

CONDITIONAL TEACHER-STUDENT LEARNING

Zhong Meng, Jinyu Li, Yong Zhao, Yifan Gong

Microsoft Corporation, Redmond, WA, USA

ABSTRACT

The teacher-student (T/S) learning has been shown to be effective for a variety of problems such as domain adaptation and model compression. One shortcoming of the T/S learning is that a teacher model, not always perfect, sporadically produces wrong guidance in form of posterior probabilities that misleads the student model towards a suboptimal performance. To overcome this problem, we propose a conditional T/S learning scheme, in which a “smart” student model selectively chooses to learn from either the teacher model or the ground truth labels *conditioned on* whether the teacher can correctly predict the ground truth. Unlike a naive linear combination of the two knowledge sources, the conditional learning is exclusively engaged with the teacher model when the teacher model’s prediction is correct, and otherwise backs off to the ground truth. Thus, the student model is able to learn effectively from the teacher and even potentially surpass the teacher. We examine the proposed learning scheme on two tasks: domain adaptation on CHiME-3 dataset and speaker adaptation on Microsoft short message dictation dataset. The proposed method achieves 9.8% and 12.8% relative word error rate reductions, respectively, over T/S learning for environment adaptation and speaker-independent model for speaker adaptation.

Index Terms— teacher-student learning, domain adaptation, speaker adaptation

1. INTRODUCTION

Teacher-student (T/S) learning [1, 2] has been widely applied to a variety of deep learning tasks in speech, language and image processing including model compression [1, 2], domain adaptation [3, 4, 5], small-footprint natural machine translation (NMT) [6], low-resource NMT [7], far-field automatic speech recognition (ASR) [8, 9], low-resource language ASR [10] and neural network pre-training [11]. T/S learning falls in the category of transfer learning, where the network of interest, as a student, is trained by mimicking the behavior of a well-trained network, as a teacher, in the presence of the same or stereo training samples. Formally, the T/S learning works by minimizing the Kullback-Leibler (KL) divergence between the output distribution of the student and teacher models, other than from the hard labels derived from the transcriptions.

Compared to using conventional one-hot hard label as the training target, the transfer of soft posteriors [1] well preserves the probabilistic relationships among different classes encoded at the output of the teacher model. Because soft labels provide more information than hard labels for the model training, the T/S learning results in better performance as reported in [1, 2, 8]. The largest benefits of using pure soft labels is learning without any hard labels, enabling the use of much larger amount of unlabeled data to improve the student model performance [1, 8].

One shortcoming of the T/S learning is that a teacher model, not always perfect, sporadically makes the incorrect predictions that

mislead the student model towards a suboptimal performance. In such a case, it may be beneficial to utilize hard labels of the training data to alleviate this effect. Hinton et. al. [2] later proposed an interpolated T/S learning called knowledge distillation, in which a weighted sum of the soft posteriors and the one-hot hard label is used to train the student model. One issue is that the simple linear combination with one-hot vectors destroys the relationships among different classes embedded naturally in the soft posteriors produced by the teacher model. Moreover, proper setting of the interpolation weight with a fixed value is known to be critical and it varies with the adaptation scenarios and the qualities of the teacher and ground truth labels.

In this paper, we propose a *conditional T/S learning* scheme, where the student model becomes smart so that it can criticize the knowledge imparted by the teacher model to make better use of the teacher and the ground truth. At the initial stage, when the student model is very weak, it blindly follows whatever knowledge infused by the teacher model and uses the soft posteriors as the solely training targets. As the student model grows stronger, it begins to selectively choose the learning source from either the teacher model or the ground truth labels *conditioned on* whether the teacher’s prediction coincides with the ground truth. That is, the student model would learn exclusively from the teacher when the teacher makes correct prediction on training samples, and otherwise from the ground truth when the teacher is wrong. With conditional T/S learning, the student makes good use of rich and correct knowledge encompassed by the teacher, while avoids receiving inaccurate knowledge generated by the teacher. Another advantage of the conditional T/S learning over the conventional T/S learning is that it forgoes tuning the interpolation weight between two knowledge sources.

