IEICE TRANS. INF. & SYST., VOL.E99–D, NO.10 OCTOBER 2016

LETTER *Special Section on Recent Advances in Machine Learning for Spoken Language Processing*

Improved End-to-End Speech Recognition Using Adaptive Per-Dimensional Learning Rate Methods

Xuyang WANG[†a)], **Pengyuan ZHANG**[†b)], ***Nonmembers***, **Qingwei ZHAO**[†], ***Member***, **Jielin PAN**[†], ***and*** **Yonghong YAN**[†,††], ***Nonmembers***

SUMMARY The introduction of deep neural networks (DNNs) leads to a significant improvement of the automatic speech recognition (ASR) performance. However, the whole ASR system remains sophisticated due to the dependent on the hidden Markov model (HMM). Recently, a new end-to-end ASR framework, which utilizes recurrent neural networks (RNNs) to directly model context-independent targets with connectionist temporal classification (CTC) objective function, is proposed and achieves comparable results with the hybrid HMM/DNN system. In this paper, we investigate per-dimensional learning rate methods, ADAGRAD and ADADELTA included, to improve the recognition of the end-to-end system, based on the fact that the blank symbol used in CTC technique dominates the output and these methods give frequent features small learning rates. Experiment results show that more than 4% relative reduction of word error rate (WER) as well as 5% absolute improvement of label accuracy on the training set are achieved when using ADADELTA, and fewer epochs of training are needed.

key words: *connectionist temporal classification, adaptive per-dimensional learning rate method, end-to-end ASR*

1. Introduction

In the last few years, deep neural networks (DNNs) have led to great improvements in automatic speech recognition (ASR). DNNs in ASR systems are commonly used for acoustic modeling based on a hidden Markov model (HMM) [1], [2]. In the hybrid HMM/DNN framework, DNNs, substituting the Gaussian mixture model (GMM), are trained to estimate the posterior probabilities of HMM states using cross-entropy criteria as the objective function. Despite the advances introduced by DNNs, the speech recognition system still remains sophisticated due to multiple training procedures of the HMM/GMM system, including the decision tree for HMM states' tying, the training of GMM models and the alignment of data for training DNNs.

Recently, Graves *et al.* [3] presents an end-to-end ASR framework, which transcribes audio data into texts directly using the connectionist temporal classification (CTC) [4] objective function. In the CTC technique, no prior knowledge of the labeling alignment is needed and the objective function is to maximize the sum of all the possible alignments of labellings with the help of a forward-backward algorithm, similar to that for HMMs. Another important feature is that the modeling units are directly phones or characters, resulting in simplification of the ASR system. The decoder integrated the word-level language model into weight finite state transducers (WFSTs) is introduced in [5], which makes the performance under the new framework comparable with that of the traditional HMM/DNN system and speeds up decoding significantly due to the context-independent (CI) phone modeling.

In [6], the role of the blank symbol is extensively discussed upon a similar challenge for handwritten recognition. Experiments and analyses suggest that the blank symbol dominate the prediction in the early stage of training since the number of the blank symbol is much larger than that of other labels. As a result, it is beneficial to give less importance to the error signal coming from the blank symbol using adaptive per-dimensional learning rates, ADAGRAD [7], for instance, during the training.

Although there are many similarities between handwritten recognition and ASR, the difference is also obvious. For example, the number of frames in a spoken utterance is always much larger than that of characters in a handwritten sentence, which may result in different configurations in the training stage. For example, the size of one mini-batch may be different due to the constraint of video memory and the number of update for ASR is larger. In this paper, we investigate to use adaptive per-dimensional learning rate methods to improve the performance of ASR based on the end-to-end framework, ADAGRAD and ADADELTA [8] included. Followed the work in [5], deep bidirectional long short-term memory (LSTM) recurrent neural networks (RNNs) are used for acoustic modeling. The RNN is trained to learn the CI phone labels given sequences of speech feature. A word-level LM, a lexicon and frame-level CTC labels are compiled into WFSTs separately, which are composed into a comprehensive search graph in the end. We find that the introduction of per-parameter learning rate methods not only accelerates the training, but also leads to a higher training accuracy rate, especially for short sentences.

