ATTENTIVE ADVERSARIAL LEARNING FOR DOMAIN-INVARIANT TRAINING

Zhong Meng, Jinyu Li, Yifan Gong

Microsoft Corporation, Redmond, WA, USA

{zhme, jinyl, ygong}@microsoft.com

ABSTRACT

Adversarial domain-invariant training (ADIT) proves to be effective in suppressing the effects of domain variability in acoustic modeling and has led to improved performance in automatic speech recognition (ASR). In ADIT, an auxiliary domain classifier takes in *equally-weighted* deep features from a deep neural network (DNN) acoustic model and is trained to improve their domain-invariance by optimizing an adversarial loss function. In this work, we propose an attentive ADIT (AADIT) in which we advance the domain classifier with an attention mechanism to *automatically weight* the input deep features according to their importance in domain classification. With this attentive re-weighting, ADDIT can focus on the domain normalization of phonetic components that are more susceptible to domain variability and generates deep features with improved domain-invariance and senone-discriminativity over ADIT. Most importantly, the attention block serves only as an *external* component to the DNN acoustic model and is not involved in ASR, so AADIT can be used to improve the acoustic modeling with any DNN architectures. More generally, the same methodology can improve any adversarial learning system with an auxiliary discriminator. Evaluated on CHiME-3 dataset, the AADIT achieves 13.6% and 9.3% relative WER improvements, respectively, over a multi-conditional model and a strong ADIT baseline.

Index Terms— adversarial learning, attention, domain-invariant training, neural network, automatic speech recognition

1. INTRODUCTION

The deep neural network (DNN) based acoustic models have been widely used in automatic speech recognition (ASR) and have achieved extraordinary performance improvement [1, 2, 3]. However, the performance of a multi-conditional acoustic model trained with speech data from a variety of environments, speakers, microphone channels, etc. is still affected by the spectral variations in each speech unit caused by the inter-domain variability [4]. Recently, adversarial learning [5] has effectively improved the noise robustness of the DNN acoustic model for ASR [6, 7, 8] or that of the deep embeddings for speaker verification [9] by using gradient reversal layer network [10] or domain separation network [11]. Similar idea has also been applied to reduce the effect of inter-speaker [12, 13], inter-language [14] and inter-dialect [15] variability in acoustic modeling that is trained with speech from multiple speakers, multiple dialects or multiple languages. We name all the above approaches *adversarial domain-invariant training (ADIT)* by referring to each speaker, environment, language, etc. that contributes to the condition-variability of the training data a *domain*.

To perform ADIT, an additional DNN domain classifier is introduced and is jointly trained with the multi-conditional DNN acoustic model to simultaneously optimize the primary task of minimizing the senone classification loss and the secondary task of mini-maximizing the domain classification loss. Through this adversarial multi-task learning procedure, a shared feature extractor is learned as the bottom layers of the DNN acoustic model that maps the input speech frames from different domains into *domain-invariant* and senone-discriminative deep hidden features, so that further senone classification is based on representations with the domain factor already normalized out. After ADIT, only the DNN acoustic model is used to generate word transcription for test data from unseen domain through *one-pass online* decoding [12]. The domain classifier is not used in ASR. In this work, we focus on improving the domain classifier in ADIT to generate deep features with increased domain-invariance *without* changing the DNN acoustic model.

In ADIT, the sequence of deep features generated by the feature extractor are weighted equally before taken as the input to the domain classifier. In fact, the deep features corresponding to different phonetic components are affected nonuniformly by domain variability and show different domain-discriminativity to the domain classifier. To improve ADIT, we introduce an *attention mechanism* to allow the domain classifier to attend to different positions in time of deep feature sequence with *nonuniform* weights. The weights are automatically and dynamically determined by the attention mechanism according to the importance of the deep features in domain classification. We call this method *attentive adversarial DIT (AADIT)*. With AADIT, the domain classifier induces attentive reversal gradients that emphasize on the domain normalization of more domain-discriminative deep features, improving the domain invariance of the acoustic model and thus the ASR performance over ADIT.

