers and CTC loss (Assael et al., 2016). Lip Net is an open source speech to text model developed at the University of Oxford (Assael et al., 2016). This approach develops frame-wise labels and then searches for alignment between the frame-wise predictions and the outputted sequence (Assael et al., 2016). While the GRID corpus data set that the LipNet model is trained on is limited, with each example following phrases following a strict syntactical pattern, it achieves a 11.4% word error rate on unseen speaker examples (Assael et al., 2016). We decided to use this model because of the availability of an open-source model and GitHub repository with training and testing architecture (Aulia Rahman Maulana, 2018).

Multi-modal Neural Networks: Multi-modal neural networks enable multiple modes of data to be channeled to train and test Deep Neural Networks. An advantage of such as system is that the variety of input data better captures spatiotemporal and contextual information to generate more robust predictions. Of the copious applications of multi-modal neural networks, we studied its application towards emotion recognition using the fusion of audio, video and text data (Ortega et al., 2019). Each modality has FC independent layers, a merge layer, a FC layer, a linear regression layer, a decimal scaling module, and an linear activation output (Ortega et al., 2019). We took inspiration from the concatenation function utilized in this implementation to merge the outputs of our independent layers.

3 Dataset and Features

We used the GRID corpus dataset to train our model, since this is the dataset that LipNet, our visual speech to text network, has been trained on. This dataset is comprised of video data with audio of a subject speaking sentences. The GRID corpus dataset itself is a compilation of high quality audio-visual recordings. In total, the dataset consists of speech segments of 34 speakers recording 1000 sentences each. The phrases spoken follow a strict syntactical pattern: "command + color + preposition + letter + digit + adverb" such as “place green at B 4 now”, where individual words are substituted but the sentence structure remains consistent.

From the GRID corpus dataset, we obtained 1000 examples from randomized speakers which were split into 800 training, 100 validation and 100 test examples. All downloaded videos were 3 seconds in length and had a frame rate of 25fps. Our prepossessing pipeline splits the video into video (mpg) and audio (wav) format. It applies a DLib face detector and the iBug face landmark predictor to the video (mpg) produce a mouth centered crop of each frame. It also computes the log of the Fourier Transform of the audio (wav) file and normalizes the feature vector using Z normalization (Assael et al., 2016).

Figure 1: One frame of GRID corpus data set mpg file with caption overlay

4 Methods

Our multi-modal model consists of an independent lipreading unit, an independent speech recognition unit, and a dependent CTC unit that processes the concatenated outputs of the independent units.

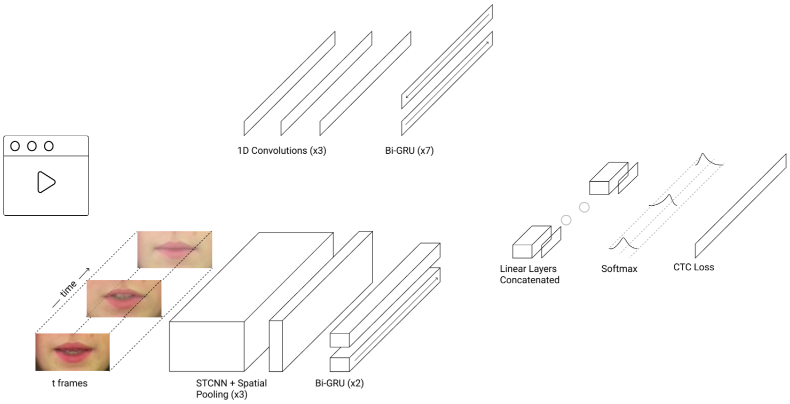

Figure 2: Multi-modal Model architecture.

Independent Lipreading Unit

A sequence of t frames is the input to our independent unit, and it is processed by three layers of Spatiotemporal Convolutional Neural Network (STCNN), each with a spatial max-pooling layer. Our STCNN from C channels to C' channels for input x and weights $w \in \mathbb{R}^{C'xCxk_txk_wxk_h}$ is able to process data temporally and spatially (Assael et al., 2016).

$$[stconv(x,w)]_{c'tij} = \sum_{c=1}^{C}\sum_{t'=1}^{k_t}\sum_{i'=1}^{k_w}\sum_{j'=1}^{k_h} w_{c'ct'i'j'} x_{c,t+t',i+i',j+j'} \tag{1}$$

