Interspeech 2024
1-5 September 2024, Kos, Greece

Transmitted and Aggregated Self-Attention for Automatic Speech Recognition

Tian-Hao Zhang[1,2], *Xinyuan Qian*[1,2,*], *Feng Chen*[3], *Xu-Cheng Yin*[1,2]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2]USTB-EEasyTech Joint Lab of Artificial Intelligence, University of Science and Technology Beijing, Beijing 100083, China
[3]EEasy Technology Company Ltd., Zhuhai 519000, China

tianhaozhang@xs.ustb.edu.cn, xyqian@ustb.edu.cn

Abstract

Transformer based models have recently achieved outstanding progress in ASR system. The attention maps are generated in self-attention to capture temporal relationships among input tokens and heavily influence transformer performance. Many works demonstrate that attention maps of different layers incorporate various contextual scopes of information. We believe that the information from diverse attention maps is valuable and complementary. This inspires us with a novel proposal, namely Transmitted and Aggregated Self-Attention (TASA), which leverages the information of attention maps in each layer to improve the overall performance. In particular, we design Residual-TASA and Dense-TASA which are distinguished by using attention maps of the previous layer or all previous layers, respectively. Extensive experiments demonstrate that the proposed method achieving up to 10.62% CER and 7.36% WER relative reduction conducted on AISHELL-1 and LibriSpeech datasets, respectively.

Index Terms: automatic speech recognition, transformer, attention map, transmitted and aggregated self-attention

1. Introduction

Transformer based models [1, 2, 3, 4, 5] have produced superior results compared to RNNs [6, 7, 8] in the field of ASR. The critical factor is attributed to the multi-head self-attention, which generates the attention maps representing the strength of the dependency between each token in the input sequence. Thus, attention maps contain essential information that determines the ability of transformer.

Many studies have been proposed to improve attention maps for Automatic Speech Recognition (ASR). Some methods [9, 10] restrict the scope of attention maps, allowing better access to local information, but lost some important global information. In [11], Xu *et al.* directly synthesize the attention map using two feed-forward layers. Nevertheless, synthesizing attention maps with fixed parameters may not be as robust as generating them by self-attention adaptively. Conformer [12] and SAN-M [13] compensate for the lack of local information by introducing additional convolutional modules. However, they do not directly optimize the attention maps despite the increased number of parameters.

We note that in the vanilla transformer [14], the attention maps in each layer are learned independently. Raganato *et al.*[15] shows that the self-attention of the lower layer tends to learn the syntax of an utterance, while the higher layer focuses more on the semantics in machine translation tasks. A similar statement is made in [16] that the self-attention of the lower layer can capture the slight difference among consecutive frames, while the higher layer captures the salient invariant features. This reveals that the contributions of different layers are distinguished, where each attention map is decisive to make the final predictions. As proofs of concept and potential usage, we believe that the diversity of attention maps from different layers is valuable and complementary. Thus, we efficiently explore the information interaction of different attention maps to boost the ASR performance.

In this paper, we propose a novel method, namely Transmitted and Aggregated Self-Attention (TASA), to simultaneously utilize information from attention maps of different layers. Specifically, Transmitted and Aggregated Self-Attention (TASA) bridges different attention maps via a convolution-based *Transmitted Module* and then aggregates the diversity information in an *Aggregated Module*, allowing a direct access and information leverage between different layers. In this way, the general pattern of inter-token dependencies is transmitted and aggregated across all layers, facilitating the generalization ability of the multi-layer transformer. To the best of our knowledge, our study is the first attempt to exploit and integrate attention maps of different layers for ASR. And our proposal includes two implementations: Residual TASA (R-TASA) and Dense TASA (D-TASA), as indicated by the blue and yellow dashed lines in Fig. 1, respectively.