We applied the proposed approach to two tasks, domain adaptation and speaker adaptation. In domain adaptation, the student model is trained using the noise corrupted data in the target domain as input and the soft target obtained from the teacher posterior computed on the corresponding clean data. We demonstrate the effectiveness of the proposed approach using CHiME-3 dataset. In speaker adaptation, it can be shown that the conventional T/S learning is equivalent to the KLD adaptation [12], where the speaker-independent model acts as a teacher and the speaker dependent model acts as a student. Similarly, we apply the conditional T/S learning to further boost the performance of the KL divergence (KLD) adaptation. We demonstrate the improvement over the KLD adaptation for supervised and unsupervised adaptation on the Microsoft Windows Phone short message dictation task.

2. TEACHER-STUDENT LEARNING

In T/S learning, a well-trained teacher network takes in a sequence of training samples $\mathbf{X}^T = \{\mathbf{x}_1^T, \dots, \mathbf{x}_N^T\}$, $\mathbf{x}_i^T \in \mathbb{R}^{D_T}$ and predicts a sequence of class labels. Here, each class is represented by an integer $c \in \{1, 2, \dots, D_C\}$ and D_C is the total number of classes in

the classification task. The goal is to learn a student network that can accurately predict the class labels for each of its input samples $\mathbf{X}^S = \{\mathbf{x}_1^S, \dots, \mathbf{x}_N^S\}$, $\mathbf{x}_i^S \in \mathbb{R}^{D_S}$ by using the knowledge transferred from the teacher network. To ensure effective knowledge transfer, the input sample sequences \mathbf{X}^T and \mathbf{X}^S need to be parallel to each other, i.e., each pair of train samples \mathbf{x}_i^T and \mathbf{x}_i^S share the same ground truth class label $c_i \in \{1, 2, \dots, D_C\}$.

2.1. T/S Learning with Soft Labels

T/S learning minimizes the Kullback-Leibler (KL) divergence between the output distributions of the teacher network and the student network given parallel data X^T and X^S are at the input to the networks [1]. The KL divergence between the teacher and student output distributions $p(c|\mathbf{x}_i^T; \theta_T)$ and $p(c|\mathbf{x}_i^S; \theta_S)$ is formulated as:

$$\mathcal{KL} \left[p(c|\mathbf{x}_i^T; \theta_T) \parallel p(c|\mathbf{x}_i^S; \theta_S) \right] = \sum_{i=1}^N \sum_{c=1}^{D_C} p(c|\mathbf{x}_i^T; \theta_T) \log \left[\frac{p(c|\mathbf{x}_i^T; \theta_T)}{p(c|\mathbf{x}_i^S; \theta_S)} \right], \quad (1)$$

i is the sample index, θ_T and θ_S are the parameters of the teacher and student networks, respectively, $p(c|\mathbf{x}_i^T; \theta_T)$ and $p(c|\mathbf{x}_i^S; \theta_S)$ are the posteriors of class c predicted by the teacher and student network given the input samples \mathbf{x}_i^T and \mathbf{x}_i^S , respectively. To learn a student network that approximates the given teacher network, we minimize the KL divergence with respect to only the parameters of the student network while keeping the parameters of the teacher network fixed, equivalent to minimizing the loss function below:¹

$$\mathcal{L}_{TS}(\theta_S) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{D_C} p(c|\mathbf{x}_i^T; \theta_T) \log p(c|\mathbf{x}_i^S; \theta_S). \quad (2)$$

