

Multi-Head Decoder for End-to-End Speech Recognition

Tomoki Hayashi¹, Shinji Watanabe², Tomoki Toda¹, Kazuya Takeda¹

¹Nagoya University, ²Johns Hopkins University

hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp, shinjiw@jhu.edu,
tomoki@icts.nagoya-u.ac.jp, kazuya.takeda@nagoya-u.ac.jp

Abstract

This paper presents a new network architecture called multi-head decoder for end-to-end speech recognition as an extension of a multi-head attention model. In the multi-head attention model, multiple attentions are calculated, and then, they are integrated into a single attention. On the other hand, instead of the integration in the attention level, our proposed method uses multiple decoders for each attention and integrates their outputs to generate a final output. Furthermore, in order to make each head to capture the different modalities, different attention functions are used for each head, leading to the improvement of the recognition performance with an ensemble effect. To evaluate the effectiveness of our proposed method, we conduct an experimental evaluation using Corpus of Spontaneous Japanese. Experimental results demonstrate that our proposed method outperforms the conventional methods such as location-based and multi-head attention models, and that it can capture different speech/linguistic contexts within the attention-based encoder-decoder framework.

Index Terms: speech recognition, end-to-end, attention, dynamical neural network

1. Introduction

Automatic speech recognition (ASR) is the task to convert a continuous speech signal into a sequence of discrete characters, and it is a key technology to realize the interaction between human and machine. ASR has a great potential for various applications such as voice search and voice input, making our lives more rich. Typical ASR systems [1] consist of many modules such as an acoustic model, a lexicon model, and a language model. Factorizing the ASR system into these modules makes it possible to deal with each module as a separate problem. Over the past decades, this factorization has been the basis of the ASR system, however, it makes the system much more complex.

With the improvement of deep learning techniques, end-to-end approaches have been proposed [2]. In the end-to-end approach, a continuous acoustic signal or a sequence of acoustic features is directly converted into a sequence of characters with a single neural network. Therefore, the end-to-end approach does not require the factorization into several modules, as described above, making it easy to optimize the whole system. Furthermore, it does not require lexicon information, which is handcrafted by human experts in general.

The end-to-end approach is classified into two types. One approach is based on connectionist temporal classification (CTC) [2–4], which makes it possible to handle the difference in the length of input and output sequences with dynamic programming. The CTC-based approach can efficiently solve the sequential problem, however, CTC uses Markov assumptions to perform dynamic programming and predicts output symbols such as characters or phonemes for each frame independently.

Consequently, except in the case of huge training data [5,6], it requires the language model and graph-based decoding [7].

The other approach utilizes attention-based method [8]. In this approach, encoder-decoder architecture [9, 10] is used to perform a direct mapping from a sequence of input features into text. The encoder network converts the sequence of input features to that of discriminative hidden states, and the decoder network uses attention mechanism to get an alignment between each element of the output sequence and the encoder hidden states. And then it estimates the output symbol using weighted averaged hidden states, which is based on the alignment, as the inputs of the decoder network. Compared with the CTC-based approach, the attention-based method does not require any conditional independence assumptions including the Markov assumption, language models, and complex decoding. However, non-causal alignment problem is caused by a too flexible alignment of the attention mechanism [11]. To address this issue, the study [11] combines the objective function of the attention-based model with that of CTC to constrain flexible alignments of the attention. Another study [12] uses a multi-head attention (MHA) to get more suitable alignments. In MHA, multiple attentions are calculated, and then, they are integrated into a single attention. Using MHA enables the model to jointly focus on information from different representation subspaces at different positions [13], leading to the improvement of the recognition performance.

Inspired by the idea of MHA, in this study we present a new network architecture called multi-head decoder for end-to-end speech recognition as an extension of a multi-head attention model. Instead of the integration in the attention level, our proposed method uses multiple decoders for each attention and integrates their outputs to generate a final output. Furthermore, in order to make each head to capture the different modalities, different attention functions are used for each head, leading to the improvement of the recognition performance with an ensemble effect. To evaluate the effectiveness of our proposed method, we conduct an experimental evaluation using Corpus of Spontaneous Japanese. Experimental results demonstrate that our proposed method outperforms the conventional methods such as location-based and multi-head attention models, and that it can capture different speech/linguistic contexts within the attention-based encoder-decoder framework.

