Interspeech 2018
2-6 September 2018, Hyderabad

Encoder Transfer for Attention-based Acoustic-to-word Speech Recognition

Sei Ueno[1,2], *Takafumi Moriya*[1], *Masato Mimura*[2], *Shinsuke Sakai*[2], *Yusuke Shinohara*[1], *Yoshikazu Yamaguchi*[1], *Yushi Aono*[1], *Tatsuya Kawahara*[2]

[1]NTT Media Intelligence Laboratories, NTT Corporation, Japan
[2]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto, Japan

ueno@sap.ist.i.kyoto-u.ac.jp, takafumi.moriya.nd@hco.ntt.co.jp

Abstract

Acoustic-to-word speech recognition based on attention-based encoder-decoder models achieves better accuracies with much lower latency than the conventional speech recognition systems. However, acoustic-to-word models require a very large amount of training data and it is difficult to prepare one for a new domain such as elderly speech. To address the problem, we propose domain adaptation based on transfer learning with layer freezing. Layer freezing first pre-trains a network with the source domain data, and then a part of parameters is re-trained for the target domain while the rest is fixed. In the attention-based acoustic-to-word model, the encoder part is frozen to maintain the generality, and only the decoder part is re-trained to adapt to the target domain. This substantially allows for adaptation of the latent linguistic capability of the decoder to the target domain. Using a large-scale Japanese spontaneous speech corpus as source, the proposed method is applied to three target domains: a call center task and two voice search tasks by adults and by elderly. The models trained with the proposed method achieved better accuracy than the baseline models, which are trained from scratch or entirely re-trained with the target domain.

Index Terms: End-to-end speech recognition, Attention-based encoder-decoder model, Adaptation, Transfer learning

1. Introduction

Deep neural networks (DNNs) have drastically improved the performance of automatic speech recognition (ASR). Conventional ASR systems adopt DNN for acoustic modeling with Hidden Markov Model (HMM) and RNN for language modeling [1, 2]. However, these conventional ASR systems have very complicated architectures. Moreover, decoding multiple times with large language models need a large computational cost and latency.

On the other hand, end-to-end speech recognition, which maps acoustic features into a target symbol sequence, has been investigated intensively. Acoustic-to-word end-to-end speech recognition, which directly maps acoustic features into a word sequence, does not need a pronunciation dictionary or a language model [3, 4, 5, 6, 7]. Therefore, the acoustic-to-word model has very low runtime latency. In a previous work [8], we showed the attention-based encoder-decoder model outperforms the CTC-based model.

The acoustic-to-word model has a drawback with regard to adaptation. The end-to-end system requires a huge amount of labeled speech data for training a number of parameters [4]. Typically we cannot assume such a large data set for a new domain to train or re-train the model. Especially, the acoustic-to-word model has a problem when adding word entries for a new domain. In this paper, we investigate an efficient transfer learning approach for acoustic-to-word model to adapt to a new target domain of low resource.

Transfer learning [9, 10, 11, 12] is used for solving low resource scenario and also for model adaptation. Transfer learning utilizes a model pre-trained with a source domain for improving the performance of a target domain. We cannot simply use a popular domain adaptation approach such as fine-tuning a pre-trained model because the acoustic-to-word model structure must be revised to add output nodes corresponding to new word entries. In this paper, we propose transfer learning with layer freezing for the attention-based acoustic-to-word encoder-decoder model [12]. The layer freezing approach is defined as the following three steps: Firstly, both the encoder and decoder networks are trained with the source domain. Secondly, the decoder network is replaced with the target domain. Finally, the parameters of the decoder network is re-trained using the target domain data while freezing the encoder network. We expect that the approach can effectively and efficiently adapt the model to the target domain of low resource.

The proposed method is applied to three target domains of real applications. They are a call center dialogue task and two voice search tasks which are spoken by adults and elderly. We compare the proposed approach with three baselines: the model trained with the source domain data, the one trained with the target data from scratch, and the one trained by transfer learning without layer freezing that trains the newly attached decoder as well as the pre-trained encoder parameters.

