INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Efficient Knowledge Distillation from an Ensemble of Teachers

Takashi Fukuda[1], *Masayuki Suzuki*[1], *Gakuto Kurata*[1], *Samuel Thomas*[2], *Jia Cui*[2], *Bhuvana Ramabhadran*[2]

[1]IBM Watson Multimodal, IBM Research, Chuo-ku Hakozaki, Tokyo, 103-8510, JAPAN
[2]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, US

{fukuda, szuk, gakuto}@jp.ibm.com, {sthomas, jiacui, bhuvana}@us.ibm.com

Abstract

This paper describes the effectiveness of knowledge distillation using teacher student training for building accurate and compact neural networks. We show that with knowledge distillation, information from multiple acoustic models like very deep VGG networks and Long Short-Term Memory (LSTM) models can be used to train standard convolutional neural network (CNN) acoustic models for a variety of systems requiring a quick turnaround. We examine two strategies to leverage multiple teacher labels for training student models. In the first technique, the weights of the student model are updated by switching teacher labels at the minibatch level. In the second method, student models are trained on multiple streams of information from various teacher distributions via data augmentation. We show that standard CNN acoustic models can achieve comparable recognition accuracy with much smaller number of model parameters compared to teacher VGG and LSTM acoustic models. Additionally we also investigate the effectiveness of using broadband teacher labels as privileged knowledge for training better narrowband acoustic models within this framework. We show the benefit of this simple technique by training narrowband student models with broadband teacher soft labels on the Aurora 4 task.

Index Terms: Speech recognition, knowledge distillation, teacher-student, CNN, VGG, LSTM, bandwidth

1. Introduction

Automatic speech recognition has been shown to benefit by combining information at multiple levels of the acoustic modeling pipeline. These strategies include combining various acoustic feature sets together at the input of acoustic models, joint training of complex acoustic models after fusing various neural network architectures [1, 2] and combination of acoustic scores predicted at the output of various acoustic models. In addition to this, significant performance improvements have been obtained by augmenting the training data with various kinds of variabilities - for example with several kinds of noises at different SNR levels - instead of just using clean training data, to train acoustic models. While these techniques can be used to train complex acoustic models, it is often the case that these models cannot be deployed for real-time decoding of streaming speech data because of constraints they pose in terms of latency and computation resources [3]. More recently to tackle this limitation, compact models have been trained via knowledge distillation or model compression.

In the knowledge distillation framework, instead of training models which had reduced computational requirements and improved latency performances directly on hard targets in a single step, training is now performed in two separate steps [4, 5, 6, 7, 8, 9, 10]. In the first step, complex teacher acoustic models are first trained by combining information at various levels as described above. Compact acoustic models or student networks are then trained on the soft outputs of teachers using training criteria that minimize the differences between the student and teacher distributions. This technique has been shown to be very successful in various settings - fully supervised [9], semi-supervised [6], multilingual [11], sequence training [12] - to train student networks perform better than training similar models from scratch using hard targets.

To improve the performance of student networks, more recent work has focused on techniques to leverage information from multiple teachers by training student networks on an ensemble of teachers. [7] uses temparature to balance different teachers. In other approaches, [13] uses oracle to select the best teacher ensemble for each utterance while in [14] both temparature and multitasking are used to combine teacher labels and original hard labels. In these approaches, ensembles of teachers are created by first combining the outputs of complimentary teacher networks trained on multiple input features/architectures/training criteria into a single output distribution and then training student networks on the combined output to learn this ensemble distribution. Although this approach allows the student network to learn better distributions that eventually led to lower error rates, the student network is not presented directly the individual complimentary teacher distributions. We hypothesize that if student networks are provided with multiple streams of information via the various teacher distributions, the student will observe various ”views” of the data and will be able to generalize better while at the same time capture complimentary information available in each of the teacher streams. To facilitate this, we combine the distillation framework with a simple data augmentation strategy. In this approach, instead of augmenting data using various kinds of signal distortions to the input acoustic features as is often done, we augment the training data by creating multiple copies of data with corresponding soft output targets from various teachers. We demonstrate the effectiveness of our approach by training compact CNN based student networks that perform better than similar models trained on combined outputs from an ensemble of teachers - an LSTM based teacher and a VGG based teacher.

