INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Knowledge Distillation Approach for Efficient Internal Language Model Estimation

Zhipeng Chen[1], *Haihua Xu*[1], *Yerbolat Khassanov*[1], *Yi He*[1], *Lu Lu*[1], *Zejun Ma*[1], *Ji Wu*[2]

[1] ByteDance
[2] Department of Electronic Engineering, Tsinghua University
{chenzhipeng.speech,haihua.xu,yerb.khass,heyi.hy,lulu.0314,mazejun}@bytedance.com,
wuji_ee@tsinghua.edu.cn

Abstract

Internal language model estimation (ILME) has demonstrated its efficacy in domain adaptation for end-to-end (E2E) ASR. However, the performance improvement is achieved at the expense of computational cost, compared with conventional shallow fusion. To estimate the internal language model prior, one should run an extra forward operation on either ASR decoder or a separate density ratio (DR) language model (LM) for each decoding utterance. In this paper, we propose to employ knowledge distillation (KD) approach to realize efficient ILME for the Listen-Attend-Spell (LAS) E2E ASR model. First, we extensively explore diverse ILME and DR methods. We find that the ILM can be approximated with a DR-LM much smaller than the original ASR decoder. Furthermore, to reach the performance of ILME, we propose to employ the estimated ILM as teacher to teach a small DR-LM by KD. In this way, we achieve the best of both worlds: comparable performance to ILME and high efficiency of DR with a small DR-LM.

Index Terms: ASR, language model, ILME, density ratio, knowledge distillation, efficiency

1. Introduction

One advantage of the end-to-end (E2E) ASR modeling framework comes from the capability of joint acoustic and language model learning via a network pipeline which is normally comprised of an encoder and a decoder [1–5]. To date, E2E ASR has revealed its edge in terms of data and model scaling aspects. Given more data, as well as larger model size, E2E ASR can always push the recognition performance to higher level, approaching human speech recognition capability [6, 7].

Although encouraging, E2E ASR still has some limitations, one of them being the domain mismatch problem. Specifically, an input utterance from an unseen domain will result in sub-optimal recognition performance of E2E ASR. In such situation, one would resort to domain adaptation techniques to alleviate the performance drop [8–10].

One of the most potent domain adaptation techniques is language model (LM) fusion. There are diverse LM fusion methods, such as component fusion [11], deep fusion [12], cold fusion [13], and shallow fusion [14]. All these methods comply with a common formula implicitly or explicitly, that is, $\hat{W} = \mathrm{argmax}_W \, P_\theta(W|X) P_{\mathrm{ELM}}(W)$ where θ represents parameters of E2E ASR and $P_\theta(W|X)$ is ASR output. W is a token sequence and X is a corresponding speech sequence. $P_{\mathrm{ELM}}(W)$ comes from external LM (ELM). Despite effectiveness, the fusion results with the above-mentioned methods are sub-optimal. This is because the formula is hard to be interpreted with Bayes rule, with which the recognition formula should be $\hat{W} = \mathrm{argmax}_W \, P_\theta(X|W) P_{\mathrm{ELM}}(W)$. Therefore, to realize a Bayes-style LM fusion approach, we need to convert the posterior $P_\theta(W|X) \equiv P_\theta(X|W) P_{\mathrm{prior}}(W)$ to a likelihood, namely $P_\theta(X|W) \equiv P_\theta(W|X)/P_{\mathrm{prior}}(W)$. As we normally perform logarithmic operation in speech recognition, such a division means a subtraction operation in terms of logarithm. Consequently, we need to estimate the prior $P_{\mathrm{prior}}(W)$, and it is the output of the so-called internal language model (ILM) which is implicitly learned from training transcripts.

