INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

SAN-M: Memory Equipped Self-Attention for End-to-End Speech Recognition

Zhifu Gao[1], *Shiliang Zhang*[1], *Ming Lei*[1], *Ian McLoughlin*[2]

[1]Speech Lab, Alibaba DAMO Academy
[2]ICT Cluster, Singapore Institute of Technology
{zhifu.gzf, sly.zsl, lm86501}@alibaba-inc.com, ian.mcloughlin@singaporetech.edu.sg

Abstract

End-to-end speech recognition has become popular in recent years, since it can integrate the acoustic, pronunciation and language models into a single neural network. Among end-to-end approaches, attention-based methods have emerged as being superior. For example, *Transformer*, which adopts an encoder-decoder architecture. The key improvement introduced by Transformer is the utilization of self-attention instead of recurrent mechanisms, enabling both encoder and decoder to capture long-range dependencies with lower computational complexity. In this work, we propose boosting the self-attention ability with a DFSMN memory block, forming the proposed memory equipped self-attention (SAN-M) mechanism. Theoretical and empirical comparisons have been made to demonstrate the relevancy and complementarity between self-attention and the DFSMN memory block. Furthermore, the proposed SAN-M provides an efficient mechanism to integrate these two modules. We have evaluated our approach on the public AISHELL-1 benchmark and an industrial-level 20,000-hour Mandarin speech recognition task. On both tasks, SAN-M systems achieved much better performance than the self-attention based *Transformer* baseline system. Specially, it can achieve a CER of 6.46% on the AISHELL-1 task even without using any external LM, comfortably outperforming other state-of-the-art systems.

Index Terms: speech recognition, end-to-end, attentional model, Transformer, san-m

1. Introduction

Conventional automatic speech recognition (ASR) systems usually adopt the hybrid architecture [1], which consists of separate acoustic, pronunciation and language models (AM, PM, LM). Recently, so-called *end-to-end* (E2E) approaches have rapidly gained prominence in the speech recognition community. End-to-end ASR systems fold the AM, PM and LM into a single neural network that dramatically simplifies the training and decoding pipelines. Two popular approaches for this are neural networks with Connectionist Temporal Classification (CTC) -like criteria [2, 3] and attention-based models [4, 5]. The CTC-based approach has demonstrated its superiority over hybrid architecture, however, it requires an external LM for good performance [6, 7]. Unlike CTC-based approaches, attention-based models generate character sequences without any unreasonable independence assumption between characters, which enables it to effectively learn an implicit language model.

A typical attention-based model could be divided into two main components; an encoder and a decoder, which are jointly trained towards maximizing the likelihood of target sequences generated from acoustic feature sequences. In early works [4,8], long short-term memory neural networks (LSTMs) were widely used to model long-term dependencies among

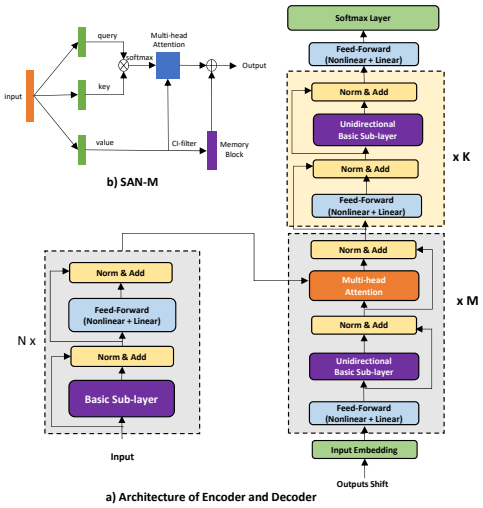

Figure 1: *Illustration of: a) the architecture of encoder and decoder. b) the SAN-M architecture (top left).*

acoustic features in the encoder and output sequences in the decoder. The attention module inside the decoder interacts between the output representations of the encoder and the hidden states of the decoder, to compute context vectors. LSTM-type networks have a strong ability to capture long-term dependencies within the sequential data using the mechanism of recurrent feedback. However, they suffer from the high computational complexity and a 'painful' training process, *i.e.*, gradient vanishing [9]. Therefore, many authors have been inspired to search for more computationally-efficient and flexible architectures for sequential modeling.