In Section 2, a review on attention-based acoustic-to-word model is given. The proposed layer freezing method is described in Section 3, and experimental evaluation are presented in Section 4, before conclusions in Section 5.

2. Acoustic-to-word End-to-End Speech Recognition

End-to-end speech recognition learns the mapping between speech and symbol sequences. There are two major approaches: one is connectionist temporal classification (CTC) [13, 14], and the other is attention-based encoder-decoder model [15, 16, 17, 18, 19, 20]. CTC allows a ”blank” symbol and repeated symbols. It marginalizes and condenses all possible frame-wise output symbol sequences. The attention-based encoder-decoder model first encodes the input into a frame-wise distributed representation with one RNN such as LSTM, and then decodes it to a target symbol sequence with another RNN. The LSTM-based decoder predicts the next symbol using a history of previous symbols, thus it substantially includes a language model. As the attention-based model shows better recognition performance than CTC-based model [8], in this paper, we focus on the attention-based model.

10.21437/Interspeech.2018-1424

2.1. Attention-based encoder-decoder model

2.1.1. Framework

The attention-based encoder-decoder model is a seq2seq model [15, 16, 17, 18, 19, 20]. This architecture has two distinct networks. One is an encoder network, which maps an acoustic feature sequence to a distributed representation of the same length T. Using this intermediate information, the decoder network predicts a symbol sequence whose length is L ($L \leq T$). The decoder network uses only a relevant portion of the encoded sequential representation for predicting a symbol at each time step using the attention mechanism. The encoder is implemented with a multi-layer bidirectional RNN such as an LSTM, and the decoder usually consists of a 1-layer of unidirectional RNN followed by a softmax output layer.

The attention-based model is formulated as follows. The encoder transforms an acoustic feature sequence $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_T)$ to intermediate representation vectors $\boldsymbol{H} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_T)$. In the following decoding step, the hidden state (memory) activation of the RNN-based decoder at the *l*-th time step is computed as:

$$\boldsymbol{s}_l = \text{Recurrency}\left(\boldsymbol{s}_{l-1}, \boldsymbol{g}_l, \boldsymbol{y}_{l-1}\right) \tag{1}$$

where $\boldsymbol{g}_l$ and $\boldsymbol{y}_{l-1}$ denote the "glimpse" at the *l*-th target label and the predicted symbol at the previous step. The glimpse $\boldsymbol{g}_l$ is a weighted sum of the encoder output sequence as:

$$\boldsymbol{g}_l = \sum_t \alpha_{l,t} \boldsymbol{h}_t \tag{2}$$

where $\alpha_{l,t}$ is an attention weight of $\boldsymbol{h}_t$. In this paper, we use a content-based attention mechanism formulated as follows:

$$e_{l,t} = \boldsymbol{w}^T \tanh(\boldsymbol{W}\boldsymbol{s}_{l-1} + \boldsymbol{V}\boldsymbol{h}_t + \boldsymbol{U}f_{l,t} + \boldsymbol{b}) \tag{3}$$

$$\boldsymbol{f}_l = \boldsymbol{F} * \boldsymbol{\alpha}_{l-1} \tag{4}$$

$$\alpha_{l,t} = \exp(e_{l,t}) / \sum_{t'=1}^{T} \exp(e_{l,t'}) \tag{5}$$

where $*$ denotes a 1-dimensional convolution. Using $\boldsymbol{g}_l$ and $\boldsymbol{s}_{l-1}$, the decoder predicts the next symbol $\boldsymbol{y}_l$ as:

$$\boldsymbol{y}_l \sim \text{Generate}\left(\boldsymbol{s}_{l-1}, \boldsymbol{g}_l\right) \tag{6}$$

where the Generate function is implemented as:

$$\boldsymbol{R}\tanh\left(\boldsymbol{P}\boldsymbol{s}_{l-1} + \boldsymbol{Q}\boldsymbol{g}_l\right) \tag{7}$$

The objective function for training the attention models is a cross entropy loss calculated between the predicted symbol sequences and the target oracle symbol sequences. In this paper, we prepare special symbols for denoting start-of-sentence ($\langle$sos$\rangle$) and end-of-sentence ($\langle$eos$\rangle$). The decoder completes the process when an $\langle$eos$\rangle$ symbol is output.

