Dynamic Encoder Transducer: A Flexible Solution For Trading Off Accuracy For Latency

Yangyang Shi, Varun Nagaraja, Chunyang Wu, Jay Mahadeokar, Duc Le, Rohit Prabhavalkar,
Alex Xiao, Ching-Feng Yeh, Julian Chan, Christian Fuegen, Ozlem Kalinli, Michael L. Seltzer

Facebook AI, USA

yyshi@fb.com

Abstract

We propose a dynamic encoder transducer (DET) for on-device speech recognition. One DET model scales to multiple devices with different computation capacities without retraining or finetuning. To trading off accuracy and latency, DET assigns different encoders to decode different parts of an utterance. We apply and compare the layer dropout and the collaborative learning for DET training. The layer dropout method that randomly drops out encoder layers in the training phase, can do on-demand layer dropout in decoding. Collaborative learning jointly trains multiple encoders with different depths in one single model. Experiment results on Librispeech and in-house data show that DET provides a flexible accuracy and latency trade-off. Results on Librispeech show that the full-size encoder in DET relatively reduces the word error rate of the same size baseline by over 8%. The lightweight encoder in DET trained with collaborative learning reduces the model size by 25% but still gets similar WER as the full-size baseline. DET gets similar accuracy as a baseline model with better latency on a large in-house data set by assigning a lightweight encoder for the beginning part of one utterance and a full-size encoder for the rest.

Index Terms: Transducer, Transformer, Layer Dropout, Collaborative Learning

1. Introduction

Due to the ubiquity of voice assistants in smartphones, smart speakers, and smart wearable devices, privacy is getting more and more attention. On-device automatic speech recognition (ASR) [1, 2, 3, 4] performs the speech recognition entirely on the device. Without streaming audio from the device to the server, on-device ASR provides better privacy protection than server ASR system.

Recurrent neural network transducer (RNN-T) [5, 3, 6] has been widely applied to on-device ASR. Like many other End-to-end (E2E) models [7, 8, 9, 10, 11, 12], RNN-T directly optimizes the transduction from the acoustic feature sequence to label sequence by combining acoustic model, pronunciation, and language model all into one neural network. RNN-T intrinsically supports online streaming speech recognition. More importantly, RNN-T models have a much lower footprint than hybrid systems, which is more suitable for on-device ASR. In this work, we used an alignment restricted transducer model [13] which gives faster training and better token emission latency.

Recently, better accuracy and real-time factors have been achieved by replacing LSTM encoder with transformer [14] and its variants (e.g. Conformer [15], Emformer [16], etc.) in sequence transducer [17, 18, 15, 19, 20]. In this work, we adopt the Emformer transducer in on-device ASR for low latency speech recognition. The emformer transducer model applies a deep structured transformer as the encoder. Each layer

in the encoder contains millions of parameters involving high computation cost and memory consumption. On-device ASR often needs accuracy, computation cost, and latency trade-off to achieve the best user experience for different devices. A high-end device can support a big model to deliver good accuracy and a smooth user experience. In contrast, a low-end device may only support a smaller model with less computation cost and memory consumption. Trade-off sometimes is also needed even in decoding a single utterance. In [21], inference computation cost reduction can be achieved by pivoting different computation pathways for different parts of the utterance.

Neural network model pruning [22, 23, 24], quantization [25, 24], low rank matrix factorization [26, 27, 28] and knowledge distillation [29, 30] have been applied to reduce the footprint and computation cost without significant accuracy loss. However, many of these methods involve finetuning or retraining, which are not flexible. In this work, we propose a dynamic encoder transducer (DET) for on-device ASR with a flexible trade-off between computation cost and accuracy.

We apply layer dropout and collaborative learning for DET training. Layer dropout [31] was proposed to stabilize and regularize deep convolution neural network training. The work [32] applied the layer dropout to train a deep transformer model for speech recognition. The work [33] used layer dropout as a structured dropout method to prune the over-parameterized transformer model for many natural language processing tasks. Collaborative learning [34] leverages the strengths of auxiliary loss, multi-task learning, and knowledge distillation to improve the deep neural classifier's generalization and robustness to label noise. Our work [35] applied collaborative learning in transformer transducer to jointly train the teacher and the multiple students all at the same time from scratch. The weight sharing among teachers and students improves the teacher model's performance and all of the student models.