We conduct experiments on two publicly available 178-hour AISHELL-1 and 960-hour LibriSpeech datasets. On the AISHELL-1 test set, the R-TASA based and D-TASA based transformer model can achieve a relative CER reduction (CERR) of 8.41% and 10.62%, respectively, over the vanilla speech transformer. For the LibriSpeech test-clean set, our D-TASA based transformer model achieves a relative WERR of 7.36% compared to the baseline transformer. Furthermore, we integrate TASA with the stronger ASR baseline Conformer and obtain a relative WERR of 5.71% with R-TASA Conformer and 9.86% with D-TASA Conformer compared to the Conformer baseline in the AISHELL-1 dataset. Our main contributions are summarized as follows.

- We propose a novel method named TASA, which leverages the valuable and complementary information of attention maps in different layers with improved ASR performance. We design TASA with two implementations i.e., R-TASA and D-TASA which use the attention maps of the previous layer and all previous layers, respectively.
- Our proposed TASA is a pluggable module that can theoretically be used for any variants of the transformer. We validate on the Conformer architecture where the significant performance improvement sets a strong proof.
- We conduct extensive experiments on two datasets with the

* Corresponding author.

10.21437/Interspeech.2024-2374

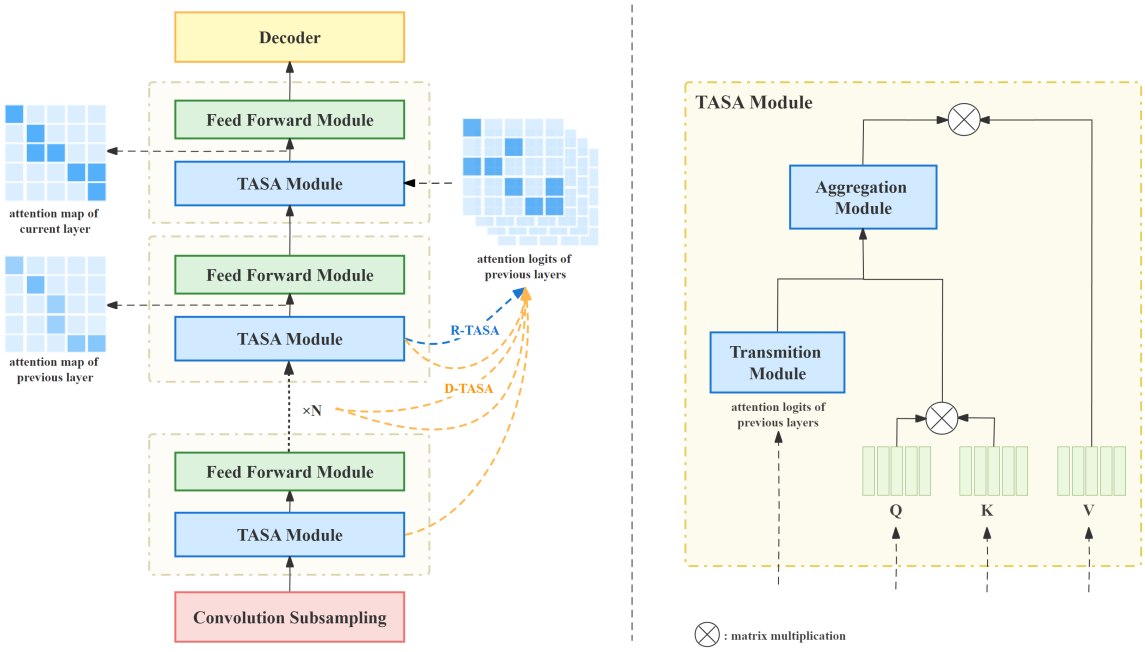


Figure 1: *TASA-Transformer model architecture. The attention logits refer to the attention maps before calculating the softmax. Residual-TASA (R-TASA) uses the attention maps of the previous layer, while Dense-TASA (D-TASA) uses the attention maps of all previous layers.*

results verifying the effectiveness of our proposal. By visualizing the generated attention maps of different layers, we observe that the attention maps of TASA exhibits strong homogeneity due to the aggregation of information from different layers.

