Ensemble methods and efficient decoding

Jeremy H. M. Wong

Department of Engineering, University of Cambridge
Trumpington Street, CB2 1PZ Cambridge, England
jhmw2@cam.ac.uk
Supervised by: Prof. *Mark J. F. Gales*

Abstract

The performance of automatic speech recognition can often be significantly improved by combining an ensemble of models together. When using ensemble methods, natural questions arise regarding the most effective methods of obtaining diversity and performing combination. Also, there are several methods that can improve the computational efficiency when decoding through the ensemble. This PhD project is focused around addressing each of these issues.

Index Terms: speech recognition, ensemble, combination, student-teacher

1. Introduction

Ensemble combinations have often been found to outperform single models in Automatic Speech Recognition (ASR) [1–4]. The performance gains are attributed to the possibility of correcting errors that occur in each model [1], reducing the likelihood of selecting a poor model, and increasing the space of possible models [5]. Furthermore, these gains are primarily dependent on the ensemble diversity and the accuracy of the individual models [6]. There are many possible methods of introducing diversity into an ensemble. Moreover, there are also many possible schemes to combine the models together. Unfortunately, the computational demand of performing hypothesis-level combinations, such as ROVER [1], scales linearly with the ensemble size. As such, more efficient schemes need to be investigated to allow fast decoding while maintaining the standard of performance. This paper summarises the research directions taken in this project to address these issues of ensemble diversity, combination, and decoding efficiency.

2. Diversity methods

The models within an ensemble need to be diverse in order to achieve significant combination gains [6]. An intuitive interpretation of this is that models with uncorrelated output errors may possibly correct for each other's mistakes. From a Bayesian perspective, having a more diverse ensemble implies marginalising over models that span a wider region of the space of possible models.

Diversity can be introduced into the ensemble by injecting randomness at any stage along the ASR pipeline. From the bottom up, diversity can be introduced into the features, trained transform and phonetic classifier, phonetic clustering, transition model, and language model. The diversity introduced can be intrinsic, as in Deep Neural Network (DNN) initialisation, bagging [8], dropout [9], and random forests [10], which rely on random sampling from a prior to achieve diversity. It can also be extrinsic, as in Adaboost [11] and negative correlation learning [12], which explicitly train the models to be different. Furthermore, it is possible to combine different model types, such as combining DNN and Recurrent Neural Network (RNN) acoustic models.

In this project, several intrinsic diversity methods that operate at different stages along the ASR pipeline are compared for the diversity that they provide and their combination gains. The novel intrinsic diversity methods of Echo State Network (ESN) [13] random projection and RNN Language Model (RNNLM) [14] random initialisation are also investigated.

2.1. Feature representation diversity

Diversity can be introduced into the feature representation by performing a random projection of the feature vectors. Ideally, the resulting information content within the projection should be different for each model within the ensemble, allowing each model to learn to classify based on these different information. One possible random projection that can achieve this is the ESN. The ESN is a single RNN layer that has random weights that remain untrained. By varying the prior distribution that the ESN weights are sampled from, it is possible to influence the time scales for which recurrent information is stored. This project investigates using the ESN projection as the input to a standard hybrid architecture.

2.2. Language model diversity

At the top end of the ASR pipeline, it is possible to obtain improved performance by re-scoring the decoding lattice using an RNNLM [14]. As with other neural network architectures, the RNN is trained starting from a random initialisation, and its training is usually a non-convex problem. It may therefore be possible to obtain significant diversity from RNNLMs that have been trained starting from different weight initialisations.

2.3. Diversity metrics

When investigating different diversity methods, it is important to be able to objectively compare the amount of diversity that each method provides. Depending on the combination method used, it may be more important to maximise the diversity at the frame or hypothesis level. At the frame level, one possible diversity metric is the KL-divergence between the frame posteriors of each model, represented by the DNN output. At the hypothesis level, it is difficult to compute the KL-divergence between word or sentence posteriors, because the hypotheses are often stored in pruned lattices, which do not encompass the space of all possible hypotheses. An alternative diversity metric is to compute the WER of one model, using the transcription generated by another model as a reference. This provides a mea-

sure of how many words are recognised differently between the two models.

