Conformer-based Hybrid ASR System for Switchboard Dataset

Mohammad Zeineldeen[1,2,*] , **Jingjing Xu**[1,*], Christoph Lüscher[1,2], Wilfried Michel[1,2], Alexander Gerstenberger[1], Ralf Schlüter[1,2], Hermann Ney[1,2]

RWTH Aachen University[1], AppTek GmbH[2]

ICASSP, May, 2022

[*] Equal contribution

Overview

1. Introduction
2. Our Proposed Conformer Acoustic Model
3. Experimental Setup
4. Experimental Results
5. Conclusion & Outlook

Introduction

Hybrid ASR system & Conformer Architecture

- Hybrid neural network (NN)-hidden Markov model (HMM) automatic speech recognition (ASR) systems [Bourlard & Morgan 93] have achieved state-of-the-art performance on different tasks [Zhou & Michel^{+} 20, Lüscher & Beck^{+} 19, Kitza & Golik^{+} 19].

Introduction

Hybrid ASR system & Conformer Architecture

- Hybrid neural network (NN)-hidden Markov model (HMM) automatic speech recognition (ASR) systems [Bourlard & Morgan 93] have achieved state-of-the-art performance on different tasks [Zhou & Michel$^+$ 20, Lüscher & Beck$^+$ 19, Kitza & Golik$^+$ 19].
- Recently, the conformer model [Gulati & Qin$^+$ 20], was proposed and achieved state-of-the-art performance on Librispeech 960h dataset [Panayotov & Chen$^+$ 15].

Introduction

Hybrid ASR system & Conformer Architecture

- Hybrid neural network (NN)-hidden Markov model (HMM) automatic speech recognition (ASR) systems [Bourlard & Morgan 93] have achieved state-of-the-art performance on different tasks [Zhou & Michel+ 20, Lüscher & Beck+ 19, Kitza & Golik+ 19].
- Recently, the conformer model [Gulati & Qin+ 20], was proposed and achieved state-of-the-art performance on Librispeech 960h dataset [Panayotov & Chen+ 15].
- The conformer architecture was investigated for different end-to-end systems such as attention encoder-decoder models [Wang & Sun+ 21, Tüske & Saon+ 21]

Hybrid ASR system & Conformer Architecture

- Hybrid neural network (NN)-hidden Markov model (HMM) automatic speech recognition (ASR) systems [Bourlard & Morgan 93] have achieved state-of-the-art performance on different tasks [Zhou & Michel$^+$ 20, Lüscher & Beck$^+$ 19, Kitza & Golik$^+$ 19].
- Recently, the conformer model [Gulati & Qin$^+$ 20], was proposed and achieved state-of-the-art performance on Librispeech 960h dataset [Panayotov & Chen$^+$ 15].
- The conformer architecture was investigated for different end-to-end systems such as attention encoder-decoder models [Wang & Sun$^+$ 21, Tüske & Saon$^+$ 21]
- Impact of conformer acoustic model for hybrid ASR has not been investigated

Hybrid ASR system & Conformer Architecture

- Hybrid neural network (NN)-hidden Markov model (HMM) automatic speech recognition (ASR) systems [Bourlard & Morgan 93] have achieved state-of-the-art performance on different tasks [Zhou & Michel^{+} 20, Lüscher & Beck^{+} 19, Kitza & Golik^{+} 19].
- Recently, the conformer model [Gulati & Qin^{+} 20], was proposed and achieved state-of-the-art performance on Librispeech 960h dataset [Panayotov & Chen^{+} 15].
- The conformer architecture was investigated for different end-to-end systems such as attention encoder-decoder models [Wang & Sun^{+} 21, Tüske & Saon^{+} 21]
- Impact of conformer acoustic model for hybrid ASR has not been investigated

⇒ **We present and evaluate a competitive conformer-based hybrid model training recipe**

Efficient Training With Time Down-/up-sampling

- The self-attention mechanism requires allocating the whole input batch sequences into memory

Introduction

Efficient Training With Time Down-/up-sampling

- The self-attention mechanism requires allocating the whole input batch sequences into memory
- The time complexity of self-attention mechanism grows quadratically with sequence length