In order to estimate the piror, [15] proposed a hybrid autoregressive transducer (HAT) method, and [16] proposed a more general density ratio (DR) approach. In DR method, one need to estimate $P_{\mathrm{prior}}(W)$ by a separate LM (DR-LM) in addition to a normal external LM to conduct LM fusion. More recently, a bunch of works attempt to estimate the prior using the ASR decoder module instead, including encoder output zeroing-out [17], encoder output averaging [18], learning based mini-LSTM [18], and the label-synchronous context vector learning (LSCL) [19]. These decoder-based prior estimation methods are called ILM estimation (ILME). It is shown that both DR and ILME can yield notable ASR performance improvement over shallow fusion for both intra- and cross-domain adaptation tasks, and ILME usually outperforms DR [17–19]. However, the computation of ILME depends on the size of ASR decoder. If ASR decoder is large, this entails much more computational cost[1] compared with the normal shallow fusion, and it is not desirable where faster inference is decisive. Our preliminary tests show that ILME is 19% slower than shallow fusion, which is measured by throughput, i.e., the number of processed audio seconds per second for a single GPU.

In this paper, we propose to employ knowledge distillation (KD) approach to realize efficient ILME for the Listen-Attend-Spell (LAS) E2E ASR model [20]. We first compare different ILME and DR methods using both perplexity and character error rate (CER) metrics. The results show that the ILM can be approximated with a small LSTM-based DR-LM. Moreover, in order to make it reach the performance of the best ILME method, we propose to employ the ILM as teacher to teach this DR-LM via knowledge distillation [21, 22]. Eventually, our KD-based ILME approach achieves comparable ASR performance to decoder-based ILME, while the number of ILM-related parameters is reduced by 95%, which suggests more efficient inference.

The main contributions of the paper are as follows. First, we have conducted a comprehensive comparison between the ILME and DR methods. Secondly, we propose to close the performance gap between ILME and DR by KD, yielding compa-

[1] In this paper, computational cost refers to memory or CPU/GPU computing resource consumption. We believe that significant reduction of ILM-related parameters will result in less memory footprint and faster inference speed under either CPU or GPU setting.

10.21437/Interspeech.2023-2479

rable performance to ILME with much reduced computational cost. Thirdly, we have conducted a thorough analysis for the proposed method.

2. Related work

To reduce the computational cost of the ILME-based LM fusion, [23] recently proposed to employ low-order n-gram LM as DR-LM to fuse the RNN-Transducer (RNN-T) ASR model. It has been shown that a bigram DR-LM can yield significant performance improvement, compared with traditional shallow fusion. Since the decoder in [23] is a single layer LSTM module, we hypothesize that the ILM is rather weak, which has also been shown in [24], and hence a lower order bigram DR-LM is sensible. In this work, we use a stronger LSTM decoder with four layers in LAS framework. Our preliminary experiments show that traditional N-gram DR-LM is inferior to decoder-based ILME. Therefore, we employ a smaller LSTM to estimate DR-LM instead.

Another related work is [25], which proposed a residual LM approach to directly model the residual factor of external and internal LMs. As a result, computational cost in inference is hugely reduced. However, residual LM depends on ASR model, which means the trained residual LM can not be reused for other ASR models. Our method keeps the internal and external LMs independent so that the external LM can be flexibly applied to other ASR models. Moreover, for each ASR model, we only have to estimate the ILM once and then it can be combined freely with any external LMs.

3. Proposed method

3.1. Internal language model estimation

As mentioned in Section 1, the LM shallow fusion process of E2E ASR complies with the following formula:

$$\hat{W} = \arg\max_{W}[\log P_\theta(W|X) + \log P_{\text{ELM}}(W)] \tag{1}$$

For domain adaptation, an external LM fused with the ASR model is decisive when more domain-specific text data is available. However, what Equation (1) conveys cannot be interpreted with Bayes rule, which advocates using the following formula to proceed with inference instead:

$$\hat{W} = \arg\max_{W}[\log P_\theta(X|W) + \log P_{\text{ELM}}(W)] \tag{2}$$

where log posterior term $\log P_\theta(W|X)$ in Equation (1) is changed to log likelihood $\log P_\theta(X|W)$ in Equation (2). According to Bayes rule, their relation can be written as:

$$\log P_\theta(W|X) \equiv \log P_\theta(X|W) + \log P_{\texttt{prior}}(W) \tag{3}$$

where $P_{\texttt{prior}}(W)$ is implicitly learned from the transcripts during E2E ASR model training. As a result, to realize a Bayes-style LM fusion method for E2E ASR model, Equation (1) should be changed to:

$$\hat{W} = \arg\max_{W}[\log P_\theta(W|X) - \log P_{\texttt{prior}}(W) + \log P_{\text{ELM}}(W)] \tag{4}$$