The rest of the paper is organized as follows. Section 2 describes the network architecture and Sect. 3 gives a brief introduction to the CTC technique. The adaptive per-

Manuscript received January 25, 2016.
Manuscript revised May 11, 2016.
Manuscript publicized July 19, 2016.

[†]The authors are with the Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, China.
[††]The author is with the Xinjiang Laboratory of Minority Speech and Language Information Processing, Chinese Academy of Sciences, China.
a) E-mail: wangxuyang@hccl.ioa.ac.cn
b) E-mail: zhangpengyuan@hccl.ioa.ac.cn (Corresponding author)

DOI: 10.1587/transinf.2016SLL0001

parameter learning rate methods are introduced in Sect. 4. We report our experiment results in Sect. 5 and conclude our work in Sect. 6.

2. Model Architecture

Unlike standard feedforward neural networks, RNNs can use their internal memory to process arbitrary sequences of inputs, which makes RNNs suitable for sequence modeling tasks, such as handwritten recognition and speech recognition. To tackle the vanishing gradient problem of training RNNs, LSTM units are served as the building blocks of RNNs. LSTM RNNs have been shown to outperform the state of the art DNNs for acoustic modeling in ASR [9]. Moreover, conventional RNNs are only able to use previous context, so the future context also deserves to be exploited and a deep bidirectional RNN (BRNN) is chosen as our acoustic model.

Given an input sequence $\mathbf{x} = (x_1, \ldots, x_T)$, a BRNN computes the forward hidden sequence $\overrightarrow{\mathbf{h}}$ from $t = 1$ to T:

$$\overrightarrow{h}_t = \sigma(\overrightarrow{W}_{hx} x_t + \overrightarrow{W}_{hh} \overrightarrow{h}_{t-1} + \overrightarrow{b}_h) \tag{1}$$

where σ is the hidden layer activation, $\overrightarrow{W}_{hx}$ denotes the weight matrix connected inputs with the hidden layer, $\overrightarrow{W}_{hh}$ is the hidden-to-hidden weight matrix, and $\overrightarrow{b}_h$ is the bias vector. A backward hidden sequence $\overleftarrow{\mathbf{h}}$ is computed from $t = T$ to 1 as well:

$$\overleftarrow{h}_t = \sigma(\overleftarrow{W}_{hx} x_t + \overleftarrow{W}_{hh} \overleftarrow{h}_{t+1} + \overleftarrow{b}_h) \tag{2}$$

At each frame t, the output of the current recurrent layer is the concatenation of forward and backward hidden outputs $[\overrightarrow{h}_t, \overleftarrow{h}_t]$, which is the input to the next recurrent layer.

As mentioned above, LSTM units are used to address the vanishing gradient problem. In a forward LSTM layer, the gates and memory cells activation are computed sequentially from $t = 1$ to T, and the backward ones from $t = T$ to 1, similarly. The forward computation at time step t is implemented as follows:

$$i_t = \sigma(W_{ix} x_t + W_{ih} h_{t-1} + W_{ic} c_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{fx} x_t + W_{fh} h_{t-1} + W_{fc} c_{t-1} + b_f) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{cx} x_t + W_{ch} h_{t-1} + b_c) \tag{5}$$

$$o_t = \sigma(W_{ox} x_t + W_{oh} h_{t-1} + W_{oc} c_t + b_o) \tag{6}$$

$$h_t = o_t \tanh(c_t) \tag{7}$$

where σ is the logistic sigmoid function, and i, f, o, c are the input gate, forget gate, output gate and cell activation vector, respectively. $W_{.x}$ weight matrices connect inputs with the units and $W_{.h}$ matrices connect previous hidden outputs with the units. $W_{.c}$ are diagonal matrices for peephole connections. $b_.$ vectors denotes the bias vectors.