We apply DET in two scenarios for accuracy and computation cost trade-off. The first one is flexibility in tuning encoder depth as a pruning method for a specific device in model deployment. The other one is dynamically applying different encoders in DET for decoding different parts in one utterance, which is similar to the scenario discussed in [21, 36]. Note the work [21, 36] is about LSTM based encoder.

2. Dynamic Encoder Transducer

This section introduces the sequence transducer modeling, the emformer based encoder, the layer dropout method, the collaborative learning method and the DET.

2.1. Sequence transducer

The sequence transducer consists of an encoder, a preditor, and a joiner. Given a sequence of acoustic feature vectors $X = \{x_1, ..., x_T\}$ with length T, the encoder f^e generates a

sequence of representations $H^e = \{h_1^e, ..., h_T^e\}$.

$$\{h_1^e, ..., h_T^e\} = f^e(X). \qquad (1)$$

Let $Y = \{y_1, ..., y_U\}$ in length U where $y_u \in \mathcal{Y}$ be the sequence of output units which are sentence pieces [37]. We define $\bar{\mathcal{Y}} = \mathcal{Y} \cup \{\phi\}$, where ϕ is the blank label. The predictor f^p generates a sequence of representation for the output sequence prefixed with the blank label as follows:

$$\{h_1^p, ..., h_u^p\} = f^p(\{\phi, y_1, ..., y_{u-1}\}). \qquad (2)$$

The joiner f^j combines the output representations from the encoder and the predictor to generate the logits $h_{t,u}$.

$$h_{t,u} = f^j(h_t^e, h_u^p). \qquad (3)$$

A softmax function is applied to produce the posterior distribution of the next label $y_u \in \bar{\mathcal{Y}}$

$$P(y_u|x_{1:t}, y_{1:u-1}) = \text{softmax}(h_{t,u}). \qquad (4)$$

Using forward-backward algorithm, the posterior of a sequence of output units $Y = \{y_1, ..., y_U\}$ where $y_u \in \mathcal{Y}$ as follows:

$$P(Y|X) = \sum_{A \in \mathcal{B}^{-1}(Y)} P(A|X), \qquad (5)$$

where A is one alignment for the sequence of output Y. Each label in A is from label set $\bar{\mathcal{Y}}$. $\mathcal{B}$ is the operation to remove the blank from the alignment. The loss is defined as

$$L^{Tr} = -log(P(Y|X)). \qquad (6)$$

2.2. Emformer based encoder

We use emformer transducer [16] as a basic model architecture for low latency streaming on-device ASR. The emformer modifies the transformer model with the block processing [38, 39] to support low latency streaming speech recognition. In training, the emformer uses an attention mask and a "right context hard copy" trick to constrain the receptive field for self-attention. In decoding, the emformer applies a cache method to optimize the efficiency optimization by saving the computation for the key and value in self-attention for the left context.

In the emformer transducer, the encoder accounts for most of the computation cost. The encoder is much deeper than the predictor (e.g., encoder usually has 10 to 20 layers, predictor only has 2 to 3 layers). More importantly, the sequence length T in $X = \{x_1, ..., x_T\}$ is much longer than U in $Y = \{y_1, ..., y_U\}$. Reducing the number of layers in the emformer based encoder saves the computation cost. DET focuses on using different encoders with various depths for flexible trade-off accuracy and computation cost in this work.

2.3. Layer dropout

The model becomes robust in decoding with missing layers by randomly masking out some layers in training. The work [33] investigated different layer pruning strategies and found that pruning every other layer performed best in inference.

In this work, we also investigate every other layer of dropout in the training phase. Rather than applying dropout to each layer at random, we apply dropout to every other layer, resulting in a more structured dropout. In decoding, only the layers dropped out in training are pruned for decoding. Every other layer dropout provides less flexibility but makes the training and decoding more consistent.