Efficient Training With Time Down-/up-sampling

- The self-attention mechanism requires allocating the whole input batch sequences into memory
- The time complexity of self-attention mechanism grows quadratically with sequence length
- Different time down-sampling techniques were introduced, mainly for end-to-end systems [Chan & Jaitly+ 16, Zeyer & Alkhouli+ 18]

Introduction

Efficient Training With Time Down-/up-sampling

- The self-attention mechanism requires allocating the whole input batch sequences into memory
- The time complexity of self-attention mechanism grows quadratically with sequence length
- Different time down-sampling techniques were introduced, mainly for end-to-end systems [Chan & Jaitly+ 16, Zeyer & Alkhouli+ 18]
- It is not straightforward to apply such down-sampling methods for models trained with frame-wise target alignment

Introduction

Efficient Training With Time Down-/up-sampling

- The self-attention mechanism requires allocating the whole input batch sequences into memory
- The time complexity of self-attention mechanism grows quadratically with sequence length
- Different time down-sampling techniques were introduced, mainly for end-to-end systems [Chan & Jaitly+ 16, Zeyer & Alkhouli+ 18]
- It is not straightforward to apply such down-sampling methods for models trained with frame-wise target alignment

$\Rightarrow$ **We apply time downsampling for efficient training and use transposed convolutions to upsample the output sequence**

Standard Conformer Architecture [Gulati & Qin+ 20]

- One conformer block consists of 3 types of modules: feed-forward (FFN) module, multi-head self-attention (MHSA) module, convolution (Conv) module

- Let x be the input sequence to conformer block i, then the equations of conformer block can be defined:

$$
\begin{aligned}
x_{FFN_1} &= x + \frac{1}{2}\mathrm{FFN}(x) \\
x_{MHSA} &= x_{FFN_1} + \mathrm{MHSA}(x_{FFN_1}) \\
x_{Conv} &= x_{MHSA} + \mathrm{Conv}(x_{MHSA}) \\
x_{FFN_2} &= x_{Conv} + \frac{1}{2}\mathrm{FFN}(x_{Conv}) \\
\mathrm{ConformerBlock}_i &= \mathrm{LayerNorm}(x_{FFN_2})
\end{aligned}
$$