In the past few years, some efficient models, *e.g.*, convolutional neural networks [10] and time-delay neural networks [11], have been employed to improve the training process. Specially, Zhang *et al*. proposed a deep feed-forward sequential memory network (DFSMN) to replace LSTM in hybrid architectures [12, 13] and in CTC-based models [14, 15]. More recently, Transformer has become popular in seq2seq tasks, *e.g.*, neural machine translation [16], ASR [17–20], and has shown very promising performance. The key improvement is the utilization of self-attention instead of recurrent models, *e.g.*, LSTM, to model feature sequences in both encoder and decoder. This enhances the ability to capture long-range dependencies with lower computational complexity and to enable more parallelizable training.

Both self-attention and DFSMN memory blocks was proposed to replace LSTM for sequential modeling. Self-attention has powerful long-term dependency modeling abilities inside the full sequence [16]. Unlike self-attention, a single DFSMN memory block layer was designed to model local-term depen-

http://dx.doi.org/10.21437/Interspeech.2020-2471

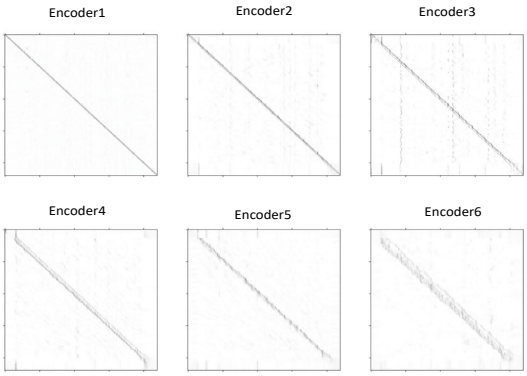

Figure 2: *Self-attention image maps from different encoder layers for a given sequence.*

dencies, with the long-term contexts captured by stacking multiple layers [12]. To some extent, the self-attention and DFSMN memory block seems complementary for each other. Thus, in this work, we aim to design a new structure that exploits the complementarity between self-attention and DFSMN memory blocks, under the framework of attention-based models. Firstly, we have made theoretical and empirical comparisons between self-attention and DFSMN memory blocks. Secondly, we have designed a new structure called memory equipped self-attention (SAN-M) to effectively combine the strength of both. You *et al*. proposed inserting self-attention layers into DFSMN for hybrid architectures [21]. In contrast, we propose incorporating these into E2E ASR models. Furthermore, the proposed SAN-M combines them both within a single basic sub-layer, in deep-fusion fashion.

We report extensive experiments on the public AISHELL-1 benchmark and an industrial-level 20,000-hour Mandarin speech recognition task. On both tasks, SAN-M based systems achieve much better performance than the self-attention based *Transformer* baseline system. Specially, achieving 6.46% CER on AISHELL-1 even without an external LM, which is the best performance on this task to date (shown later in Table.2).

2. The proposed methods

2.1. Overview

Transformer was first proposed for neural machine translation [16], where it obtained state-of-the-art results on many tasks. It was then introduced into speech processing tasks, *e.g.*, ASR [17, 18] and text-to-speech [22].

As shown in Fig. 1 a), our network follows the overall architecture of Transformer [16], which consists of an encoder and a decoder. The former maps an input sequence $\mathbf{X}$ to a sequence of hidden representations $\mathbf{Z}$ and consists of N blocks of basic sub-layer and feed-forward sub-layer. The decoder, meanwhile, generates one element of output sequence $\mathbf{Y}$ at each time step, consuming representations $\mathbf{Z}$. As an auto-regressive decoder, it consumes the previously produced characters as additional inputs when producing the next character at each step [23]. It consists of three components. The first components is M blocks which each consist of a feed-forward sub-layer, a unidirectional basic sub-layer and a multi-head attention sub-layer. Then K blocks which each comprise a feed-forward and a unidirectional basic sub-layer. The last component is a single feed-forward sub-layer to output characters.

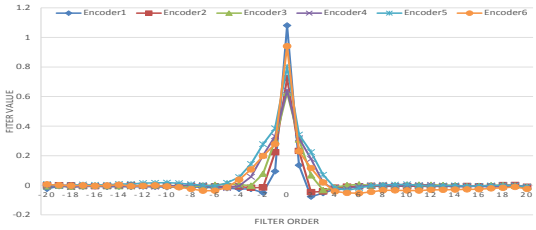

Figure 3: *Visualization of the learned filters(averaged by filter order) of DFSMN memory blocks in different encoder layers.*

In this paper, we will firstly give a brief review of self-attention and DFSMN memory block. Then we will incorporate them into the (unidirectional) basic sub-layer respectively. Theoretical analysis and empirical result comparisons will determine the strength of this approach. Furthermore, we present the proposed new structure, memory equipped self-attention (SAN-M) as the (unidirectional) basic sub-layer to effectively combine the strength of self-attention and DFSMN.