2.4. Collaborative learning

Figure 1 gives one DET model with two encoders E_1 and E_2. The encoder E_2 share all its layers with E_1 except its last layer. Given $X = \{x_1, ..., x_T\}$, multiple sequence of acoustic rep-
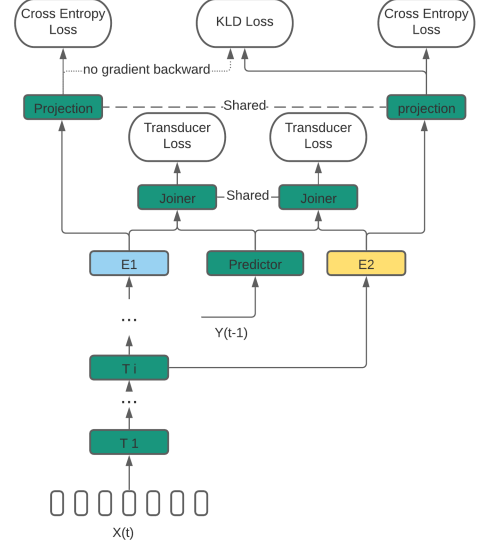

Figure 1: *Collaborative learning for sequence transducer model for on-device ASR. Two encoders with different depths are used. E_2 shares all its layers with E_1 except the last layer. All the shared nodes by these two encoder are in green.*

resentations $H^i = \{h_1^i, ..., h_T^i\}$ are generated from different encoders i. In Figure 1, $i \in \{1, 2\}$.

$$\{h_1^i, ..., h_T^i\} = f^i(X). \qquad (7)$$

Combining with the exact predictor's output $\{h_1^p, ..., h_u^p\}$, the DET model generates multiple transducer losses L^{Tr_i} by passing the different encoders' output through the same joiner.

Similar to the work [40], the collaborative learning [35] and this work use the auxiliary supervision based on chenones [41] prediction. Let $C = \{c_1, ..., c_T\}$ be the forced alignment label sequence for X. The posterior distribution of c_t from encoder i is obtained as follows:

$$P^i(c_t|x_t) = \text{softmax}(f^{pro}(h_t^i)), \qquad (8)$$

where f^{pro} is a multi-layer perception shared by all encoders. Given the posterior distribution, we can get the sum of the cross-entropy losses for all the encoders.

$$L^{CE} = -\frac{1}{T} \sum_i \sum_T log(P^i(c_t|x_t)). \qquad (9)$$

The auxiliary cross-entropy loss introduces the forced alignment information into DET training and constructs the bridge to do the knowledge distillation from the full-size encoder e_1 to the lightweight encoders e_i where $i > 1$. As shown in Figure 1, Kullback–Leibler divergence loss is used.

$$L^{KLD} = -\frac{1}{T} \sum_{i>1} \sum_T P^{e_1}(c_t|x_t) log(\frac{P^{e_1}(c_t|x_t)}{P^i(c_t|x_t)}). \qquad (10)$$

where e_1 the deepest encoder in DET is used as a teacher for all the rest encoders. Note the gradient from L^{KLD} is not back-propagated to e_1. The final loss L to train DET model is

$$L = \alpha L^{CE} + \beta L^{KLD} + \sum_i L^{Tr_i}, \qquad (11)$$

where α and β are the interpolation weight for the cross-entropy loss and the Kullback–Leibler divergence loss, respectively.

2.5. Applications of dynamic encoder transducer

Figure 2 illustrates two application scenarios of the DET model. The figure's middle part shows a typical pruning process where one specific depth encoder is selected from the DET. This process happens when we need to optimize the on-device ASR model computation cost for a specific device. The right part shows the decoding uses two different encoders e_1 and e_2 jointly to get the representation for acoustic features:

$$\{h_1^e, ..., h_{T1}^e\} = f^{e_1}(\{x_1, ..., x_K\}), \quad (12)$$

$$\{h_{K+1}^e, ..., h_T^e\} = f^{e_2}(\{x_{K+1}, ..., x_T\}). \quad (13)$$

Decoding using multiple encoders jointly on one utterance provides a flexible solution to trade-off accuracy and computation cost for a selected device. It is helpful to deal with the one-shot assistant queries with wake words. When a user invokes a voice assistant with a wake word and then speaks a query, the on-device ASR often needs to process the user's query as well as the wake word. Note *audio bursting* happens, where the audio corresponding to the wake word is available for the ASR engine all at once. The user-perceived latency will be improved if a smaller encoder processes the audio bursting with less computation cost. Additionally, there is a cold-start latency when users start to interact with the voice assistant in a new session. The device needs to load the ASR model into memory. Loading a small encoder initially and gradually loading the rest of the encoder layers in parallel reduces the on-device ASR engine cold-start latency.
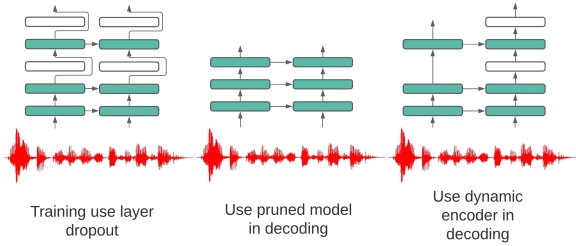