In practice, the general implementation of Equation (4) is:

$$\hat{W} = \arg\max_{W}[\log P_\theta(W|X) - \lambda_{\texttt{ILM}} \log P_{\texttt{prior}}(W) + \lambda_{\text{ELM}} \log P_{\text{ELM}}(W)] \tag{5}$$

where $\lambda_{\texttt{ILM}}$ and λ_{ELM} are scaling factors which are normally tuned from the `dev` data set. Equation (5) describes an ILME-based LM fusion method, and our focus is on how to estimate the term $\log P_{\texttt{prior}}(W)$.

As introduced in Section 1, there are two categories of approaches to estimate the term $\log P_{\texttt{prior}}(W)$. The simpler one is the so-called density ratio method. The other one is to estimate the prior from the ASR decoder directly, denoted as ILME in this work. For the ILME category, we employ the encoder output zeroing-out [17] and LSCL method [19].

3.2. Reducing computational cost for ILME

As observed from Equation (5), we need to estimate the extra term $\lambda_{\texttt{ILM}} \log P_{\texttt{prior}}(W)$ for each decoding step, compared with conventional shallow fusion. Suppose such a term is estimated via an independent LM with the same size as the decoder of the ASR model. Both DR and ILME require the same extra computational cost. In this paper, we attempt to enjoy the performance gain from ILME but with the minimal computational cost. Basically, we follow density ratio framework since the DR-LM is standalone. We can change the model size of DR-LM directly to control computational cost, while that of ILME methods depends on the size of the ASR decoder.

In practice, assuming that an ILME method has already obtained desirable ASR results, we fix it and start with building an LSTM-based DR-LM with similar size as our ASR decoder (LSTM with 4-layer-1024-neuron). Then we reduce the size of the DR-LM step-by-step, and bring it down to as small as a 2-layer-256-neuron LSTM, which is much smaller than the ASR decoder. We found that such a small DR-LM can still achieve notable performance improvement over the conventional shallow fusion method but a little bit performance drop compared with the ILME method. To make such a small DR-LM achieve comparable results to ILME, we utilize the estimated ILM distribution from ILME to teach the DR-LM, i.e., employing knowledge distillation approach [21] to proceed with training loss as follows:

$$\mathcal{L}(\theta_{\text{DR}}) = \alpha\mathcal{L}_{\text{GT}}(\theta_{\text{DR}}) + (1-\alpha)\mathcal{L}_{\texttt{ILME-decoder}}(\theta_{\text{DR}}) \tag{6}$$

where $\mathcal{L}_{\text{GT}}(\theta_{\text{DR}})$ is the cross-entropy loss between DR-LM output and ground truth labels, $\mathcal{L}_{\texttt{ILME-decoder}}(\theta_{\text{DR}})$ is the cross-entropy loss between the output distributions of DR-LM and ILME, and α is a hyper-parameter balancing these two losses during training.