2.2. T/S Learning with Interpolated Labels

However, in T/S learning, the knowledge from the teacher is not accurate when the teacher's classification decision is incorrect. To deal with this, Hinton et. al. [2] later suggested an interpolated T/S method which uses a weighted sum of the soft posteriors and the one-hot hard label to train the student model. Assuming that the sequence of one-hot ground truth class labels that both X^T and X^S are aligned with is $\mathbf{C} = \{c_1, \dots, c_N\}$, The interpolated T/S learning aims to minimizing the loss function below:

$$\mathcal{L}_{ITS}(\theta_S) = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{D_C} \left[(1 - \lambda) \mathbb{1}[c = c_i] + \lambda p(c|\mathbf{x}_i^T; \theta_T) \right] \log p(c|\mathbf{x}_i^S; \theta_S), \quad (3)$$

where $0 \leq \lambda \leq 1$ is the weight for the class posteriors and $\mathbb{1}[\cdot]$ is the indicator function which equals to 1 if the condition in the squared bracket is satisfied and 0 otherwise. Note that the interpolated T/S learning becomes soft T/S when $\lambda = 1.0$ and becomes standard cross-entropy training with hard labels when $\lambda = 0.0$. Although interpolated T/S compensates for the imperfection in knowledge transfer, the linear combination of soft and hard labels destroys the correct relationships among different classes embedded naturally in the soft class posteriors and deviates the student model parameters from the optimal direction. Moreover, the search for the best student model is subject to the heuristic tuning of λ between 0 and 1.

¹In some cases, the senone posteriors generated by the teacher network are flattened by a temperature $T > 1$ before serving as the soft labels [2]. But in speech area, T is normally fixed at 1 [9, 13, 14]. We obtain the best performance when $T = 1$ and the same conclusion is also reported in [15, 6].

3. CONDITIONAL TEACHER-STUDENT LEARNING

Instead of blindly combining the soft and hard labels, the student network needs to be critical about the knowledge infused by the teacher network, i.e., to judge whether the class posteriors are accurate or not before learning from them. One natural judgment is that the teacher's knowledge is deemed accurate when it correctly predicts the ground truth given the input samples, and deemed inaccurate otherwise. Therefore, the training target for the student model should be *conditioned* on the correctness of the teacher's prediction, i.e., the student network exclusively uses the soft posteriors from the teacher as the training target when the teacher is correct and uses the hard label instead when the teacher is wrong as shown in Fig. 1.

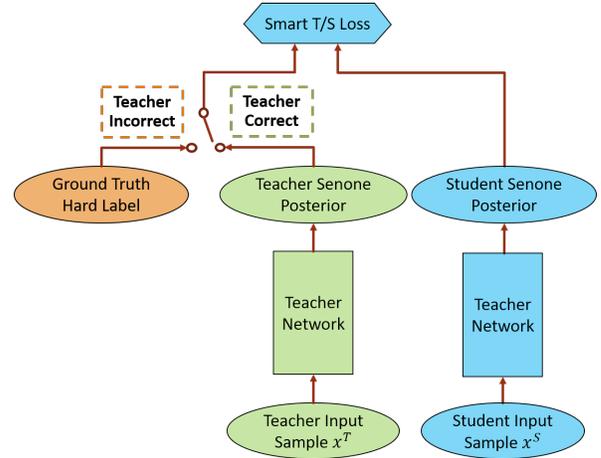


Fig. 1. The framework of conditional T/S learning.

In other words, assuming $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$, $\mathbf{y}_i \in \mathbb{R}^{D_C}$ to be the sequence of *conditional* class label vectors used as the target to train the student network, the c^{th} element of \mathbf{y}_i becomes

$$y_{i,c} = \begin{cases} p(c|\mathbf{x}_i^T; \theta_T), & \arg \max_{k \in \{1, \dots, D_C\}} p(k|\mathbf{x}_i^T; \theta_T) = c_i \\ \mathbb{1}[c = c_i], & \text{otherwise,} \end{cases} \quad (4)$$