2. Attention-Based End-to-End ASR

The overview of attention-based network architecture is shown in Fig. 1. The attention-based method directly estimates a posterior $p(\mathbf{C}|\mathbf{X})$, where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\}$ represents a sequence of input features, $\mathbf{C} = \{c_1, c_2, \dots, c_L\}$ represents a sequence of output characters. The posterior $p(\mathbf{C}|\mathbf{X})$ is factor-

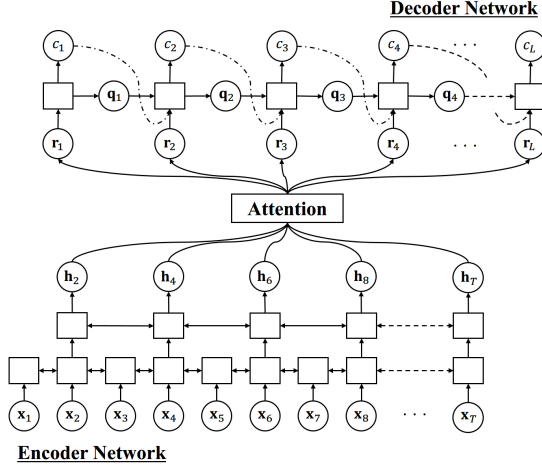


Figure 1: Overview of attention-based network architecture.

ized with a probabilistic chain rule as follows:

$$p(\mathbf{C}|\mathbf{X}) = \prod_{l=1}^L p(c_l | c_{1:l-1}, \mathbf{X}), \quad (1)$$

where $c_{1:l-1}$ represents a subsequence $\{c_1, c_2, \dots, c_{l-1}\}$, and $p(c_l | c_{1:l-1}, \mathbf{X})$ is calculated as follows:

$$\mathbf{h}_t = \text{Encoder}(\mathbf{X}), \quad (2)$$

$$a_{lt} = \begin{cases} \text{DotProductAttention}(\mathbf{q}_{l-1}, \mathbf{h}_t), \\ \text{AdditiveAttention}(\mathbf{q}_{l-1}, \mathbf{h}_t), \\ \text{LocationAttention}(\mathbf{q}_{l-1}, \mathbf{h}_t, \mathbf{a}_{l-1}), \\ \text{CoverageAttention}(\mathbf{q}_{l-1}, \mathbf{h}_t, \mathbf{a}_{1:l-1}), \end{cases} \quad (3)$$

$$\mathbf{r}_l = \sum_{t=1}^T a_{lt} \mathbf{h}_t, \quad (4)$$

$$p(c_l | c_{1:l-1}, \mathbf{X}) = \text{Decoder}(\mathbf{r}_l, \mathbf{q}_{l-1}, c_{l-1}), \quad (5)$$

where Eq. (2) and Eq. (5) represent encoder and decoder networks, respectively, a_{lt} represents an attention weight, \mathbf{a}_l represents an attention weight vector, which is a sequence of attention weights $\{a_{l0}, a_{l1}, \dots, a_{lT}\}$, $\mathbf{a}_{1:l-1}$ represents a subsequence of attention vectors $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_{l-1}\}$, \mathbf{h}_t and \mathbf{q}_l represent hidden states of encoder and decoder networks, respectively, and \mathbf{r}_l represents the letter-wise hidden vector, which is a weighted summarization of hidden vectors with the attention weight vector \mathbf{a}_l .

The encoder network in Eq. (2) converts a sequence of input features \mathbf{X} into frame-wise discriminative hidden states \mathbf{h}_t , and it is typically modeled by a bidirectional long short-term memory recurrent neural network (BLSTM):

$$\text{Encoder}(\mathbf{X}) = \text{BLSTM}(\mathbf{X}). \quad (6)$$

In the case of ASR, the length of the input sequence is significantly different from the length of the output sequence. Hence, basically outputs of BLSTM are often subsampled to reduce the computational cost [8, 14].