4. Experiments and results

4.1. Experimental setup

We conduct experiments on our in-house anonymized Mandarin speech, which is extracted from short-video covering diverse speaking styles, topics and background music/noise. All ASR models for the experiments are LAS with transformer encoder, $\{$layer, dim$\}$=$\{18, 512\}$, and LSTM decoder with different configurations.The primary one is trained with 100k hours tran-

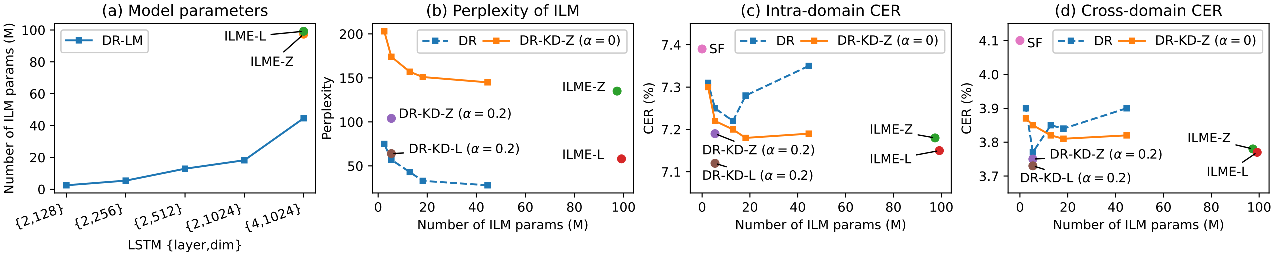

Figure 1: *Effects of ILM distillation from model parameter reduction and CER performance perspectives for both intra- and cross-domain LM fusion tasks*

scribed speech (*zh-100kh*). To verify the efficacy of our proposed method on limited-resource scenarios, we extract two subsets from the full training set and train another two ASR models, with 10k hours (*zh-10kh*) and 1k hours (*zh-1kh*) respectively. We reduce the size of decoders properly for *zh-10kh* and *zh-1kh* to avoid overfitting issue. Details of all ASR models are listed in Table 1.

Table 1: *Overview of ASR models with the same Transformer encoder configurations, {layer, dim}={18, 512}, but different decoder configurations ({layer, dim}), depending on training data amount.*

ASR Model	Train data (kh)	Decoder	#param
zh-100kh	100.0	{4, 1024}	175M
zh-10kh	10.0	{2, 1024}	142M
zh-1kh	1.0	{2, 256}	109M

For language modeling, we mix transcripts of Mandarin 100kh ASR training data and 60GB web-crawled Mandarin text to train a Long Short-Term Memory Projected (LSTMP) [26] LM, configured with {layer, dim, proj}={3, 4096, 1024} (130M parameters). The resulting LM is used as the external LM for intra-domain fusion. The corresponding intra-domain testset contains 3610 utterances (denoted as *Zh*). As for cross-domain testing, we train a separate medical domain LM with {layer, dim, proj}={2, 2048, 512} LSTMP, using the same 100kh transcripts and another 2GB medical domain text, and then evaluate the ASR performance on a medical testset containing 2407 utterances (denoted as *Med*). During DR-LM training and ILM knowledge distillation, we use only the corresponding ASR training transcripts. Distillation is performed for 60k steps using Adam optimizer [27] and a tri-stage exponential scheduler with 5e-4 peak learning rate.

4.2. Baseline results

Table 2 shows the baseline results using *zh-100kh* model. Here we consider two decoder-based ILME methods, **ILME-Z** and **ILME-L**. ILME-Z denotes the method in [17], where the attention context vector is simply zeroed out. ILME-L refers to the LSCL method [19]. We can see that both ILME-Z and ILME-L achieve the overall best performance. Our goal is to achieve performance as close to ILME as possible, but with little extra computational cost over SF. As recommended in [23], we also build a 2-gram LM with ASR training transcripts and use it as DR-LM (**DR-2gram**), but it doesn't achieve comparable results to the ILME or DR-LSTM methods[2]. Thus, we turn to approximating the ILM by a separate small neural network based LM in what follows.

