Deep Scattering End-to-End Speech Recognition

Iyalla John Alamina[1]; David Wilson, PhD[1]; Andrew Crampton, PhD[1]
[1]University of Huddersfield, England, United Kingdom

Abstract

End-to-end discriminative neural network speech models have now become a well established method in Automatic Speech Recognition.

Our Bi-directional Recurrent neural network (Bi-RNN) end-to-end system, is augmented by features derived from a deep scattering network as opposed to the standard Mel Frequency Cepstral Coefficients(MFCC) features used in state of the art acoustic models. These specialised deep scattering features, consumed by the Bi-RNN, model a light-weight convolution network. This work shows that it is possible to build a speech model from a combination of deep scattering features and a Bi-RNN. There has been no record of deep scattering features being used in end-to-end bi-RNN speech models as far as we are aware.

Introduction

CTCC [3] models currently have been developed using standard MFCC features. The model developed in this work employs deep scattering features which compared to MFCC posses a greater number of features of a higher dimension (152 compared to 39). These deep scattering vectors have been shown to perform well on music genre classification[1] and TIMIT phone recognition[6].

Data Set and CTCC Model

GPU training of the speech model architecture for this research (figure 3) was developed from Mozilla deepspeech [5] CTC bi-directional RNN[4] implementation along with the accompanying Mozilla Common voice dataset [2] . The Common Voice Dataset project consists of voice samples in short recordings of an average 5 of seconds per recording. The complete dataset is about 250 hours of recording divided into training, test and development subsets.

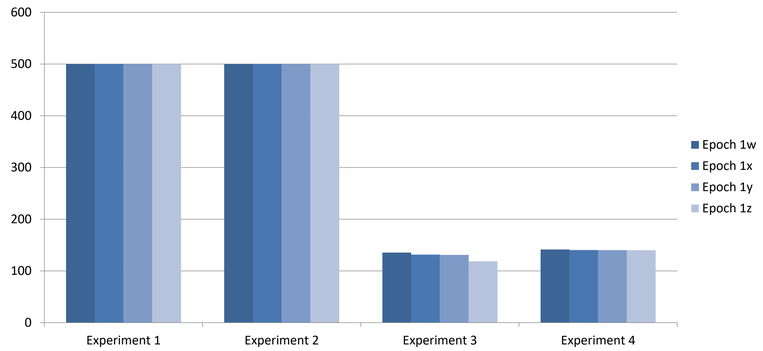

Figure 1. Training Loss, where w < x < y < z are taken arbitrarily across the total number of epochs

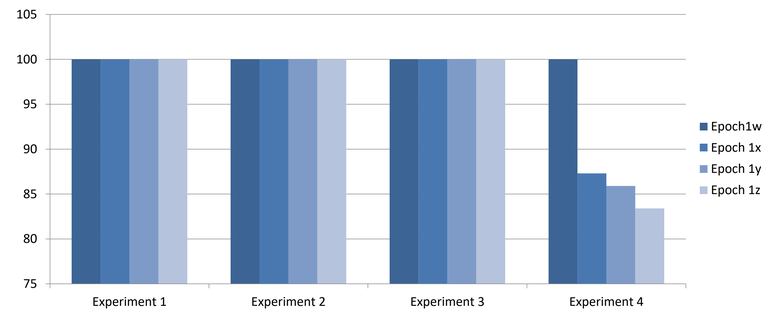

Figure 2. WER, where w < x < y < z are taken arbitrarily across the total number of epochs

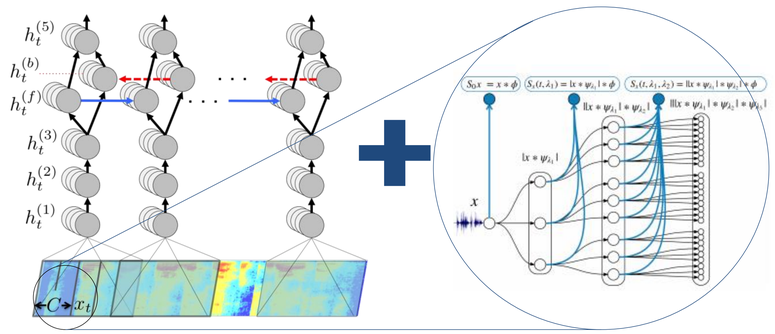

Figure 3. Deep Scattering Bi-RNN CTCC Model

Results

Figures 1, 2 and Table 1. The output of the training produced mostly gibberish when trained in both configurations using only just one hour of training data Training loss reduced significantly once the data was increased to ten hours of training. However word error rates (WER) only showed improvement on the 40 hours dataset.

Table 1. CTCC Model Training Times.

Experiment	Hours of speech	Total Training Time	Estimated Training
1. 2xGPU 10GB RAM	1	7 days	Complete
2. 2xGPU 10GB RAM	10	150 days+	300 days
3. 5xGPU 15GB RA	10	17 hours	Complete
4. 5xGPU 15GB RAM	40	2 days+	10 days

Discussion

The results showed that the training of the model was heading towards a very slow convergence as indicated by the slow decrements in training loss. However, we perceive that given the complete dataset to train the model will not only converge but also show improvements in word error rates.

The next phase in this research will obtain results from MFCC feature-based Bi-RNN speech models to serve as the baseline. Researchers seek partnership(s) to facilitate this.

Conclusion

We show in this work that Deep Scattering features derived from wavelet filter operations on audio data produce viable feature candidates for end-to-end training of Automatic speech recognition models.

Contact

I. John Alamina
University of Huddersfield
Email:john.alamina@hud.ac.uk
Website:www.hud.ac.uk
Phone:07459136287

References

1. Anden, Joakim, and Stephane Mallat. “Deep Scattering Spectrum.” IEEE Transactions on Signal Processing, vol. 62, no. 16, 2014, pp. 4114–4128., doi:10.1109/tsp.2014.2326991.
2. “Common Voice by Mozilla.” Common Voice, voice.mozilla.org/.
3. Graves, Alex. “Connectionist Temporal Classification.” Studies in Computational Intelligence Supervised Sequence Labelling with Recurrent Neural Networks, 2012, pp. 61–93., doi:10.1007/978-3-642-24797-2_7.
4. Hannun, Awni Y., et al. "First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns." arXiv preprint arXiv:1408.2873 (2014).
5. Mozilla. “Mozilla/DeepSpeech.” GitHub, 18 June 2019, github.com/mozilla/DeepSpeech.
6. Zeghidour, Neil, et al. “Learning Filterbanks from Raw Speech for Phone Recognition.” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, doi:10.1109/icassp.2018.8462015.