Our Proposed Conformer Acoustic Model

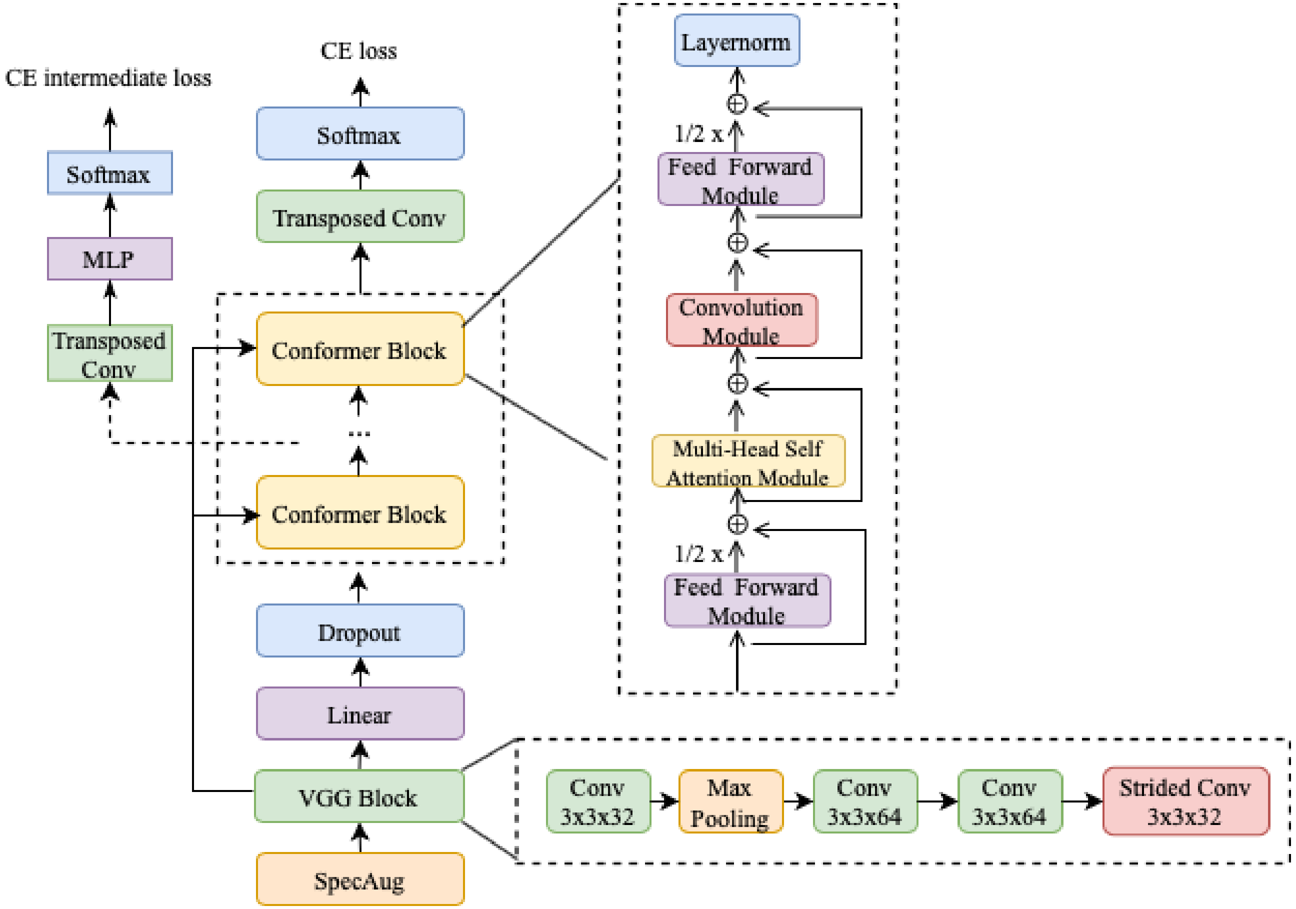
Time down-/up-sampling

- Use a strided convolution as part of the VGG network for downsampling.
- Use transposed convolution to upsample again to the frame-wise target alignment length before output

Intermediate loss[Tjandra & Liu+ 20]: add intermediate losses at different layers

LongSkip: connect the output of the VGG network to the input of each conformer block [Huang & Liu+ 17]

Parameter sharing: share the parameters of intermediate loss layers and the ones of transposed convolution layers



Experimental Setup

Data

- Switchboard 300h dataset (English telephone speech)
- Hub5'00 (Switchboard + CallHome) as development set and Hub5'01 as test set.

Baseline

- Input: 40-dimensional Gammatone features
- Output units: 9001 state-tied CART (Classification and Regression Tree) labels
- Target: frame-wise alignment generated from a HMM-GMM system
- Model size: 12 conformer blocks
- Model dimension: attention dimension of each MHSA module is 512 with 8 attention heads, dimension of the feed-forward module is 2048
- Regularization: dropout with 10%, focal loss with factor 2

Experimental Setup

Language Model

- Use 4-gram count-based language model (LM) and LSTM LM in single pass decoding
- Use transformer (Trafo) LM for rescoring

LM	PPL (word-level) on Hub5'00
4-gram	79.5
LSTM	51.3
Transformer	48.1

Sequence Discriminative Training

- Lattice-based state-level minimum Bayes risk (sMBR) criterion
- Lattices generated using a bigram LM
- sMBR loss scale 0.9 and CE loss scale 0.1

Experimental Results

Kernal size & Number of Conformer Blocks (with 4-gram LM)

- Kernel size has a significant effect on WER
- Using smaller kernel size for depth-wise convolution is better
- We gain performance as we use deeper network

• Kernel size for depth-wise convolution

Kernel size	WER [%]		
	Hub5'00		
	SWB	CH	Total
6	8.4	17.1	12.8
8	**8.1**	**16.8**	**12.5**
16	8.2	17.6	12.9
32	8.4	18.0	13.2

