A state of the art end-to-end speech recognition algorithm with a homogenous structure.

Anonymous Author(s)
Affiliation
Address
email

Abstract

TODO(chanw.com) Revise the abstract. This paper proposes a new end-to-end speech recognition system referred to as Recurrent neural network Sequence Classification (RSC). Recently, end-to-end speech recognition systems have been gaining more attention from researchers and several different structures have been proposed. Typical examples include Connectionist Temporal Classification (CTC), Recurrent Neural Network Transducer (RNN-T), and Sequence-to-Sequence Modeling using the attention mechanism. The CTC-based approaches are using the CTC loss assuming the conditional independence property. Although the structure of CTC is simpler than other end-to-end systems, the performance is worse due to this conditional independence assumption. RNN-T and attention-based approaches have distinct components such as the prediction network and encoder in RNN-T, and the encoder, the decoder, and the attention layer in the attention-based model. Compared to these models, FSC has a homogeneous structure from the bottom to the top layer. LSTM and max-pooling layer are repeated followed by the top softmax layer. The embedded softmax output is fed back as input of each RNN layer. The training starts with flat initialization and alignment is made after every epoch. The loss is the Cross Entropy (CE) loss using this alignment result. In our experimental results, FSC algorithm has shown better performance than more complicated attention-based end-to-end speech recognition system. Another major advantage is training is significantly faster.

1 Introduction

Recently, there has been tremendous improvements in speech recognition systems fueled by advances in deep neural networks [8, 13, 16, 19, 20, 22].

TODO(chanw.com) Talk about the AI speakers.

TODO(chanw.com) Talk about the advancements of the end-to-end systems.

Recently, it has been frequently observed that if sequence-to-sequence end-to-end ASR systems are trained on sufficiently large amounts of acoustic training data, they can outperform conventional HMM-DNN/RNN hybrid systems [6, 1].

Recently, we observed that training with large-scale noisy data generated by a *Room Simulator* [5] improves speech recognition accuracy dramatically. This system has been successfully employed for training acoustic models for Google Home or Google voice search [5].

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

2 Review on sequence-to-sequence speech recognition algorithms

In this section, we review well-known *sequence-to-sequence* algorithms used in speech recognition. [7, 9, 12, 17]. A *sequence-to-sequence* speech recognizers maps a sequence of input acoustic features into a sequence of graphemes [7, 4] or words [17]. We denote a sequence of input acoustic feature vectors by $\mathcal{X}_0^{M-1}$ and a sequence of target labels in *one-hot vector* representation by $\mathcal{Y}_0^{L-1}$ as shown below:

$$\mathcal{X}_0^{M-1} = \left\{ \vec{x}[m] \middle| 0 \leq m \leq M-1,\ \vec{x}[m] \in \mathbb{R}^d \right\}, \tag{1a}$$

$$\mathcal{Y}_0^{L-1} = \left\{ \vec{y}_l \middle| 0 \leq l \leq L-1,\ \vec{y}_l \in \mathbb{V} \right\}, \tag{1b}$$

where M is the number of frames in the input feature sequence, d in (1a) is the dimension of the input feature. L is the number of labels in the output target sequence, and $\mathbb{V}$ in (1b) is the set of output labels, which may be graphemes [7, 4], subword units [24, 6], and words [21]. Note that the sequence index in (1a) is a frame index, whereas the sequence index in (1b) is label index. Depending on whether the neural network generates the inference output for every frame or not, there are following categories:

- **Label-synchronous inference** In this category, the neural network generates the inference output $\widehat{\vec{y}_l}$ only when a new label is expected. The loss function $\mathbb{L}\left(\mathcal{X}_0^{M-1}, \mathcal{Y}_0^{L-1}\right)$ is defined as follows:

$$\mathbb{L}\left(\mathcal{X}_0^{M-1}, \mathcal{Y}_0^{L-1}\right) = \sum_{l=0}^{L-1} y_l \odot \log\left(\circ\, \widehat{\vec{y}_l}\right) \tag{2}$$

  where $\odot$ and $\log(\circ)$ denote element-wise product (a.k.a *Hadamard product*) and element-wise log, respectively.

