Lehrstuhl Informatik 6
Human Language Technology and Pattern Recognition

RWTH AACHEN UNIVERSITY

Paper ID: 4938 Session: SPE-50.3

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]

Motivation

Hybrid ASR system & Conformer Architecture

- Success of using conformer architecture for end-to-end ASR system
- Impact of conformer acoustic model for hybrid ASR never investigated

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

Efficient Training with Time Down-sampling

- 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+ 2016][Zeyer+ 2018]
- It is not straightforward to apply such down-sampling methods for models trained with frame-wise target alignment

⇒ We apply time down-sampling for efficient training and use transposed convolutions to upsample the output sequence

Conformer Architecture

Standard Conformer Architecture [Gulati+ 20]

- Conformer block consists of: feed-forward (FFN) module, multi-head self-attention (MHSA) module, convolution (Conv) module
- Let x be the input sequence to conformer block i, equations are

$$x_{FFN_1} = x + \frac{1}{2}\text{FFN}(x)$$

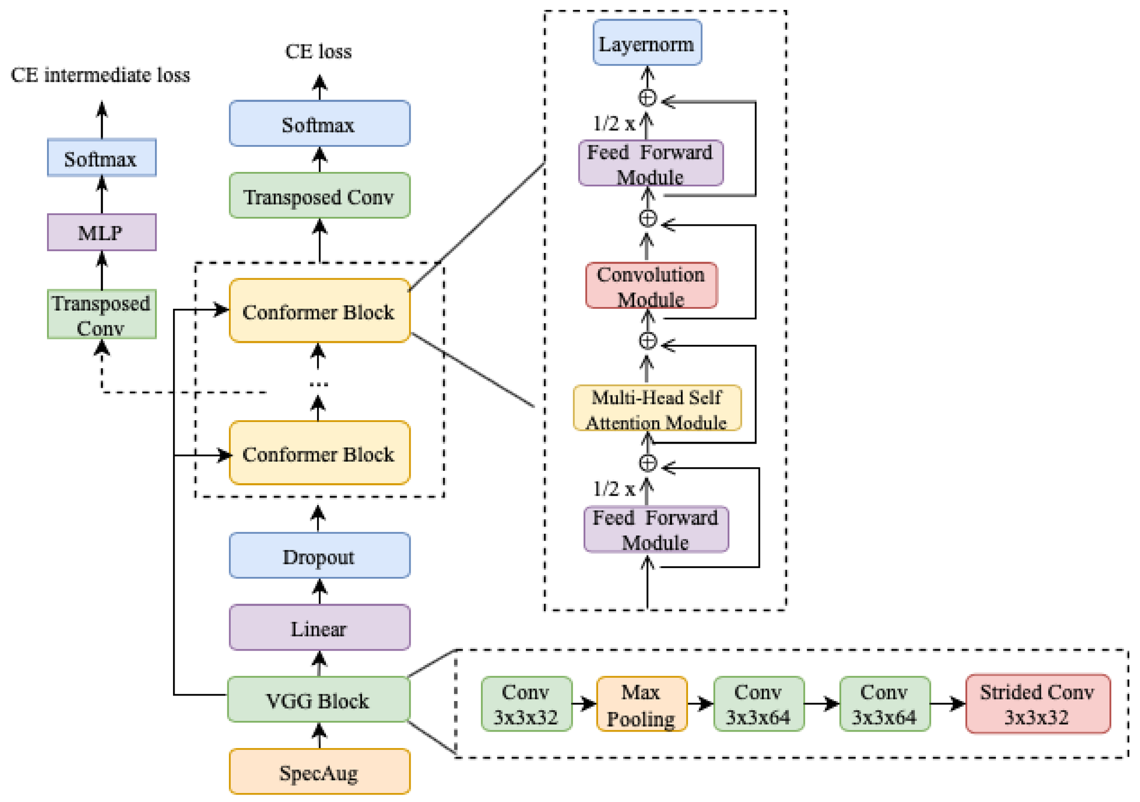
$$x_{MHSA} = x_{FFN_1} + \text{MHSA}(x_{FFN_1})$$

$$x_{Conv} = x_{MHSA} + \text{Conv}(x_{MHSA})$$

$$x_{FFN_2} = x_{Conv} + \frac{1}{2}\text{FFN}(x_{Conv})$$

$$\text{ConformerBlock}_i = \text{LayerNorm}(x_{FFN_2})$$

Our Proposed Conformer AM Architectur



Acknowledgements

European Research Council
Established by the European Commission

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement nº 694537, project "SEQ-CLAS"). The work reflects only the authors' views and the European Research Council Executive Agency (ERCEA) is not responsible for any use that may be made of the information it contains. This work was partially supported by the project HYKIST funded by the German Federal Ministry of Health on the basis of a decision of the German Federal Parliament (Bundestag) under funding ID ZMVI1-2520DAT04A.

