CS224 Final Project

Re-Alignment Improvements for Deep Neural Networks on Speech Recognition Systems

Firas Abuzaid

Abstract

The task of automatic speech recognition has traditionally been accomplished by modeling the problem as a Hidden Markov Model (HMM)[1]. Gaussian Mixture Models (GMM) have been used to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. However, recent research has shown that current speech recognition systems which use Deep Neural Networks (DNN) with many hidden layers have been shown to outperform GMMs on a variety of speech recognition benchmarks, sometimes by a large margin[2],[3]. We investigate further possible improvements of the DNN-HMM hybrid, by examining the role of forced alignments in the training algorithm.

Introduction

In the context of speech recognition, the GMM-HMM hybrid approach has a serious shortcoming – GMMs are statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space. Because speech is typically produced by modulating a relatively small number of parameters, it has been hypothesized that the true underlying structure is much lower-dimensional than is immediately apparent in a window that contains lots of coefficients.

Deep Neural Networks, on the other hand, have the potential to learn much better models of data that lie on or near a non-linear manifold. Many studies have confirmed this hypothesis, with DNN systems outperform GMMs by approximately 6% for Individual Word Error Rates (IWER)[2]. The networks are trained to optimize a given training objective function using the standard error backpropagation procedure. In a DNN-HMM hybrid system, the DNN is trained to provide posterior probability estimates for the different HMM states. Typically, cross-entropy is used as the objective and the optimization is done through stochastic gradient descent (SGD)[4],[5]. For any given objective, the important quantity to calculate is its gradient with respect to the activations at the output layer. The gradients for all the parameters of the network can be derived from this one quantity based on the back-propagation procedure.

One significant hurdle in training speech recognition systems is determining the appropriate alignment between word sequences and acoustic observations. Typically, the acoustic data is divided into frames, with a frame size that's approximately 25 ms, which then must be aligned with the sequence of words in the training set[1]. Usually, to determine these alignments, we start by obtaining an initial

"forced" alignment, and then improving the alignment through multiple passes through the DNN. Specifically, we start by running the Viterbi algorithm over an initial pass of the training data, which produces our initial set of reference state labels. The Viterbi algorithm, however, is "forced" to pass through particular word sequences, those which most likely seem to initially match the acoustic observations in our training data. This algorithm leverages the HMM as well, which is initialized to have random probabilities in its state transition matrices, as is custom in most E-M algorithms.

Therefore, we effectively "force" an alignment of the acoustics with the word transcripts in our training set, and we run the Viterbi algorithm to pass through those specific word sequences. This forced alignment now represents the single best state path corresponding to the training observation sequence, and these initial alignments between HMM states and acoustic observations can then be fed into our DNN as part of our back-propagation algorithm.

It has been shown that using the forced alignment approach – with the initially randomized HMM – is the accepted method for HMM-based Speech Recognition systems. An interesting area of research, then, has focused on whether this same procedure can be used throughout multiple epochs of the DNN. Specifically, we're investigating whether recomputing these forced alignments after every pass through the DNN, we'll lead to a further increase in accuracy. Recent research – see Vesely et al.'s (2013) study[6] – demonstrates that computing these re-alignments could indeed be effective in improving the accuracy of speech recognition systems.

Experiments

We ran two sets of experiments to test our hypothesis. For our first set of experiments, we started with a smaller training set, simply to get an early indication of whether our hypothesis was correct.We trained the DNN on 60 hours of telephone conversation recordings from the Switchboard corpus, and we configured our neural network to have 5 hidden layers with 1200 nodes. Finally, we trained the DNN for 4 epochs.

To test our hypothesis, we computed a re-alignment after each epoch, using the Viterbi algorithm that's used initially to compute the "forced" alignment before the first epoch. We simultaneously ran a separate DNN that did not compute these re-alignments after each epoch. We then compare the accuracy for both recognizers by evaluating them on the training set and the test set provided in the Kaldi open source toolkit.

Our second experiment is identical to the first, except we train on a much larger corpus of training data – 300 hours from Switchboard. We also modified the neural network to contain 2048 nodes instead of 1200.