The extracted features are then processed by two Bi-GRUs where each time-step of the GRU output is processed a linear layer. GRU's, or gated recurrent units, allow recurrent networks to retain long term memory by passing forward essential information from previous layers to later layers via a hidden state. Unlike LSTMs, GRUs rely on two gates, an update gate and a reset gate, to determine which previous information to pass forward and which to forget.

$$[u_t, r_t]^T = sigm(W_z z_t + W_h h_{t-1} + b_g) \tag{2}$$

$$\tilde{h_t} = tanh(U_z z_t + U_h(r_t \odot h_{t-1}) + b_h) \tag{3}$$

$$h_t = (1 - u_t) \odot h_{t-1} + u_t \odot \tilde{h_t} \tag{4}$$

$z := \{z_1, ..., z_T\}$ is the input sequence to the RNN. The Bi-GRU ensures that h_t depends on $z_{t'}$ for all t' (Assael et al., 2016).

Independent Speech Recognition Unit

Our independent speech recognition unit relies on conversion of the input audio sequence to a spectrogram. The spectrogram is passed as input to a 1D convolutional layer with 1026 nodes, followed by three Gated Recurrent Units (GRUs – described above in Independent Lipreading Unit section) and a fully connected dense layer. Batch normalization is used after each layer to normalize the outputs, ensuring that no activations are too high or low and allowing use of higher learning rates without risk of exploding or vanishing gradients.

Dependent CTC Classification Unit

Our dependent CTC (Connectionist Temporal Classification) unit consists of a fully connected layer to process the concatenated outputs of the previous independent units (lipreading and speech recognition units), followed by a softmax classification layer. CTC allows us to work with temporal data without needing to specifically annotate the timestamp of each character in the input, a time-consuming process. Both LipNet and DeepSpeech (the original networks upon which we based our approach) rely on CTC to effectively work with temporal speech and video data. This motivated us to use CTC in tandem with our final classification layers rather than simply implementing softmax or logistic regression.

The output neurons of the CTC delineate the distribution over whole character sequences $c \in \{A, B, C, ..., blank, space\}$ where x is the time-step encoded feature vector from the GRU.

$$P(c|x) = \prod_{i=1}^{N} P(c_i|x) \tag{5}$$

From the distribution, interpret the mapping over possible transcriptions y using where $\beta(c)$ is a mapping over y.

$$P(y|x) = \sum_{c:\beta(c)=y} P(c|x) \tag{6}$$

Network parameters θ are updated to maximize the probability of $y*$, the correct label (Coates, 2016).

$$\theta^* = \underset{\theta}{\text{argmax}} \sum_{c:\beta(c)=y^{*(i)}} P(c|x^{(i)}) \tag{7}$$

5 Experiments/Results/Discussion

We conducted an ablation study in order to test whether our multi-modal model improved upon our individual speech recognition and Lipnet model on the GRID corpus.

Our evaluation metrics were Word Error Rate (WER) and Character Error Rate (CER) since they robust metrics that encompass word and character level accuracy.

$$WER = \frac{S_w + D_w + I_w}{N_w} \tag{8}$$

where S_w is the number of word substitutions, D_w is the number of word deletions, I_w is the number of word insertions, N_w is the number of words in the ground truth.

$$CER = \frac{S_c + D_c + I_c}{N_c} \tag{9}$$

where S_c is the number of character substitutions, D_c is the number of character deletions, I_c is the number of character insertions, N_c is the number of character in the ground truth.

Evaluation		
Model	**WER**	**CER**
DeepSpeech	1.143	0.842
LipNet	0.114	0.064
Multi-modal	?	?

While DeepSpeech is a robust model that conducts end-to-end speech recognition, it performed poorly on a sentence level because it was not trained on the GRID corpus data set and therefore was unable to make good predictions on sentences with strict syntactical rules. It performed slightly better on a character level as the model was able to recognize phonemes and decode some of them.

LipNet performance metrics were obtained from the Lipnet paper in its tests on Unseen Speakers. Our reported LipNet WER metrics are based upon published results, representative of a large training set of 17111 utterances of 261 speakers for training (about 34.9 hours).