under conditional T/S learning. That is to say, the conditional class label \mathbf{y}_i is a soft vector of class posteriors if the teacher is correct and a hard one-hot vector if the teacher is wrong. The loss function to be minimized is formulated as the cross-entropy between the conditional class labels and the class posteriors generated by the student network as follows:

$$\begin{aligned} \mathcal{L}_{CTS}(\theta_S) &= -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^{D_C} y_{i,c} \log p(c|\mathbf{x}_i^S; \theta_S) \\ &= -\frac{1}{N} \sum_{i=1}^N \left\{ \left[\sum_{c=1}^{D_C} p(c|\mathbf{x}_i^T; \theta_T) \log p(c|\mathbf{x}_i^S; \theta_S) \right] \right. \\ &\quad \left. \mathbb{1} \left[\arg \max_{k \in \{1, \dots, D_C\}} p(k|\mathbf{x}_i^T; \theta_T) = c_i \right] \right. \\ &\quad \left. + \log p(c_i|\mathbf{x}_i^S; \theta_S) \mathbb{1} \left[\arg \max_{k \in \{1, \dots, D_C\}} p(k|\mathbf{x}_i^T; \theta_T) \neq c_i \right] \right\}. \quad (5) \end{aligned}$$

The student network parameters are optimized through standard back propagation with stochastic gradient decent. With conditional T/S learning, the student can learn from only the selected *accurate* knowledge generated by the teacher while simultaneously take

advantage of the well-preserved probabilistic relationships among different classes and is thus expected to achieve improved performance in classification tasks.

4. CONDITIONAL T/S LEARNING FOR ACOUSTIC MODEL ADAPTATION

With the advent of deep acoustic models, the performance of ASR has been greatly improved [16, 17, 18]. A deep acoustic model takes the speech frames as the input and predicts the corresponding senone posteriors at the output layer. To achieve robust ASR over different domains and speakers, we apply conditional T/S learning to the domain and speaker adaptation of deep acoustic models. In these tasks, both teacher and student networks represent deep acoustic models, \mathbf{X}^T and \mathbf{X}^S are sequences of input speech frames, and c denotes one senone in the set of all possible senones $\{1, \dots, D_C\}$ predicted by the teacher and student acoustic models.

4.1. Conditional T/S Learning for Domain Adaptation

ASR suffers from performance degradation when a well-trained acoustic model is applied in a new domain [19]. T/S learning [3, 8, 9] and adversarial learning [20, 21, 22, 23, 24] are two effective approaches that can suppress this domain mismatch by adapting a source-domain acoustic model to target-domain speech. T/S learning is more suited for the situation where unlabeled parallel data is available for adaptation,² in which a sequence of source-domain speech features is fed as the input to a source-domain teacher model and a parallel sequence of target-domain features is at the input to the target-domain student model to optimize the student model parameters by minimizing the T/S loss in Eq. (2).

To further improve T/S learning, we introduce the conditional T/S learning by using the ground truth hard labels \mathbf{C} of the adaptation data and propose the following steps for domain adaptation.

1. Use a well-trained source-domain acoustic model as the teacher network and initialize the student network with the parameters of the teacher.
2. Use *paralleled* source and target domain adaptation data as \mathbf{X}^T and \mathbf{X}^S , respectively. All pairs of \mathbf{x}_i^T and $\mathbf{x}_i^S, \forall i \in \{1, \dots, N\}$ are frame-by-frame synchronized.
3. Perform T/S learning [1] to train the student network by minimizing $\mathcal{L}_{TS}(\theta_S)$ in Eq. (2).
4. After Step 3, the student network has performed reasonably well on target-domain data. Conduct *conditional T/S* learning with conditional senone labels \mathbf{Y} defined in Eq. (4) to train the student network by minimizing $\mathcal{L}_{CTS}(\theta_S)$ in Eq. (5).
5. Use the optimized student network as the adapted acoustic model for decoding test utterances in the target domain.