Figure 2: *Two usage scenarios of the DET. The middle part shows that one specific depth encoder is selected. The right part shows that different depth encoders are used for decoding.*

3. Experiments

To evaluate the performance of DET, we carry out two sets of experiments *one encoder decoding* and *dynamic encoder decoding*. *One encoder* simulates the situation that one specific encoder needs to be selected to meet a device's computation capacity. *Dynamic encoder decoding* uses multiple encoders in DET for decoding one utterance.

3.1. Data

The experiments use LibriSpeech corpus [42] and a large in-house dataset. LibriSpeech is an open-source speech corpus that contains 1000 hours of speech derived from audiobooks in the LibriVox project. Approximately 30 hours of data is used for development and evaluation, which are split into *clean* subsets and *other* subsets.

The in-house dataset consists of data from *Voice Assistant* and *Open Domain*. All data is anonymized with personally identifiable information (PII) removed. The *Voice Assistant* data contains 68k hours of data from human transcribed data from 20K crowd-sourced workers recorded via mobile devices, 1K hours voice commands, sampled from production traffic and 23k hours of voice command data generated from an in-house TTS system. The *Open Domain* data includes 13K hours of data

from public social media English videos that are anonymized with PII removed and annotator transcribed and 1.5M hours from the same source with a large offline model's transcriptions.

In evaluation, we use *assi* and *dict* dataset. The *assi* is 13.6K manually transcribed de-identified utterances from in-house volunteer employees, which begin with a wake word. The *dict* is 8 hours open domain dictation from crowd-sourced workers recorded via mobile devices. All the evaluation data is anonymized with personally identifiable information (PII) removed. For both Librispeech and in-house dataset, we generate forced alignment using a context and positional dependent graphemes (i.e., chenones) [41] based hybrid system.

3.2. Experiment Setting

In all the experiments, we use 80-dimensional log Mel filter bank features at a 10ms frame rate. To increase the training robustness, *SpecAugment* [43] without time warping are used. For the Librispeech experiment, we map the 80-dimensional features by a linear layer to 128 dimension vectors. A 512-dimensional vector is formed by concatenating four continuous 128-dimensional vectors, which is the input to Emformer. For a large in-house dataset, a 480-dimensional superframe is formed by concatenating six continuous features. A linear layer maps the superframe to a 512-dimensional vector.

Each Emformer uses eight heads of self-attention with input dimension 512. The output from the self-attention goes through a feed-forward layer with dimensionality 2048. We set dropout 0.1 for all layers across all experiments. The segment sizes are 160ms in LibriSpeech experiment and 300ms for in-house data experiments. The left context sizes are 1.2s for Librispeech and 1.8s for in-house data. The 40ms and 60ms look-ahead contexts are used for LibriSpeech and in-house experiments, respectively. The 512-dimensional output from the stack of Emformer layers goes through a layer norm followed by a linear layer. Finally, the output from the encoder in the transducer model is a 1024-dimensional vector.

The predictor consists of a 256-dimensional embedding layer, three LSTM layers with 512 hidden nodes, and a linear projection layer with 1024 output nodes. The combined 1024 dimensional embeddings from the encoder and the predictor go through a Tanh activation and then another linear projection to the final output with 4096 sentence pieces [37].

All models are trained with the adam optimizer [44] using warming-up updates and the learning rate 1e-3. In Librispeech experiments, the last checkpoint from 120 epoch training is for evaluation. For large in-house data, we use the checkpoint from 800K updates for evaluation. In collaborative training, we set both α and β to 0.5. In layer dropout, the optimal dropout rate for each layer is 0.1 obtained by grid search.

All the model training uses 32 Nvidia V100 GPUs. We evaluate the latency by the real-time factors (RTFs) and speech engine perceived latency (SPL). SPL measures the time from speech engine gets the last word from user utterance to speech engine transcribes the last word and gets the endpoint signals. For in-house data experiment, the SPL evaluation uses the static endpointer, neural endpointer, and transducer's end-of-sequencing symbol. We use low-end android device for latency evaluation. We sample 100 utterances from *clean* and *assi* for latency evaluation for Librispeech and in-house data, respectively.