2.1.2. Label smoothing

Label smoothing [21] is a regularization method to prevent the model from the over-fitting. When calculating the cross-entropy, we do not simply use the grand-truth label 1.0, but discount it and assign a small value to all other symbols with a uniform distribution. In this paper, we followed the same design as [21].

2.2. Acoustic-to-word model

The conventional end-to-end systems mostly base on subword units, such as phones, syllables and characters. They still need a pronunciation lexicon and a language model for transducing into a word sequence. Recently, word-level end-to-end speech recognition has been investigated. The remarkable advantages of acoustic-to-word model include a very simple architecture and much fast decoding speed.

However, the acoustic-to-word model has a serious problem in training especially with a small amount of data. The number of output nodes in the acoustic-to-word model is much larger than that of subword-based models. Moreover, the occurrence distribution of word entries is much unbalanced than that of subword entries. Therefore, the acoustic-to-word model needs a huge amount of training data. It is usually difficult to prepare a sufficient data set for new domains and particular user populations.

Another problem with the acoustic-to-word model is that it cannot recognize words that are not included in the training data, and moreover, it cannot add or change word entries unlike subword-based systems. Therefore, there are inevitably many out-of-vocabulary words in the acoustic-to-word model and the recognition of these unknown words is very difficult. In many speech recognition applications such as voice search, there are many new words such as named entities. This would be a serious problem in deployment of the acoustic-to-word model. In [8, 22], in order to handle out-of-vocabulary words, the character-level model is used when the word-level model detects an unknown word, but perfect recovery of unknown words is not easy.

3. Encoder Transfer of acoustic-to-word attention model

To improve the portability of the acoustic-to-word model, we propose adaptation based on the transfer learning framework. Transfer learning has been investigated not only in domain adaptation but also cross-lingual model learning, in which a baseline model is trained with a rich-resource language and then transferred to low-resource languages. A simple method is to just transfer the baseline model to a new language and fine-tune it using its data set of a small size [10]. In this case, the entire model parameters are updated though some regularization can be applied. Instead of updating all parameters with a small amount of training data, we can freeze some layers (typically the lower part) of the network which can be shared across languages, and update the language-dependent layers (typically the upper part including the output layer) of the network [12]. This method is referred to as layer freezing.

Inspired by these studies, we propose encoder transfer of the attention-based acoustic-to-word model. We firstly train a source domain model whose training data is large enough to converge. After that, we transfer the encoder network of the source domain model to the target. In training the target domain model, we do not update the transferred encoder. We presume that the encoder learns common intermediate representation of the acoustic features independent of the domains since the acoustic features cover many speakers and recording environment. We expect that the model maintains generality and should be frozen rather than updating with a small amount of new data. On the other hand, the language information strongly depends on the domain. The vocabulary is different and there are many new lexical entries which occur on the specific task.

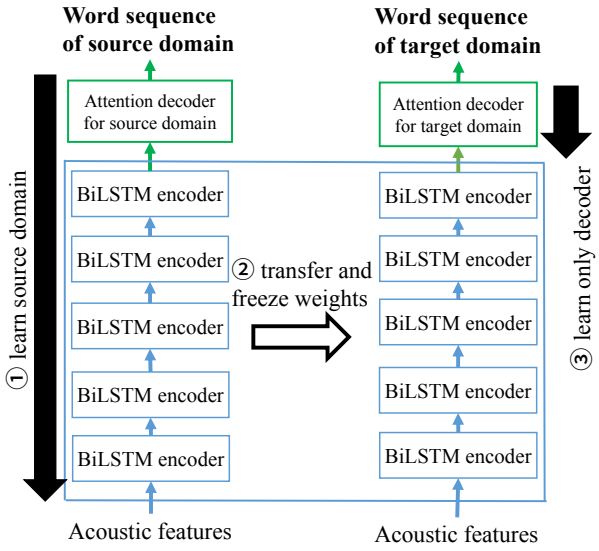


Figure 1: *The architecture of our proposed transfer learning with layer freezing. (1) we learn both encoder and decoder for the source domain, (2) we transfer the encoder of the source domain to the target domain, (3) we learn the target domain model by updating decoder only.*

Therefore, we need to learn the decoder which maps the encoded information to a word sequence.