Self-attention [16] is a new technique to improve encoder-decoder end-to-end models [17, 18, 19]. It is used in [20] to enhance the performance of the generative adversarial network [5] by generating images based on cues from all feature locations. To the best of our knowledge, we introduce, for the first time, the attention mechanism only as an *auxiliary* component to the *external* of a DNN acoustic model to reduce the domain variability and improve ASR performance. Note that, similar to the domain classifier, the auxiliary attention block does not participate in the ASR decoding, so the proposed AADIT framework can be widely applied to acoustic model with any DNN architectures. AADIT can also improve the robustness in knowledge transferring for T/S learning [21, 22] as in [8]. More generally, the same methodology can be used to enhance the capability of the discriminators in any generative adversarial network [5] or gradient reversal layer network [10] for improved domain [10, 23] and speaker [24] adaptation, speech enhancement [25, 26, 27], speech synthesis [28, 29], voice conversion [30], speaker verification [9], image generation [5, 31] and translation [32], etc.

We perform AADIT to suppress speaker and environment variabilities of the DNN acoustic model and thus to improve ASR. We explore two types of local attention mechanisms: additive attention

and dot-product attention for AADIT and investigate the effect of attention window size, key/query dimension, positional encoding and multi-head attention on the ASR performance. Evaluated on CHiME-3 dataset, AADIT of DNN acoustic model achieves 13.6% and 9.3% relative word error rate (WER) improvements over the multi-conditional model and ADIT, respectively.

2. ADVERSARIAL DOMAIN-INVARIANT TRAINING

ADIT aims at reducing the variances of hidden and output unit distributions of the DNN acoustic model that are caused by the inherent inter-domain variability in the speech signal. To achieve domain-robustness, one solution is to learn a *domain-invariant* and *senone-discriminative* deep hidden feature in the DNN acoustic model through adversarial multi-task learning and make senone posterior predictions based on the learned deep feature. In order to do so, we need a sequence of speech frames $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}, \mathbf{x}_t \in \mathbb{R}^{r_x}, t = 1, \ldots, T$, a sequence of senone labels $\mathbf{Y} = \{y_1, \ldots, y_T\}$, $y_t \in \mathbb{R}$ aligned with $\mathbf{X}$ and a sequence of domain labels $\mathbf{D} = \{d_1, \ldots, d_T\}, d_t \in \mathbb{R}$ aligned with $\mathbf{X}$. We view the first few layers of the acoustic model as a feature extractor network M_f with parameters θ_f that maps input speech frames X from different domains to intermediate deep hidden features $\mathbf{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_T\}, \mathbf{f}_t \in \mathbb{R}^{r_f}$ and the upper layers of the acoustic model as a senone classifier M_y with parameters θ_y that maps the deep features $\mathbf{F}$ to the senone posteriors $p(s|\mathbf{f}_t; \theta_y), s \in \mathbb{S}$ as follows:

$$M_y(\mathbf{f}_t) = M_y(M_f(\mathbf{x}_t)) = p(s|\mathbf{x}_t; \theta_f, \theta_y). \quad (1)$$

We further introduce a domain classifier network M_d which maps the deep features $\mathbf{F}$ to the domain posteriors $p(u|\mathbf{f}_t; \theta_d)$, $u \in \mathbb{U}$ as follows:

$$M_d(M_f(\mathbf{x}_t)) = p(u|\mathbf{x}_t; \theta_f, \theta_d), \quad (2)$$

where u is one domain in the set of all domains $\mathbb{U}$. To make the deep features $\mathbf{F}$ domain-invariant, the distributions of $\mathbf{F}$ from different domains should be as close to each other as possible. Therefore, we jointly train M_f and M_d with an adversarial objective, in which θ_f is adjusted to *maximize* the frame-level domain classification loss $\mathcal{L}_{\text{domain}}$ while θ_d is adjusted to *minimize* $\mathcal{L}_{\text{domain}}$ below:

$$\mathcal{L}_{\text{domain}}(\theta_f, \theta_d) = -\frac{1}{T} \sum_{t=1}^{T} \log p(d_t|\mathbf{f}_t; \theta_d)$$
$$= -\frac{1}{T} \sum_{t=1}^{T} \sum_{u \in \mathbb{U}} \mathbb{1}[u = d_t] \log M_d(M_f(\mathbf{x}_t)), \quad (3)$$

where $\mathbb{1}[\cdot]$ is the indicator function which equals to 1 if the condition in the squared bracket is satisfied and 0 otherwise. This minimax competition will first increase the discriminativity of M_d and the domain-invariance of the deep features generated by M_f, and will eventually converge to the point where M_f generates extremely confusing deep features that M_d is unable to distinguish.