We extend this proposed technique to the generalized distillation framework, where in addition to distillation of information from teacher networks, privileged information available only during training is also factored in. To illustrate the efficacy of our approach we show how an improved narrow band CNN based acoustic model can be trained by using privileged information from outputs of broadband models, instead of training the student network on only narrow band teacher models. Privileged information is presented to the student network not only via both an ensemble of teachers but also by data augmentation of training data as described earlier. The training data

http://dx.doi.org/10.21437/Interspeech.2017-614

for this task is created using soft targets from an ensemble of narrow band teachers, an ensemble of broad band teachers and also hard targets. We show that while the proposed approach performs better than conventional training with hard targets, it improves over training on teacher ensembles which have been combined to form a single teacher stream and also benefits from privileged knowledge.

In section 2 we describe the teacher-student distillation framework and training procedure and introduce improvements to this framework via data augmentation of teacher ensemble outputs. Section 3 outlines various experiments on the Aurora 4 task that demonstrate the usefulness of our proposed techniques. The paper concludes with a discussion in section 4.

2. Ensembles of multiple teachers

Various algorithms have been proposed for transferring knowledge from teacher models to student models as described earlier. Instead of using the ground truth labels, the teacher-student training approach defines the loss function as

$$\mathcal{L}(\theta) = -\sum_i q_i \log p_i, \tag{1}$$

where q_i is the so-called soft label of the teacher model, which works as a pseudo label. p_i is output probability of the class of the student model. In q_i, the competing classes will have small but nonzero posterior probabilities for each training example. In a conventional method using multiple teacher models, soft labels q_i are created by weighted ensembles of posteriors from each teacher model as

$$q_i = \sum_k w_k q_{ik}, \tag{2}$$

where $w_k \in [0, 1]$ is the interpolation weight. q_{ik} is the soft label of k-th teacher. This technique is described below.

Algorithm 1 intepolated-training

```
for all minibatches in training data do
    pick minibatch $i$;
    for all teachers in pool of teachers do
        use teacher $j$ to provide soft-targers for minibatch $i$;
    end for
    combine soft-targets from all teachers with preassinged weights $w_j$ for each teacher;
    update neural network model with minibatch $i$;
end for
```

Though this is one of reasonable methods to use multiple teachers [13], the interpolation method weakens the complimentariness obtained by multiple models. Dissimilarities between acoustic models should be more explicitly maintained/leveraged to make student model represent various characteristics.

In this paper, we propose two additional techniques to train better student models. In the first method that we call *switched-training*, for each minibatch, soft-targets corresponding to the data are derived from a randomly selected teacher. The weights of the student model hence updated by switching teacher labels q_{ik} at the minibatch level as shown in algorithm 2. In contrast to the *intepolated-training* technique no predetermined weights are used to combine the outputs from multiple teachers.

In our second method we extend the *switched-training* technique with data augmentation by producing multiple copies of the data with soft-targets from multiple teachers and training on all the created data. This technique allows the network to train on multiple data views of the data and differs from conventional data augmentation where the input features are transformed keeping the labels the same. The *augmented-training* method is illustrated in algorithm 3.

Algorithm 2 switched-training

```
for all minibatches in training data do
    pick minibatch $i$;
    randomly select a teacher $j$ from the pool of teachers to provide to provide soft-targets for minibatch $i$;
    update neural network model with minibatch $i$;
end for
```

Algorithm 3 augmented-training

```
for all minibatches in training data do
    pick minibatch $i$;
    for all teachers in pool of teachers do
        use teacher $j$ to provide soft-targets for minibatch $i$;
        update neural network model with minibatch $i$;
    end for
end for
```

3. Experiments

3.1. Baselines

The proposed training techniques are evaluated using a series of experiments using several public data sets. In our first set of experiments (Section 3.2) we explore the training of various student models using multiple teachers trained on a medium size speech corpus. In a second set of experiments (Section 3.3) we investigate how ASR performance can be improved with priviledged information using NN acoustic models trained on different spectrum bandwidths.