3. Combination schemes

Model combination is a crucial aspect of ensemble methods. Part of this project is involved in analysing the relationships between the multitude of existing combination schemes. From a Bayesian perspective, each of the combination schemes can be viewed as applying a different set of assumption when marginalising over possible models. Combination schemes at the hypothesis level, such as Minimum Bayes' Risk (MBR) combination decoding [4], confusion network combination [2], and ROVER [1], aim at obtaining a better hypothesis posterior. Frame level schemes, such as linear ensembles [3] and joint decoding [7] aim to obtain a better frame posterior. There are various advantages and disadvantages in performing each of these combination methods, in terms of the diversity they utilise, their efficiency, and the constraints they enforce.

4. Efficient decoding

Although ensemble methods may be able to provide performance gains over single models, the computational demand of standard combination methods tends to scale linearly with the ensemble size. This can especially hinder real-time applications. This project investigates methods to improve the computational efficiency of decoding. One method that can partially alleviate this demand is to perform a frame-level combination, such as joint decoding, instead of a hypothesis-level combination. This requires the generation and processing of only a single decoding lattice for the whole ensemble.

4.1. Student-teacher training

It is possible to train a single student model to emulate the performance of a teacher ensemble [15]. Decoding then only needs to be performed through the single student model. Existing work on student-teacher training have largely investigated training at the frame level, where the student DNN is trained to match the outputs of the teacher ensemble DNNs, thereby emulating their frame posteriors. Sequence training criteria have been shown to outperform frame-level criteria. This project investigates incorporating sequence training into the student-teacher framework. One possibility of achieving this is to train the student model to emulate the hypothesis posteriors of the teacher ensemble. A potential criteria is to minimise the KL-divergence between the student and teacher ensemble hypothesis posteriors, which can be seen as a generalisation of the Maximum Mutual Information (MMI) criterion. Since MBR criteria have been shown to outperform MMI, it may prove beneficial to investigate methods of incorporating aspects of MBR criteria into student-teacher training.

4.2. Methods for random forest ensemble

The random forest method is capable of providing a relatively large amount of diversity. However, it is not a trivial task in trying to improve its ensemble decoding efficiency, because each model has a different set of phonetic clusters. Frame-level combination and student-teacher methods therefore need to be modified to accommodate such a diversity. This project shall investigate such modifications.

5. Conclusion

This paper has presented a summary of the research directions taken in an investigation into ensemble methods. The main foci of this project are to investigate the diversity methods, combination schemes, and techniques to improve decoding efficiency while maintaining the standard of performance of the ensemble.

6. References

[1] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recogniser output voting error reduction (ROVER)," in *ASRU*, Santa Barbara, USA, Dec 1997, pp. 347–354.

[2] G. Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Speech Transcription Workshop*, vol. 27, 2000.

[3] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *INTERSPEECH*, Singapore, Sep 2014, pp. 1915–1919.

[4] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, vol. 25, no. 4, pp. 802–828, Oct 2011.

[5] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, Cagliari, Italy, Jun 2000, pp. 1–15.

[6] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, Oct 1990.

[7] H. Wang, A. Ragni, M. J. F. Gales, K. M. Knill, P. C. Woodland, and C. Zhang, "Joint decoding of tandem and hybrid systems for improved keyword spotting on low resource languages," in *INTERSPEECH*, Dresden, Germany, Sep 2015, pp. 3660–3664.

[8] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug 1996.

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, Jun 2014.

[10] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of ASR systems using randomized decision trees," in *ICASSP*, Philadelphia, USA, Mar 2005, pp. 197–200.

[11] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," AT&T Bell Laboratories, Tech. Rep., 1995.

[12] Y. Liu and X. Yao, "Ensemble learning via negative correlation," *Neural Networks*, vol. 12, no. 10, pp. 1399–1404, Dec 1999.

[13] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks," Fraunhofer Institute for Autonomous Intelligent Systems, Tech. Rep., Jan 2010.

[14] T. Mikolov, M. Karafiát, L. Burget, J. H. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH*, Makuhari, Japan, Sep 2010, pp. 1045–1048.

[15] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *KDD*, Philadelphia, USA, Aug 2006, pp. 535–541.