At the same time, we want to make $\mathbf{F}$ senone-discriminative by minimizing the cross-entropy senone classification loss between the predicted senone posteriors and the senone labels below:

$$\mathcal{L}_{\text{senone}}(\theta_f, \theta_y) = -\frac{1}{T} \sum_{t=1}^{T} \log p(y_t|\mathbf{x}_t; \theta_f, \theta_y)$$
$$= -\frac{1}{T} \sum_{t=1}^{T} \sum_{s \in \mathbb{S}} \mathbb{1}[s = y_t] \log M_y(M_f(\mathbf{x}_t)). \quad (4)$$

In ADIT, the acoustic model network and the condition classifier network are trained to jointly optimize the primary task of senone classification and the secondary task of domain classification with an adversarial objective function.

3. ATTENTIVE ADVERSARIAL DOMAIN-INVARIANT TRAINING

In ADIT, the mini-maximization of the domain classification loss (Eq. (3)) plays an important role in normalizing the intermediate deep feature $\mathbf{F}$ against different domains. However, the domain classification loss is still computed from a sequence of *equally-weighted* deep features. In fact, not all deep features are equally affected by domain variability and provide identical domain-discriminative information to the domain classifier. For example, deep features extracted from voiced frames are more domain-discriminative than those from the silence; deep features aligned with vowels are in general more susceptible to domain variability than those with consonants. To address this problem, we introduce an *attention mechanism* to dynamically and automatically adjust the weights for the deep features in order to put more emphasis on the domain normalization of more domain-discriminiative deep features and therefore enhance the overall domain-invariance of the deep features. The acoustic model with such a domain-invariant deep feature is expected to achieve improved ASR performance over ADIT.

In the proposed AADIT, we use soft local (time-restricted) [33, 34] self-attention [16] because it is more suitable for ASR where the input sequence consists of a relatively large number of speech frames. The local attention we adopt selectively focuses on a small window of context centered at the current time and can jointly attend different points in time with different weights. Specifically, for each deep feature $\mathbf{f}_t$ at time t, the keys are the projections of deep features in an r_a dimensional space within the attention window of size $L + R + 1$, i.e., $\mathbf{K}_t = \{\mathbf{k}_{t-L}, \ldots, \mathbf{k}_t, \ldots, \mathbf{k}_{t+R}\}$, and the query $\mathbf{q}_t$ is the projection of $\mathbf{f}_t$ in the r_a dimensional space, i.e.,

$$\mathbf{k}_t = \mathbf{W}_k \mathbf{f}_t \quad (5)$$
$$\mathbf{q}_t = \mathbf{W}_q \mathbf{f}_t, \quad (6)$$

where $\mathbf{W}_k$ is a r_a by r_f key projection matrix, $\mathbf{W}_q$ is a r_a by r_f query projection matrix and L and R are the length of left and right context, respectively in the attention window. The attention probability $\mathbf{a}_t$ of each current frame $\mathbf{f}_t$ against all the context deep features in the attention window, i.e., $\mathbf{V}_t = \{\mathbf{f}_{t-L}, \ldots, \mathbf{f}_t, \ldots, \mathbf{f}_{t+R}\}$, is computed by normalizing the similarity scores $e_{t,\tau} \in \mathbb{R}$ between the query $\mathbf{q}_t$ and each key $\mathbf{k}_\tau$ in the window $\mathbf{K}_t$, i.e.,

$$a_{t,\tau} = \frac{\exp(e_{t,\tau})}{\sum_{\tau'=t-L}^{t+R} \exp(e_{t,\tau'})}, \quad (7)$$

where $\tau = t - L, \ldots, t, \ldots, t + R$ and $a_{t,\tau} \in \mathbb{R}$ is the $[\tau - (t - L)]^{\text{th}}$ dimension of the attention probability vector $\mathbf{a}_t \in \mathbb{R}^{L+R+1}$. The similarity scores $e_{t,\tau}$ can be computed in two different ways according to the type of attention mechanism applied:

- Dot-product attention

$$e_{t,\tau} = \frac{\mathbf{k}_\tau^\top \mathbf{q}_t}{\sqrt{r_a}}, \quad (8)$$

- Additive attention

$$e_{t,\tau} = \mathbf{g}^\top \tanh(\mathbf{k}_\tau + \mathbf{q}_t + \mathbf{b}), \quad (9)$$

where $\mathbf{g} \in \mathbb{R}^{r_a}$ is a column vector, $\mathbf{b} \in \mathbb{R}^{r_a}$ is the bias column vector.