For the first set of experiments, neural network based acoustic models are trained on 500 hours of audio data. 50% of this training corpora is clean audio from three public corpora - 100 hours from broadcast news, 100 hours from Mixer 6 [15], and 20 hours from the AMI corpus [16] and 30 hours of private speech data. The corpora is further augmented with realistic environmental noises from the JEIDA corpus [17] and impulse responses from RWCP [18] at various SNRs between 5 to 20 dB. CNN based acoustic models are trained on this multi-condition training set with 40 dimensional log Mel-frequency spectra augmented with Δ and $\Delta\Delta$s as inputs. The log Mel-frequency spectra are extracted by first applying mel scale integrators on power spectral estimates in short analysis windows (25 ms) of the signal followed by the log transform. Each frame of speech is also appended with a context of 11 frames after applying a speaker independent global mean and variance normalization. The CNN systems use two convolutional layers with 128 and 256 hidden nodes each in addition to four fully connected layers with 2048 per layer to estimate posterior probabilities of 9300 output targets. All of the 128 nodes in the first feature extracting layer are attached with 9×9 filters that are two dimensionally convolved with the input log Mel-filterbank representations. The second feature extracting layer with 256 nodes has a similar set of 3×4 filters that processes the non-linear activations after max pooling from the preceding layer. The non-

Table 1: *Comparing CNN trained with hard targets and student CNN learned from VGG and LSTM on the Aurora 4 task.*

Model	Target	AVG WER
CNN	hard	13.3
VGG	hard	10.5
LSTM	hard	11.7
CNN: VGG	soft	11.6
CNN: LSTM	soft	12.8
CNN: VGG+LSTM	soft/interpolation	11.4
CNN: VGG+LSTM	soft/switching	11.2

Table 2: *Comparing Compact CNN trained with hard targets and student CNN learned from VGG and LSTM.*

Model	Target	Aurora 4
Compact CNN	hard	15.1
Compact CNN: VGG	soft	13.6
Compact CNN: VGG+LSTM	soft/switching	13.2

linear outputs from the second feature extracting layer are then passed onto the subsequent fully connected layers. All the layers use the sigmoid non-linearity.

In addition to these baseline CNN models we attempt to also train compact yet accurate acoustic models with a good decoding speed. We try to construct compact CNN acoustic models for a quick turnaround of decodings by using the knowledge distillation framework with multiple teacher models. The compact CNN acoustic models have two convolutional layers with 64 and 128 hidden nodes followed by two fully connected layers with 768 hidden units in each layer. Both the classes of CNN models are trained on both hard targets and also with soft targets using the student-teacher framework. We evaluate the effectiveness of our proposed methods primarily on the Aurora 4 task using the task-standard WSJ0 bigram language model.

3.2. Training with an ensemble of teachers

Student CNN models are learned from two teacher NNs with methods described in Section 2. One of the teachers is a VGG model comprising 10 convolutional layers, with a max-pooling layer inserted after every 3 convolutional layers, followed by 4 fully connected layers. All hidden layers have ReLU non-linearity. Batch normalization is also applied to every fully connected layers. The second teacher is LSTM model consisting of 4 bidirectional LSTM layers with 512 units per direction and a linear bottleneck layer with 256 units. The two teacher models were sequence trained after the models were constructed with a cross entropy criterion. The training data for the teachers are the same as baselines. The WER performance of two teachers and the baseline CNN model using hard labels are shown in the first part of Table 1.