The attention weight a_{lt} in Eq. (3) represents a soft alignment between each element of the output sequence c_l and the encoder hidden states \mathbf{h}_t .

- DotProductAttention(\cdot) in Eq. (3), which is the most simplest attention [15], is calculated as follows:

$$e_{lt} = \mathbf{q}_{l-1}^T \mathbf{W}_a \mathbf{h}_t, \quad (7)$$

$$\mathbf{a}_l = \text{Softmax}(\mathbf{e}_l), \quad (8)$$

where \mathbf{W}_a represents trainable matrix parameters, and \mathbf{e}_l represent a sequence of energies $\{e_{l1}, e_{l2}, \dots, e_{lT}\}$

- AdditiveAttention(\cdot) in Eq. (3) is additive attention [16], and the calculation of the energy in Eq. (7) is replaced with following equation:

$$e_{lt} = \mathbf{g}^T \tanh(\mathbf{W}_q \mathbf{q}_{l-1} + \mathbf{W}_h \mathbf{h}_t + \mathbf{b}), \quad (9)$$

where \mathbf{W}_* represents trainable matrix parameters, \mathbf{g} and \mathbf{b} represent trainable vector parameters.

- LocationAttention(\cdot) in Eq. (3) is location-based attention [8], and the calculation of the energy in Eq. (7) is replaced with following equations:

$$\mathbf{F}_l = \mathbf{K} * \mathbf{a}_{l-1}, \quad (10)$$

$$e_{lt} = \mathbf{g}^T \tanh(\mathbf{W}_q \mathbf{q}_{l-1} + \mathbf{W}_h \mathbf{h}_t + \mathbf{W}_f \mathbf{f}_{lt} + \mathbf{b}), \quad (11)$$

where \mathbf{F}_l consists of the vectors $\{\mathbf{f}_{l1}, \mathbf{f}_{l2}, \dots, \mathbf{f}_{lT}\}$, and \mathbf{K} represents trainable one-dimensional convolution filters.

- CoverageAttention(\cdot) in Eq. (3) is coverage mechanism [17], and the calculation of the energy in Eq. (7) is replaced with following equations:

$$\mathbf{v}_l = \sum_{l'=1}^{l-1} \mathbf{a}_{l'}, \quad (12)$$

$$e_{lt} = \mathbf{g}^T \tanh(\mathbf{W}_q \mathbf{q}_{l-1} + \mathbf{W}_h \mathbf{h}_t + \mathbf{w}_v \mathbf{v}_{lt} + \mathbf{b}), \quad (13)$$

where \mathbf{w} represents trainable vector parameters.

The decoder network in Eq. (5) estimates the next character c_l from the previous character c_{l-1} , hidden state vector of itself \mathbf{q}_{l-1} and the letter-wise hidden state vector \mathbf{r}_l , similar to RNN language model (RNNLM) [18]. It is typically modeled using LSTM as follows:

$$\mathbf{q}_l = \text{LSTM}(c_{l-1}, \mathbf{q}_{l-1}, \mathbf{r}_l), \quad (14)$$

$$\text{Decoder}(\cdot) = \text{Softmax}(\mathbf{W} \mathbf{q}_l + \mathbf{b}), \quad (15)$$

where \mathbf{W} and \mathbf{b} represent trainable matrix and vector parameters, respectively.

Finally, the whole of above networks are optimized using back-propagation through time (BPTT) [19] to minimize the following objective function:

$$\begin{aligned} \mathcal{L} &= -\log p(\mathbf{C}|\mathbf{X}) \\ &= -\log \left(\prod_{l=1}^L p(c_l | c_{1:l-1}^*, \mathbf{X}) \right), \end{aligned} \quad (16)$$

where $c_{1:l-1}^* = \{c_1^*, c_2^*, \dots, c_{l-1}^*\}$ represents the ground truth of the previous characters.

3. Multi-Head Decoder

The overview of our proposed multi-head decoder (MHD) architecture is shown in Fig. 2. In MHD architecture, multiple attentions are calculated with the same manner in the conventional multi-head attention (MHA) [13]. We first describe the conventional MHA, and extend it to our proposed multi-head decoder (MHD).