Table 2: *ASR Model zh-100kh CER (%) results of diverse settings: baseline (No LM), shallow fusion (SF), LSTM-based DR (DR-LSTM), n-gram-based DR (DR-2-gram) and two ILME methods. The numbers in brackets are relative CER reduction over SF.*

CER	Zh (intra-domain)	Med (cross-domain)
No LM	7.61	5.07
SF	7.39	4.10
ILME-Z	7.18 (2.8%)	3.78 (7.8%)
ILME-L	7.15 (3.2%)	3.77 (8.0%)
DR-LSTM	7.25 (1.9%)	3.77 (8.0%)
DR-2gram	7.36 (0.4%)	3.97 (3.2%)

4.3. Internal language model distillation

Experiments in this section are conducted on ASR model *zh-100kh*. We consider LSTM-LMs with five diverse configurations. The number of parameters for each configuration, as well as ILME-Z and ILME-L, is presented in Figure 1 (a). Note that our LAS decoder is also a {4, 1024} LSTM but the number of ILM-related parameters (97.4M) is much greater than that of a separate {4, 1024} LSTM-LM (44.6M). This is because in LAS decoder, we concatenate the attention context vector (2048 dimension, projected from 512) to the output of each LSTM layer (1024 dimension) and then feed it to the next layer as input.

To begin with, we use ILME-Z, the most straightforward ILME method, as the teacher, and fix $\alpha = 0.0$ in Equation (6) to perform knowledge distillation (**DR-KD-Z**). As a comparison, we also conduct naive density ratio (**DR**) fusion with the same DR-LM configurations to see how much benefit we can obtain from distillation. As shown in Figure 1 (b) (c) (d), naive DR is able to get CER close to ILME-Z if DR-LM is relatively small ({2, 256} or {2, 512}), but degrades CER greatly when DR-LM is as large as {4, 1024}, although the perplexity is getting lower. We hypothesize that this is due to over-training that makes DR-LM a better LM but its behavior has already deviated too much from the real ILM. On the contrary, DR-KD-Z does not suffer from larger DR-LM since the learning target is an ILM (ILME-Z).

Besides, DR-KD-Z gets slightly better results than naive

[2] We also try even higher N-gram LM, but no further performance gains are obtained.

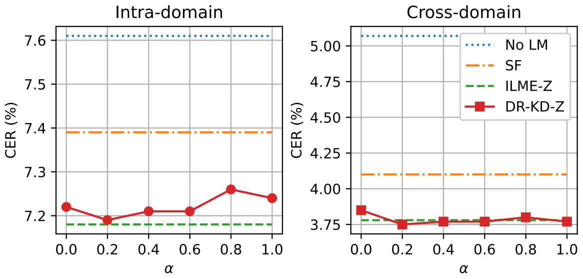

Figure 2: *CER (%) versus α in Equation (6) with {2, 256} DR-LM*

Table 3: *CER (%) results of using different ILME teachers for KD training, where the student is {2, 256} for zh-100kh and zh-10kh, {2, 128} for zh-1kh, and $\alpha = 0.2$. DR has the same DR-LM configuration as the KD student's.*

ASR Model	zh-100kh		zh-10kh		zh-1kh	
Testset	Zh	Med	Zh	Med	Zh	Med
No LM	7.61	5.07	8.98	6.87	16.93	14.90
SF	7.39	4.10	8.79	5.58	16.33	12.72
DR	7.25	3.77	8.64	4.88	15.42	10.17
ILME-Z	7.18	3.78	8.69	4.99	16.10	9.76
DR-KD-Z	7.19	3.75	8.64	4.89	15.74	9.64
ILME-L	7.15	3.77	**8.54**	4.84	**15.35**	**9.27**
DR-KD-L	**7.12**	**3.73**	**8.54**	**4.76**	15.41	9.36

DR though its perplexity is much higher than DR's, as shown in Figure 1 (b). Therefore, we conclude that the perplexity of ILM is not that critical to get a good ASR performance. What does matter is the estimated ILM should behave like what the ASR decoder has learned from the training transcripts.