Posteriors of top 50 most likely labels for each prediction of the teacher are then used to train student CNN networks using the *interpolated-training* and *switched-training* methods described earlier. In these experiments we do not interpolate the teacher's label with the original labels. The KL-divergence criterion used for training the student model is equivalent to minimizing the cross entropy of the soft target labels. The student models of both standard and compact CNNs start from random initialization, learning from the soft labels provided by the VGG and LSTM models.

Table 1 shows the experimental results of training various student networks. As seen in the table, accuracies of the student models were improved by using soft labels generated with VGG and LSTM models. Comparing the standard size student CNN with baseline CNN, the word error rate was down from 13.3% to 11.6% with a single VGG teacher. By applying LSTM second teacher, additional 3.4% relative improvement with the *switched-training* strategy was obtained. We observe similar improvements with the compact student NN as shown in Table 2. This model achieves comparable performance to standard-size baseline CNN model with a RTF reduction of 23%. The proposed *switched-training* method performs better than the conventional *interpolated-training* and effectively distills knowledge from the strong VGG and LSTM teachers even with compact networks. In the *switched-training* scheme one of two teachers at the minibatch is randomly selected.

Table 3: *Performance of narrowband and broadband baseline CNN and teacher models on the Aurora 4 task.*

Models	A	B	C	D	AVG
Matched CNN-8k	4.6	11.5	6.4	17.7	12.1
Matched CNN-16k	3.9	7.8	6.0	17.5	11.6
VGG-8k	6.3	12.5	7.5	16.2	13.3
VGG-16k	4.8	8.4	6.2	14.3	10.5
LSTM-8k	6.3	11.8	7.3	15.1	12.5
LSTM-16k	4.8	9.3	7.4	16.0	11.7

3.3. Training with priviledged information

In this section we explore how the proposed training techniques can be used to train student networks with additional privileged information. To illustrate the efficacy of our approach we show how an improved narrowband CNN based acoustic model can be trained by using privileged information from outputs of broadband models, instead of training the student network on only narrowband teacher models. Privileged information is presented to the student network not only via both an ensemble of teachers but also by data augmentation of training data using the *augmented-training* method described earlier. The training data for this task is created using soft targets from an ensemble of narrowband teachers, an ensemble of broadband teachers and also hard targets.

The training data defined in Aurora 4 contains both clean and various noisy speech. Test results are reported on 4 subsets commonly referred to as clean (test set A), noisy (test set B), clean with channel distortion (test set C) and noisy with channel distortion (test set D). The performance of the baseline CNN and two teacher networks are tabulated in Table 3.

The "Matched CNN-8k" and "Matched CNN-16k" systems are both trained with Aurora 4 training data set using hard labels. The "Matched CNN-8k" narrowband baseline system is trained on audio data downsampled from 16kHz to 8kHz. We explore how the performance of the baseline narrowband model at a WER AVG of 12.1% can be further improved via knowledge distillation by employing four teachers models to train various student networks. The broadband teacher models (VGG-16k and LSTM-16k) were described earlier in Section 3.2. The narrowband teacher models (VGG-8k and LSTM-8k) are trained with 500 hours of Switchboard corpus after 250 hours from the corpus was augmented with JEIDA environmental noises and RWCP impulse responses used for the broadband systems. Since generic teacher models are not trained with matched training data to Aurora 4, the VGG-8k and LSTM-8k are worse than baseline CNN-8k.

Various student narrowband models are compared in Table

Table 4: *Performance on student narrowband models.*

Teachers	Target	A	B	C	D	AVG
VGG8k	soft	6.4	12.4	8.3	17.0	13.6
VGG16k	soft	6.0	11.4	7.8	16.3	12.9
VGG8k + VGG16k	soft/augment	5.7	11.1	7.2	15.6	12.4
Hard + VGG8k	soft/augment	3.9	8.9	5.5	13.6	10.3
Hard + VGG16k	soft/augment	4.0	8.6	5.5	13.7	10.2
Hard + VGG8k + VGG16k	soft/switching	4.3	8.6	5.6	13.2	10.1
Hard + VGG8k + VGG16k	soft/augment	3.8	8.2	5.5	12.9	9.7
Hard + VGG8k + VGG16k + LSTM8k + LSTM16k	soft/switching	4.5	8.8	5.9	13.3	10.2
Hard + VGG8k + VGG16k + LSTM8k + LSTM16k	soft/augment	3.8	8.3	5.3	12.8	9.7