2. Methods

2.1. Overview

Our proposed Transmitted and Aggregated Self-Attention Transformer (TASA-Transformer) is illustrated in Fig. 1. The encoder consists of a *Convolution Subsampling* layer followed by a stack of TASA-Transformer blocks. Each block contains the *TASA Module* and the same *Feed Forward Module* as the vanilla transformer. To enable the current layer to draw on the knowledge from the previous layers, the attention logits from the previous layers are transmitted to the current layer through a *Transmition Module*, which evolves the attention logits to fit the current layer better. Then, both the evolved and current attention logits are aggregated with the *Aggregation Module* in the current layer to produce more valid attention maps.

We propose two kinds of TASA-based architectures. One is the Residual-TASA (R-TASA) in the form of residual connection, which only considers the attention knowledge of the previous layer, as shown by the blue dashed line in Fig. 1. The other one is designed as the dense connection, namely Dense-TASA(D-TASA). As shown by the yellow dashed line in Fig. 1, the attention knowledge of all previous layers is transmitted and aggregated in the current layer, allowing the information of each previous layer to be accessed more directly.

2.2. Residual Transmitted and Aggregated Self-Attention

Given the input embedding sequence $\mathbf{X} \in \mathbf{R}^{\mathrm{T}} \times \mathrm{D}$, where T is the sequence length and D is the model dimension size, we first map $\mathbf{X}$ to a query sequence $\mathbf{Q}_i$ and a key sequence $\mathbf{K}_i$, respectively.

$$\mathbf{Q}_i = \mathbf{X}\mathbf{W}_i^{\mathrm{Q}},\ \mathbf{K}_i = \mathbf{X}\mathbf{W}_i^{\mathrm{K}} \tag{1}$$

where $\mathbf{W}_i^{\mathrm{Q}} \in \mathbf{R}^{\mathrm{D}\times\mathrm{D}_q}$, $\mathbf{W}_i^{\mathrm{K}} \in \mathbf{R}^{\mathrm{D}\times\mathrm{D}_k}$ are the parameters of linear projections. In order to leverage different attending representations jointly, multi-head self-attention is applied. $i \in \{1, ..., \mathrm{H}\}$ denotes generating individual query and key under i-th self-attention head. Normally, $\mathrm{D}_q = \mathrm{D}_k = \mathrm{D}/\mathrm{H}$. Then, the attention logits $\mathbf{M}$ can be computed in a dot-product way as follows.

$$\begin{aligned} \mathbf{M}_i &= \mathbf{Q}_i\mathbf{K}_i^{\top} \\ \mathbf{M} &= \mathrm{Concat}(\mathbf{M}_1, \mathbf{M}_2, ..., \mathbf{M}_h) \end{aligned} \tag{2}$$

where $\mathbf{M}_i$ is a $\mathrm{T} \times \mathrm{T}$ matrix, whose element $\alpha_{m,n}^i$ represents the attention weight of the query at position m, with the key at position n. As described previously, vanilla transformer generates the attention maps independently in each layer, which causes range-limited attention maps since there is no information sharing with other layers. We propose a convolution-based TASA module to solve this problem, which enables the current layer to draw on the inter-token correlation information from the previous layers.

Due to the multi-head attention, as described in Eq. 2, the attention logits is concatenated by the output of H attention heads, i.e., $\mathbf{M} \in \mathbf{R}^{\mathrm{H}\times\mathrm{T}\times\mathrm{T}}$. Hence, we can consider the attention logits as a $\mathrm{T} \times \mathrm{T}$ 2D-image with H channels. Based on this assumption, we use the 2D-convolution layer as the transmission module to evolve the attention logits of the previous layer.

$$\mathbf{M}_t = \mathrm{Transmit}(\mathbf{M}_{\mathrm{pre}}) = \mathrm{CNN}_t(\mathbf{M}_{\mathrm{pre}}) \tag{3}$$

where $\mathbf{M}_{\mathrm{pre}}$ is the attention logits of the previous layer. $\mathbf{M}_t$ is the transmitted attention logits whose output channel is set to H, so that the attention logits of all heads can be jointly computed and aligned with the one in current layer.