3.3. One encoder decoding

In one encoder decoding, both layer dropout and collaborative learning are used as prune methods to select one encoder that

Table 1: *WER on LibriSpeech data for DET. "base pruned" prunes the full size encoder. "random" dropout each layer randomly. "group" dropout layer 1, 4, 7, 10, 13 and 16 as a group randomly. "6-16:2" dropouts every 2 layers from layer 6 to layer 16. "1-16:3" dropouts every 3 layers from layer 1 to layer 16. "cl" denotes the DET trained using collaborative learning.*

arch	# layers	*clean*	*other*	#params
base	20	3.62	9.86	77M
base pruned	14	31.42	49.50	58M
base	14	3.87	10.35	58M
random	20	3.75	9.38	77M
6-16:2	14	4.80	12.45	58M
1-16:3	14	4.35	11.56	58M
group	20	3.97	10.09	77M
1-16:3	14	4.32	11.38	58M
cl	20	**3.54**	**9.04**	77M
cl	14	3.66	**9.60**	58M

Table 2: *WER, RTFs and SPL on in-house data for DET trained by layer dropout and collaborative learning. "random" dropout each layer randomly. "1-16:3" dropouts every 3 layers from layer 1 to layer 16. "cl" denotes the DET trained using collaborative learning.*

arch	#layers	*dict*	*assi*	RTFs	SPL
baseline	20	16.40	3.83	0.47	756
baseline	14	17.56	4.04	0.40	697
random	20	16.21	3.95	0.47	766
1-16:3	14	19.29	6.74	0.39	688
cl	20	17.41	4.01	0.45	732
cl	14	18.87	4.53	0.38	685

meets one specific device's requirement. Table 1 and Table 2 gives the word error rate (WER) for DET on Librispeech data and in-house data, respectively. Table 1 shows that the 20-layers encoder trained by the layer dropout and collaborative learning reduces the WER of 14-layer encoder on *other* by 5% and 8%, respectively. "base pruned" shows that directly pruning the 20 layers baseline model to 14 layers does not work. In Table 1, the best-pruned model with 14 layers is from collaborative learning, which gets the on-par accuracy as the full-size model. The pruned models from both the "random" and "group" dropout method are worse than re-training the same size model from scratch. Note the results from collaborative learning in Table 1 is different with [35], in this paper, each student encoder shares its layers except the last one.

Table 2 shows that the pruned model trained by collaborative learning is more accurate than the one trained from layer dropout. However, neither of the pruned models outperforms the same size baseline model trained separately. For in-house data, the full-size encoder trained by layer dropout performs similarly as same size baseline with a slight improvement on *dict* and slight degradation on *assi*.

3.4. Dynamic encoder in decoding

Table 3 and Table 4 show the experiment results using multiple encoders in decoding one utterance on Librispeech data and in-house data, respectively. For Librispeech data, we compare the layer dropout and the collaborative learning. For in-house data, we only use the layer dropout method. The full-size encoder in DET trained by the layer dropout is more accurate than the same-size encoder trained by collaborative learning.

Both tables show that by adjusting the audio's length to be decoded by the small encoder in DET, we can do a trade-off between accuracy and latency. Table 3 shows that decoding using DET trained by collaborative learning gives better WER and RTFs trade-off than layer dropout. Using a small encoder for the beginning 0.8s audio, the DET gets 6% relative WER reduction and over 10% RTFs reduction.

Table 3: *WER and RTFs on LibriSpeech data for DET trained by layer dropout and collaborative learning. "time" column denotes the length of the beginning part of audio decoded by the small encoder. "drop" and "cl" denotes the DET trained using layer dropout and collaborative learning, respectively.*

arch	time	*clean*	*other*	#layers	RTFs
base	-	3.62	9.86	20	0.63
base	-	3.87	10.35	14	0.47
drop	6.4s	4.18	11.07	14/20	0.51
drop	3.2s	4.04	10.45	14/20	0.52
drop	1.6s	3.90	10.02	14/20	0.57
drop	0.8s	3.84	9.49	14/20	0.60
cl	6.4s	3.75	9.63	14/20	0.51
cl	3.2s	3.76	9.52	14/20	0.54
cl	1.6s	3.72	9.45	14/20	0.55
cl	0.8s	**3.62**	**9.27**	14/20	0.56

Table 4 shows that using multiple encoders for one utterance decoding achieves advantages from each encoder. In the experiment, we use multiple encoders decoding for *assi* which has a wake word at the beginning for each utterance. We use the full-size encoder for *dict* which does not have wake word. The bottom line in Table 4 shows that using a small encoder for the beginning 1.5s audio gives slight WER degradation with better RTFs and SPL.

