Scalable Recurrent Neural Network Language Models for Speech Recognition

Xie Chen

University of Cambridge Engineering Department, Cambridge, U.K.
xc257@eng.cam.ac.uk
Supervised by Prof. Mark Gales

Abstract

Language model is a vital part in modern state-of-art ASR systems. N-Gram LMs have been dominating the area of language modelling during last several decades. Recently, recurrent neural network language models (RNNLMs) present promising performance in many areas and various tasks [1, 2, 3, 4, 5, 6, 7, 8]. In this thesis, we aim to further explore recurrent neural network in the application of automatic speech recognition from the aspect of language models. RNNLMs have good ability of generalization and modelling long term history. However, the success of RNNLMs was mainly reported on small train corpus and N-best rescoring is used for speech recognition in previous research, which limits its potential application. It is not easy to do lattice rescoring and parallel training of RNNLMs due to its long-term history character. Efficient lattice rescoring and training are studied for speech recognition. We then study the efficient adaptation of RNNLMs. Topic model based adaptation of RNNLMs is presented in Chapter 4, which makes good use of the context information and improvements on both perplexity and WER are obtained on state-of-art ASR system. The next chapter will investigate the integration of metadata in RNNLM for ASR. The last chapter draws the conclusion and discuss the future work.

1. Thesis Structure

1.1. Automatic Speech Recognition

Chapter 1 gives an general introduction of automatic speech recognition (ASR) systems. Acoustic and language models in ASR system are described. Acoustic model (Gaussian Mixture Model and Deep Neural Network) based on Hidden Markov Model and related adaptation are discussed, followed by language model and its adaptation. Lastly, search (decoding) on LVCSR system is also briefly described.

1.2. Previous work on Recurrent Neural Network Language Models

Chapter 2 introduces recurrent neural network language models (RNNLMs). RNNLMs become increasingly popular choice in many areas such as speech recognition, machine translation due to their promising performance. RNNLMs are able to model long term history and realize implicit parameter sharing by project each word into a low and continues space, which give good generalization and works quite well with combination of N-Gram language model. Previous work on RNNLMs will be discussed. The long-term history characters of RNNLMs also constrain its potential application. e.g. difficult for lattice rescoring and long train time. Besides, the adaptation of RNNLMs is not easy. These issues will be studied in this thesis in the following section.

1.3. Efficient Lattice Rescoring of RNNLMs

In most previous work, RNNLMs were used for N-Best rescoring due to its intrinsic long-term history character. Although promising improvements are obtained compared to N-Gram using Viterbi decoding. lattice is still very useful and important in many applications. e.g. confusion network (CN) decoding is able to be applied to further improve ASR performance. Nevertheless, lattice is important for key word spotting and confidence score calculation. Two algorithms are proposed for efficient lattice rescoring in Chapter 3 [9]. The first method uses an n-gram style clustering of history contexts. The second approach directly exploits the distance measure between recurrent hidden history vectors. Both methods produced 1-best performance comparable with a 10k-best rescoring baseline RNNLM system. Significant lattice size compression and consistent recognition performance improvements after confusion network (CN) decoding were also obtained over the prefix tree structured N-best rescoring approach.

1.4. Efficient Training and Evaluation of RNNLMs

Chapter 4 describes efficient training and evaluation of RNNLMs. In previous work, class based RNNLMs are
widely used on CPUs as it could reduce computation significantly. However, it is difficult to parallel and normally be trained one sample by one sample, which limits the amount of corpus to be used. An alternative approach that allows RNNLNs to be efficiently trained on GPUs with bunch mode is proposed [10]. This enables larger quantities of data to be used, and networks with an unclustered, full output layer to be trained. One potential issue associated with this form of RNNLM with unclustered, full output layer is its heavy computation, which becomes a bottleneck in test time on CPU. To solve this problem, two strategies are explored to avoid normalisation on output layer for fast RNNLM evaluation [11, 12]. The first way introduces variance regularisation into training of RNNLNs by explicitly constrain the variance of normalisation term. As an alternative choice, noise contrastive estimation (NCE) is also investigated for RNNLM training, which constrains the variance of normalisation term implicitly. Both of these two methods allow RNNLM to be used without normalisation.

1.5. RNNLNs Adaptation

Chapter 5 investigates the topic model based adaptation during the training and testing of RNNLNs. In previous research RNNLNs have normally been trained on well-matched in-domain data. The adaptation of RNNLNs remains an open research area to be explored. In this thesis, topic based RNNLM adaptation techniques are investigated. A number of techniques including Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation and Hierarchical Dirichlet Processes are used to extract show level topic information. These were then used as additional input to the RNNLM during training, which can facilitate unsupervised test time adaptation. Experiments show that both improvements on perplexity and word error rate are obtained [13]. Motivated by the success of topic model based RNNLM adaptation, the incorporation of meta data during meeting is also investigated. In the task of meeting transcription, besides the recording, there are many types of useful materials which are easily accessed and beneficially to utilize, such as the agenda, slides, and related documents which associated with the content of meetings. How to make good use of these metadata to adapt RNNLNs, further improving ASR performance has practical value. Besides, it is also possible to detect the slides according to agenda. By anchoring the slides according to the agenda, which could further improve the ASR performance. These factors will be explored in this thesis as well.

1.6. Investigation of interpolation methods between RNNLNs and N-Gram

Chapter 6 gives an analysis of the RNNLNs and N-Gram LMs. A novel back off based interpolation scheme is proposed. Linear interpolation is widely used to combine RNNLNs and N-Gram LMs and good performance is achieved. However, there is few investigation to explore the interpolation between RNNLMs and N-Gram LM. In order to fully exploit the detailed n-gram level complementary attributes between the two found in previous research, a backoff based compact representation of n-gram dependent interpolation weights is proposed in this paper to allow robust weight parameter estimation. Experimental results are reported on the various tasks with varying amounts of training data. Small and consistent improvements in both perplexity and WER were obtained using the proposed interpolation approach over the baseline fixed weighting based linear interpolation.

1.7. Experiments

Experimental results on RNNLNs are given on Chapter 7. Firstly, experiments of efficient RNNLNs lattice rescoring are reported on CTS and meeting tasks. Efficient RNNLM training and evaluation are evaluated on CTS and Multi-Genre Broadcast (MGB) challenging tasks [1]. The adaptation of RNNLNs are conducted on MGB challenging and meeting tasks. The interpolation of RNNLNs and N-Gram LMs are studied and compared on three corpora, they are Penn Treebank corpus, Babel data and MGB task respectively.

1.8. Conclusions and Future Work

Chapter 9 concludes the thesis and discusses the future direction.

2. Conclusions

This thesis aims to apply RNNLM on large amount of training data to improve the performance of state-of-art ASR system. Several key issues are studies and solutions are proposed in this issue. The lattice rescoring and efficient GPU based training of RNNLMs are address in Chapter 3 and Chapter 4. The adaptation of RNNLMs is investigated in Chapter 5. The combination of RNNLMs and N-Gram LMs are explored in Chapter 6. The experiments in Chapter 7 are reported on many tasks to demonstrate the effect of these techniques.

---

1The detail of MGB challenge could be found from http://www.mgb-challenge.org/
3. References