3. Connectionist Temporal Classification

Neural networks in the hybrid framework are typically trained with the cross-entropy (CE) criterion given frame-level alignments. However, this recipe depends on HMM/GMMs and the alignments of speech data are irrelevant to most speech recognition tasks, where word-level transcriptions matter. Connectionist temporal classification (CTC) is an objective function that maximizes the likelihood of a target sequence by effectively summing over all possible alignments of the input sequence. Hence, neural networks trained with the CTC objective function do not need any prior alignments between the input and target sequences. Assume that the output layer of a neural network in the hybrid framework contains K unique labels (characters, phonemes etc.). A blank symbol is added and useful for the alignment.

Given a length T input sequence $\mathbf{x}$, the output vector y_t is normalized with a softmax function and represents the posterior probabilities of emitting labels at time t. A CTC alignment $p = (p_1, \ldots, p_T)$ is a frame-level sequence of labels with a length of T. The probability of this alignment is the production of posterior probabilities at each time:

$$Pr(\mathbf{p}|\mathbf{X}) = \prod_{t=1}^{T} y_t^{p_t} \tag{8}$$

A given transcription sequence $\mathbf{z}$ can be mapped to many CTC alignments due to the blank symbol. The set of CTC alignments corresponding to $\mathbf{z}$ is denoted as $\Phi(\mathbf{z})$. The likelihood of $\mathbf{z}$ given the input sequence $\mathbf{x}$ is then calculated as the sum of the probabilities of all CTC alignments:

$$Pr(\mathbf{z}|\mathbf{x}) = \sum_{\mathbf{p} \in \Phi(\mathbf{z})} Pr(\mathbf{p}|\mathbf{x}) \tag{9}$$

The sum operation in the equation above is computationally intractable. Therefore, a forward-backward algorithm is introduced to recursively solve the problem. The transcription sequence $\mathbf{z}$ is firstly transformed into an augmented label sequence $\mathbf{l} = (l_1, \ldots, l_{2U+1})$ after insert a blank symbol into every pair of the original labels. A forward variable α_t^u represents the probabilities of all CTC alignments which end with label u at time t and can be computed from α_{t-1}^u and α_{t-1}^{u-1}. Similarly, the probabilities of all alignments which start with label u at time t and reach the end time T is denoted a backward variable β_t^u. So $Pr(\mathbf{z}|\mathbf{x})$ is computed as:

$$Pr(\mathbf{z}|\mathbf{x}) = \sum_{u=1}^{2U+1} \frac{\alpha_t^u \beta_t^u}{y_t^{l_u}} \tag{10}$$

The objective $\ln(Pr(\mathbf{z}|\mathbf{x}))$ now is differentiable to the RNN output $\mathbf{y}_t$, and the derivative is derived as follows:

$$\frac{\partial ln(Pr(\mathbf{z}|\mathbf{x}))}{\partial y_t^k} = \frac{1}{Pr(\mathbf{z}|\mathbf{x})} \frac{1}{(y_t^k)^2} \sum_{u \in \Gamma(\mathbf{l},k)} \alpha_t^u \beta_t^u \tag{11}$$

where $\Gamma(\mathbf{l}, k)$ returns the elements whose labels are k. These errors are back-propagated through the softmax layer and further into the RNN to update the model parameters.

4. Adaptive Per-Dimension First Order Methods

As the majority of the likelihood $Pr(\mathbf{z}|\mathbf{x})$ derives from the

blank symbol in the CTC training, which will not appear in the recognition results and contributes to the peaks of phone or character posterior probabilities [6], [10], emphases on phone or character labels seems to be beneficial for training. Adaptive per-dimension first order method, ADAGRAD and ADADELTA, are adopted to improve the recognition.

ADAGRAD is proposed in [7] and investigated in [8], [11] for ASR, which demonstrates that this method fails to improve the recognition compared to the momentum, a simplest extention to the stochastic gradient descent (SGD), with the CE objective function. However, a characteristic of ADAGRAD that the informativeness of rare features is emphasized during training is attractive in the CTC training. The update rules for ADAGRAD is:

$$\Delta x_t = -\frac{\eta}{\sqrt{\sum_{\tau=1}^{t} g_\tau^2}} g_t \tag{12}$$

$$x_{t+1} = x_t + \Delta x_t \tag{13}$$

where x_t is the parameter of the model, η is a global learning rate and Δx_t is the value to be incremented to x_t and g_t is the gradient of parameters at the t-th iteration. As shown in Eq. (12), large gradients have small learning rates and vice versa, owing to the accumulation of history gradient magnitudes in the denominator. Meanwhile, a disadvantage of ADAGRAD is that the learning rates continue to decrease and will become very small at the end of the training.