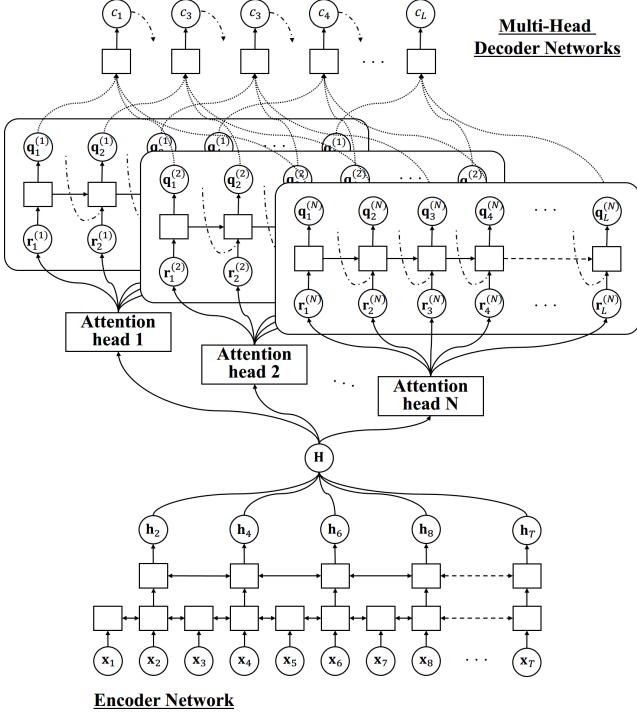


Figure 2: Overview of multi-head decoder architecture.

3.1. Multi-head attention (MHA)

The layer-wise hidden vector at the head n is calculated as follows:

$$a_{lt}^{(n)} = \text{Attention}(\mathbf{W}_Q^{(n)} \mathbf{q}_{l-1}^{(n)}, \mathbf{W}_K^{(n)} \mathbf{h}_t, \dots), \quad (17)$$

$$\mathbf{r}_l^{(n)} = \sum_{t=1}^T a_{lt}^{(n)} \mathbf{W}_V^{(n)} \mathbf{h}_t, \quad (18)$$

where $\mathbf{W}_Q^{(n)}$, $\mathbf{W}_K^{(n)}$, and $\mathbf{W}_V^{(n)}$ represent trainable matrix parameters, and any types of attention in Eq. (3) can be used for $\text{Attention}(\cdot)$ in Eq. (17).

In the case of MHA, the layer-wise hidden vectors of each head are integrated into a single vector with a trainable linear transformation:

$$\mathbf{r}_l = \mathbf{W}_O \left[\mathbf{r}_l^{(1)\top}, \mathbf{r}_l^{(2)\top}, \dots, \mathbf{r}_l^{(N)\top} \right]^\top, \quad (19)$$

where \mathbf{W}_O is a trainable matrix parameter, N represents the number of heads.

3.2. Multi-head decoder (MHD)

On the other hand, in the case of MHD, instead of the integration at attention level, we assign multiple decoders for each head and then integrate their outputs to get a final output. Since each attention decoder captures different modalities, it is expected to improve the recognition performance with an ensemble effect. The calculation of the attention weight at the head n in Eq. (17) is replaced with following equation:

$$a_{lt}^{(n)} = \text{Attention}(\mathbf{W}_Q^{(n)} \mathbf{q}_{l-1}^{(n)}, \mathbf{W}_K^{(n)} \mathbf{h}_t, \dots). \quad (20)$$

Instead of the integration of the letter-wise hidden vectors $\{\mathbf{r}_l^{(1)}, \mathbf{r}_l^{(2)}, \dots, \mathbf{r}_l^{(N)}\}$ with linear transformation, each letter-wise hidden vector $\mathbf{r}_l^{(n)}$ is fed to n -th decoder LSTM:

$$\mathbf{q}_l^{(n)} = \text{LSTM}^{(n)}(c_{l-1}, \mathbf{q}_{l-1}^{(n)}, \mathbf{r}_l^{(n)}). \quad (21)$$