4.2. Conditional T/S Learning for Speaker Adaptation

Speaker adaptation aims at learning a set of speaker-dependent (SD) acoustic models by adapting an speaker-independent (SI) acoustic model to the speech of target speakers. Different from domain adaptation, speaker adaptation has only access to limited adaptation data from target speakers and has no access to the source-domain data.

Many techniques have been proposed for speaker adaptation of deep acoustic models, such as regularization-based [25, 26, 27], transformation-based [28, 29], singular value decomposition-based

[30, 31], subspace-based [32, 33] and adversarial learning-based [34, 35] approaches. Among these, KL divergence (KLD) regularization [25] is one of the most popular methods to prevent the adapted model from overfitting the limited speaker data. This regularization is realized by augmenting the training criterion with the KLD between the output distributions of the SD and SI models.

Apparently, the KLD adaptation is a special case of the interpolated T/S learning [12], in which the SI model acts as a teacher, the SD model acts as a student, and both take the adaptation data as input. The teacher network is more like a regularizer that constrains the student network from straying too far away from the teacher network. However, the linear combination between soft posteriors and hard labels does not make full use of two knowledge sources, and the best regularization weight λ is subject to heuristic tuning. We apply the conditional T/S learning to further improve the KLD adaptation. That is, when the SI model makes the right predictions, the SD model exclusively learns from the SI model; when the SI model is wrong, the adaptation target backs off to the hard labels.

Note that since the SD model grows from the SI model, the adaptation can be interpreted as a self-taught learning process. In the step of learning from the SI model, the SD model basically reviews what it has already known once again, which sounds not quite informative. However, if we remove this step, i.e., adapt the SD model only when the SI model makes a mistake, the performance degrades. This is because using partial training set leads to catastrophic forgetting and skews the estimation of the senone distributions for the target speaker towards those samples the teacher model makes mistakes on and there is no guarantee that the student model can work well on those samples the teacher model is good at.

The conditional T/S learning for speaker adaptation consists of the following steps.

1. Use a well-trained SI acoustic model as the teacher network and initialize the student network with the parameters of the teacher.
2. Use adaptation data from a target speaker as both \mathbf{X}^T and \mathbf{X}^S .
3. Perform *conditional T/S* learning with conditional senone labels \mathbf{Y} defined in Eq. (4) to train the student network by minimizing $\mathcal{L}_{CTS}(\theta_S)$ in Eq. (5).
4. Use the optimized student network as the SD acoustic model for this target speaker.

For unsupervised speaker adaptation, we use the SI model to generate the hard labels \mathbf{C} to judge the SI model itself. Since the recognition hypotheses are generated through the cooperation of the SI acoustic model along with the language model, the derived hard labels are expected to be more accurate than the senone classification decisions generated by only the SI model at the frame level.

5. EXPERIMENTS

5.1. Domain Adaptation

As a major category of domain adaptation, we first verify conditional T/S learning with environment adaptation experiments. Specifically, we adapt a well-trained clean acoustic model to the noisy training data of CHiME-3 [36] using different methods. The CHiME-3 dataset incorporates Wall Street Journal (WSJ) corpus sentences spoken in challenging noisy environments, recorded using a 6-channel tablet. The real far-field noisy speech from the 5th microphone channel in CHiME-3 development data set is used for testing. A standard WSJ 5K word 3-gram language model (LM) is used for decoding.

²The parallel data can be either recorded or simulated as in [3].

System	BUS	CAF	PED	STR	Avg.
Unadapted	43.47	45.93	30.43	36.13	38.96
Hard Label	24.92	20.63	15.96	18.01	19.84
Soft T/S	22.46	19.10	14.88	16.47	18.20
IT/S ($\lambda = 0.2$)	24.84	19.79	15.55	18.36	19.60
IT/S ($\lambda = 0.5$)	22.61	18.94	14.52	18.43	18.59
IT/S ($\lambda = 0.8$)	23.51	19.10	14.49	16.56	18.37
Conditional T/S	20.72	17.46	12.52	15.09	16.42

Table 1. The WER (%) performance of environment adaptation using one-hot hard label, T/S, interpolated T/S (IT/S) and conditional T/S learning on the real noisy test set of CHiME-3.