Therefore, a context vector $\mathbf{c}_t$ is formed at each time t as a weighted sum of the deep features in the attention window $\mathbf{V}_t$ with the attention vector $\mathbf{a}_t$ serving as the combination weights, i.e.,

$$\mathbf{c}_t = \sum_{\tau=t-L}^{t+R} a_{t,\tau} \mathbf{f}_\tau. \qquad (10)$$

As shown in Fig. 1, we view the entire attention process described in Eq.(5) to Eq.(10) as a single attention function $M_a(\cdot)$ with parameters $\theta_a = \{\mathbf{W}_k, \mathbf{W}_q, \mathbf{g}, \mathbf{b}\}$ which takes in the sequence of deep features $\mathbf{F}$ as the input and outputs the sequence of context vectors $\mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_T\}$, $\mathbf{c}_t \in \mathbb{R}^{r_a}$, i.e., $\mathbf{c}_t = M_a(\mathbf{f}_t)$.
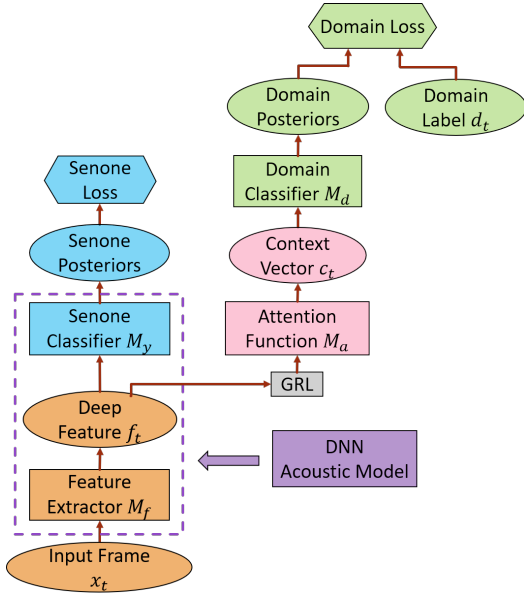

Fig. 1. The framework of attentive adversarial domain-invariant training (AADIT) of the acoustic models. Only the DNN acoustic model consisting of M_f and M_y (on the left) are used for ASR on test data. M_a and M_d are discarded after AADIT.

The domain classifier M_d then takes in the context vector $\mathbf{c}_t$ as the input to predict the frame-level domain posteriors for $u \in \mathbb{U}$.

$$p(u|\mathbf{x}_t; \theta_f, \theta_a, \theta_d) = M_d(\mathbf{c}_t) = M_d(M_a(\mathbf{f}_t))$$
$$= M_d(M_a(M_f(\mathbf{x}_t))). \qquad (11)$$

To make the deep features $\mathbf{F}$ domain-invariant, M_f, M_a and M_d are jointly trained with an adversarial objective, in which θ_f is adjusted to *maximize* the frame-level domain classification loss $\mathcal{L}_{\text{domain}}(\theta_f, \theta_a, \theta_d)$ while θ_a and θ_d are adjusted to *minimize* $\mathcal{L}_{\text{domain}}(\theta_f, \theta_a, \theta_d)$ below:

$$\mathcal{L}_{\text{domain}}(\theta_f, \theta_a, \theta_d) = -\frac{1}{T} \sum_{t=1}^{T} \log p(d_t|\mathbf{x}_t; \theta_f, \theta_a, \theta_d)$$
$$= -\frac{1}{T} \sum_{t=1}^{T} \sum_{u \in \mathbb{U}} \mathbb{1}[u = d_t] \log M_d(M_a(M_f(\mathbf{x}_t))). \quad (12)$$