4. Unlike experiments in the previous section, hard targets are also used together with soft labels estimated from the VGG and LSTM teacher models. As seen in the table, using only soft labels is not effective because the VGG-8k and LSTM-8k teachers perform lower than the baseline. In contrast, combining hard label with soft labels significantly improves performance. When broadband teachers are used to train narrowband systems, very clear additional gains are observed in addition to the combination of hard and VGG-8k model, leading to a 5.8% error reduction (10.3% to 9.7%). These results clearly highlight the value of the proposed training techniques for ensemble of teachers and the additional information that is being distilled from the priviledged information made available via broadband soft targets. It can also be seen that while the *switched-training* technique saturates in WER performance when 5 teachers including hard target is used, augmenting the training data by increasing the target in each frame using the *augmented-training* method provides better improvement over the randomly selected teachers in each frame. To further characterize how the student networks are performing we conduct an "oracle" experiment using hand tuned weights to combine the output scores various teacher networks during test time. Table 5 shows the system combination results of various teacher networks. It can be seen that the student networks trained using the proposed techniques achieves parity performance with the hand tuned systems.

Table 5: *System combination performance of various teacher networks.*

Teachers	A	B	C	D	AVG
VGG8k+LSTM8k	5.8	10.7	13.6	6.7	11.3
VGG16k+LSTM16k	4.7	8.2	6.2	13.9	10.3
VGG8k+VGG16k + LSTM8k+LSTM16k	4.8	8.0	5.7	13.0	9.7

The Aurora 4 task is a medium vocabulary task, primarily used to evaluate noise robust algorithms. Having demonstrated the impact of the proposed teacher-student schemes on this task, we explore the effectiveness of the same on two well-know large vocabulary continuous speech recognition (LVCSR) tasks: Aspire [19] and Broadcast news. The training data available for both these tasks is significantly higher (greater than 20 orders of magnitude) than the Aurora task. To keep the experiments manageable and to evaluate the generalization of these techniques on other tasks, we chose to train student models on the same data as the teacher models (described in Section 3.1). The same VGG and LSTM models are used as teachers. Table 6 presents student CNN models trained using the proposed strategies. It can be seen that the VGG teacher is significantly better than the LSTM teacher on both these tasks (Row 2). This very knowledge is transferred to the student as well (Row 4) resulting in WERs of 37.9% and 15.4%, resulting in the VGG-student beating the performance of the LSTM-teacher. When a student is trained using the *interpolated-training* scheme using the VGG and LSTM as teachers, significant reductions in WER can be seen on both tasks (Row 6). Similar gains can be seen when the student model is trained using the *switched-training* scheme. The *switched-training* scheme is comparable to the *interpolated-training* scheme for the BN-dev04f test set, while the *interpolated-training* scheme is slightly better on the Aspire test set. Table 7 illustrates the performance of the compact CNN on these tasks. Compared to the baseline, we still see significant reductions in WER. Furthermore, the performance of this student trained with the proposed *switched-training* scheme stays better than a student trained with a VGG teacher only.