6 Conclusion/Future Work

We conclude that, in theory, concatenation of the outputs of a visual and an auditory network should indeed improve the accuracy of end to end speech transcription. This conclusion is based upon previous work by Google Deepmind, which relies on a multi-modal deep learning audio-visual audio visual network to transcribe speech (Ephrat et al., 2018). Although the approach wasn't end to end, it achieved significantly improved Signal to Distortion ratio. Although SDR is not directly comparable to WER, both refer to accuracy in some context (one in a single speaker context and the other in a high background noise context). As a result, we believe that our multi-modal model would achieve higher accuracy (lower WER) relative to the baseline purely speech based transcription in high noise settings. This is evident in contexts with multiple speakers. This matches our original hypothesis that lipreading provides context and is a valuable form of noise reduction. However, since the GRID Corpus dataset we use for training consists of only clean speech from a single speaker, we could not test this hypothesis directly.

In the future, we hope to train the final classification layers for our model, to allow us to validate our prediction that the multimodal model will produce improved WER results relative to baseline purely audio or purely visual networks, based on previous experimentation in a non end to end trained model (which attained an improvement in the absolute speech recognition rate up to 3.10% for multi-modal relative to purely audio based approaches (Ivanko et. al. 2018). Our work seeks to demonstrate this effect in a fully end to end trained audio-visual model.

Furthermore, we hope to train more noise-robust ASR systems through a multi-modal approach. Our hypothesis is based on use of lipreading by hearing impaired persons as a means of increasing understanding in high noise and is supported by previous work (Ivanko et. al. 2018).

7 Contributions

Hikaru - He took a lead on developing a pipeline that splits video data into mpg wav formats and pipes it to our independent modules. He also set up the speech recognition independent module (*by solving dependency issues) and not only evaluated it on the GRID corpus but also developed the pipeline to retrieve intermediate outputs (Linear) of the model to train our multi-modal network. He conceptualized and formalized the multi-modal model architecture and developed the training pipeline for the model and was responsible for the completion of the final write-up.

Tom - Tom took the lead on extracting intermediate outputs from the LipNet Grid Corpus files as well as local setup of the Lipnet architecture. He applied the pipeline to convert LipNet .mpg files to .wav files developped by Hikaru and modified it to work with the Grid Corpus dataset. He produced a document that correlated training files to labels that allowed testing of the DeepSpeech network on the Grid Corpus dataset to establish baseline speech transcription accuracy. Tom took the lead on the poster and assisted with the writeup.

References

[1] ARM, Muhammad Rizki. LipNet, 26 June 2018, github.com/rizkiarm/LipNet.

[2] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates: "Deep Speech: Scaling up end-to-end speech recognition", 2014; [http://arxiv.org/abs/1412.5567 arXiv:1412.5567].

[3] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski: "State-of-the-art Speech Recognition With Sequence-to-Sequence Models", 2017; [http://arxiv.org/abs/1712.01769 arXiv:1712.01769].

[4] Coates, Adam, and Vinay Rao. "Speech Recognition and Deep Learning." ba$_{d}ls_{s}peech2016, 2016.$

[5] Ivanko, D., Karpov, A., Fedotov, D., Kipyatkova, I., Ryumin, D., Ivanko, D., … Zelezny, M. (2018). Multimodal speech recognition: Increasing accuracy using high speed video data. Journal on Multimodal User Interfaces, 12(4), 319–328. https://doi.org/10.1007/s12193-018-0267-1

[6] Juan D. S. Ortega, Mohammed Senoussaoui, Eric Granger, Marco Pedersoli, Patrick Cardinal: "Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition", 2019; [http://arxiv.org/abs/1907.03196 arXiv:1907.03196].

[7] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman: "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation", 2018, ACM Trans. Graph. 37(4): 112:1-112:11 (2018); [http://arxiv.org/abs/1804.03619 arXiv:1804.03619]. DOI: [https://dx.doi.org/10.1145/3197517.3201357 10.1145/3197517.3201357].

[8] Ozair, Sherjil. "Ctc." Sherjilozair, 2015, github.com/sherjilozair/ctc.

[9] Research, Baidu. "Ba-Dls-Deepspeech." Baidu-Research, 2017, github.com/baidu-research/ba-dls-deepspeech.

[10] Research, Baidu. "Warp-Ctc." Baidu-Research, 7 July 2018, github.com/baidu-research/warp-ctc.

[11] Triantafyllos Afouras, Joon Son Chung: "Deep Lip Reading: a comparison of models and an online application", 2018; [http://arxiv.org/abs/1806.06053 arXiv:1806.06053].

[12] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson: "LipNet: End-to-End Sentence-level Lipreading", 2016; [http://arxiv.org/abs/1611.01599 arXiv:1611.01599].