2.2. Multi-Head Attention

Multi-head attention was proposed to jointly attend information from different representation subspaces at different positions [16]. It could be formulated as:

$$\mathrm{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = [\mathrm{head}_1,...,\mathrm{head}_\mathrm{h}]\mathbf{W}^O \quad (1)$$

$$\mathrm{head}_\mathrm{i} = \mathrm{Attention}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) \quad (2)$$

$$(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) = (\mathbf{H}\mathbf{W}_i^Q,\mathbf{X}\mathbf{W}_i^K,\mathbf{X}\mathbf{W}_i^V) \quad (3)$$

Where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ are queries, keys and values respectively. The projections are parameter matrices $\mathbf{W}_i^Q \in \mathbb{R}^{d_{model}\times d_k}$, $\mathbf{W}_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $\mathbf{W}_i^V \in \mathbb{R}^{d_{model}\times d_v}$ and $\mathbf{W}^O \in \mathbb{R}^{hd_v\times d_{model}}$. h is the number of heads, d_{model} is the model dimension and d_k is the key dimension. $\mathbf{X} \in \mathbb{R}^{T\times d_{model}}$ and $\mathbf{H} \in \mathbb{R}^{T'\times d_{model}}$ are the inputs. For each head, "scaled dot-product attention" [16] was adopted as the attention mechanism. Given that, the outputs are formulated as:

$$\mathrm{Attention}(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i) = \mathrm{softmax}\left\{\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_k}}\right\}\mathbf{V}_i \quad (4)$$

2.3. Memory Block

DFSMN [13] improved on the FSMN architecture by introducing skip connections and memory strides. It consists of three components: a linear projection, a memory unit and a weight connection from memory unit to the next hidden sub-layer. The key elements in DFSMN are the learnable FIR-like memory blocks, which are used to encode long-context information into a fixed-size representation. As a result, DFSMN is able to model long-term dependencies in sequential data without using recurrent feedback. The operation in the l-th memory block takes the following form:

$$\mathbf{h}_t^\ell = \max(\mathbf{W}^\ell\mathbf{m}_t^{\ell-1} + \mathbf{b}_t^\ell, 0) \quad (5)$$

$$\mathbf{p}_t^\ell = \mathbf{V}_t^\ell\mathbf{h}_t^\ell + \mathbf{v}_t^\ell \quad (6)$$

$$\mathbf{m}_t^\ell = \mathbf{m}_t^{\ell-1} + \mathbf{p}_t^\ell + \sum_{i=0}^{N_1^\ell}\mathbf{a}_i^\ell \odot \mathbf{p}_{t-s_1*i}^\ell + \sum_{j=1}^{N_2^\ell}\mathbf{c}_j^\ell \odot \mathbf{p}_{t+s_2*j}^\ell \quad (7)$$

$$\mathbf{M}^\ell = [\mathbf{m}_1^\ell, \mathbf{m}_2^\ell, ..., \mathbf{m}_T^\ell] \quad (8)$$

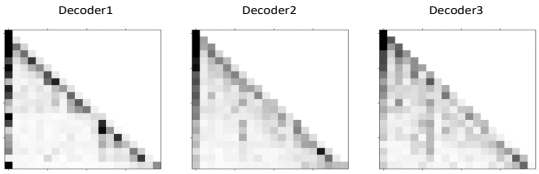

Figure 4: *Image maps from three decoder layers to illustrate self-attention from a given sequence.*

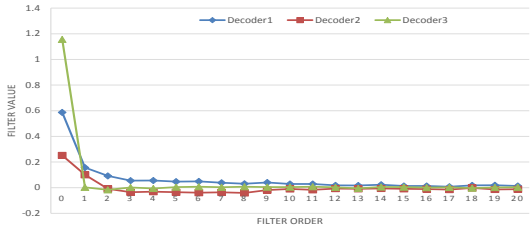

Figure 5: *Visualization of the learned filters(averaged by filter order) in DFSMN memory blocks from different decoder layers(mirrored).*

Here, $\mathbf{M}^\ell$ is the memory block. $\mathbf{h}_t^\ell$ and $\mathbf{p}_t^\ell$ denote the outputs of the ReLU layer and linear projection layer respectively. $\mathbf{m}_t^\ell$ denotes the output of the ℓ-th memory block. N_1^ℓ and N_2^ℓ denote the look-back and lookahead order of the ℓ-th memory block, respectively, while s_1 and s_2 are their respective stride factors.