In order to adaptively fuse $\mathbf{M}_t$ and $\mathbf{M}$, we first concatenate them in the self-attention head dimension and then use the 2D-convolution layer as the aggregation module to redistribute them dynamically.

$$\begin{aligned}\mathbf{M}_a &= \mathrm{Aggregate}(\mathrm{Concat}(\mathbf{M}_t, \mathbf{M})) \\ &= \mathrm{CNN}_a(\mathrm{Concat}(\mathbf{M}_t, \mathbf{M}))\end{aligned} \tag{4}$$

The input channel and the output channel of CNN_a are set to 2H and H, respectively. Thus, $\mathbf{M}_a$ maintains the original number of self-attention heads. The convolutional kernel size of CNN_t and CNN_a are both 3×3 in the default settings of our experiments. Moreover, we do not use TASA in the first layer of the encoder. After that, the normal self-attention computation continues.

$$\mathbf{A} = \mathrm{Softmax}(\frac{\mathbf{M}_a}{\sqrt{\mathrm{D}}}) \tag{5}$$

The attention maps $\mathbf{A}$ are obtained with the softmax function, which is used to normalize $\mathbf{M}$ to be positive with a sum of one. To prevent results from entering microscopic regions, the scaled factor $\sqrt{\mathrm{D}}$ is divided before the softmax. The output feature of the TASA module is obtained via:

$$\begin{aligned}\mathbf{V}_i &= \mathbf{X}\mathbf{W}_i^{\mathrm{V}} \\ \mathbf{V} &= \mathrm{Concat}(\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_h) \\ \mathbf{O} &= \mathbf{A}\mathbf{V}\mathbf{W}^{\mathrm{O}}\end{aligned} \tag{6}$$

where $\mathbf{W}_i^{\mathrm{V}} \in \mathbf{R}^{\mathrm{D} \times \mathrm{D}_v}$ denotes the parameter of value projection, and the value $\mathbf{V}$ is concatenated by $\mathbf{V}_i$ from different heads, which is generated like query $\mathbf{Q}_i$ and key $\mathbf{K}_i$. After the dot-product of $\mathbf{A}$ and $\mathbf{V}$, the output is obtained by the linear prediction, with parameter $\mathbf{W}^{\mathrm{O}} \in \mathbf{R}^{\mathrm{D} \times \mathrm{D}}$.

Finally, the output of the block is derived after a fully connected feed-forward network, which consists of two linear layers with a ReLU activation between them.

$$\mathrm{Block}(\mathbf{X}) = \mathrm{Max}(0, \mathbf{O}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2 \tag{7}$$

where $\mathbf{W}_1 \in \mathbf{R}^{\mathrm{D} \times \mathrm{D}_f}$, $\mathbf{W}_2 \in \mathbf{R}^{\mathrm{D}_f \times \mathrm{D}}$, $\mathbf{b}_1 \in \mathbf{R}^{\mathrm{D}_f}$, $\mathbf{b}_2 \in \mathbf{R}^{\mathrm{D}}$. D_f is the feature dimensionality of the inner layer.

2.3. Dense Transmitted and Aggregated Self-Attention

To obtain the information from the previous layers more directly, D-TASA was proposed. Compared with the R-TASA, the attention logits of all previous layers are transmitted to the current layers of the D-TASA. Therefore, Eq. 3 is reformulated as:

$$\mathbf{M}_t^l = \mathrm{Transmit}(\mathbf{M}^l) = \mathrm{CNN}_t^l(\mathbf{M}^l) \tag{8}$$

where l denotes the l-th layer so that the aggregated $\mathbf{M}_a^l$ at layer l is obtained by:

$$\begin{aligned}\mathbf{M}_a^l &= \mathrm{Aggregate}(Concat(\mathbf{M}_t^1, \mathbf{M}_t^2, ..., \mathbf{M}_t^{l-1}, \mathbf{M}^l)) \\ &= \mathrm{CNN}_a^l(\mathrm{Concat}(\mathbf{M}_t^1, \mathbf{M}_t^2, ..., \mathbf{M}_t^{l-1}, \mathbf{M}^l))\end{aligned} \tag{9}$$

where the input channel of CNN_a^l is set to $l * h$. The other processes and settings remain the same as the R-TASA.

