Multimodal pipeline for end to end speech transcription

Hikaru Hotta, Tom Pritsky

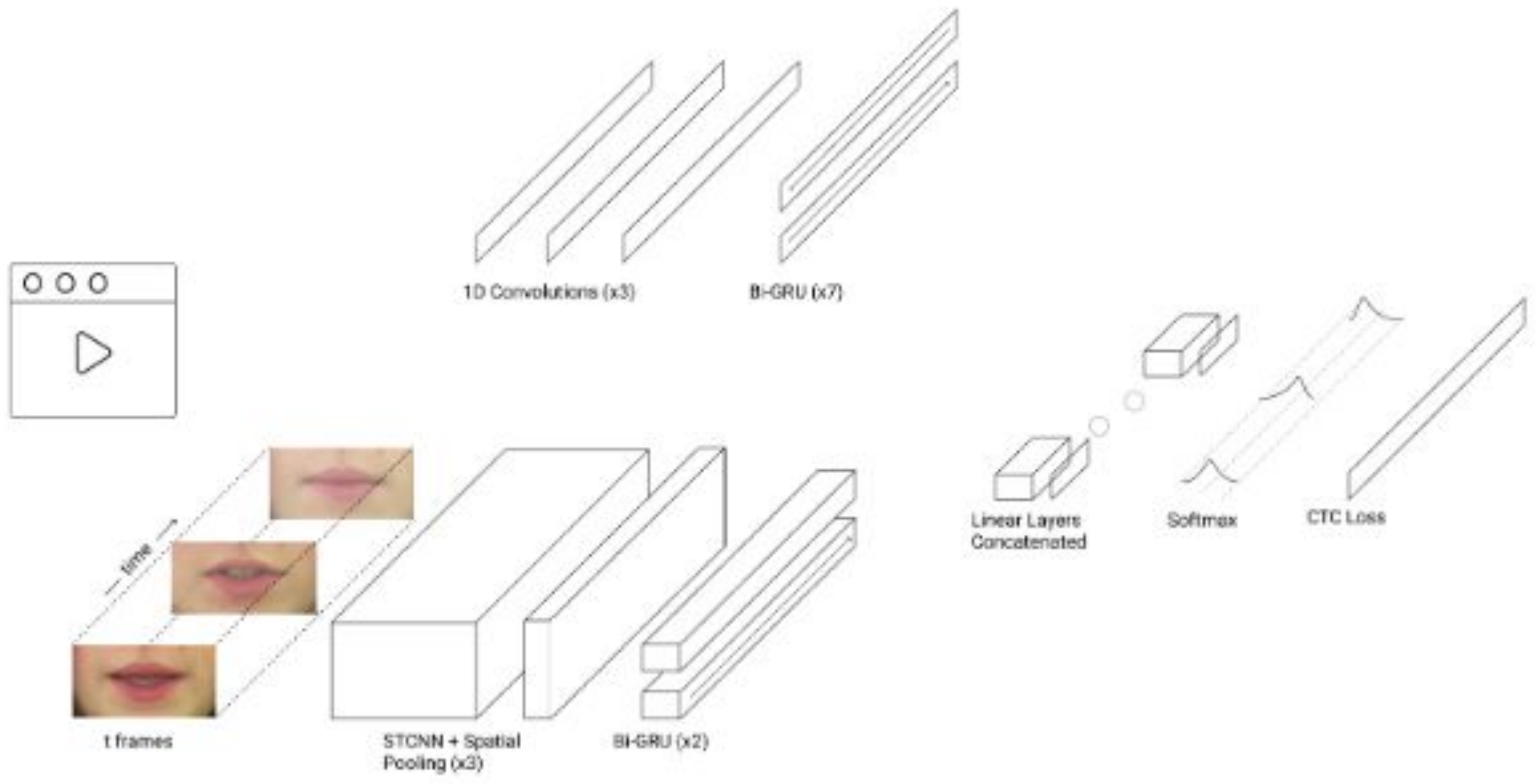
Introduction

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 to success. Real time audio-based captioning is a highly valuable accommodation; however, accuracy in real world settings is limited due to factors such as high noise and distance to speaker. We hope to develop a multi-modal model for captioning that integrates both auditory and visual cues for higher captioning accuracy. Our approach is unique since it relies on fully end to end models, with the goal of improving end to end trained ASR performance through the integration of lip motion inputs in tandem with audio. Such a model would be particularly useful as a noise reduction technique, allowing a single speaker to be isolated from a crowd's background noise. This approach is rooted in lip-reading, a technique commonly used by the hearing impaired to better understand speech.