As a source-domain acoustic model, a clean long short-term memory (LSTM)- recurrent neural networks (RNN) [37, 38, 39] is trained with 9137 clean training utterances of CHiME-3 dataset by using cross-entropy criterion. The 29-dimensional log Mel filterbank features together with 1st and 2nd order delta features (totally 87-dimensional) for both the clean and noisy utterances are extracted by following the process in [40]. The features are fed as the input of the LSTM after global mean and variance normalization. The LSTM has 4 hidden layers with 1024 hidden units for each layer. A 512-dimensional projection layer is inserted on top each hidden layer to reduce the number of parameters. The output layer of the LSTM has 3012 output units corresponding to 3012 senone labels. There is no frame stacking, and the output HMM senone label is delayed by 5 frames. Senone-level forced alignment of the clean data is generated using a Gaussian mixture model-HMM system. The clean CHiME-3 LSTM acoustic model achieves 7.43% and 38.96% WERs on clean and real noisy test data of CHiME-3, respectively. The clean LSTM acoustic model serves as the teacher network in the subsequent T/S learning methods. Trained with noisy and clean data using their one-hot hard labels, the multi-style LSTM acoustic model achieves 19.84% WER on the noisy test data.

For domain adaptation in [3], parallel data consisting of 9137 pairs of clean and noisy utterances in the CHiME-3 training set are used as the adaptation data for T/S learning. In order to make the student model invariant to environments, the training data for student model should include both clean and noisy data. Therefore, we extend the original T/S learning work by also including 9137 pairs of the clean and clean utterances in CHiME-3 for adaptation as in [4]. As shown in Table 1, soft T/S learning achieves 18.20% average WERs after environment adaptation, which is 51.3% relative improvement over the clean model. To further improve the student model, we perform conditional T/S learning with the help of hard labels as described in Section 4.1. As a comparison, we conduct interpolated T/S learning [2] with different weights for soft labels. The conditional T/S learning achieves 16.42% average WERs with 9.8% and 11.7% relative improvements over soft T/S learning and the best performed interpolated T/S ($\lambda = 0.5$), respectively.

Note that we can get a better student model if we have a better teacher model. Then, we did a quick experiment by using a 375 hour-trained Cortana model which was used in [3] as the teacher model to learn the student model with the same CHiME-3 parallel data. The soft T/S model gets 13.56% WER which is significantly better the one in Table 1, and the conditional T/S can reach 11.13% WER, which stands for 17.9% relative improvement over soft T/S.

5.2. Speaker Adaptation

We further perform speaker adaptation on a Microsoft internal Windows Phone short message dictation (SMD) task. The test set consists of 7 speakers with a total number of 20,203 words. A separate

System	Supervised	Unsupervised
SI	13.95	
Hard Label	13.20	13.77
KLD ($\lambda = 0.2$)	12.61	13.65
KLD ($\lambda = 0.5$)	12.54	13.55
KLD ($\lambda = 0.8$)	13.17	13.72
Conditional T/S	12.17	13.21

Table 2. The WER (%) performance of speaker adaptation using one-hot hard label, KLD and conditional T/S learning on Microsoft SMD task. The SI LSTM model is trained with 2600 hours Microsoft live US English data.

adaptation set of 200 sentences per speaker is used for model adaptation. We train an SI LSTM acoustic model with 2600 hours of Microsoft internal live US English data. This SI model has 4 hidden LSTM layers with 1024 units in each layer and the output size of each LSTM layer is reduced to 512 by linear projection. The acoustic feature is 80-dimensional log Mel filterbank. The output layer has a dimension of 5980. The LSTM-RNN is trained to minimize the frame-level cross-entropy criterion. There is no frame stacking, and the output HMM state label is delayed by 5 frames. A trigram LM is used for decoding with around 8M n-grams. This SI LSTM acoustic model achieves 13.95% WER on the SMD test set.

