NEURAL NETWORK ACOUSTIC MODELS WITH SUPERVISED HIDDEN LAYERS FOR AUTOMATIC SPEECH RECOGNITION

Jiapei Huang, Yuning Yang, Xiangang Li, Xihong Wu

Speech and Hearing Research Center
Key Laboratory of Machine Perception (Ministry of Education)
Peking University, Beijing, 100871

ABSTRACT

Supervising information, usually, would only appear in the output of deep neural networks (DNNs). This paper focuses on utilizing one or more than one types of supervising information in all output layer and hidden layers of DNNs for automatic speech recognition (ASR). Such scheme shows stable and significant improvement over both traditional single- and multi-task DNNs. This work also discusses performance between different controlling schemes of hidden layers' supervising weights.

Index Terms— Deep Neural Networks, Automatic Speech Recognition, Supervising Information

1. INTRODUCTION

In recent years, deep neural networks have boosted performance of ASR. [1] is the earliest work on DNNs acoustic modeling for ASR, and utilizes deep belief networks to initialize parameters of DNNs so as to improve effectiveness of its discriminative training process. It stacks multiple layers of Restricted Boltzmann Machines (RBMs), which are trained with a greedy layer-wise unsupervised learning algorithm called Contrastive Divergence (CD) pretraining, as the initialization of model's weights and then finetunes the whole neural network on a classification task with standard back-propagation (BP) process. Another famous and similar method is stacked denoising autoencoders [2]. The above methods use no supervising classification information while pretraining and hence belong to unsupervised pretraining methods. Though this type of method adds constraints to weights of hidden layers at the pretraining stage, it only exploits classification information when finetuning the overall model. And there is other type of pretraining methods belonging to supervised pretraining. Such methods include greedy layer-wise supervised training [3] and discriminative pretraining [4]. The former one firstly trains a one-hidden-layer neural network using labels with error BP discriminatively, then discard its output layer, and add another hidden layer randomly initialized on top of the previously trained hidden layer along with a new output layer, train the layers newly added, and go on with the same process until a desired neural network is obtained. The latter one differs from the former one in that it updates the whole neural network including both the layers newly added and the layers previously trained each time. These methods, in the stage of pretraining, like the way this paper adopts, update hidden layers' weights with direct participation of supervising information. But they, unlike this paper's method, make use of labels directly from only output layer while finetuning.

Apart from the above pretraining techniques, there are also lots of studies researching on how different structures affect performance of ASR. Convolutional neural networks (CNNs) [5] [6] provides significant improvements over DNNs with shared weights and locally connected edges. Some structure adjustment works based on CNNs include using heterogeneous pooling [7] and deepening convolution kernels which strengthen convolutional kernels' feature extraction ability [8] [9] [10]. Another novel DNNs structure for acoustic modeling is recurrent neural networks [11], which incorporates long term history of acoustic observations. No matter how these structure varies, none of them consider directly incorporating classification information for hidden layers.

The most recent works having hidden layers supervised include GoogLeNet [12], deeply-supervised nets [13] and autoencoders without layer-wise training [14]. GoogLeNet won ImageNet 2014 detection and classification challenges and is a 22-layers CNNs with supervised hidden layers. It shows that even deeper is even better. It is the auxiliary supervising information added to hidden layers that do the most work to make it possible to go deeper, increasing gradient signal and providing additional regularization. While GoogLeNet focuses on more layers, the deeply-supervised nets emphasize its ability to alleviate common exploding and vanishing gradients problem. And autoencoders without layer-wise training work has similar thought of combining objective for all output and hidden layers while it applies the idea into deep autoencoders and its global objective is composed of reconstructing error instead of classification error.

This work makes use of one or more than one types of supervising information in all hidden layers, and verifies it-

s effectiveness. In the case of multiple types of supervising information, it can be also treated as combining supervising hidden layers and multi-task DNNs [15] [16] [17]. Multi-task learning, needing to be explained before going on, is a technique that learns multiple relative prediction tasks simultaneously and helps every single task to perform better. For each task, the rest of the tasks can provide regularization effect. And this work makes contributions in that 1) it verifies the idea, supervising hidden layers's effectiveness for ASR, which is far more complicated than classification task since its performance is not only decided by acoustic model but also language model and Hidden Markov models (HMMs) for sequence constraint; 2) it looks into different types of supervising information weighting schemes; 3) it incorporates more than one kinds of supervising information, and shows the resulted model provides even better regularization and much better performance.