Models

Approach: We planned to concatenate intermediate outputs from two end to end models (below) to train final classification layers

1. **Lipnet:** End to End trained automated lipreading network (ALR)
2. **DeepSpeech:** End to end trained automated speech recognition (ASR)
3. **Overall Network Architecture:**
   a. ASR and ALR intermediate inputs
   b. Final Softmax classification with CTC loss



Formulas:

1. STCNN: Allows 3D convolutions

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

2. Bi-GRUs: Long term memory retention

$$[u_t, r_t]^T = sigm(W_z z_t + W_h h_{t-1} + b_g)$$
$$h_t = tanh(U_z z_t + U_h(r_t \odot h_{t-1}) + b_h)$$
$$h_t = (1-u_t)\odot h_{t-1} + u_t \odot \tilde{h}_t$$

3. CTC: Accounts for variable input timing

A.

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

B.

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

C.

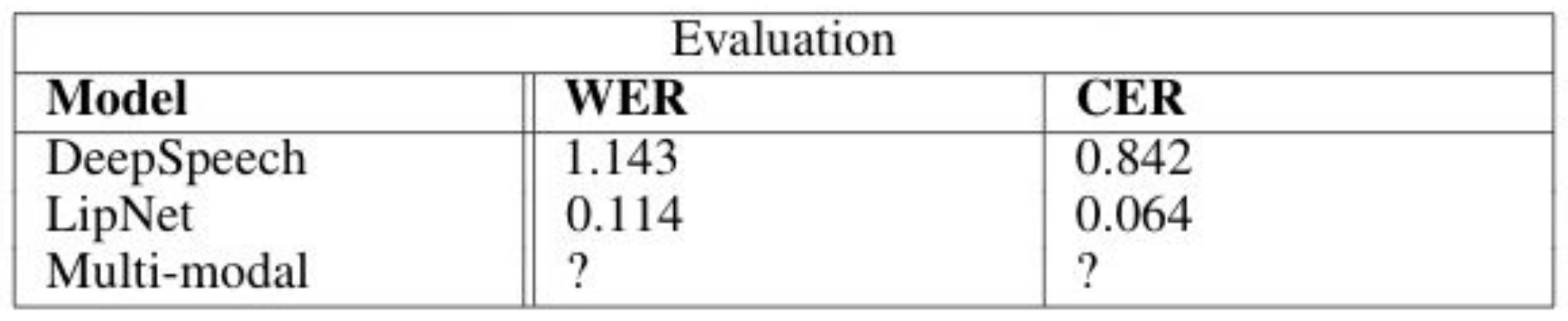
$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{c:\beta(c)=y^{*(i)}} P(c|x^{(i)})$$

Data

- For training data, we relied on the Lombard-Grid corpus, a high quality audio-visual dataset
- Due to use of two end to end networks, we had two data inputs:
  - ASR inputs: 1000 .wav audio files
  - ALR inputs: 1000 corresponding .mpg short, single subject, color video files
- All data came from a clean dataset with no background noise
- Inputs were labeled with timestamp delimited speech transcripts (align files)
- We concatenated intermediate outputs from ASR and ALR to use as inputs for training of final classification layers

Results

- Unable to complete training due to dependency challenges
- High WER for DeepSpeech due to lack of training on GRID Corpus
- Lipnet achieves lower WER and CER due to training and test sets from same distribution; larger training set



Evaluation

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

Future Directions

1. Complete final classification layers of model and implement concatenation
2. Train final classification layers on larger, multi-speaker dataset (GRID Corpus)
3. Adapt model for real time function
   a. Provide intermediate inputs to model as speech is perceived, and update past predictions as new speech is heard
   b. Use of faster hardware

Features

- Raw videos (.mpg) were processed to 75 individual 100x50px mouthcrop frames
  - ALR inputs: [img_channels, frames_n, img_w, img_h] = [3, 75, 100, 50] for video files
- Raw Audio (.wav) files
  - Max amplitude = 1; 25 khz; variable length (~1 sec)

Discussion

- Expect significant improvement by multimodal network in noisy data, due to improved Signal to Distortion Ratio for multi-modal models published in previous literature
- Low noise settings benefit less, due to lower impact of noise reduction
- Direct concatenation of ALR and ASR intermediate outputs may lead to challenges
  - Must normalize inputs and rescale
  - The two ALR and ASR outputs could be desynchronized
- Use of logistic regression on concatenated ALR and ASR outputs could allow us to establish a baseline accuracy

Sources

[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_dls_speech2016, 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].