Table 4: *WER, RTFs and SPL on in-house data for DET trained by layer dropout. "1-16:3" dropouts every 3 layers from layer 1 to layer 16. "drop" denotes the DET trained using layer dropout. "time" column denotes the length of the beginning part of audio decoded by the small encoder.*

arch	time	#layers	*dict*	*assi*	RTFs	SPL
baseline	-	20	16.40	3.83	0.47	756
baseline	-	14	17.56	4.04	0.40	697
drop	3.0s	14/20	16.21	4.04	0.41	700
drop	1.5s	14/20	16.21	3.99	0.44	714

4. Conclusions

This paper proposed a dynamic encoder transducer as a flexible on-device ASR model for accuracy and latency trade-off. Two application scenarios were discussed: pruning encoder to fit device computation limit and decoding different parts of one utterance using different encoders. We applied layer dropout and collaborative learning for DET model training. The layer dropout method that randomly masked some encoder layers in training makes DET flexible and robust in dropping layers in decoding. The collaborative learning jointly trained multiple encoders with different depths using weight sharing, auxiliary tasks, and knowledge distillation. On Librispeech data, experiments showed that the DET model outperformed the baseline in accuracy and latency for both scenarios. In-house data experiments showed that DET gets on par accuracy as baseline model but with better RTFs and latency.

5. References

[1] X. Lei, A. Senior, A. Gruenstein, and J. Sorensen, "Accurate and compact large vocabulary speech recognition on mobile devices," in *Proc. INTERSPEECH*, 2013.

[2] I. McGraw, R. Prabhavalkar, R. Alvarez, M. G. Arenas, and Others, "Personalized speech recognition on mobile devices," in *Proc. ICASSP*, 2016.

[3] Y. He, T. N. Sainath, R. Prabhavalkar, and Others, "Streaming End-to-end Speech Recognition for Mobile Devices," in *Proc. ICASSP*, 2019.

[4] J. Li, R. Zhao, H. Hu, and Y. Gong, "Improving RNN transducer modeling for end-to-end speech recognition," in *Proc. ASRU*, 2019.

[5] A. Graves, "Sequence Transduction with Recurrent Neural Networks," *arXiv preprint arXiv:1211.3711*, 2012.

[6] K. Rao, H. Sak, and R. Prabhavalkar, "Exploring architectures, data and units for streaming end-to-end speech recognition with RNN-transducer," in *Proc. ASRU*, 2018.

[7] A. Graves and N. Jaitly, "Towards End-To-End Speech Recognition with Recurrent Neural Networks," in *Proc. JMLR*, 2014.

[8] D. Amodei, R. Anubhai, E. Battenberg, C. Case, and Others, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

[9] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proc. ASRU*, 2016.

[10] S. Zhang, M. Lei, Y. Liu, and W. Li, "Investigation of Modeling Units for Mandarin Speech Recognition Using Dfsmn-ctc-smbr," in *Proc. ICASSP*, 2019.

[11] Y. Shi, M.-Y. Hwang, and X. Lei, "End-To-End Speech Recognition Using A High Rank LSTM-CTC Based Model," in *Proc. ICASSP*, 2019.

[12] A. Das, J. Li, R. Zhao, and Y. Gong, "Advancing Connectionist Temporal Classification With Attention Modeling," in *Proc. ICASSP*, 2018.

[13] J. Mahadeokar, Y. Shangguan, D. Le, and Others, "Alignment restricted streaming recurrent neural network transducer," in *Proc. SLT*, 2021.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017.

[15] A. Gulati, J. Qin, C. C. Chiu, and Others, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. INTERSPEECH*, 2020.

