Conversational Speech Transcription Using Context-Dependent Deep Neural Networks

Anonymous Authors[1]

Abstract

Context-Dependent Deep-Neural-Network HMMs, or CD-DNN-HMMs, combine the classic artificial-neural-network HMMs with traditional context-dependent acoustic modeling and deep-belief-network pre-training. CD-DNN-HMMs greatly outperform conventional CD-GMM (Gaussian mixture model) HMMs: The word error rate is reduced by up to one third on the difficult benchmarking task of speaker-independent single-pass transcription of telephone conversations.

1. Introduction

Context-dependent deep-neural-network HMMs (CD-DNN-HMMs) are a recently proposed acoustic-modeling technique for HMM-based speech recognition [1, 2] that combines three techniques: the hybrid approach of modeling HMM state emission densities through scaled likelihoods from an MLP [3]; traditional acoustic co-articulation modeling of speech through context-dependent phoneme models (crossword triphones with tied states); and deep networks, leveraging Hinton's deep-belief-network (DBN) pre-training procedure. The power of this model was first shown through a 16relative recognition error reduction over conventional CD-GMM-HMMs on a business search task [1, 2]. This work describes our subsequent efforts [4] on scaling it up in terms of training-data size (from 24 hours to 309), model complexity (from 761 output classes to 9304), depth (up to 9 hidden layers), and task (from voice queries to speech-to-text transcription). The model achieves a one-third word-error reduction on the publicly available benchmark of phone-call transcription (Switchboard 2000 NIST Hub5/RT03S-FSH).

2. The Context-Dependent Deep Neural Network HMM

In HMM-based large-vocabulary speech recognition, speech is modeled by hidden Markov models (HMMs), where each word's HMM is decomposed into phoneme HMMs. These are commonly three-state left-to-right HMMs, where each state's emission probability is a mixture of Gaussians (GMM). Co-articulation is modeled by context-dependent (CD) phonemes, such as triphones. Due to data scarcity, triphone states are commonly tied with similar other states. A limitation of GMMs is their difficulty to use high-dimensional features, such as multiple consecutive frames of short-term spectral features. To address this, it was proposed in the early 90's to replace GMMs with artificial neural networks (ANNs). The ANNs are trained to classify observation vectors into HMM state labels [3], and state posteriors are converted to scaled likelihoods for use as HMM state emissions. However, these early attempts were limited to shallow models (1–2 hidden layers) and monophone states as ANN outputs (even when CD phones were modeled) [5, 6]. The CD-DNN-HMM extends these hybrid ANN-HMMs two-fold: First, we model tied triphone states directly. It was long assumed that thousands of triphone states were too many to be accurately modeled by an MLP, but [1] has shown that it works very well. Secondly, we use a deep MLP with with many hidden layers. Many layers of simple non-linearities can model complicated non-linearities and are more efficient in representing structures since lower-layer feature detectors can be reused by the higher-layer feature detectors. Also, each layer is constrained by the adjacent layers and so it is less likely to cause over-fitting (although it is more likely to cause under-fitting). The key enablers to the training of these were the deep belief network (DBN) pre-training algorithm proposed by Hinton [7], as well as the advent of affordable, massively parallel computing devices (GPGPUs). Algorithm 1 summarizes the training procedure [4]. First, a conventional CD-GMM-HMMs is trained. Secondly, the DNN, after initialization as a DBN, is trained as a frame classifier, where the class labels are state labels assigned to each input frame through forced alignment using the CD-GMM-HMM. Midway, the alignment is updated once using the DNN model.

acoustic model and training	recognition mode	RT03S	Hub5'00	voice mails	teleconf
CD-GMM 40-mix, SWB 309h	single-pass SI	FSH=27.4 SW=37.6	23.6	30.8	33.9
CD-DNN 7 layers x 2048, SWB 309h (this work) single-pass SI 18.5 27.5 16.1 22.9 24.4 (rel. change CD-GMM → CD-DNN)	single-pass SI	FSH=18.5 SW=27.5	16.1	22.9	24.4
CD-GMM 72-mix, Fisher 2000h	multi-pass adaptive	FSH=18.6 SW=25.2	17.1	-	-

Table 1. Standard CD-GMM-HMM vs. CD-DNN-HMM for single-pass speaker-independent recognition on five speech-to text test sets (word-error rates in %), and for comparison our group's best-ever CD-GMM-HMM result for three set.

3. Experimental Results

We evaluate the effectiveness of CD-DNN-HMMs on speech-to-text transcription of telephone conversations, a considerably difficult task. We use the publicly available 309-hour 'SWBD-I' training set and associated benchmark sets, as well as two in-house sets. Recognition is single-pass without speaker adaptation. Table 1 shows that compared to our discriminatively trained CD-GMM-HMM baseline, the word-error rate (WER) on the 'RT03S-FSH' benchmark drops from 27.4 % to 18.5%—a rather significant one-third reduction. Much of the gain carries over to less well-matched sets (voicemail, teleconferences). The 309h CD-DNN-HMM system also reaches our best multi-pass system (18.6%, last row), which uses 6 times as much acoustic training data and speaker adaptation. Further experiments show that the deep network is indeed critical—a shallow 1-hidden-layer network using the same number of parameters as the 7-hidden-layer one leads to five percentage points worse word-error rate. We also find that as an alternative to DBN pretraining, it is possible to discriminatively pre-train the model in a supervised layer-growing fashion

4. Conclusion

By using CD-DNN-HMMs, a one-third word-error reduction has been achieved on a difficult benchmark task, compared to a discriminatively trained conventional CD-GMM-HMM [4]. Recent improvements on smaller tasks [1, 2] do carry over to larger corpora and speech-to-text transcription. The remarkable accuracy gains are due to three factors: direct modeling of tied triphone states through the DNN; effective exploitation of neighbor frames by the DNN; and the efficient and effective modeling ability of deeper networks.

References

[1] D. Yu et al., "Roles of Pretraining and Fine-Tuning in Context-Dependent DNN-HMMs for Real-World Speech Recognition," Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2010.

[2] G. Dahl et al., "Context-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition", IEEE Trans. Speech and Audio Proc., Special Issue on Deep Learning for Speech Processing

[3] S. Renals et al., "Connectionist Probability Estimators in HMM Speech Recognition," IEEE Trans. Speech and Audio Proc., January 1994.

[4] F. Seide, G. Li, and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," Interspeech, 2011.

[5] H. Franco et al., "Context-Dependent Connectionist Probabilty Estimatation in a Hybrid Hidden Markov Model–Neural Net Speech Recognition System," Computer Speech and Language, vol. 8, 1994.

[6] J. Fritsch et al., "ACID/HNN: Clustering Hierarchies of Neural Networks for Context-Dependent Connectionist Acoustic Modeling," Proc. ICASSP, May 1998.

[7] G. Hinton et al, "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, vol. 18, 2006.

[8] D. Rumelhart et al., "Learning Representations By Back-Propagating Errors," Nature, vol. 323, 1986