ADADELTA is presented in [8] to improve ADAGRAD. A decay constant ρ is added to prevent accumulating the sum of squared gradients over all time. In addition, the update Δx_{t-1} also contributes to the current one Δx_t referring with the Hessian approximation. The calculations are:

$$E[g^2]_t = \rho E[g^2]_{t-1} + (1-\rho) g_t^2 \tag{14}$$

$$RMS[g]_t = \sqrt{E[g^2]_t + \epsilon} \tag{15}$$

$$\Delta x_t = -\frac{RMS[\Delta x]_{t-1}}{RMS[g]_t} g_t \tag{16}$$

$$E[\Delta x^2]_t = \rho E[\Delta x^2]_{t-1} + (1-\rho)\Delta x_t^2 \tag{17}$$

where ϵ is added to better condition the denominator, $E[g^2]_0$ and $E[\Delta x^2]_0$ are initialized with 0.

5. Experiments

The experiments are conducted on a subset of Switchboard corpus with a open-source toolkit Eesen [5]. The first 4000 utterances are selected for cross validation and our training set contains about 110 hours of transcribed speech, which are the first one hundred thousand utterances exclude the validation set. The test set is the Switchboard (SWB) part of Hub5'00. Deep LSTM RNNs are applied as the acoustic model, which involves 4 bi-directional LSTM layers and each layer has 320 memory cells in both the forward and backward sub-layers. Inputs of the acoustic model are 40-dimensional filterbank features together with their first and second orders, amounting to 120-dimensional ones per frame. 46 CTC labels, including 45 phonemes and the blank symbol, are chosen as the outputs of the softmax layer. Specifically, the training utterances are sorted by their lengths for the parallel. The termination of the training is as same as that described in [5], depending on the the improvement of the average label accuracy (LAC) on the validation set between two successive epochs. Since the labels used in the CTC framework are phonemes, the average label accuracy is equivalent to the average phoneme accuracy (PAC). The baseline system are trained with SGD and the learning rate starts with 0.00004 as well as the momentum is set to 0.9. The global learning rate for ADAGRAD and ADADELTA is 0.01. The decay rate ρ and constant ϵ in ADADELTA are 0.9 and 0.001, respectively. A trigram language model (LM) is trained with the transcription of the training speech and the decoding framework is based on WFSTs.

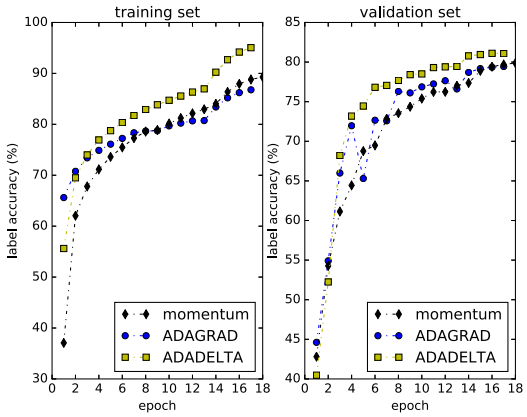

Fig. 1 Phoneme accuracy of each epoch on the training and validation set

The average phoneme accuracies of each epoch on the training and validation set are firstly shown in Fig. 1. A sharp increase at the initial stage using adaptive per-dimensional methods is observed and can be attributed to the parameter-specific learning rates. However, the PAC of ADAGRAD rises slowly after several epochs and is surpassed by the baseline system. ADADELTA achieves higher PAC all the time during training and similar phenomenon is shown on the validation set after the first two epochs.