- **Frame-synchronous inference** In this category, from the original target label in (1b), we either explicitly or implicitly obtain the label boundaries. In the framewise Cross Entropy (CE) training [14, 5, 3], *Viterbi alignment* [15, 18] is usually employed. Using these alignment algorithms, we obtain corresponding *frame-synchronous* label sequences $\vec{z}[m] \in \mathbb{V}$ for each frame index m. This latent label sequences are often called a *path*.

$$\mathcal{Z}_0^{M-1} = \left\{ \vec{z}[m] \middle| 0 \leq m \leq M-1,\ \vec{z}[m] \in \mathbb{V} \right\}, \tag{3}$$

  Fig. 2 shows one such example. Such an alignment is usually called a *path*.

$$l = \sum_{l=0}^{l=L-1} y_l \log\left(\hat{y}_l\right) \tag{4}$$

Moreover, in almost all the large speech training database, the temporal boundray of each label is not marked. In the attention-based model described in Sec. 2.3 and shown in Fig. 1c, lack of the temporal alignment information in the training database is not a problem, since the *decoder* portion of the attention-model directly generates the output label sequences. The loss function is also defined for each labels. ======= A sequence-to-sequence model maps a sequence of input acoustic features into a sequence of graphemes or words [17]. We denote a sequence of input acoustic feature vectors by $\mathcal{X}_0^{M-1}$ and a sequence of target labels by $\mathcal{Y}_0^{L-1}$:

$$\mathcal{X}_0 = \left\{ \vec{x}[0],\ \vec{x}[1],\ \cdots,\ \vec{x}[M-1] \right\}, \tag{5a}$$

$$\mathcal{Y}_0 = \left\{ y_0,\ y_1,\ \cdots,\ y_{L-1} \right\}, \tag{5b}$$

where M is the number of frames in the input feature sequence and L is the number of labels in the output target sequence. In Fig. 1a $\sim$ 1c, we show block diagrams of CTC [10], RNN-T [9, 11], and the attention-based model [7, 4] respectively. CTC in Fig. 1a and RNN-T in Fig. 1b are *frame-synchronous*, which means that the ASR system generates the output target for each input frame.

2.1 Connectionist Temporal Classification

Fig. 1a shows the entire structure of the Connectionist Temporal Classification (CTC) [10]. In CTC, we use the following CTC loss [10, 12]:

$$P\left(\mathcal{Y}_0^{L-1}|\mathcal{X}_0^{M-1}\right) = \sum_{\mathcal{Z}_0^{M-1}\in\mathcal{A}_{CTC}\left(\mathcal{X}_0^{M-1},\mathcal{Y}_0^{L-1}\right)} \Pi_{m=0}^{M-1} P\left(\vec{z}[m]|\mathcal{X}_0^{M-1}\right), \tag{6}$$

where $\mathcal{A}_{CTC}\left(\mathcal{X}_0^{M-1},\mathcal{Y}_0^{L-1}\right)$ correspond to frame-level alignments of length M such that removing blanks and repeated symbols from $\mathcal{Z}_0^{M-1}$ yields $\mathcal{Y}_0^{L-1}$.

2.2 RNN-Transducer

Fig. 1b shows the entire structure of the RNN-T structure. The loss function used in RNN-T is given as follows:

$$P\left(\mathcal{Y}_0^{L-1}|\mathcal{X}_0^{M-1}\right) = \sum_{\mathcal{Z}_0^{M-1}\in\mathcal{A}_{CTC}\left(\mathcal{X}_0^{M-1},\mathcal{Y}_0^{L-1}\right)} \Pi_{m=0}^{M-1} P\left(\vec{z}[m]|\mathcal{X}_0^{M-1}\right), \tag{7}$$

2.3 Attention-based model

3 Feedback Sequence Classifier

3.1 Neural Network Structure

Fig. 1d shows the entire structure of the FSC. As shown in this figure, LSTM layers and max-pool layers are interleaved from the bottom layer up to the third max-pool layer.

In Fig. 1d, speech features $\vec{x}[m]$ are sampled at every 10 *ms*, which is the usual frame period in speech recognition at every 10 *ms*, which is the usual speech feature frame rate in speech recognition [15, 2], Using three two-to-one max-pool layers, $\vec{z}[p]$ has a frame period of 80 *ms*. The relationship between the temporal index p in $\vec{z}[p]$ and m in $\vec{x}[m]$ is given by:

$$p = \lceil m/8 \rceil \tag{8}$$

$$\mathcal{Z}_0^{M-1} = \left\{ \vec{z}[m] \middle| 0 \leq m \leq M-1,\ \vec{z}[m] \in \mathbb{V} \right\}, \tag{9}$$

The reason for using the one dimensional max-pool layer is to make the neural network output have a similar rate to the that of each *phone* in utterances. Note that a *phone* is any distinct speech sound serving as a phonetic unit.

It has been observed that the average phone duration is between 50 *ms* ~100 *ms* [23, 25].