Note that each LSTM has its own hidden state $\mathbf{q}_l^{(n)}$ which is used for the calculation of the attention weight $a_{lt}^{(n)}$, while the input character c_l is the same among all of the LSTMs. Finally, all of the outputs are integrated as follows:

$$p(c_l | c_{1:l-1}, \mathbf{X}) = \text{Softmax} \left(\sum_{n=1}^N \mathbf{W}^{(n)} \mathbf{q}_l^{(n)} + \mathbf{b} \right), \quad (22)$$

where $\mathbf{W}^{(n)}$ represents a trainable matrix parameter, and \mathbf{b} represents a trainable vector parameter.

3.3. Heterogeneous multi-head decoder (HMHD)

As a further extension, we propose heterogeneous multi-head decoder (HMHD). Original MHA methods [12, 13] use the same attention function such as dot-product or additive attention for each head. On the other hand, HMHD uses different attention functions for each head. We expect that this extension enables to capture the further different context in speech within the attention-based encoder-decoder framework.

4. Experimental Evaluation

To evaluate the performance of our proposed method, we conducted experimental evaluation using Corpus of Spontaneous Japanese (CSJ) [21], including 581 hours of training data, and three types of evaluation data. To compare the performance, we used following dot, additive, location, and three variants of multi-head attention methods:

- Dot-product attention-based model (Dot),
- Additive attention-based model (Add),
- Location-aware attention-based model (Loc),
- Multi-head dot-product attention model (MHA-Dot),
- Multi-head additive attention model (MHA-Add),
- Multi-head location attention model (MHA-Loc).

We used the input feature vector consisting of 80 dimensional log Mel filter bank and three dimensional pitch feature, which is extracted using open-source speech recognition toolkit Kaldi [22]. Encoder and decoder networks were six-layered BLSTM with projection layer [23] (BLSTMP) and one-layered LSTM, respectively. In the second and third bottom layers in the encoder, subsampling was performed to reduce the length of utterance, yielding the length $T/4$. For MHA/MHD, we set the number of heads to four. For HMHD, we used two kind of settings: (1) dot-product attention + additive attention + location-based attention + coverage mechanism attention (Dot+Add+Loc+Cov), and (2) two location-based attentions + two coverage mechanism attentions ($2 \times \text{Loc} + 2 \times \text{Cov}$). The number of distinct output characters was 3,315 including Kanji, Hiragana, Katakana, alphabets, Arabic number and sos/eos symbols. In decoding, we used beam search algorithm [10] with beam size 20. We manually set maximum and minimum lengths of the output sequence to 0.1 and 0.5 times the length of the subsampled input sequence, respectively, and the length penalty to 0.1 times the length of the output sequence.

Table 1: Experimental conditions.

# training	445,068 utterances (581 hours)
# evaluation (task 1)	1,288 utterances (1.9 hours)
# evaluation (task 2)	1,305 utterances (2.0 hours)
# evaluation (task 3)	1,389 utterances (1.3 hours)
Sampling rate	16,000 Hz
Window size	25 ms
Shift size	10 ms
Encoder type	BLSTMP
# encoder layers	6
# encoder units	320
# projection units	320
Decoder type	LSTM
# decoder layers	1
# decoder units	320
# heads in MHA	4
# filter in location att.	10
Filter size in location att.	100
Learning rate	1.0
Initialization	Uniform $[-0.1, 0.1]$
Gradient clipping norm	5
Batch size	30
Maximum epoch	15
Optimization method	AdaDelta [20]
AdaDelta ρ	0.95
AdaDelta ϵ	10^{-8}
AdaDelta ϵ decay rate	10^{-2}
Beam size	20
Maximum length	0.5
Minimum length	0.1

All of the networks were trained using end-to-end speech processing toolkit ESPnet [24] with a single GPU (Titan X pascal). Character error rate (CER) was used as a metric. The detail of experimental condition is shown in Table 1.