In AADIT, the acoustic model network, the condition classifier network and attention function are trained to jointly optimize the primary task of senone classification and the secondary task of domain

classification with an adversarial objective function as follows

$$(\hat{\theta}_f, \hat{\theta}_y) = \underset{\theta_y, \theta_f}{\arg\min}\, \mathcal{L}_{\text{senone}}(\theta_f, \theta_y) - \lambda \mathcal{L}_{\text{domain}}(\theta_f, \hat{\theta}_a, \hat{\theta}_d), \quad (13)$$

$$(\hat{\theta}_a, \hat{\theta}_d) = \underset{\theta_a, \theta_d}{\arg\min}\, \mathcal{L}_{\text{domain}}(\hat{\theta}_f, \theta_a, \theta_d), \quad (14)$$

where λ controls the trade-off between $\mathcal{L}_{\text{senone}}$ and $\mathcal{L}_{\text{domain}}$, and $\hat{\theta}_y, \hat{\theta}_f, \hat{\theta}_a$ and $\hat{\theta}_d$ are the optimized parameters.

The parameters are updated as follows via back propagation with stochastic gradient descent:

$$\theta_f \leftarrow \theta_f - \mu \left[\frac{\partial \mathcal{L}_{\text{senone}}}{\partial \theta_f} - \lambda \frac{\partial \mathcal{L}_{\text{domain}}}{\partial \theta_f} \right], \quad (15)$$

$$\theta_a \leftarrow \theta_a - \mu \frac{\partial \mathcal{L}_{\text{domain}}}{\partial \theta_a}, \quad (16)$$

$$\theta_d \leftarrow \theta_d - \mu \frac{\partial \mathcal{L}_{\text{domain}}}{\partial \theta_d}, \quad (17)$$

$$\theta_y \leftarrow \theta_y - \mu \frac{\partial \mathcal{L}_{\text{senone}}}{\partial \theta_y}, \quad (18)$$

where μ is the learning rate. Note that the negative coefficient $-\lambda$ in Eq. (15) induces *attentive* reversal gradient that maximizes $\mathcal{L}_{\text{domain}}$ in Eq. (3) to make the deep feature domain-invariant. For easy implementation, a gradient reversal layer is introduced as in [10], which acts as an identity transform in the forward propagation and multiplies the gradient by $-\lambda$ during the backward propagation.

Note that only the optimized DNN acoustic model consisting of M_f and M_y on the left side of Fig. 1 is used for ASR on test data. The attention block M_a and domain classifer M_d (on the right) are discarded after AADIT.

We further extend the keys, queries and values with a one-hot encoding of the relative positions versus the current time in an attention window as in [34] and compute the attention vectors based on the extended representations. We also introduce a multi-head attention as in [16] by projecting the deep features H times to get H keys and queries in H different spaces. Note that the dimension of projection space for each attention head is one H^{th} of that in single-head attention to keep the number of parameters unchanged.

4. EXPERIMENTS

We perform AADIT of a multi-conditional acoustic model to suppress the speaker variability (AADIT-S) and environment variability (AADIT-E) for robust ASR.

4.1. Baseline System

As the baseline system, we first train a long short-term memory (LSTM)- hidden Markov model (HMM) acoustic model [35, 36, 37] using multi-conditional training data of CHiME-3 [38]. The CHiME-3 dataset incorporates Wall Street Journal (WSJ) corpus sentences spoken under four challenging noisy environments, i.e, on buses, in cafes, in pedestrian areas, at street junctions and one clean environment, i.e., in booth, 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 is used for decoding.

We train the baseline LSTM acoustic model with 9137 clean and 9137 noisy 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 as in [39]. The features are fed as the input to the LSTM after global mean and variance normalization. The LSTM has four 1024-dimensional hidden layers. Each hidden layer is followed by a 512-dimensional projection layer. The output layer of the LSTM has 3012 output units corresponding to 3012 senone labels. The multi-style LSTM acoustic model achieves 19.23% WER on the noisy test data.[1]

4.2. Adversarial Domain-Invariant Training

We further perform ADIT to reduce the speaker variability (ADIT-S) and environment variability (ADIT-E) of the baseline multi-conditional LSTM with 9137 noisy training utterances in CHiME-3. The M_f is initialized with the first P layers of the LSTM and M_y is initialized with the rest $(7 - P)$ hidden layers plus the output layer. P indicates the position of the deep hidden feature in the acoustic model. For ADIT-S, the speaker classifier M_d is a feedforward DNN with 3 hidden layers and 512 hidden units for each layer. The output layer of M_d has 87 units predicting the posteriors of 87 speakers in the training set. For ADIT-E, M_d is an environment classifier with the same architecture as in ADIT-S except for the 5 output units predicting 5 environments in CHiME-3. M_f, M_y and M_d are jointly trained with an adversarial multi-task objective as described in Section 2. P and λ are fixed at 4 and 0.5 in our experiments. The ADIT-S and ADIT-E LSTM acoustic models achieve 18.40% and 18.31% WER on the real test data, respectively, which are 4.3% and 4.8% relative improvements over the multi-conditional baseline.