In general, DR-KD-Z achieves comparable performance to ILME-Z, using a DR-LM as small as {2, 256} LSTM. The number of ILM-related parameters is reduced by 95%, from 97.4M to 5.4M. Moreover, we find it gets better results if this small DR-LM is trained with $\alpha > 0$. Figure 2 plots the CER versus α in Equation (6) for KD training on both intra- and cross-domain tasks. We can see that when the cross-entropy loss against ground truth labels is considered, the performance can be further improved. For instance, DR-KD-Z gets even better cross-domain results than the teacher ILME-Z when $\alpha = 0.2$.

4.4. Knowledge distillation with distinct teachers

In addition to *zh-100kh*, we also apply DR-KD on the models *zh-10kh* and *zh-1kh* to verify the generalization capacity of our proposed method. Moreover, we replace the teacher ILME-Z with ILME-L which has demonstrated better performance in this and prior works [19]. As we can see in Table 3, ILME-L outperforms ILME-Z, and it is verified that a better teacher does lead to a better student. Besides, the proposed DR-KD always achieves comparable results to the ILME teacher with significantly less computation for inference. Moreover, comparing with the naive DR which expends the same amount of computation, DR-KD-L achieves consistently better performance. Particularly, it obtains 8.0% relative reduction on CER in the limited-resource (*zh-1kh*) cross-domain scenario. Finally, results of *zh-100kh* are also indicated in Figure 1 for better clarity.

4.5. Robustness property

As shown in Equation (5), there are two fusion weights, namely, $\lambda_{\texttt{ILM}}$ and $\lambda_{\texttt{ELM}}$, to tune for both DR and ILME based LM fusion. Here, we are curious about how these weights affect the performance. Interestingly, Figure 3 illustrates that the naive DR is much more sensitive, while both ILME-L and the proposed DR-KD-L methods are robust, particularly to $\lambda_{\texttt{ILM}}$. From Figure 1 (b), we see that the estimated ILM of the ILME method is a much weaker LM, as the perplexity is rather higher than that of the DR-LM. We hypothesize that for ILME-L and DR-KD-L, the overall distribution across the entire token set is rather smoother, and hence it is less sensitive to the change of $\lambda_{\texttt{ILM}}$ dynamics, compared with the naive DR method of which the DR-LM is a real LM.

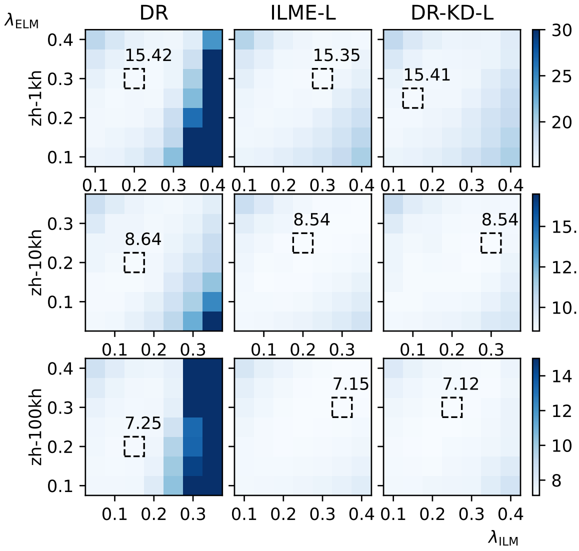

Figure 3: *Comparison of DR, ILME-L and the proposed DR-KD-L on the robustness to fusion weights, across three ASR models, where left-vertical axis refers to λ_{ELM}, bottom-horizontal axis refers to λ_{ILM}, and the bars at the side of right-vertical axis represent CER dynamics, of which darker color means higher CER.*

5. Conclusion

In this paper, we proposed a knowledge distillation approach to improve the efficiency of internal language model estimation for language model fusion based domain adaptation. In particular, we first compare recognition results between diverse ILME and density ratio based LM fusion methods, and then select ILME as teacher. In order to reduce computational cost of ILME and avoid performance drop at the same time, we let the ILME teach a smaller LSTM-based density ratio LM via knowledge distillation such that the smaller LM yields comparable results to the corresponding teacher with 95% fewer ILM-related parameters. We verified the efficacy of the proposed method on ASR models trained with 100k, 10k and 1k hours of Mandarin data, and the results demonstrated the feasibility and robustness of the proposed method on both intra- and cross-domain adaptation tasks. As future work, we will try to apply this approach to multilingual ASR and multilingual LM.