Table 6: *Comparing CNN trained with hard targets and student CNN learned from VGG and LSTM.*

Model	Target	ASpIRE	BN-dev04f
CNN	hard	41.3	18.4
VGG	hard	35.1	14.3
LSTM	hard	38.7	16.3
CNN: VGG	soft	37.9	15.4
CNN: LSTM	soft	40.4	17.1
CNN: VGG+LSTM	soft/interpolation	37.1	15.0
CNN: VGG+LSTM	soft/switching	37.5	15.1

Table 7: *Comparing Compact CNN trained with hard targets and student CNN learned from VGG and LSTM.*

Model	Target	ASpIRE	BN-dev04f
CNN	hard	44.2	20.5
CNN: VGG	soft	41.7	17.8
CNN: VGG+LSTM	soft/switching	41.3	17.7

4. Conclusions

In this paper we have proposed two new strategies for knowledge distillation using multiple teachers. The proposed schemes yield gains on both low resource and large resource settings with the *switched-training* scheme doing better in low-resource conditions. We hypothesize that this can be attributed to the random selection/order of the teachers playing a dominant role with reduced training data. Our experiments on the Aurora task show that teachers trained on completely different domains can still provide significant amounts of knowledge to the student. We have demonstrated that simpler and compact student models can achieve comparable recognition accuracy to more complex teacher models such as the VGG based models.

5. References

[1] H. Soltau, G. Saon, and T. N. Sainath, “Joint training of convolutional and non-convolutional neural networks,” *Proc. IEEE ICASSP*, pp. 5609–5613, 2014.

[2] T. Fukuda, O. Ichikawa, G. Kurata, R. Tachibana, S. Thomas, and B. Ramabhadran, “Effective joint training of denoising featuer space transforms and neural network based acoustic models,” *Proc. IEEE ICASSP*, pp. 5190–5194, 2017.

[3] T. Fukuda, O. Ichikawa, and M. Nishimura, “Long-term spectro-temporal and static harmonic features for voice activity detection,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 834–844, 2010.

[4] J. Ba and R. Caruana, “Do deep nets really need to be deep?” in *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc, 2014, pp. 2654–2662.

[5] V. Vapnik and R. Izmailov, “Learning using privileged information: Similarity control and knowledge transfer,” *Machine Learning Research*, vol. 16, pp. 2023–2049, 2015.

[6] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, “Learning small-size DNN with output-distribution-based criteria,” *Proc. Interspeech*, pp. 1910–1914, September 2014.

[7] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *arXiv:1503.02531v1*, 2015.

[8] W. Chan, N. R. Ke, and I. Lane, “Transferring knowledge from a RNN to a DNN,” *Proc. Interspeech*, pp. 3264–3268, 2015.

[9] K. J. Geras, A.-R. Mohamed, R. Caruana, G. Urban, S. Wang, O. Aslan, M. Philipose, M. Richardson, and C. Sutton, “Blending LSTMs into CNNs,” in *ICLR Workshop*, 2016.

[10] Z. Tang, D. Wang, and Z. Zhang, “Recurrent neural network training with dark knowledge transfer,” *Proc. IEEE ICASSP*, pp. 5900–5904, 2016.

[11] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, and A. Rosenberg, “Knowledge distillation across ensembles of multilingual models for low-resource languages,” *Proc. IEEE ICASSP*, pp. 4825–4829, 2017.

[12] J. H. M. Wong and M. J. F. Gales, “Sequence student-teacher training of deep neural networks,” *Proc. Interspeech*, pp. 2761–2765, 2016.

[13] Y. Chebotar and A. Waters, “Distilling knowledge from ensembles of neural networks for speech recognition,” *Proc. Interspeech*, pp. 3439–3443, 2016.

[14] K. Markov and T. Matsui, “Robust speech recognition using generalized distillation framework,” *Proc. Interspeech*, pp. 2364–2368, 2016.

[15] L. Brandchain, “The mixer 6 corpus: Resource for cross-channel and text independent speaker recognition,” *LREC*, 2010.

[16] J. Carletta, “Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus,” *Language Resources and Evaluation*, vol. 41, no. 1, pp. 181–190, 2007.

[17] S. Itahashi, “Recent speech database projects in japan,” *Proc. ICSLP*, 1990.

[18] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” *LREC*, 2000.

[19] M. Harper, “The automatic speech recognition in reverberant environments (ASpIRE) challenge,” *Proc. IEEE ASRU*, pp. 547–554, 2015.