4.3. Attentive Adversarial Domain-Invariant Training

We further perform AADIT-S and AADIT-E with the same training data as ADIT. M_f, M_y and M_d have exactly the same architectures as the ones for ADIT in Section 4.2 expect that M_d has only one hidden layer to keep the number of parameters similar as that of ADIT. P and λ are also fixed at 4 and 0.5 for all the experiments.

$L + R + 1$	15	21	25	31
WER	17.89	17.63	17.89	18.07

Table 1. The ASR WER (%) of AADIT-S with addative attention of LSTM acoustic models for different sizes of attention window $(L + R + 1)$ on real development set of CHiME-3.

We first investigate the impact of attention window size $L+R+1$ on the ASR WER via AADIT-S with additive attention in Table 1. In this work, we only use symmetric attention window with $L = R$. We fix the key and query dimension r_a at 512. The WER begins to decrease when window size is larger than 21, so we choose $L = R = 10$ for the following experiments. Then we explore the effect of different key and query dimensions r^a on the ASR performance through AADIT-S with addative attention in Table 2. The WER first decreases until reaching the minimum at $r^a = 512$ and then increases as r^a grows larger. Therefore, we fix r^a at 512 for the following experiments. Note that, by setting $r^a = 512$, the number of learnable parameters are kept roughly the same as in ADIT.

Further, we perform AADIT-S and AADIT-E with both additive and dot-product attentions and summarize the WER results for different type of domains in Table 3. We see that AADIT-S achieves

r^a	256	512	1024	2048
WER	18.53	17.63	18.01	18.19

Table 2. The ASR WER (%) of AADIT-S with addative attention of LSTM acoustic models for different dimensions of keys and values (r^a) on real development set of CHiME-3.

17.63% WER with additive attention which is 8.3% and 4.2% relatively improved over baseline multi-conditional LSTM and ADIT-S, respectively. AADIT-E performs significantly better than AADIT-S with a WER of 16.61% when using dot-product attention, which is 13.6% and 9.3% relatively improved over baseline multi-conditional model and ADIT-E, respectively. Dot-product attention performs similar to additive attention for AADIT.

System	Attention Type	Domain	
		Speaker	Environment
MC	-	19.23	19.23
ADIT	-	18.40	18.31
AADIT	AD	17.63	16.82
	DP	17.67	16.61
AADIT + PE	AD	17.57	16.68
	DP	17.37	16.94
MH AADIT	AD	17.33	17.10
	DP	17.25	16.97

Table 3. The ASR WERs (%) of multi-conditional (MC) LSTM acoustic models, ADIT, single-head AADIT, single-head AADIT with positional encoding (PE) and multi-head (MH) AADIT on real development set of CHiME-3. Both the additive (AD) and dot-product (DP) attentions are used for each AADIT system.

We also investigate the effect of positional encoding on AADIT. In Table 3, positional encoding does not consistently improve the AADIT, so we do not use it for the following experiments. We further perform AADIT with multi-head additive and dot-product attentions. The number of heads is fixed at 8 and the key/query dimension for each head is $512/8 = 64$. We observe that the multi-head AADIT-S only slightly improves the WER of single-head one, and multi-head AADIT-E does not further improve the WER. Considering the significantly better WER with less computational cost, we suggest using single-head AADIT-E for robust ASR.

5. CONCLUSIONS

We advance the domain classifier of ADIT with an attention mechanism to re-weight the deep features in a DNN acoustic model according to their importance in domain classification. With AADIT, the deep features more susceptible to domain variability are normalized with more emphasis and therefore, the overall domain-invariance of the acoustic model is greatly enhanced. The attention mechanism only serves as an auxiliary component to the external of the acoustic model that does not participate in ASR and thus can improve the DNN acoustic model with any architectures.