To implement our DNN Speech Recognizer, we used the Kaldi-Stanford codebase, a variant of the Kaldi open source toolkit[7], maintained by Professor Andrew Ng's AI research group.[1]

[1] Note: This project was done in conjunction with Andrew Ng's research group, specifically with graduate students Andrew Maas and Christopher Lengerich. The Kaldi-Stanford codebase is closed-source, and I, unfortunately, did not have permission to include it as part of my submission. However, all analysis code was created individually on

Results

Word Error Rate (WER) and Individual Word Error Rate (IWER)

We measured the WER and IWER[8] for the DNN on both the training set and test set. Our results for the two different experiments are as follows:

Small Dataset – 60 hours of Training Data

Epoch	**Training Set** Re-alignment (WER / IWER)	**Training Set** No re-alignment (WER / IWER)	**Test Set** Re-alignment (WER / IWER)	**Test Set** No re-alignment (WER / IWER)
1	24.7 / 25.5	27.3 / 30.0	28.8 / 29.7	30.5 / 31.4
2	21.2 / 22.1	23.5 / 24.2	26.9 / 26.7	28.4 / 29.2
3	19.5 / 20.2	21.7 / 22.4	25.6 / 26.3	26.8 / 27.5
4	18.6 / 19.3	20.6 / 21.2	24.3 / 25.0	25.3 / 26.0

Large Dataset – 300 hours of Training Data

Epoch	**Training Set** Re-alignment (WER / IWER)	**Training Set** No re-alignment (WER / IWER)	**Test Set** Re-alignment (WER / IWER)	**Test Set** No re-alignment (WER / IWER)
1	15.7 / 16.3	16.8 / 17.4	18.8 / 19.4	17.9 / 18.5
2	15.5 / 16.2	16.5 / 17.0	18.0 / 18.5	17.7 / 18.2
3	15.4 / 15.9	16.4 / 16.8	17.6 / 18.0	17.1 / 17.7
4	15.4 / 15.9	16.3 / 16.8	17.4 / 17.9	17.1 / 17.5

Frame Errors

For the experiments in which we calculated the re-alignments, we computed the frame accuracy with the reference frame labels as provided in the Stanford-Kaldi codebase. The results are shown below:

Small Dataset

Epoch	**Frame Accuracy**

Large Dataset

Epoch	**Frame Accuracy**

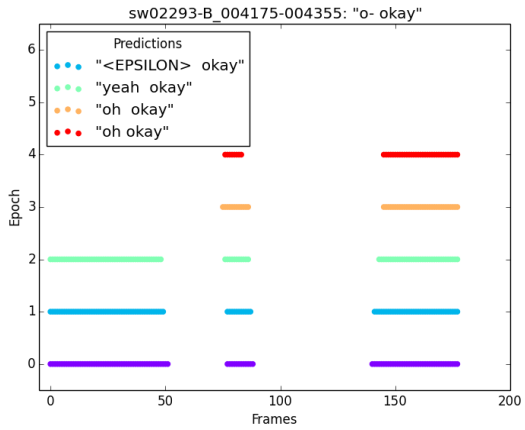
my own, and is included as part of my submission.

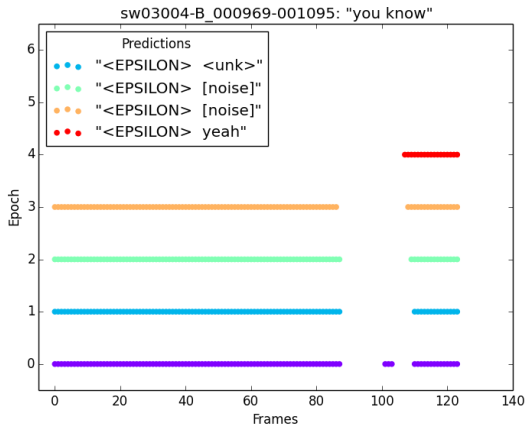
1	41.85%	1	52.85%
2	47.23%	2	56.71%
3	49.11%	3	58.28%
4	50.18%	4	57.04%

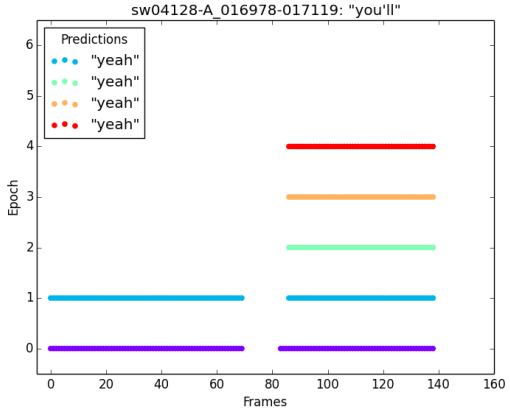
Alignments

Another key metric that we focused in our analysis was alignment variation between epochs. In the computed alignments, each frame gets assigned to a certain phone, and analyzing the shift in phones between epochs is crucial in determining whether the re-alignments lead to a substantive increase in accuracy.

In particular, the `silence` phone (denoted as 'sil') is most significant, as these frames effectively become the hypothesis demarcations between lexical tokens in our hypothesis. We generated alignment plots for the various utterances in our corpus to visually track how the alignment of the `silence` phone was shifting between epochs – below are three example plots taken from the three utterances that exhibited the most variation in their alignments, along with the corresponding hypotheses for epochs 1 through 4 (note this only for the small training set):







Part-of-Speech Tagging

As a further metric to analyze our results, we wanted to measure the grammatical accuracy of our speech recognizers – how often were the erroneous hypotheses being made by our recognizer grammatically incorrect (e.g. ‘too’ vs. ‘to’ vs. ‘two’, ‘money’ vs. ‘many’, etc.)?