[16] Y. Shi, Y. Wang, C. Wu, C.-F. Yeh, and Others, "Emformer: Efficient Memory Transformer Based Acoustic Model For Low Latency Streaming Speech Recognition," in *Proc. ICASSP*, 2021.

[17] 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," in *Proc. ICASSP*, 2020.

[18] C.-F. Yeh, J. Mahadeokar, K. Kalgaonkar, Y. Wang, D. Le, M. Jain, K. Schubert, C. Fuegen, and M. L. Seltzer, "Transformer-Transducer: End-to-End Speech Recognition with Self-Attention," *arXiv preprint arXiv:11910.12977*, 2019.

[19] C. F. Yeh, Y. Wang, Y. Shi, C. Wu, F. Zhang, and Others, "Streaming attention-based models with augmented memory for end-to-end speech recognition," in *Proc. SLT*, 2020.

[20] Y. Wang, Y. Shi, F. Zhang, and Others, "Transformer in action: a comparative study of transformer-based acoustic models for large scale speech recognition applications," *Proc. ICASSP*, 2020.

[21] J. Macoskey, G. P. Strimel, and A. Rastrow, "Bifocal Neural Asr: Exploiting Keyword Spotting For Interence Optimization," *Proc. ICASSP*, 2021.

[22] M. H. Zhu and S. Gupta, "To prune, or not to prune: Exploring the efficacy of pruning for model compression," *arXiv preprint arXiv:1710.01878*, 2017.

[23] Y. Shangguan, J. Li, L. Qiao, R. Alvarez, and I. McGraw, "Optimizing speech recognition for the edge," *arXiv preprint arXiv:1909.12408*, 2019.

[24] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *Proc. NIPS*, 2015.

[25] Y. Guo, "A survey on methods and theories of quantized neural networks," *arXiv preprint arXiv:1808.04752*, 2018.

[26] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohamadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. INTERSPEECH*, 2018.

[27] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition," in *Proc. INTERSPEECH*, 2013.

[28] R. Prabhavalkar, O. Alsharif, A. Bruguier, and L. McGraw, "On the compression of recurrent neural networks with an application to LVCSR acoustic modeling for embedded speech recognition," in *Proc. ICASSP*, 2016.

[29] S. Kim, M. L. Seltzer, J. Li, and R. Zhao, "Improved training for online end-to-end speech recognition systems," *arXiv preprint arXiv:11711.02212*, 2017.

[30] V. Manohar, P. Ghahremani, D. Povey, and S. Khudanpur, "A Teacher-Student Learning Approach for Unsupervised Domain Adaptation of Sequence-Trained ASR Models," in *Proc. STL*, 2019.

[31] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," *Lecture Notes in Computer Science*, vol. 9908 LNCS, pp. 646–661, 2016.

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

[33] A. Fan, E. Grave, and A. Joulin, "Reducing transformer depth on demand with structured dropout," *arXiv preprint arXiv:1909.11556*, 2019.

[34] G. Song and W. Chai, "Collaborative learning for deep neural networks," in *Proc. NIPS*, 2018.

[35] V. Nagaraja, Y. Shi, G. Venkatesh, and Others, "Collaborative Training of Acoustic Encoders for Speech Recognition," in *Proc. INTERSPEECH*, 2021.

[36] A. Graves, "Adaptive Computation Time for Recurrent Neural Networks," *arXiv preprint arXiv:1603.08983*, 2016.

[37] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *Proc. EMNLP*, 2018.

[38] C. Wu, Y. Shi, Y. Wang, and C.-F. Yeh, "Streaming Transformer-based Acoustic Modeling Using Self-attention with Augmented Memory," in *Proc. INTERSPEECH*, 2020.

[39] L. Dong, F. Wang, and B. Xu, "Self-attention Aligner: A Latency-control End-to-end Model for ASR Using Self-attention Network and Chunk-hopping," *Proc. ICASSP*, 2019.

[40] C. Liu, F. Zhang, D. Le, S. Kim, Y. Saraf, and G. Zweig, "Improving RNN transducer based ASR with auxiliary tasks," in *Proc. SLT*, 2021.

[41] D. Le, X. Zhang, W. Zheng, and Others, "From Senones to Chenones: Tied Context-Dependent Graphemes for Hybrid Speech Recognition," *arXiv preprint arXiv:1910.01493*, 2019.

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[43] D. S. Park, W. Chan, Y. Zhang, and Others, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

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