In general, transfer learning can be used for a target domain which cannot use the same output layer of the source domain such as cross-lingual tasks. In the encoder-decoder model, the decoder is usually different from one domain to another since a set of lexical entries are different and the acoustic-to-word model cannot train the words which do not appear in the training data. Therefore, we always need to train the decoder. Figure 1 shows our proposed framework of transfer learning with layer freezing.

We also extend our method by re-training the seed model using both source domain data and target domain data. In this scenario, the entire network including the encoder and the decoder are first trained with the combined data set of the source domain and the target domain, and then the decoder part is further fine-tuned with the target domain data. Note that this method can be applied only when we can access to the training data set, and the entire re-training takes much time.

4. Experimental evaluations

4.1. Source domain and target domains

We learned the source domain model using a standard large-scale Japanese corpus: the Corpus of Spontaneous Japanese (CSJ) [23]. CSJ includes 525 hours of oral presentations. There are three real recorded tasks as target domains: call center dialogue and voice search by elderly and by adults. All target domains are Japanese and much different from the source domain. Table 1 shows OOV rates when the source and target domain models are used. The sizes of the target domain training data are different (call center has 44 hours, voice search by elderly has 93 hours, and voice search by adults has 112 hours).

Table 1: *The out-of-vocabulary (OOV) rate (%). The "target domain" row shows OOV rates when the domain-only model is used.*

	call center	voice search elderly	voice search adult
CSJ (source domain)	5.52%	10.38%	10.39%
target domain	1.37%	1.49%	1.67%

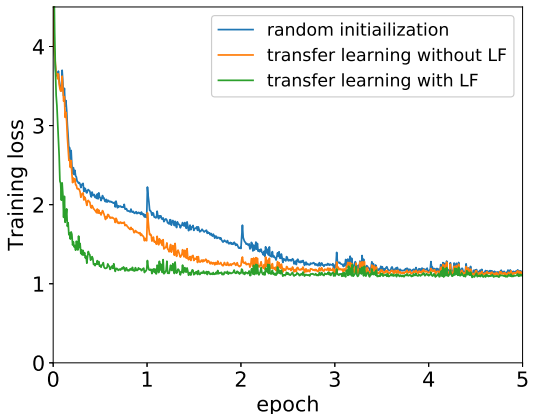


Figure 2: *The learning curve of losses of three methods on voice search by adult. These losses are normalized by the minibatch size.*

4.2. System configuration

We used a 120-dimensional feature vector of 40-channel log mel-scale filter bank (lmfb) outputs. Non-overlapping frame stacking [3] was applied to these features, in which we stack and skip three frames to make a new super-frame.

The encoder has five layers of bidirectional LSTMs with 320 cells. The dropout [24] rate was 0.2 for each BiLSTM layer. The attention-based decoder consists of one-layer LSTM with 320 cells, a hidden layer with 320 tanh nodes, and a softmax output layer for word entries. All networks apply label smoothing described in Section 2.1.2.

We optimized the parameters with Adam [25]. We set gradient clipping with a threshold of 5.0. The minibatch size was 50 at first and decreased to 40 and 10 by input time frame sizes. All parameters were initialized with random values with a uniform distribution with a range (-0.1, 0.1). To prevent from slow convergence, the input data were sorted by the length of frames before creating minibatches. We used PyTorch [26] to train the networks. In decoding with the acoustic-to-word attention model, we applied a simple beam search with the beam width of 4. We compared our proposed method with a random initialization and transfer learning without layer freezing. In these cases, we trained both encoder and decoder networks.