3. Experiments

We perform experiments on two publicly available datasets, AISHELL-1 [17] and LibriSpeech [21]. Both of them are divided into training, development and test sets with non-overlapping speakers. The AISHELL-1 dataset includes about 150 hours of training set, 10 hours of development set, and 5 hours of test set, respectively. The LibriSpeech dataset has approximately 1000 hours, of which the training set contains about 960 hours, the development set and the test set have about 10 hours of each.

Table 1: *Comparative experiments of different methods on the AISHELL-1 dataset. The LM indicates whether an additional language model is used during decoding.*

Model	LM	CER(%)	
		Dev	Test
TDNN-LFMMI [17]	✓	6.44	7.62
SA-T [18]	✗	8.30	9.30
LAS [19]	✓	-	8.71
ESPnet(Transformer) [20]	✓	6.00	6.70
Vanilla Transformer	✗	6.10	6.78
R-TASA-Transformer	✗	5.69	6.21
D-TASA-Transformer	✗	**5.57**	**6.06**
Vanilla Conformer	✗	5.05	5.78
R-TASA-Conformer	✗	4.96	5.45
D-TASA-Conformer	✗	**4.84**	**5.21**

3.1. Experimental Setup

For all experiments, we represent input vectors as a sequence of 80-dim log-Mel filter bank with 3-dim pitch features [22]. As usual, we set the frame length and shift to 25 ms and 10 ms sequentially. The features are normalized using Global CMVN. To promote the robustness of the training, the speed perturbation [23] and SpecAugment [24] are used before training on the AISHELL-1 dataset. Considering the training cost, only SpecAugment is used for the LibriSpeech dataset.

We build our TASA-Transformer based on ESPnet [25]. In our implementation, the encoder consists of 12 transformer blocks with the TASA module and two convolutional layers at the beginning for downsampling. For the AISHELL-1 dataset, the decoder consists of 6 conventional transformer decoder blocks. And for the LibriSpeech dataset, the decoder is the CTC decoder. For all transformer blocks, there are four heads in multi-head attention. The model dimension size d is 256, and the hidden dimension size d_f in the feed-forward network is 2048. Moreover, the Adam [26] optimizer is used with warm-up steps 25000. We train our baseline and TASA-Transformer for 50 epochs with two NVIDIA GeForce RTX 3090 GPUs on the AISHELL-1 dataset and 100 epochs on the LibriSpeech dataset. To improve the training efficiency, the batch size is determined dynamically by the input sequence length during the training process. We use the character error rate (CER) to evaluate the performance of different models.

3.2. Comparative Experiments

In Table 1, we compare the proposed TASA-Transformer with other popular methods on the AISHELL-1 dataset. Our baseline is a transformer implemented by ESPnet with all identical setups to TASA-Transformer except for the TASA module. It achieves 6.78% CER in the test set without loss of LM and CTC and performs better than TDNN-LFMMI [18] with 11.02% relative CERR. Our proposed R-TASA-Transformer significantly

Table 2: *Comparative experiments of different methods on the LibriSpeech dataset.*

Models	WER(%)			
	Dev Clean	Dev Other	Test Clean	Test Other
Vanilla Transformer	5.87	14.02	6.11	14.13
R-TASA-Transformer	5.60	13.95	5.70	13.79
D-TASA-Transformer	**5.58**	**13.58**	**5.66**	**13.59**

improved over the baseline, achieving 8.41% relative CERR. When we further enhance the acquisition of information with D-TASA-Transformer, the test set can obtain a relative CERR of 10.62%.

To further improve the performance and explore the generality of our methods, we apply TASA on conformer. As a result, R-TASA-Conformer achieves 5.71% relative CERR compared to the Conformer baseline on the AISHELL-1 test set. Moreover, The D-TASA-Conformer achieves 9.86% relative CERR, indicating that the proposed TASA module is pluggable and effectively improves the performance of transformer variants.