6. References

[1] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML*, vol. 32, 2014, pp. 1764–1772.

[2] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[3] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *ICASSP*. IEEE, 2016, pp. 4945–4949.

[4] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*. IEEE, 2016, pp. 4960–4964.

[5] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *Proc. of ICASSP*. IEEE, 2018.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[7] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[8] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end ASR with language model fusion," in *ICASSP*. IEEE, 2019, pp. 6096–6100.

[9] A. Mani, S. Palaskar, N. V. Meripo, S. Konam, and F. Metze, "ASR error correction and domain adaptation using machine translation," in *ICASSP*. IEEE, 2020, pp. 6344–6348.

[10] D. Hwang, A. Misra, Z. Huo, N. Siddhartha, S. Garg, D. Qiu, K. C. Sim, T. Strohman, F. Beaufays, and Y. He, "Large-scale ASR domain adaptation using self- and semi-supervised learning," in *ICASSP*. IEEE, 2022, pp. 6627–6631.

[11] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Component fusion: Learning replaceable language model component for end-to-end speech recognition system," in *Proc. ICASSP*. IEEE, 2019, pp. 5361–5635.

[12] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[13] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," in *Interspeech*. ISCA, 2018, pp. 387–391.

[14] F. Stahlberg, J. Cross, and V. Stoyanov, "Simple fusion: Return of the language model," in *WMT*. ACL, 2018, pp. 204–211.

[15] E. Variani, D. Rybach, C. Allauzen, and M. Riley, "Hybrid autoregressive transducer (hat)," in *ICASSP*. IEEE, 2020, pp. 6139–6143.

[16] E. McDermott, H. Sak, and E. Variani, "A density ratio approach to language model fusion in end-to-end automatic speech recognition," in *ASRU*. IEEE, 2019, pp. 434–441.

[17] Z. Meng, S. Parthasarathy, E. Sun, Y. Gaur, N. Kanda, L. Lu, X. Chen, R. Zhao, J. Li, and Y. Gong, "Internal language model estimation for domain-adaptive end-to-end speech recognition," in *Proc. SLT*. IEEE, 2021, pp. 243–250.

[18] M. Zeineldeen, A. Glushko, W. Michel, A. Zeyer, R. Schlüter, and H. Ney, "Investigating methods to improve language model integration for attention-based encoder-decoder ASR models," in *Interspeech*. ISCA, 2021, pp. 2856–2860.

[19] Y. Liu, R. Ma, H. Xu, Y. He, Z. Ma, and W. Zhang, "Internal language model estimation through explicit context vector learning for attention-based encoder-decoder ASR," in *Interspeech*. ISCA, 2022, pp. 1666–1670.

[20] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*. IEEE, 2016, pp. 4960–4964.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[22] Y. Shi, M.-Y. Hwang, X. Lei, and H. Sheng, "Knowledge distillation for recurrent neural network language modeling with trust regularization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7230–7234.

[23] H. Zheng, K. An, Z. Ou, C. Huang, K. Ding, and G. Wan, "An empirical study of language model integration for transducer based speech recognition," in *Interspeech*. ISCA, 2022, pp. 3904–3908.

[24] M. Ghodsi, X. Liu, J. Apfel, R. Cabrera, and E. Weinstein, "Rnn-transducer with stateless prediction network," in *Proc. of ICASSP*. IEEE, 2020.

[25] E. Tsunoo, Y. Kashiwagi, C. P. Narisetty, and S. Watanabe, "Residual language model for end-to-end speech recognition," in *Interspeech*. ISCA, 2022, pp. 3899–3903.

[26] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *INTERSPEECH*. ISCA, 2014, pp. 338–342.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.