Training Methods

Time Down-/Up-sampling
Use strided/transposed convolution for time down-/up-sampling

Intermediate Loss
Add intermediate losses at 4th and 8th conformer block

Parameter Sharing
Share paramters of the intermediate loss layers as well as the ones of transposed convolution

LongSkip
Connect output of the VGG network to input of each conformer block

Experiments and Results

Setup

- Dataset: switchboard 300h as train set, Hub5'00 as development set and Hub5'01 as test set.
- Input: 40-dimensional Gammatone features
- Target: alignments from a triphone CART-based GMM
- AM: 12-blocks conformer model with 512 attention dimension (8 heads) and 2048 feed-forward dimension
- Recognition: 4-gram count-based language model (LM)

Depthwise Convolution Kernel Size

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

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

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

- VGG-layerX: strided convolution as X^{th} layer of VGG network

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**

- BLSTM+maxpool: one BLSTM layer with 512 units followed by time max-pooling layer

Ablation Study of Training Methods

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**

- SpecAugment is the most important and gives 20% relative improvement
- Intermediate loss helps better convergence and achieves 7% relative improvement

Further Results

Comparison of BLSTM and Conformer AM architectures

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**

- BLSTM AM: 6 layers and with SpecAugment
- With comparable number of parameters, conformer AM outperforms BLSTM AM by around 9% relative

Overall Results

Work	#Epochs	Approach	AM	LM	seq. train	WER [%] Hub5'00	WER [%] Hub5'01
[Kitza+ 2019]	-	Hybrid	LSTM	4-gram	yes	13.9	-
				LSTM		11.7	
[Zhou+ 2021]	100	RNN-T	LSTM	LSTM	no	11.5	11.5
				Trafo		11.2	11.2
[Tüske+ 2020]	250	LAS	LSTM	LSTM	no	9.8	10.1
[Tüske+ 2021]	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**

- LSTM LM single pass + Transformer rescoring
- Lattice-based version of state-level minimum Bayes risk (sMBR) as sequence discriminative training (seq.train)

Conclusion

Efficient and Competitive Conformer Acoustic Model

- For the first time a training recipe for a conformer-base hybrid model is evaluated
- We combined different training methods from the literature that boosted the WER
- We applied time down-sampling using strided convolution to speed up training and used transposed convolution as a simple method to upsample again
- Our model outperforms the BLSTM-based hybrid model significantly
- Further improvement possible with speed perturbation, speaker adaptation and longer training

References

- [Kitza+ 2019] M. Kitza, P. Golik, R. Schlu ter, H. Ney. Cumulative Adaptation for BLSTM Acoustic Models. INTERSPEECH 2019, pp. 754–758
- [Zhou+ 2021] W. Zhou, S. Berger, R. Schlu ter, H. Ney. Phoneme Based Neural Transducer for Large Vocabulary Speech Recognition. ICASSP 2021, pp. 5644–5648
- [Tüske+ 2020] Z. Tu ske, G. Saon, K. Audhkhasi, B. Kingsbury. Single Headed Attention Based Sequence-to-Sequence Model for State-of-the-Art Results on Switchboard. INTERSPEECH 2020, pp. 551–555
- [Tüske+ 2021] Z. Tu ske, G. Saon, B. Kingsbury. On the Limit of English Conversational Speech Recognition. INTERSPEECH 2021, pp. 2062-2066
- [Gulati+ 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. INTERSPEECH 2020, pp. 5036–5040
- [Zeyer+ 2018] A. Zeyer, T. Alkhouli, H. Ney. RETURNN as a Generic Flexible Neural Toolkit with Application to Translation and Speech Recognition. Annual Meeting of the Assoc. for Computational Linguistics 2018
- [Chan+ 2016] 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