2.4. Comparing Self-Attention and Memory Blocks

In this section, we will make an in-depth comparison between self-attention and DFSMN memory blocks. Self-attention is an attention mechanism where the $queries$, $keys$ and $values$ are from the same sequence in Eq. (4). Then the attention vector $\mathbf{c}_t$ is calculated as:

$$\mathbf{c}_t = \sum_{j=0}^{T} \alpha_{t,j}\mathbf{h}_j = \sum_{i=0}^{t} \alpha_{t,i}\mathbf{h}_i + \sum_{j=t+1}^{T} \alpha_{t,j}\mathbf{h}_j \qquad (9)$$

Where $\boldsymbol{\alpha}_t = (\alpha_{t,1}, ..., \alpha_{t,T})$ are the weights of self-attention from a sequence at time t. In terms of formulation, Eq. (9) is similar to the scalar FSMN memory block proposed in [12]. If we take multi-head attention into consideration, Eq. (1) is similar to the vectorized FSMN memory block defined in Eq. (8). To summarise, the outputs of both DFSMN memory block and self-attention are computed by weighting and then summing the feature vectors. The important difference is how to derive the weights.

As for self-attention, weights are calculated dynamically depending on the features themselves, which could be viewed as context-dependent (CD) coefficients. This could learn time dependencies inside the full sequence. However, it may not be efficient since it must compute every time pair in the full sequence. The computational complexity is thus $\boldsymbol{O}(n^2 \cdot d)$ [1].

In terms of DFSMN memory block, weights are context-independent (CI) coefficients, and we could view this as learning the statistical average distribution of the whole dataset. As defined in Eq. (7), the range of context dependencies is controlled by N_1^ℓ and N_2^ℓ, which means it is more computationally efficient and flexible. The computational complexity is $\boldsymbol{O}((N_1^\ell + N_2^\ell) \cdot n \cdot d)$. Though the receptive field of a single layer is small, it can still model long-range dependencies by stacking multiple layers.

[1] n and d are the length and dimension of a sequence respectively.

Table 1: *Performance comparison of three basic sub-layer types on AISHELL-1.*

Encoder	Decoder	Parameter(M)	CER(%)	
			Dev	Test
SAN	SAN	46	6.58	7.33
DFSMN	DFSMN	37	5.92	6.81
SAN-M	DFSMN	43	5.74	6.46

Table 2: *State-of-the-art comparison on AISHELL-1.*

Model	E2E	LM	CER(%)	
			Dev	Test
TDNN-LFMMI [24]	N	Y	6.44	7.62
SA-T [19]	Y	N	8.30	9.30
LAS [25]	Y	Y	-	8.71
Joint CTC/attention [26]	Y	Y	6.00	6.70
Proposed **SAN-M**	Y	N	**5.74**	**6.46**

We plot the CD-coefficients of self-attention in different encoder layers for a given sequence in Fig. 2. A strong diagonal component is evident, which grows more diffuse and wider as we progress through deeper layers. This reveals that learned features are mainly locally dependent, even though self-attention is able to model long-term dependencies over the full sequence. In Fig. 3, showing the CI-coefficients of DFSMN memory for the same encoder blocks, we see a shape resembling a tower whose width increases as we progress through deeper layers. For comparison, we also plot the self-attention matrix weights and DFSMN memory block vectors for the same decoder layers in Figs. 4 and 5. These show that self-attention has learned much longer-range dependencies than the DFSMN memory block. Our investigations have found that in practice, self attention for acoustic features in the encoder is often dominated by short-term dependencies. Consequently, it may therefore not be effectively capturing longer-term dependencies.

From the discussion above, we can briefly summarise: (a) Self-attention has the ability to learn long-range dependencies inside the full sequence, yet the learned features are not necessarily always long-term dependent, particularly in the encoder. (b) DFSMN memory blocks tend to learn local dependencies. Meanwhile they are more computationally efficient, and flexible, than self-attention. (c) While self-attention learns long-term context dependencies focusing on single features, DFSMN memory blocks learn local-term dependencies from the statistical average distribution over the whole dataset, meaning that they may well be more robust in practice.