• Number of Conformer blocks comparison

L	Params. [M]	WER [%]		
		Hub5'00		
		SWB	CH	Total
6	42	8.5	18.0	13.3
8	59	**8.1**	17.3	12.7
12	88	**8.1**	**16.8**	**12.5**

Time Downsampling Factors and Variants (with 4-gram LM)

- Set the filter size and the stride of the transposed convolution as time reduction factor
- Choose down-samping factor 3 by considering tradeoff between speed and performance
- Strided convolution applied at the end of the VGG network works best

- Time downsampling factor comparison

Factor	Train time [h]	WER [%] Hub5'00 SWB	CH	Total
2	1.28	8.3	16.4	12.4
3	0.92	8.1	16.8	12.5
4	0.86	8.4	17.9	13.2
5	0.73	8.7	18.6	13.7

- Time downsampling variants comparison

Method	WER [%] Hub5'00 SWB	CH	Total
BLSTM+maxpool	8.2	17.0	12.7
VGG-layer2	8.4	17.7	13.1
VGG-layer4	**8.1**	**16.8**	**12.5**

* VGG-layerX refers to strided convolution as X^{th} layer of VGG network, BLSTM+maxpool refers to one BLSTM layer with 512 units followed by time max-pooling layer

Ablation Study of Training Methods (with 4-gram LM)

- SpecAugment is the most important method giving 20% relative improvement
- Using intermediate loss is important for better convergence and gives 7% relative improvement in WER
- Sharing parameters between transposed convolutions helps
- Other training methods have marginal improvements

Training method	WER [%]		
	Hub5'00		
	SWB	CH	Total
Baseline	**8.1**	**16.8**	**12.5**
- SpecAugment	9.8	21.5	15.7
- Intermediate loss	8.9	18.1	13.5
- Share transp. conv params.	8.5	17.3	12.9
- LongSkip	**8.1**	17.2	12.7
- Focal Loss	**8.1**	17.0	12.6
+ Share MLP params.	8.2	16.9	**12.5**

Comparison between Conformer and BLSTM AM (with 4-gram LM)

- The BLSTM-based model consists of 6 BLSTM layers following a well-optimized setup as here [Kitza & Golik+ 19]
- With comparable number of parameters, conformer AM outperforms BLSTM AM by around 9% relative

AM	LSTM dim.	Params. [M]	Hub5'00
BLSTM	500	41	14.2
	600	57	13.8
	700	76	13.8
	800	96	13.7
	1000	146	13.3
Conformer	-	88	**12.5**

Overall Results

- 8.5% relatively better on Hub5'00 compared to BLSTM hybrid system with LSTM LM
- Outperforms a well-trained RNN-T model with much fewer epochs.
- On par with a well-optimized BLSTM attention system [Tüske & Saon+ 20] on Hub5'01 test set
- The state-of-the-art conformer attention-based system trains much longer and uses cross-utterance LM

Work	#Epochs	Approach	AM	LM	seq. train	WER [%] Hub 5'00	WER [%] Hub 5'01
[Kitza & Golik+ 19]	-	Hybrid	LSTM	4-gram	yes	13.9	-
				LSTM		11.7	
[Zhou & Berger+ 21]	100	RNN-T	LSTM	LSTM	no	11.5	11.5
				Trafo		11.2	11.2
[Tüske & Saon+ 20]	250	LAS	LSTM	LSTM	no	9.8	10.1
[Tüske & Saon+ 21]	250	LAS	Conf.	-	no	9.9	10.1
				LSTM		8.6	8.5
				Trafo		8.4	8.5
ours	27	Hybrid	Conf.	4-gram	no	12.5	12.1
				LSTM		11.3	10.5
				4-gram	yes	11.9	11.4
				LSTM		10.7	10.1
				Trafo		**10.3**	**9.7**

Conclusion & Outlook

Summary

- For the first time, a training recipe for a conformer-based hybrid model is evaluated
- We combined different training methods from the literature that boosted the word-error-rate
- We applied time down-sampling using strided convolution to speedup training and used transposed convolution as a simple method to upsample again
- We observed SpecAugment and intermediate loss layers are necessary to achieve good performance
- Our model outperforms the BLSTM-based hybrid model significantly

Follow up work

- We extend this training recipe as well as use speaker adaptation to improve the WER 11% relative i.e. from 10.3 to 9.2 on Hub5'00 with Transformer LM [Zeineldeen & Xu+]

Thank you for your attention

Any questions?