3.2 Training of the Feedback Sequence Classifier

In this section, we describe how to train FSC. As mentioned in Sec. , the FSC operates in the frame -synchronous way. Since we do not have the alignment information

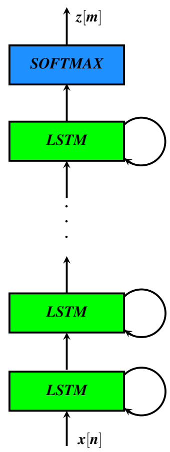
The first step is making an In this section, we describe how to train FSC. The first step is making an »»»> cf67f8b3ff96936469f82b1916daebe02799ed40 alignment using the N-best alignment. The second step is updating the neural network parameters using the Cross Entropy (CE) criterion.

3.3 Training procedure

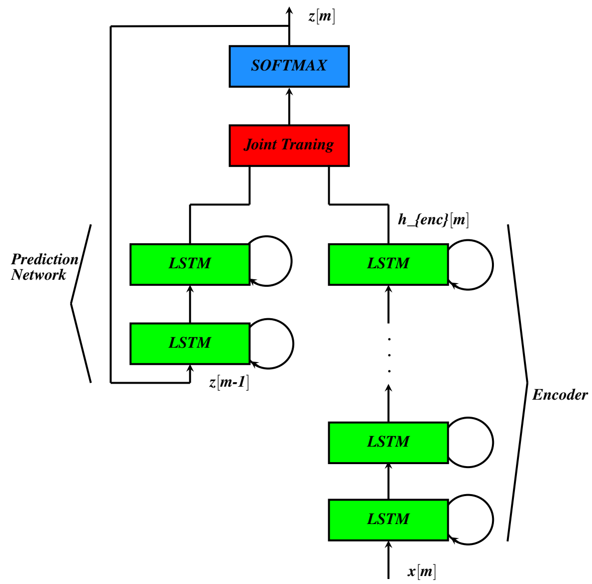
The first step is finding the optimal frame-synchronous sequence $\mathcal{Z}_0^{M-1}$. $\mathcal{Z}$

$\mathcal{Z} = A(\mathcal{X}|\Theta)$

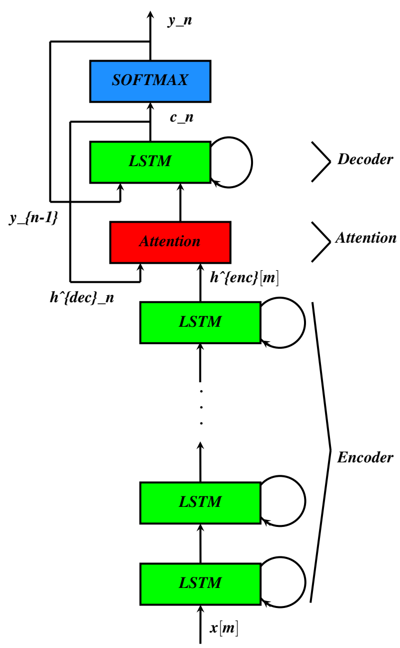
Updates the model.



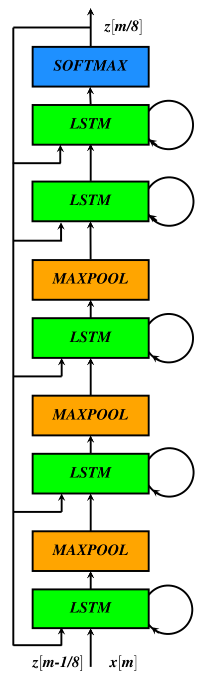
(a) Connectionist Temporal Classification (CTC) Model.



(b) RNN-T Model.



(c) Attention-based Model.



(d) Feedback Sequence Classifier (FSC) Model.

Figure 1: Comparison of block diagrams of different sequence-to-sequence speech recognition approaches. The proposed Feedback Sequence Classifier (FSC) is shown in Fig. 1d.

3.4 Viterbi alignment and N-best alignment

In CTC in Sec. 2.1 and in RNN-T in Sec. 2.2, the forward-backward algorithm is employed to update parameters. TODO(chanw.com) cite HMM tutorial paper.

For example in CTC, the forward algorithm is given by the following equation:

In conventional frame-wise CE-training, Viterbi alignment (a.k.a forced alignment) has been performed to obtain the frame-level acoustic unit boundaries TODO(chanwcom). Even though the Viterbi