2.5. Memory Equipped Self-Attention

From Section 2.4, we found that self-attention tends to learn CD-coefficients within a single feature whereas DFSMN memory blocks tend to learn CI-coefficients from the statistical average distribution of whole dataset. We think that the two structures might therefore be complementary to each other. Following that insight, we designed memory equipped self-attention (SAN-M) to combine the strengths of both approaches. As shown in Fig. 1 b), a DFSMN filter has been added on the $values$ inside the $Multi\text{-}Head\ Attention$ to output memory

Table 3: *Comparison of models on the 20000-hour Mandarin speech recognition task.*

Models	CTC1	CTC2	EXP1	EXP2	EXP3	EXP4	EXP5
Encoder	DFSMN	DFSMN	SAN	DFSMN	SAN-M	SAN-M	SAN-M
Decoder	-	-	SAN	DFSMN	DFSMN	DFSMN	DFSMN
d_{basic} - d_{ffn}	-	-	512-2048	512-2048	512-2048	256-1024	320-1280
N	-	-	10	10	10	40	40
M	-	-	6	6	6	6	6
K	-	-	0	0	0	6	6
Parameter (M)	25	45	59	47	55	42	63
Common Set (CER%)	11.6	9.9	9.8	10.2	9.4	9.0	8.3
Far-field Set (CER%)	20.3	17.7	15.0	16.7	14.3	13.7	12.5

block. The memory content is then added to the output of the $Multi\text{-}Head\ Attention$, which could be formulated as:

$$\mathbf{Y} = \mathrm{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \mathbf{M}(\mathbf{V}) \tag{10}$$

Where $\mathbf{Y}$ denotes the output of SAN-M. Unidirectional SAN-M means that both self-attention and DFSMN memory blocks themselves are unidirectional.

3. Experiments

3.1. Experimental Setup

We conduct extensive experiments to evaluate the performance of self-attention, DFSMN memory block and the combined SAN-M on Mandarin speech recognition tasks. We report results on the 170-hour AISHELL-1 released in [24] and an industrial-level 20000-hour-task described in [15], collected from multiple domains including news, sport, tourism, game, literature, education etc. It is divided into a training set and a development set in the ratio of 95% to 5%. A far-field set consisting of about 15 hours data, and a common set consisting of about 30 hours data, are used to evaluate the performance. Acoustic features are 80-dimensional energy-based log-mel filter-banks (FBK), computed on a window of 25ms with 10ms shift. A low frame rate (LFR) is made by stacking consecutive frames into a size 7 context window (3+1+3) and then down-sampling the input frame rate to 60ms. Acoustic modeling units are Chinese characters, which are 4233 and 9000 for AISHELL-1 and the 20,000-hour tasks respectively. For the E2E system, all models are trained to output characters directly, without using any external LM.

All E2E experiments are conducted with the OpenNMT [27] toolkit. We adopt the LazyAdamOptimizer with $\beta_1 = 0.9$, $\beta_2 = 0.998$, and a *noam_decay_v2* learning rate strategy with $d = 512, warmup\_n = 8000$, and $k = 1$ [16]. Label smoothing and dropout regularization of 0.1 are included to prevent over-fitting.

3.2. AISHELL-1 Task

We first evaluate the performance on AISHELL-1. For all system, we set $N = 6, M = 3, K = 0$. The basic sub-layer output dimension, denoted d_{basic}, and feed-forward sub-layer d_{ffn}, are set to 512 and 2048 respectively. SpecAugment [28] is employed to augment the dataset.

From Table 1, we see that incorporating SAN-M in a basic sub-layer obtains the best performance, compared to self-attention and DFSMN memory block. Specially, SAN-M achieves 11.8% relative improvement over self-attention. From the results, it is clear that incorporating DFSMN memory blocks can boost the performance of self-attention.

We also compared the proposed SAN-M with other the popular systems in Table 2. The "LM" column denote whether an external LM is added when decoding. TDNN-LFMMI is a popular baseline reported by the dataset releaser [24]. SA-T was proposed to replace the RNN with self-attention in RNN-T to obtain a performance improvements [19]. LAS extended the attention-based model with an LM when decoding [25]. Shigeki *et al.* [26] proposed jointly training CTC and attention-based models to achieve state-of-the-art performance. Yet the proposed SAN-m system obtained slightly better performance even without using an external LM (and being more elegant).