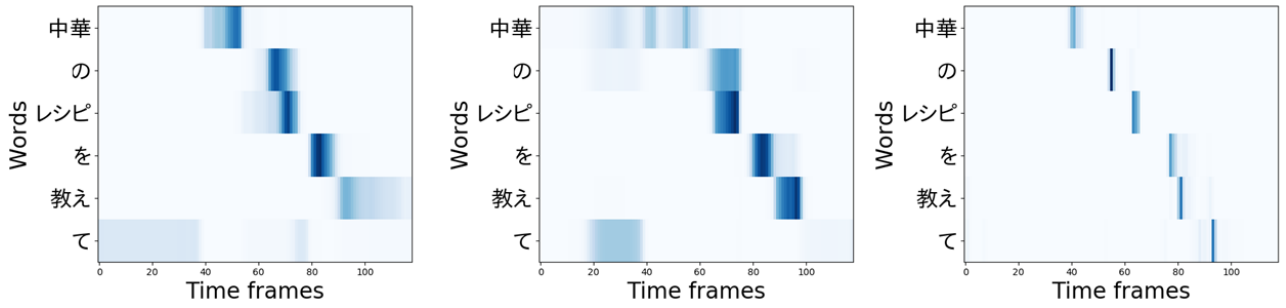
We also built a DNN-HMM hybrid system using each data set not including the CSJ for reference. The DNN-HMM system has six hidden layers with 2048 sigmoidal nodes and a softmax output layer with 3072 nodes. The language model was 3-gram and trained using a 1M Japanese web text corpus. We used VoiceRex [27, 28] for decoding with this system.

4.3. Results

Table 2 shows the ASR performance in word error rate (WER) for three target domains. At first, we observed that source domain model cannot recognize all target domains very well

Table 2: *The training curve of performance (WER (%)) on target domains: call center domain and voice search domains. The amount of training is given for each domain. LF means layer freezing. "Source domain + target domain" means the model trained by mixing the source domain and the target domain. "+ with LF" means our proposed transfer learning trained with the target domain.*

model	call center (44 hours)	voice search by elderly (93 hours)	voice search by adult (112 hours)
DNN-HMM	27.41	12.37	9.78
random initialization	29.29	11.11	10.07
source domain model	39.91	65.63	65.92
transfer learning without LF	34.23	11.46	9.18
transfer learning with LF [proposed]	**18.77**	**10.97**	**8.22**
source domain + target domain	16.65	9.70	7.61
+ with LF [proposed]	**14.59**	10.00	8.03



(a) Random initialization (b) Transfer learning without LF (c) Transfer Learning with LF

Figure 3: *Attention weights on some data of voice search by adults. The column is the ground truth word sequence and arranged from top to bottom. The row is the time frames and arranged from left to right. Dark blue means a word strongly corresponds to the time frame.*

because the word entries of source domain model are different from those of the target domains as shown in Table 1. Especially, voice search tasks include many names of actors/actresses and TV programs which CSJ does not cover. We also observed that training with random initialization or re-training the entire model do not provide good performance for the call center domain because the training data size is very small (44 hours). The proposed transfer learning with layer freezing effectively improved the performance by using the source encoder as it is. In both voice search tasks, the proposed method achieved better performance than the other three baselines.

In these evaluations, we used the source domain model as a seed model, assuming that we cannot re-train the model using the source domain data. Next, we conduct additional experiments in which we re-train the model by combining the target domain data with the source domain data. This realizes simple domain adaptation though it takes much time. Then, the proposed method is applied to this re-trained model by additional fine-tuning of the decoder network. The results are listed in the lower two rows in Table 2. It is confirmed that the proposed method is still effective for the call center domain, but there is no additional improvement for the voice search domains (there is no significant difference). These results can be explained by the data size of the adaptation domain. When we have 100-hour data for the target domain, simple re-training by combining the training data sets is sufficient. But it is not so common that we can get this scale of data for a new domain.