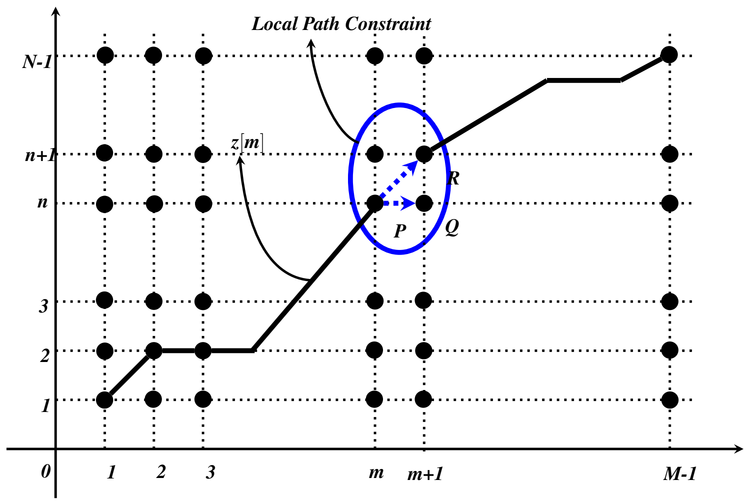


Figure 2: An example of a path $\mathcal{Z}_0^{M-1}$ and the path movement constraint shown by TODO(chanw.com)[Check the latex arrow symbol] $P->Q$ and $P->R$.

algorithm has advantages in simplicity and efficiency, it is based on the conditional independence assumption.

In TFSC, we propose the following N-best alignment approach rather than the Viterbi alignment algorithm to find the alignment information.

for $m = 0, ..., M-1$ **do**
    for $l = 0, ..., N_b - 1$ **do**
        $\pi^{(l)}[m]$
    end for
end for
if $i \geq maxval$ **then**
    $i \leftarrow 0$
else
    if $i + k \leq maxval$ **then**
        $i \leftarrow i + k$
    end if
end if

4 Experimental Results

Acknowledgments

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Remember that you can go over 8 pages as long as the subsequent ones contain *only* cited references.**

References

[1] Chanwoo Kim, Minkyu Shin, Abhinav Garg, and Dhananjaya N. Gowda. Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system. Sept. 2019 (submitted).

[2] Chanwoo Kim and Richard. M. Stern. Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, pages 1315–1329, July 2016.

[3] B. Li, T. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K-C Sim, R. Weiss, K. Wilson, E. Variani, Chanwoo Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon. Acoustic modeling for Google Home. In *INTERSPEECH-2017*, pages 399–403, Aug. 2017.

[4] William Chan, Navdeep Jaitly, Quoc Le, and Oriol 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)*, pages 4960–4964, March 2016.

[5] Chanwoo Kim, A. Misra, K.K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani. Generation of simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home. In *INTERSPEECH-2017*, pages 379–383, Aug. 2017.

[6] Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778, April 2018.

[7] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 577–585. Curran Associates, Inc., 2015.

[8] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide. Feature learning in deep neural networks - studies on speech recognition tasks. In *Proceedings of the International Conference on Learning Representations*, 2013.

[9] Alex Graves. Sequence transduction with recurrent neural networks. *CoRR*, abs/1211.3711, 2012.

[10] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 369–376, New York, NY, USA, 2006. ACM.

[11] Alex Graves, Abdel rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, May 2013.

[12] Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo yiin Chang, Kanishka Rao, and Alexander Gruenstein. Streaming end-to-end speech recognition for mobile devices. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6381–6385, May 2019.

[13] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, Nov 2012.

[14] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, Nov 2012.

[15] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2001.

[16] M. Seltzer, D. Yu, and Y.-Q. Wang. An investigation of deep neural networks for noise robust speech recognition. In *Int. Conf. Acoust. Speech, and Signal Processing*, pages 7398–7402, 2013.

[17] Rohit Prabhavalkar, Kanishka Rao, Tara N. Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. A comparison of sequence-to-sequence models for speech recognition. In *Proc. Interspeech 2017*, pages 939–943, 2017.

[18] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[19] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, Ananya Misra, and Chanwoo Kim. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):965–979, May 2017.

[20] Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Arun Narayanan, Michiel Bacchiani, Bo Li, Ehsan Variani, Izhak Shafran, Andrew Senior, Kean Chin, Ananya Misra, and Chanwoo Kim.

[21] Hagen Soltau, Hank Liao, and Haşim Sak. Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition. pages 3707–3711, 2017.

[22] Vincent Vanhoucke, Andrew Senior, and Mark Z. Mao. improving the speed of neural networks on cpus. In *Deep Learning and Unsupervised Feature Learning Workshop, NIPS 2011*, 2011.

[23] X. Wang, L. C. W. Pols, and L. F. M Bosh. Integration of context-dependent durational knowledge into hmm-based speech recognition. In *ICSLP-1996*, pages 1073–1076, Oct. 1996.

[24] A. Zeyer, Kazuki Irie, Ralf Schlüter, and H. Ney. Improved training of end-to-end attention models for speech recognition. In *INTERSPEECH-2018*, pages 7–11, 2018.

[25] B. Ziółko and M. Ziółko. Time durations of phonemes in polish language for speech and speaker recognition. In Zygmunt Vetulani, editor, *Human Language Technology. Challenges for Computer Science and Linguistics*, pages 105–114, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.