3.3. 20,000-hour Tasks

We extend our experiments to evaluate on the 20,000-hour dataset. The configuration of different systems and their results are shown in Table 3. For the CTC-based systems [15], we trained two DFSMN-CTC-sMBR systems with 10 and 20 DFSMN-layers, denoted CTC1 and CTC2 respectively. d_{basic} and d_{ffn} are the same as described in Section 3.2.

Let us first compare EXP1 and EXP2. The $Common\ Set$ mainly contains near-field short duration records, and the DFSMN memory block shows comparable performance with self-attention on this task. $Far\text{-}field$, which mainly contains long-duration records, highlights the superiority of self-attention at long-distance modeling.

Now comparing EXP 1 and EXP 3, we see performance improves in the system with fewer parameters, in accord with Section 3.2. This confirms that self-attention and DFSMN memory blocks are complementary, and SAN-M is able to effectively combine their strengths. When we further explore configurations, we find that 'thinner' and 'deeper' structures achieve more performance improvements, as shown in EXP 4 and 5. Compared to the EXP1 baseline, EXP5 obtains 15.3% and 17% relative improvements on $Common\ Set$ and $Far\text{-}field$ tasks respectively, yet only increases the model size by less than 7%.

4. Conclusions

In this work, we proposed memory equipped self-attention (SAN-M) to combine the strength of self-attention and DFSMN memory blocks for end-to-end speech recognition. Our theoretical analysis and empirical comparisons concur in demonstrating the complementarity of the techniques. This is confirmed by extensive experiments on two Mandarin ASR tasks. On the AISHELL-1 task, SAN-M obtains a 11.8% relative improvement and matches other state-of-the-art systems yet does not require an external LM. Meanwhile on a 20,000-hours Mandarin ASR task, SAN-M outperforms the self-attention based Transformer baseline by over 10%.

5. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.

[2] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[3] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[4] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[5] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*. IEEE, 2017, pp. 4835–4839.

[6] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[7] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," *arXiv preprint arXiv:1507.06947*, 2015.

[8] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[9] Y. Bengio, P. Simard, P. Frasconi *et al.*, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[10] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[11] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[12] S. Zhang, C. Liu, H. Jiang, S. Wei, L. Dai, and Y. Hu, "Feedforward sequential memory networks: A new structure to learn long-term dependency," *arXiv preprint arXiv:1512.08301*, 2015.

[13] S. Zhang, L. Ming, Z. Yan, and L. Dai, "Deep-FSMN for large vocabulary continuous speech recognition," pp. 5869–5873, 2018.

[14] S. Zhang and M. Lei, "Acoustic modeling with DFSMN-CTC and joint CTC-CE learning." in *Interspeech*, 2018, pp. 771–775.

[15] S. Zhang, M. Lei, Y. Liu, and W. Li, "Investigation of modeling units for mandarin speech recognition using DFSMN-CTC-sMBR," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7085–7089.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[17] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Muller, and A. Waibel, "Very deep self-attention networks for end-to-end speech recognition," *arXiv preprint arXiv:1904.13377*, 2019.

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

[19] Z. Tian, J. Yi, J. Tao, Y. Bai, and Z. Wen, "Self-attention transducers for end-to-end speech recognition," *Proc. Interspeech 2019*, pp. 4395–4399, 2019.

[20] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. Mcdermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," *arXiv preprint arXiv:2002.02562*, 2020.

[21] Z. You, D. Su, J. Chen, C. Weng, and D. Yu, "DFSMN-SAN with persistent memory model for automatic speech recognition," *arXiv preprint arXiv:1910.13282*, 2019.

[22] S. M. Lakew, M. Cettolo, and M. Federico, "A comparison of transformer and recurrent neural networks on multilingual neural machine translation," *arXiv preprint arXiv:1806.06957*, 2018.

[23] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[24] 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 *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[25] 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 *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5361–5635.

[26] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang *et al.*, "A comparative study on Transformer vs RNN in speech applications," *arXiv preprint arXiv:1909.06317*, 2019.

[27] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "Opennmt: Open-source toolkit for neural machine translation," *arXiv preprint arXiv:1701.02810*, 2017.

[28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.