2. MODEL

For usual DNNs, ground truth labels appear in the output layer of a model, and after calculating loss error signal would propagate back from the output layer down to its first hidden layer. In the above scenario, when BP process is applied for training, gradient signal would easily explode or vanish as neural networks go deep. That is because gradient calculation is in a iterative form

$$\frac{\partial Loss}{\partial z_i} = \frac{\partial Loss}{\partial z_{i+1}} * \frac{\partial z_{i+1}}{\partial z_i} \quad (1)$$

where z_i is the i-th layer's output value of activation function. We can see that gradient changes, vanishes or explodes, in an exponential speed as it passes down networks if $\frac{\partial z_{i+1}}{\partial z_i}$ less or larger than 1 for most i. That's one intuition indicating why gradient back propagation source should be added to hidden layers, to compensate the hard-to-control exponentially changing gradient signal. Another intuition is that though each hidden layer is learning better representation for the final classification, it does not know directly what the final classification task should be. Would it do better if it is aware of what the final goal is? Would it be better if each hidden layer learns representation by also considering discriminative power of it? That is the second intuition.

DNNs with supervised hidden layers look like the Figure 1, where all "Supervision" layers contain ground truth labels the same as the output layer. Then the final objective function would be changed into

$$Loss(X, W, W') = l(W, X) + \sum_i \alpha_i l(W_{0...i}, W'_i, X) \quad (2)$$

where W is the weight corresponding to usual DNNs and W' denotes the weight added between every hidden layer and its corresponding "Supervision" layer generating additional outputs. And W'_i is that weight for the i-th hidden layer. $Loss$ is the total loss function while l is the individual loss from the output layer or each "Supervision" layer. And α_i is the supervising weight for the i-th hidden layer, X is the input. When gradient is propagating through hidden layers, what only needs to be done is to add additional gradient signal from corresponding "Supervision" layer.

In spite of single source ground truth labels, this work also extends the idea of adding supervising information into hidden layers to multiple supervising information sources, in this work's situation, senone and additional gender. This extension is just like the extension from single-task to multi-task learning. By introducing other relative supervising information, a model would be able to learn the distribution for one task better by knowing other factors that have effects over that task. For instance, in this work for the multiple types of labels occasion all supervision layers contain two types of output, senone and gender, and when supervised by these two types of labels, the representation would be discriminative for both tasks. And since the model knows that current observation is more likely coming from a female , for the representation learned contains this information, it can predict senones as if there are different decision boundaries for different genders and it uses the one for female. In this way, the distribution learned by the model would be like a combination of two sub-distributions, one for male and one for female. So that it would not be easily confused by speaking difference caused by different genders. That is the intuition explaining why bother adding additional gender task. In this work, an output layer keeps the same labels as "Supervision" layers. And while training, as what multi-task learning do, update parameters according to different sets of labels iteratively, that is first senone, then gender, and senone, gender until the training procedure ends. And notice that in this work, all outputs for different supervision sources share most parameters except the one layer of weights just before output layer and "Supervision" layers.

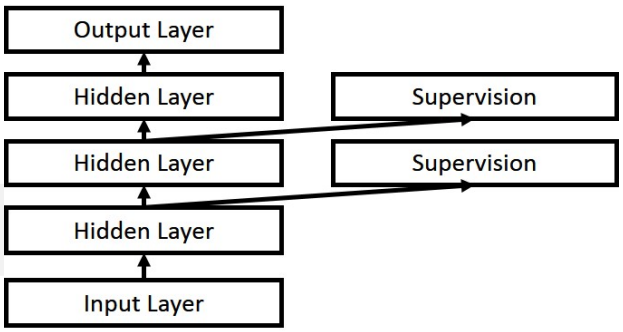

Fig. 1. Structure of neural networks with supervised hidden layers

3. EXPERIMENTS

The experiments evaluate this paper's method with hub4, a Mandarin Chinese news audio corpus of about 30 hours. And all recordings were done using one single channel, 16-bit quantization and 16-KHz sample frequency. And the speech in the original hub4 training set with bad quality was excluded and about 20 hours of training data is used for the experiments. The resulting data contains speech from 234 speakers and the speech duration from different genders distributes approximately even, about 10.4 hours for female and about 9.8 hours for male. And approximate 20 minutes of the above data was separated randomly as development set. There are no same utterance appearing in both training set and development set. And the experiments use the original hub4 test set, which contains about 40 minutes of speech, for evaluation.