The experiment results on the LibriSpeech are shown in Table 2. Since our TASA module focuses on the encoder side, such for LibriSpeech dataset we used a simple CTC decoder for a fair comparison, and thus it may be a little short of the best results so far. For vocabulary, we use 5000 sub-word types based on byte-pair encoding [27], and all models use greedy search as the decoding method. Compared with the baseline model, the proposed method can achieve 7.36% relative WERR on the test-clean set and also improvement in other sets.

3.3. Ablation and Visualization Analysis

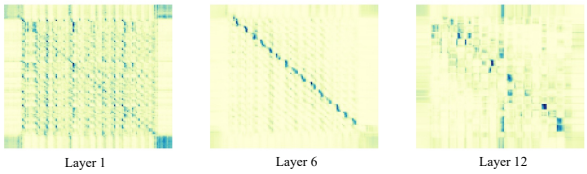
As shown in Table 3, the convolutional module is crucial for the TASA-Transformer to achieve promising performance. While removing the convolutional aggregated module and replacing it with gate fusion used in [28, 29], which uses a manually set hyperparameter α to balance the attention maps of different layers and add them together for fusion, the performance drops significantly to 6.46% CER. Note that this is the best performance by taking several different α. We suggest that the main reason for this difference is that the convolutional module enables dynamic information aggregation for different layers, so that the most valuable information from previous layers is retained.

When we remove the convolutional transmitted module and instead feed the information from the previous layer directly to the current layer, there is also a slight performance degradation for both the R-TASA-Transformer and D-TASA-Transformer. Although the performance decreases significantly when only the gate fusion is used compared to the R-TASA-Transformer, there is still an improvement compared to the vanilla transformer, which indicates that it is beneficial to leverage the information from the attention maps of different layers for accurate prediction.

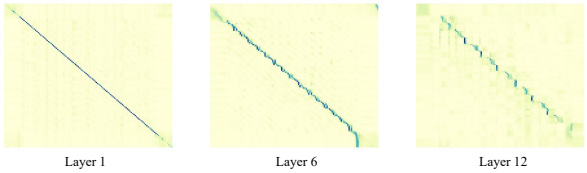
We also perform a visualization analysis of attention maps. Fig. 2(a) shows the bottom, middle, and top layers of the baseline model, respectively. Compared to the R-TASA-Transformer with the same input in Fig. 2(b), we find that attention maps in R-TASA have better consistency, which proves that the information from the previous layers indeed plays a role in the current layer. Furthermore, we notice that the bottom layer of the vanilla transformer does not successfully capture the dependencies between tokens as its attention map is very messy. In contrast, the bottom layer of the R-TASA-Transformer enables each token to focus on its local information, providing a better basis for the attention map prediction in the succeeding layers. We consider this because the parameters of the bottom layer can be optimized directly in R-TASA through the transmission module, thus better learning contextual relations of the input audio. The attention maps of D-TASA have similar visualizations as R-TASA. To avoid redundancy, we don't elaborate them here.

Table 3: *Ablation study of the convolution-based cross-layer connection module. ΔParams means the increase parameter number.*

Models	CER(%)	ΔParams(K)
Vanilla Transformer	6.78	**0**
R-TASA-Transformer	6.21	5.26
w/o aggregated conv	-	-
w/ gate fusion	6.46	1.78
w/o transmitted conv	6.23	3.50
w/ gate fusion	6.62	**0**
D-TASA-Transformer	**6.06**	20.90
w/o transmitted conv	6.11	11.13



(a) Attention maps of vanilla transformer of encoder layers 1, 6, and 12.



(b) Attention maps of R-TASA-Transformer of encoder layers 1, 6, and 12.

Figure 2: *Attention maps visualization for vanilla transformer and R-TASA-Transformer.*

4. Conclusion

In this work, we propose a novel TASA method to leverage the complementary information of attention maps in each layer. Our proposed method transmits and aggregates attention maps from different layers to enable direct access to the information with improved ASR performance. Experimental results show that we achieve at most a 10.62% relative CERR compared to the baseline model on the AISHELL-1 dataset and 7.36% relative WERR on the LibriSpeech dataset. From the visualization results, we can prove the effectiveness of the proposal. For future work, we plan to further explore the TASA mechanism to the decoding process and other transformer variants.