Using the Stanford NLP group’s Part-Of-Speech Tagger[9], we tagged each word in both the Reference utterance and the Hypothesis utterance, and then calculated the fraction of substitution errors that also differed in the corresponding part-of-speech tag. The results are as follows (note that this was only done for the small training set):

Small Training Set – Re-alignment Experiments

Epoch	POS Error Rate
1	84.04%
2	83.26%
3	82.46%
4	81.40%

Analysis

As the results indicate, our hypothesis was correct for the smaller training set, but was incorrect for the larger training set. This is a somewhat surprising result; one potential explanation for this result could be the hyperparameters chosen for the larger neural network, particularly the number of nodes. This could lead to overfitting and higher variance in the Acoustic Model generated by the DNN – further experiments need to be run that tune these hyperparameters more appropriately.

Moreover, the expectation that the frame accuracy and Word Error Rate would be highly correlated turned out not to be true; although we saw frame accuracy improve, this did not necessarily translate to a lower Word Error Rate. (A separate, but related, concern is the small decrease in frame accuracy in the last epoch of the large training set experiment.) One would expect that an improved frame error rate would lead to a better Word Error Rate; however, it is possible that we are observing a trivial increase in frame accuracy: the frames that are being correctly classified don’t have a significant effect on improving the Acoustic Model in our system. The total number of frames that are correctly labeled with the proper senone may be increasing, but these frames could be at the margin of our utterances, for example. Or, even more likely, these frames’ senones map to phones that don’t directly influence the DNN’s ability to determine the temporal boundaries between words.

This, is why, in our alignments analysis, we focused on the `silence` phone. Our intuition was that an alignment for a particular utterance that undergoes a significant shift in its labeling of the `silence` phone should lead to a more accurate prediction of the speech recognizer. This would confirm the

theory that it's not how many frames are accurately labeled – it's which frames are accurately labeled.

Based on the analysis we ran, this theory is partially correct. The first example above certainly corroborates this hypothesis – as the alignments change, we get a much more accurate prediction of the utterance. But the other two examples don't provide confirmation; in the last example, the prediction doesn't change at all, despite the drastic shift in alignment. So, although the alignment has changed dramatically, it hasn't approached the optimal alignment of the silence phones for that utterance. One possibility worth exploring is continuing to train the neural network on more epochs, and seeing if the alignment shifts yet again. Broadly speaking, it's possible that our neural network requires more training epochs to maximize its objective function, and this intuition applies to the alignments as well.

Lastly, the Part-Of-Speech analysis demonstrates that, although our experiments focus on addressing the Acoustic Model, perhaps our attention should shift to improving the Language Model. In particular, this could easily address the confusion of homonyms in our recognizer. This error from our Speech Recognizer is particularly instructive:

Utterance ID: sw02125-A_026122-026613
Reference: "we've been wanting to start camping again this year **too** uh my oldest"
Hypothesis: "we've been wanting to start camping again this year **to** uh my oldest"

In these situations, the Acoustic Model simply cannot distinguish between these two words – if we improve the Language Model, however, we could be able to address these errors.

Future Work

In the future, we'd like to further examine the effect of further epochal training for our DNN, as well as fine-tuning the hyperparameters of the neural network. Lastly, we'd like to explore alternative language models – the work of Mikolov et al.[10] demonstrates that the language model for DNNs can be improved, so coupling those improvements with re-alignments could yield a further increase in accuracy.

References

[1] M.J.F. Gales and S.J. Young (2008). The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing.
[2] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, 29(November), 82–97.
[3] Dahl, G., Yu, D., Deng, L., & Acero, A. (2011). Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, Special Issue on Deep Learning for Speech and Langauge Processing, 1–13.
[4] Povey, D., & Woodland, P. C. (2002). Minimum Phone Error and I-Smoothing for Improved Discriminative Training. ICASSP.

[5] Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., & Visweswariah, K. (2008). Boosted MMI for model and feature-space discriminative training. ICASSP (pp. 4057–4060). IEEE.
[6] “Sequence-discriminative training of deep neural networks”, K. Vesely, A. Ghoshal, L. Burget and D. Povey, to appear in: Interspeech 2013
[7] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Veselý, K., Goel, N., et al. (2011). The kaldi speech recognition toolkit. ASRU.
[8] Sharon Goldwater, Dan Jurafsky, Christopher D. Manning, Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates, Speech Communication, Volume 52, Issue 3, March 2010, pp. 181-200.
[9] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.
[10] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky,Sanjeev Khudanpur: Recurrent neural network based language model, In: Proc. INTERSPEECH 2010