All experiments use phones with tone as the acoustic unit, and they are clustered into about 3k senones. The output layer and "Supervision" layers use softmax and the loss functions for them are cross entropy. While training, learning rate keeps fixed and starts to half when improvement of neighboring iterations' validation loss is less than curtain specified small value. All experiments are using the same setup, fully connected structure, 4 hidden layers and 2000 nodes for each layer, sigmoid activation function, learning rate of 0.008, and conducted without pretraining.

Firstly, experiments were conducted for normal single-task DNNs. Refer to Table 1 for the experimental results. These experiments compare multiple supervising weighting schemes.

The first two cases assigns the same supervision weight for different hidden layers. The former one keeps the supervising factor fixed for each iterations, and the latter one scales supervision weight as training goes. We can see that even weighting does not improve performance much.

Then we may consider that supervising weight for different hidden layers should be different, and the one nearer to output layer should have higher weight. Based on this consideration, the last two experiments use the following weighting scheme

$$\alpha_i = \alpha * p^{|i-c|}, \quad 0 < p < 1, 0 < \alpha < 1 \tag{3}$$

where p denotes the scaling factor, α denotes base weight, i denotes the hidden layer's position, and c denotes hidden layer position with the peak weight. In this way, the c-th layer's supervision weight is highest, and as the distance between i-th layer and c-th layer increase, the weight decreases exponentially. Then this paper reports two more experiments of different supervision weight controlling schemes. The former experiment adopts a static scheme, and sets parameter c to the position of output layer and fixes it for each iteration. While the latter one adopts a moving peak scheme, and increases c by 1 from 0 to infinite for every two iterations. Both of them set parameters α to 1, p to 0.5. Setting $alpha$ and p to other reasonable values may report comparable improvements since in the experiments there are no much work on searching for optimal values for these two parameters. All four schemes make improvements while the last one most significant, reducing word error rate (WER) from 15.32% to 14.66%, about 4.31% relative improvement. We can see from these results that supervising weight controlling scheme is quite important.

Then consider the experiments with additional supervising gender labels. These experiments treat additional gender labels as assisting labels to help regularize the main classification task, and the gender class prediction is discarded after training procedure is over. The labels for output layer and "Supervision" layers contain both the same senone and gender information. This work updates parameters according to the gradient from the same set of labels for all output layer and hidden layers. And the gender task has different learning rate with the senone task, and in this work, its learning rate keeps a fixed percentage of the senone task's. And while choosing that percentage, this work selects the one from a certain range for both setup, with or without "Supervision" layers, and reports the best results respectively for fairness. And the percentage for 'DNNs with additional gender task' is 0.2 and the one for 'DNNs with additional gender task and SHL, moving peak' is 0.4. Refer to Table 2 for experiment results. We can see that adding additional gender task help improve the performance comparing to single-task DNNs for ASR. And combining it with supervised hidden layers, it further reduces WER. It obtains about 4.06% relative improvement over the one with additional gender task DNNs acoustic model and about 5.94% relative improvement over the one with standard single-task DNNs acoustic model.