Acknowledgements: The research is supported by National Science and Technology Major Project (2020AAA0109701), National Science Fund for Distinguished Young Scholars(62125601), National Natural Science Foundation of China (62076024).

5. References

[1] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2018, pp. 5884–5888.

[2] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese," in *19th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 791–795.

[3] N. Pham, T. Nguyen, J. Niehues, M. Müller, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," in *20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 66–70.

[4] N. Moritz, T. Hori, and J. L. Roux, "Streaming automatic speech recognition with the transformer model," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2020, pp. 6074–6078.

[5] C.-F. Zhang, Y. Liu, T.-H. Zhang, F. Chen, and X.-C. Yin, "Non-autoregressive transformer with unified bidirectional decoder for automatic speech recognition," *arXiv preprint arXiv:2109.06684*, 2021.

[6] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Annual Conference on Neural Information Processing Systems, NIPS*, 2015, pp. 577–585.

[7] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, J. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *33rd International Conference on Machine Learning, ICML*, 2016, pp. 173–182.

[8] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2017, pp. 4835–4839.

[9] P. Daniel, H. Hossein, G. Pegah, K. Li *et al.*, "A time-restricted self-attention layer for asr," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2018, pp. 5874–5878.

[10] C. Liang, M. Xu, and X.-L. Zhang, "Transformer-based end-to-end speech recognition with residual gaussian-based self-attention," *arXiv preprint arXiv:2103.15722*, 2021.

[11] M. Xu, S. Li, and X. Zhang, "Transformer-based end-to-end speech recognition with local dense synthesizer attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2021, pp. 5899–5903.

[12] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *21th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 5036–5040.

[13] Z. Gao, S. Zhang, M. Lei, and I. McLoughlin, "SAN-M: memory equipped self-attention for end-to-end speech recognition," in *21th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 6–10.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.*, "Attention is all you need," in *Annual Conference on Neural Information Processing Systems, NIPS*, 2017, pp. 5998–6008.

[15] A. Raganato and J. Tiedemann, "An analysis of encoder representations in transformer-based machine translation," in *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP*, 2018, pp. 287–297.

[16] Y. Shi, Y. Wang, C. Wu, C. Fuegen, F. Zhang *et al.*, "Weak-attention suppression for transformer based speech recognition," in *21th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 4996–5000.

[17] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline," in *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, O-COCOSDA*, 2017, pp. 1–5.

[18] Z. Tian, J. Yi, J. Tao, Y. Bai, and Z. Wen, "Self-attention transducers for end-to-end speech recognition," in *20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 4395–4399.

[19] C. Shan, C. Weng, G. Wang, D. Su, M. Luo *et al.*, "Component fusion: Learning replaceable language model component for end-to-end speech recognition system," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2019, pp. 5631–5635.

[20] S. Karita, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang *et al.*, "A comparative study on transformer vs RNN in speech applications," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU*, 2019, pp. 449–456.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2015, pp. 5206–5210.

[22] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal *et al.*, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2014, pp. 2494–2498.

[23] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *16th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 3586–3589.

[24] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph *et al.*, "Specaugment: A simple data augmentation method for automatic speech recognition," in *20th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 2613–2617.

[25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, Y. Unno *et al.*, "Espnet: End-to-end speech processing toolkit," in *19th Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 2207–2211.

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

[27] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*, 2016.

[28] J. B. M. Z. Yujing Wang, Yaming Yang *et al.*, "Predictive attention transformer: Improving transformer with attention map prediction," *https://openreview.net/pdf?id=YQVjbJPnPc9*, 2021.

[29] Y. Wang, Y. Yang, J. Bai, M. Zhang, J. Bai *et al.*, "Evolving attention with residual convolutions," in *Proceedings of the 38th International Conference on Machine Learning, ICML*, M. Meila and T. Zhang, Eds., vol. 139, 2021, pp. 10 971–10 980.