We perform conditional T/S learning as in Section 4.2 to adapt the SI LSTM with 200 utterances in the adaptation set for each test speaker. For supervised adaptation, the hard labels come from the human transcription though forced alignment. For unsupervised adaptation, we use the SI model to generate the hypothesis. As a comparison, the standard adaptation with hard labels and KLD adaptation [25] with regularization weights λ of 0.2, 0.5 and 0.8 are also conducted to adapt the SI LSTM. Note that the adaptation with hard labels is equivalent to KLD adaptation with $\lambda = 0$. As in Table 2, the KLD adaptation produces its best WERs of 12.54% and 13.55% for supervised and unsupervised adaptation at $\lambda = 0.5$, respectively. The conditional T/S learning outperforms the KLD adaptation. It achieves 12.17% WER for supervised adaptation, which is 12.8% and 3.0% relative gain over the SI model and the best performed KLD adaptation ($\lambda = 0.5$). For unsupervised adaptation, the conditional T/S learning achieves 13.21% WER, which is 5.3% and 2.5% relative gain over the SI acoustic model and KLD adaptation.

6. CONCLUSION

We proposed a conditional T/S learning method, in which the student network selectively learns from either the soft posteriors generated by the teacher network or the one-hot hard label *conditioned on* whether the teacher makes correct decisions or not. Instead of blindly following whatever knowledge the teacher infuses as in the conventional T/S learning, the conditional T/S learning pursues the most trustworthy knowledge throughout the training, eliminating the burden tuning interpolation weights. We applied conditional T/S learning to domain adaptation and obtained 9.8% relative WER improvement over a strong T/S learning baseline on the CHiME-3 dataset. For speaker adaptation, the conditional T/S learning outperformed the KLD adaptation, which is equivalent to the interpolated T/S learning. It achieved 12.8% and 5.3% relative WER gains for supervised and unsupervised adaptations, respectively, over a well-trained SI LSTM model.