To further check the performance of adaptive per-dimensional methods, the average phoneme accuracies of the last epoch on the training set are reported in Fig. 2, which are counted every one thousand utterances. The result using ADAGRAD or ADADELTA significantly outperforms that using the momentum method at beginning. Since the training set are sorted by length, this may be ascribed to the fact that the percentage of the blank symbol is larger when the utterance is relatively short. Nevertheless, the average phoneme accuracies of ADAGRAD constantly decline with the length of training utterances becomes longer, which maybe results from the fact that ADAGRAD makes the actual learning rate become smaller and smaller. As a re-

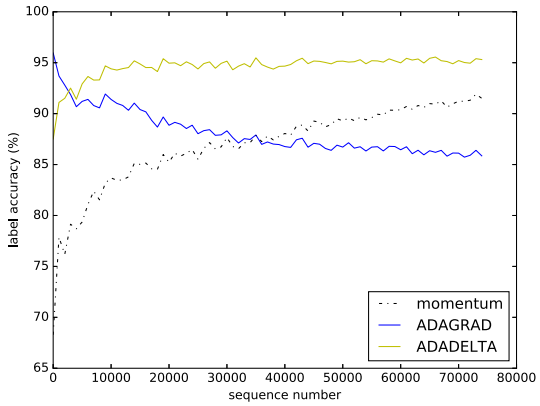

Fig. 2 Average phoneme accuracy of the final epoch on the training set

Table 1 The training and recognition performance of each learning rate method

Method	PAC%		WER%
	Train	Valid	
momentum	89.55	79.94	20.5
ADAGRAD	86.78	79.44	21.5
ADADELTA	95.05	81.07	19.6

sult, the model updates little with the training goes on in an epoch and the PAC of in an epoch decreases constantly. The decay rate introduced in ADADELTA remedies the problem and improve the training performance, resulting in an overall higher phoneme accuracies.

Finally, the word error rates (WER) and average phoneme accuracies of the last epoch on the training and validation set are listed in Table 1. Compared to the baseline system, 4.4% relative reduction of WER is achieved using ADADELTA, in contrast to the fact that using ADAGRAD leads to a poor performance. Besides, 5.5% and 1.1% absolute improvements of phoneme accuracy on the training and validation set are obtained when ADADELTA is utilized.

6. Conclusion

In this paper, we investigate to use adaptive per-dimensional learning rate methods to improve the recognition performance based on an end-to-end ASR framework. Deep RNNs are trained with CTC objective function for acoustic modeling and the blank symbol introduced by CTC dominates the output. Instead of momentum, ADAGRAD and ADADELTA are exploited to give frequent features small learning rates and emphasize infrequent ones. Experiment results show that the ADAGRAD deteriorates the recognition, but that ADADELTA leads to improvement of both WER and phoneme accuracy on the training set compared with the baseline system, while ADADELTA proves to be a robust learning rate method and makes little contribution to WER in the traditional HMM/DNN system.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Nos. 11461141004, 91120001, 61271426), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), the National 863 Program (No. 2012AA012503) the CAS Priority Deployment Project (No. KGZD-EW-103-2) and the Key Science and Technology Project of the Xinjiang Uygur Autonomous Region (No. 201230118-3).

References

[1] A.-R. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," Audio, Speech, and Language Processing, IEEE Transactions on, vol.20, no.1, pp.14–22, 2012.

[2] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," Audio, Speech, and Language Processing, IEEE Transactions on, vol.20, no.1, pp.30–42, 2012.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp.1764–1772, 2014.

[4] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," Proceedings of the 23rd international conference on Machine learning, pp.369–376, ACM, 2006.

[5] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," arXiv preprint arXiv:1507.08240, pp.167–174, 2015.

[6] T. Bluche, H. Ney, J. Louradour, and C. Kermorvant, "Framewise and ctc training of neural networks for handwriting recognition," Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pp.81–85, 2015.

[7] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," J. Machine Learning Research, vol.12, pp.2121–2159, 2011.

[8] M.D. Zeiler, "Adadelta: An adaptive learning rate method," arXiv preprint arXiv:1212.5701, 2012.

[9] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH), 2014.

[10] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," arXiv preprint arXiv:1507.06947, 2015.

[11] A. Senior, G. Heigold, M. Ranzato, and K. Yang, "An empirical study of learning rates in deep neural networks for speech recognition," Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp.6724–6728, 2013.