Experimental results are shown in Table 2. First, we focus on the results of the conventional methods. Basically, it is known that location-based attention yields better performance than additive attention [11]. However, in the case of Japanese sentence, its length is much shorter than that of English sentence, which makes the use of location-based attention less effective. In most of the cases, the use of MHA brings the improvement of the recognition performance. Next, we focus on the effectiveness of our proposed MHD architecture. By comparing with the MHA-Loc, MHD-Loc (proposed method) improved the performance in Tasks 1 and 2, while we observed the degradation in Task 3. However, the heterogeneous extension (HMHD), as introduced in Section 3.3, brings the further improvement for the performance of MHD, achieving the best performance among all of the methods for all test sets.

Finally, Figure 3 shows the alignment information of each head of HMHD ($2 \times \text{Loc} + 2 \times \text{Cov}$), which was obtained by visualizing the attention weights. Interestingly, the alignments of the right and left ends seem to capture more abstracted dynamics of speech, while the rest of two alignments behave like normal alignments obtained by a standard attention mechanism. Thus, we can see that the attention weights of each head have a

Table 2: Experimental results.

	CER [%]		
	Task 1	Task 2	Task 3
Dot	12.7	9.8	10.7
Add	11.1	8.4	9.0
Loc	11.7	8.8	10.2
MHA-Dot	11.6	8.5	9.3
MHA-Add	10.7	8.2	9.1
MHA-Loc	11.5	8.6	9.0
MHD-Loc	11.0	8.4	9.5
HMHD (Dot+Add+Loc+Cov)	11.0	8.3	9.0
HMHD ($2 \times \text{Loc} + 2 \times \text{Cov}$)	10.4	7.7	8.9

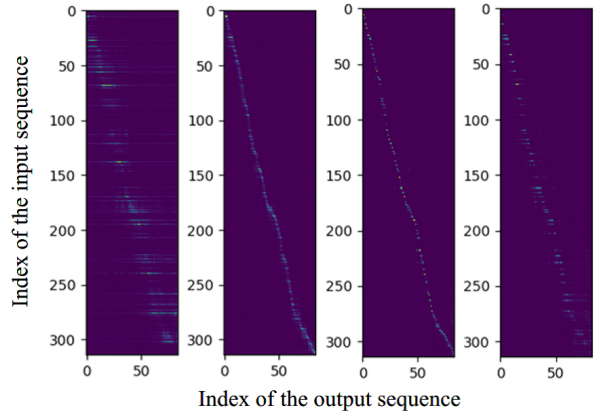


Figure 3: Attention weights of each head. Two left figures represent the attention weights of the location-based attention, and the remaining figures represent that of the coverage mechanism attention.

different tendency, and it supports our hypothesis that HMHD can capture different speech/linguistic contexts within its framework.

5. Conclusions

In this paper, we proposed a new network architecture called multi-head decoder for end-to-end speech recognition as an extension of a multi-head attention model. Instead of the integration in the attention level, our proposed method utilized multiple decoders for each attention and integrated their outputs to generate a final output. Furthermore, in order to make each head to capture the different modalities, we used different attention functions for each head. To evaluate the effectiveness of our proposed method, we conducted an experimental evaluation using Corpus of Spontaneous Japanese. Experimental results demonstrated that our proposed methods outperformed the conventional methods such as location-based and multi-head attention models, and that it could capture different speech/linguistic contexts within the attention-based encoder-decoder framework.

In the future work, we will combine the multi-head decoder architecture with Joint CTC/Attention architecture [11], and evaluate the performance using other databases.

6. References

- [1] F. Jelinek, "Continuous speech recognition by statistical methods," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.
- [2] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.
- [3] 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.
- [4] 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.
- [5] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.
- [6] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition," *arXiv preprint arXiv:1610.09975*, 2016.
- [7] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.
- [8] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.
- [9] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
- [11] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.
- [12] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," *arXiv preprint arXiv:1712.01769*, 2017.
- [13] 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. 6000–6010.
- [14] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.
- [15] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.
- [16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [17] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.
- [18] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
- [19] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.
- [20] M. D. Zeiler, "Adadelata: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.
- [21] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese." in *LREC*. Citeseer, 2000.
- [22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.
- [23] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth annual conference of the international speech communication association*, 2014.
- [24] "ESPnet," <https://github.com/espnet/espnet>.