Figure 2 shows the training curve of the loss function by the three methods. The loss function is cross-entropy. Transfer learning with layer freezing converges much faster than the other methods since the number of updated parameters is very small.

Figure 3 shows an alignment by the attention weights on some data of voice search by adults. The column indicates a ground truth word sequence which means "Tell me a Chinese food recipe". In (a) random initialization and (b) transfer learning without layer freezing, we observe that the last word alignment is backed left. On the other hand, (c) transfer learning with layer freezing has clearly learned a left-to-right alignment. Moreover, the range of (a) and (b) alignment is much wider than that of (c) alignment. This suggests that the model learns the precise mapping from acoustic features to the word sequence.

5. Conclusions

We have proposed transfer learning with layer freezing for attention-based acoustic-to-word encoder-decoder speech recognition. Using layer freezing for encoder transfer, the number of parameters the model should learn in a new domain is significantly reduced. The model training can be done even when the amount of training data is small. In evaluations with three domains applied, the proposed method achieved better accuracy than training with random initialization and re-training without layer freezing. The proposed method is particularly effective the target domain for which has less training data. Since the method is simple and straightforward, it can be easily applied to many new domains.

6. References

[1] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[2] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-l. Lim, B. Roomi, and P. Hall, "English Conversational Telephone Speech Recognition by Humans and Machines," in *Interspeech 2017*, 2017.

[3] 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.

[4] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition," in *Interspeech2017*, 2017, pp. 3707–3711.

[5] L. Lu, X. Zhang, and S. Renals, "On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5060–5064.

[6] K. Audhkhasi, B. Ramabhadran, G. Saon, M. Picheny, and D. Nahamoo, "Direct acoustics-to-word models for English conversational speech recognition," in *Interspeech2017*, 2017, pp. 959–963.

[7] Z. Chen, Q. Liu, H. Li, and K. Yu, "On modular training of neural acoustics-to-word model for lvcsr," *arXiv preprint arXiv:1803.01090*, 2018.

[8] S. Ueno, H. Inaguma, M. Mimura, and T. Kawahara, "Acoustic-to-word attention-based model complemented with character-level CTC-based model," in *Acoustics, Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*, 2018, pp. 5804–5808.

[9] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7304–7308.

[10] A. Das and M. Hasegawa-Johnson, "Cross-lingual transfer learning during supervised training in low resource scenarios," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] M. Suzuki, R. Tachibana, S. Thomas, B. Ramabhadran, and G. Saon, "Domain adaptation of CNN based acoustic models under limited resource settings." in *Interspeech*, 2016, pp. 1588–1592.

[12] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, "Transfer learning for speech recognition on a budget," *arXiv preprint arXiv:1706.00290*, 2017.

[13] A. Graves, S. Fernandez, 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*, 2006, pp. 369–376.

[14] A. Graves and N. Jaitly, "Towards End-To-End speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 1764–1772.

[15] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent NN: first results," in *NIPS: Workshop Deep Learning and Representation Learning Workshop*, 2014.

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

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

[18] 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)*, 2016, pp. 4945–4949.

[19] 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.

[20] R. Prabhavalkar, T. N. Sainath, B. Li, K. Rao, and N. Jaitly, "An analysis of attention in sequence-to-sequence models," in *Interspeech 2017*, 2017, pp. 3702–3706.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[22] J. Li, G. Ye, R. Zhao, J. Droppo, and Y. Gong, "Acoustic-to-word model without oov," *arXiv preprint arXiv:1711.10136*, 2017.

[23] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese." in *in International Conference on Language Resources and Evaluation (LREC)*, 2000, pp. 947–9520.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, pp. 1929–1958, 2014.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, pp. 1–15, 2014.

[26] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[27] H. Masataki, D. Shibata, Y. Nakazawa, S. Kobashikawa, A. Ogawa, and K. Ohtsuki, "VoiceRex - spontaneous speech recognition technology for contact-center conversations," *NTT Technical Review*, vol. 55, no. 1, pp. 22–27, 2007.

[28] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 4, pp. 1352–1365, 2007.