Ryan Hartsfield & Garrett Lewellen

CSCI/LING 5832

2015-03-12

Final Project Proposal

For the final project in this class, our group will investigate a variety of current approaches to Automatic Speech Recognition (ASR). Our focus will be on deep learning techniques and how they are utilized for acoustic models. We have not settled on our three primary papers, but each will likely focus on a different model, with additional supporting papers being used to fill in relevant gaps. We have tentatively agreed to divide the paper and work along the following lines:

Our first section (1) of the literature review will aim to provide some historical context and to summarize current research. Ryan will begin by laying out the traditional approach to ASR, using Gaussian mixture models (GMM), and the problems that accompany them. He will then discuss the rise in deep learning techniques, and how they have impacted the field of ASR. We will need to explore a variety of journals and conferences for recent trends in order to make this section fairly comprehensive.

The following section (2) will be broken up in order to discuss three alternatives to the use of GMMs for the acoustic model: deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN). A basic overview of each model will be given explaining their overall differences. We will then move on to discuss how each takes into account linguistic details. Garrett will focus on DNNs and RNNs while Ryan will explore CNNs.

In the next section (3), Garrett will present and discuss the algorithms used to train the different neural network models. Decoding will be passed over as that is part of the traditional GMM-HMM setup.

In our fourth section, Garrett will look at published results on word error rates for all four methods on common test sets such as Switchboard, TIMIT, etc. Finally, Garrett will summarize what we have covered and make predictions on which direction looks most promising for further research.

Kayla Stearns has decided to withdraw from the course due to personal circumstances and will not be participating in the group.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *22*(10), 1533-1545.

Adami, A. G. (2010). Automatic speech recognition: From the beginning to the Portuguese language. In *The Int. Conf. on Computational Processing of Portuguese (PROPOR). Rio Grande do Sul: Porto Alegre*.

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, *20*(1), 30-42.

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (pp. 6645-6649). IEEE.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, *29*(6), 82-97.

Hinton, G., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.

Jaeger, H. (2002). *Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the" echo state network" approach*. GMD-Forschungszentrum Informationstechnik.

Larochelle, H., Bengio, Y., Louradour, J., & Lamblin, P. (2009). Exploring strategies for training deep neural networks. *The Journal of Machine Learning Research*, *10*, 1-40.

Seide, F., Li, G., & Yu, D. (2011, August). Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. In *Interspeech* (pp. 437-440).

Yu, D., & Deng, L. (2015). *Automatic Speech Recognition*. Springer.