To further analysis the effect of adding one or multiple supervising information to hidden layers, we can check the validation loss's situation as training goes. Refer to Figure 2 for these information. Firstly, there is a second sharp drop at around 6-th iterations. That sharp drop comes from the fact that learning rate is halved since the validated loss improvement is little. Secondly adding supervising information to normal DNNs seems not providing better classification ability while introducing better recognition results. This phenomenon may comes from the reality that in ASR, labels for senone are never strictly correct. In other words, you can never tell that certain frame belongs surely to certain senone since speech signal is changing smoothly as the content changes and you can never be able to find a boundary of each phone. In this way, stronger supervision information may classify worse or not so better while provide better recognition result. Thirdly DNNs with additional gender task generalizes better than single-task DNNs and its validation loss is less, though not very much, than the single-task DNNs for each iterations. That implies the positive impact of choosing gender as an additional set of labels for ASR. And lastly combining supervising hidden layer technique and adding additional

Model	WER (%)
DNNs	15.32
DNNs with SHL, static	15.15
DNNs with SHL, scaling	15.16
DNNs with SHL, static peak	15.25
DNNs with SHL, moving peak	**14.66**

Table 1. Results for single-task DNNs, SHL stands for supervised hidden layers

Model	CER (%)
DNNs with additional gender task	15.02
DNNs with additional gender task and SHL, moving peak	**14.41**

Table 2. Results for DNNs with additional gender task

supervising source provides better regularization and not only improves classification ability but also recognition performance. The supervising hidden layer technique strengthens the supervising gender information's effect. So this method obtains benefit from both techniques.

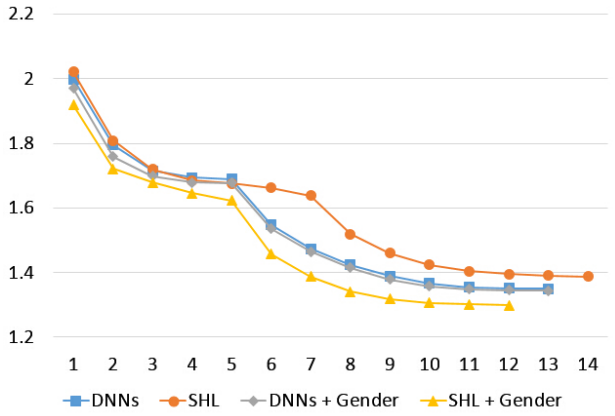

Fig. 2. Validating losses for each iterations

4. CONCLUSIONS

This work verifies the effectiveness of adding supervising information for hidden layers in ASR with DNNs acoustic model, a 4.31% relative improvement over traditional fully-connected single-task DNNs. And it compares various schemes of controlling the factor for hidden layer supervising loss, and the one with increasing peak position is the best in this work. It extends supervising information from one source to multiple sources and verified its effectiveness too, a 5.94% relative improvement comparing to single-task DNNs and 4.06% over multi-task DNNs. Then this work analyzes how supervising hidden layers and adding multiple supervising source affect DNNs' generalization ability.

5. REFERENCES

[1] A. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *NIPS Workshop on Deep Learning for Speech Recognition and Related Applications*, 2009.

[2] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[3] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, pp. 153, 2007.

[4] D. Yu, L. Deng, F.T.B. Seide, and G. Li, "Discriminative pretraining of deep neural networks," May 30 2013, US Patent App. 13/304,643.

[5] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, 1995.

[6] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4277–4280.

[7] L. Deng, O. Abdel-Hamid, and D. Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6669–6673.

[8] M. Lin, Q. Chen, and S.C. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[9] L. Tóth, "Phone recognition with deep sparse rectifier neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6985–6989.

[10] L. Tóth, "Combining time-and frequency-domain convolution in convolutional neural network-based phone recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014.

[11] A. Graves, *Supervised sequence labelling with recurrent neural networks*, Ph.D. thesis, Technische Universität at München, 2008.

[12] C. Szegedy, W. Liu, Y.Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[13] C.Y. Lee, S.N. Xie, P. Gallagher, Z.Y. Zhang, and Z.W. Tu, "Deeply-supervised nets," *arXiv preprint arXiv:1409.5185*, 2014.

[14] Y.B. Zhou and Govindaraju V., "Learning deep autoencoders without layer-wise training," *arXiv preprint arXiv:1405.1380*, 2014.

[15] R. Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Machine Learning: Proceedings of the Tenth International Conference*, 1993, pp. 41–48.

[16] R. Caruana, *Multitask learning*, Springer, 1998.

[17] M.L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6965–6969.

[18] A.B. Smith, C.D. Jones, and E.F. Roberts, "Article title," *Journal*, vol. 62, pp. 291–294, January 1920.