References

[Bourlard & Morgan 93] H. A. Bourlard, N. Morgan.
Connectionist Speech Recognition: A Hybrid Approach.
Kluwer Academic Publishers, USA, 1993.

[Chan & Jaitly+ 16] W. Chan, N. Jaitly, Q. Le, O. Vinyals.
Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition.
In *ICASSP*, pp. 4960–4964, May 2016.

[Gulati & Qin+ 20] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang.
Conformer: Convolution-augmented Transformer for Speech Recognition.
In *INTERSPEECH*, pp. 5036–5040, Shanghai, China, Oct. 2020.

[Huang & Liu+ 17] G. Huang, Z. Liu, K. Q. Weinberger.
Densely Connected Convolutional Networks.
In *CVPR*, pp. 2261–2269, Los Alamitos, CA, USA, jul 2017.

References

[Kitza & Golik+ 19] M. Kitza, P. Golik, R. Schlüter, H. Ney.
Cumulative Adaptation for BLSTM Acoustic Models.
In *INTERSPEECH*, pp. 754–758, Graz, Austria, Sept. 2019.

[Lüscher & Beck+ 19] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schlüter, H. Ney.
RWTH ASR Systems for LibriSpeech: Hybrid vs Attention.
In *INTERSPEECH*, pp. 231–235, Graz, Austria, Sept. 2019.

[Panayotov & Chen+ 15] V. Panayotov, G. Chen, D. Povey, S. Khudanpur.
LibriSpeech: An ASR Corpus Based on Public Domain Audio Books.
In *ICASSP*, pp. 5206–5210, South Brisbane, Australia, April 2015.

[Tjandra & Liu+ 20] A. Tjandra, C. Liu, F. Zhang, X. Zhang, Y. Wang, G. Synnaeve, S. Nakamura, G. Zweig.
DEJA-VU: Double Feature Presentation and Iterated Loss in Deep Transformer Networks.
In *ICASSP*, pp. 6899–6903, Barcelona, Spain, May 2020.

References

[Tüske & Saon+ 21] Z. Tüske, G. Saon, B. Kingsbury.
On the Limit of English Conversational Speech Recognition.
CoRR, Vol. abs/2105.00982, May 2021.

[Tüske & Saon+ 20] Z. Tüske, G. Saon, K. Audhkhasi, B. Kingsbury.
Single Headed Attention Based Sequence-to-Sequence Model for State-of-the-Art Results on Switchboard.
In *INTERSPEECH*, pp. 551–555, Shanghai, China, Sept. 2020.

[Wang & Sun+ 21] X. Wang, S. Sun, L. Xie, L. Ma.
Efficient Conformer with Prob-Sparse Attention Mechanism for End-to-EndSpeech Recognition.
CoRR, Vol. abs/2106.09236, June 2021.

[Zeineldeen & Xu+] M. Zeineldeen, J. Xu, C. Lüscher, R. Schlüter, H. Ney.
Improving the Training Recipe for a Robust Conformer-based Hybrid Model.
Submitted to INTERSPEECH 2022.

References

[Zeyer & Alkhouli+ 18] A. Zeyer, T. Alkhouli, H. Ney.
RETURNN as a Generic Flexible Neural Toolkit with Application to Translation and Speech Recognition.
In *Annual Meeting of the Assoc. for Computational Linguistics*, Melbourne, Australia, July 2018.

[Zhou & Berger+ 21] W. Zhou, S. Berger, R. Schlüter, H. Ney.
Phoneme Based Neural Transducer for Large Vocabulary Speech Recognition.
In *ICASSP*, pp. 5644–5648, June 2021.

[Zhou & Michel+ 20] W. Zhou, W. Michel, K. Irie, M. Kitza, R. Schlüter, H. Ney.
The RWTH ASR System for TED-LIUM Release 2: Improving Hybrid HMM with SpecAugment.
In *ICASSP*, pp. 7839–7843, Barcelona, Spain, May 2020.