Evaluated on CHiME-3 dataset, the single-head AADIT achieves 13.6% and 9.3% relative WER gains over a multi-conditional LSTM acoustic model and a strong ADIT baseline, respectively. AADIT-E performs significantly better than AADIT-S. Teh additive and dot-product attentions achieve similar ASR performance. WERs of AADIT do not improve significantly with additional positional encoding and multi-head self-attention.

---

[1]Note that our experimental setup does not achieve the state-of-the-art performance on CHiME-3 (e.g., we did not perform beamforming, sequence training or use recurrent neural network language model for decoding.) since our goal is to simply verify the improved capability of AADIT in reducing inter-domain variability over ADIT.

6. REFERENCES

[1] 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.

[2] Dong Yu and Jinyu Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[3] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. INTERSPEECH*, 2011.

[4] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4, pp. 745–777, April 2014.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative adversarial nets," in *Proc. NIPS*, pp. 2672–2680. 2014.

[6] Yusuke Shinohara, "Adversarial multi-task learning of deep neural networks for robust speech recognition.," in *INTERSPEECH*, 2016, pp. 2369–2372.

[7] D. Serdyuk, K. Audhkhasi, P. Brakel, et al., "Invariant representations for noisy speech recognition," in *NIPS Workshop*, 2016.

[8] Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang Juang, "Adversarial teacher-student learning for unsupervised domain adaptation," in *Proc. ICASSP*. IEEE, 2018, pp. 5949–5953.

[9] Zhong Meng, Yong Zhao, Jinyu Li, and Yifan Gong, "Adversarial speaker verification," in *Proc. ICASSP*, 2019.

[10] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. ICML*, 2015.

[11] K. Bousmalis, G. Trigeorgis, N. Silberman, et al., "Domain separation networks," in *Proc. NIPS*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., pp. 343–351. Curran Associates, Inc., 2016.

[12] Z. Meng, J. Li, Z. Chen, et al., "Speaker-invariant training via adversarial learning," in *Proc. ICASSP*, 2018.

[13] G. Saon, G. Kurata, T. Sercu, et al., "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[14] J. Yi, J. Tao, Z. Wen, et al., "Adversarial multilingual training for low-resource speech recognition," 2018.

[15] S. Sun, C. Yeh, M. Hwang, et al., "Domain adversarial training for accented speech recognition," *arXiv preprint arXiv:1806.02786*, 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[17] K. Cho, B. Van Merriënboer, C. Gulcehre, et al., "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[18] J. Chorowski, D. Bahdanau, D. Serdyuk, et al., "Attention-based models for speech recognition," in *Proc. NIPS*, 2015, pp. 577–585.

[19] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*. IEEE, 2016, pp. 4960–4964.

[20] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[21] 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.

[22] Zhong Meng, Jinyu Li, Yong Zhao, and Yifan Gong, "Conditional teacher-student learning," in *Proc. ICASSP*, 2019.

[23] Z. Meng, Z. Chen, V. Mazalov, J. Li, and Y. Gong, "Unsupervised adaptation with domain separation networks for robust speech recognition," in *Proc. ASRU*, 2017.

[24] Zhong Meng, Jinyu Li, and Yifan Gong, "Adversarial speaker adaptation," in *Proc. ICASSP*, 2019.

[25] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "Segan: Speech enhancement generative adversarial network," in *Interspeech*, 2017.

[26] Zhong Meng, Jinyu Li, and Yifan Gong, "Cycle-consistent speech enhancement," *Interspeech*, 2018.

[27] Zhong Meng, Jinyu Li, and Yifan Gong, "Adversarial feature-mapping for speech enhancement," *Interspeech*, 2018.

[28] T. Kaneko, H. Kameoka, N. Hojo, et al., "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. ICASSP*, 2017, vol. 2017, pp. 4910–4914.

[29] W. Hsu, Y. Zhang, R. J. Weiss, et al., "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," *Proc. ICASSP*, 2019.

[30] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

[31] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[32] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[33] Minh-Thang Luong, Hieu Pham, and Christopher D Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[34] D. Povey, H. Hadian, P. Ghahremani, et al., "A time-restricted self-attention layer for asr," *ICASSP 2018*, 2018.

[35] F. Beaufays H. Sak, A. Senior, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech*, 2014.

[36] 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*. IEEE, 2017, pp. 271–275.

[37] 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.

[38] 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.

[39] 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.