7. REFERENCES

- [1] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria.," in *Proc. INTERSPEECH*, 2014, pp. 1910–1914.
- [2] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015.
- [3] J. Li, M. L. Seltzer, X. Wang, et al., "Large-scale domain adaptation via teacher-student learning," in *INTERSPEECH*, 2017.
- [4] Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang (Fred) Juang, "Adversarial teacher-student learning for unsupervised domain adaptation," in *Proc. ICASSP*, 2018.
- [5] L. Mošner, M. Wu, A. Raju, et al., "Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning," *arXiv preprint arXiv:1901.02348*, 2019.
- [6] Yoon Kim and Alexander M. Rush, "Sequence-level knowledge distillation," in *EMNLP*, 2016, pp. 1317–1327.
- [7] Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li, "A teacher-student framework for zero-resource neural machine translation," in *Proc. ACL*, 2017, pp. 1925–1935.
- [8] J. Li, R. Zhao, Z. Chen, et al., "Developing far-field speaker system via teacher-student learning," *arXiv preprint arXiv:1804.05166*, 2018.
- [9] S. Watanabe, T. Hori, J. Le Roux, et al., "Student-teacher network learning with enhanced features," in *Proc. ICASSP*, 2017.
- [10] J. Cui, B. Kingsbury, B. Ramabhadran, et al., "Knowledge distillation across ensembles of multilingual models for low-resource languages," in *Proc. ICASSP*. IEEE, 2017.
- [11] Z. Tang, D. Wang, and Z. Zhang, "Recurrent neural network training with dark knowledge transfer," in *Proc. ICASSP*, 2016.
- [12] T. Asami, R. Masumura, Y. Yamaguchi, et al., "Domain adaptation of dnn acoustic models using knowledge distillation," in *Proc. ICASSP*, 2017, pp. 5185–5189.
- [13] W. Chan, N. R. Ke, and I. Lane, "Transferring knowledge from a rnn to a dnn," in *INTERSPEECH*, 2015.
- [14] T. Tan, Yanmin Q., and D. Yu, "Knowledge transfer in permutation invariant training for single-channel multi-talker speech recognition," in *ICASSP*, 2018.
- [15] L. Lu, M. Guo, and S. Renals, "Knowledge distillation for small-footprint highway networks," in *ICASSP*, 2017.
- [16] N. Jaitly, P. Nguyen, A. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *Proc. INTERSPEECH*, 2012.
- [17] G. Hinton, L. Deng, D. Yu, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [18] Li Deng, Jinyu Li, Jui-Ting Huang, et al., "Recent advances in deep learning for speech research at Microsoft," in *ICASSP*, 2013.
- [19] J. Li, L. Deng, Y. Gong, et al., "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2014.
- [20] S. Sun, B. Zhang, L. Xie, and Y. Zhang, "An unsupervised deep domain adaptation approach for robust speech recognition," *Neurocomputing*, 2017.
- [21] Z. Meng, Z. Chen, V. Mazalov, J. Li, and Y. Gong, "Unsupervised adaptation with domain separation networks for robust speech recognition," in *Proceeding of ASRU*, 2017.
- [22] Z. Meng, J. Li, and Y. Gong, "Attentive adversarial learning for domain-invariant training," in *Proc. ICASSP*, 2019.
- [23] Z. Meng, J. Li, and Y. Gong, "Cycle-consistent speech enhancement," *Interspeech*, 2018.
- [24] Z. Meng, J. Li, and Y. Gong, "Adversarial feature-mapping for speech enhancement," *Interspeech*, 2018.
- [25] D. Yu, K. Yao, H. Su, et al., "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. ICASSP*, 2013.
- [26] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. ICASSP*, May 2013, pp. 7947–7951.
- [27] Z. Huang, J. Li, S. Siniscalchi, et al., "Rapid adaptation for deep neural networks through multi-task learning," in *Interspeech*, 2015.
- [28] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, Dec 2011, pp. 24–29.
- [29] P. Swietojanski, J. Li, and S. Renals, "Learning hidden unit contributions for unsupervised acoustic model adaptation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1450–1463, Aug 2016.
- [30] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition.," in *Interspeech*, 2013, pp. 2365–2369.
- [31] Y. Zhao, J. Li, and Y. Gong, "Low-rank plus diagonal adaptation for deep neural networks," in *Proc. ICASSP*, 2016.
- [32] S. Xue, O. Abdel-Hamid, H. Jiang, et al., "Fast adaptation of deep neural network based on discriminant codes for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Dec 2014.
- [33] L. Samarakoon and K. C. Sim, "Factorized hidden layer adaptation for deep neural network based acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2241–2250, Dec 2016.
- [34] Zhong Meng, Jinyu Li, and Yifan Gong, "Adversarial speaker adaptation," in *Proc. ICASSP*, 2019.
- [35] Z. Meng, J. Li, Z. Chen, et al., "Speaker-invariant training via adversarial learning," in *Proc. ICASSP*, 2018.
- [36] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. ASRU*, 2015, pp. 504–511.
- [37] F. Beaufays H. Sak, A. Senior, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech*, 2014.
- [38] Z. Meng, S. Watanabe, J. R. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *ICASSP*, 2017.
- [39] H. Erdogan, T. Hayashi, J. R. Hershey, et al., "Multi-channel speech recognition: Lstms all the way through," in *CHiME-4 workshop*, 2016, pp. 1–4.
- [40] J. Li, D. Yu, J.-T. Huang, and Y. Gong, "Improving wide-band speech recognition using mixed-bandwidth training data in CD-DNN-HMM," in *Proc. SLT*. IEEE, 2012, pp. 131–136.