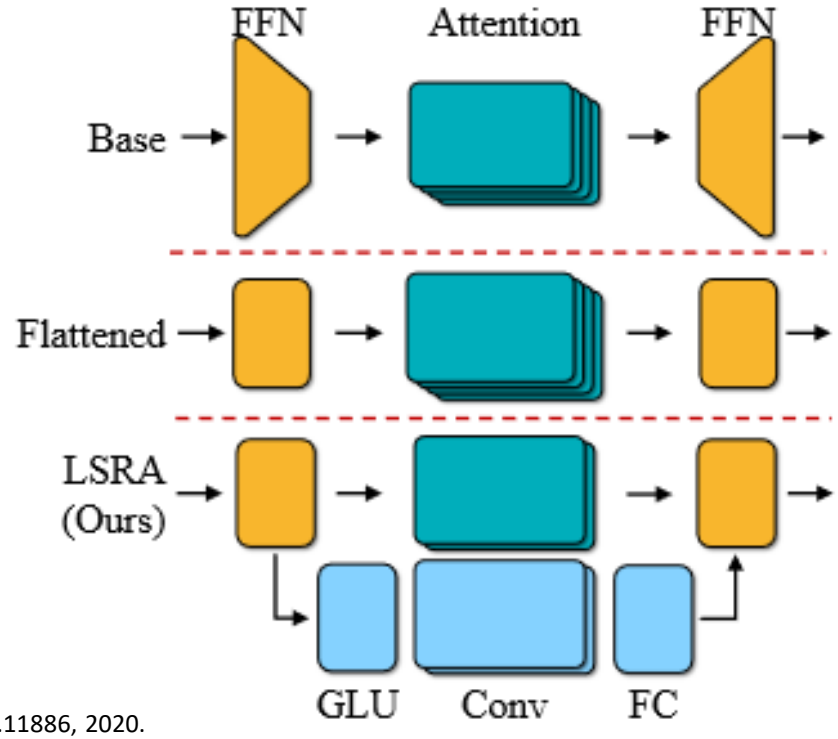
Conformer: Convolution-augmented Transformer for Speech Recognition

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang (Google Inc.)

Presenter：Wenxin Hou

2020.08.07

- Transformer is good at modeling long-range global context but weak at extracting local feature patterns, right opposite to Convolutional neural networks (CNN)
- Recent work has shown that combining convolution and self-attention brings improvements, e.g. LSRA [1]

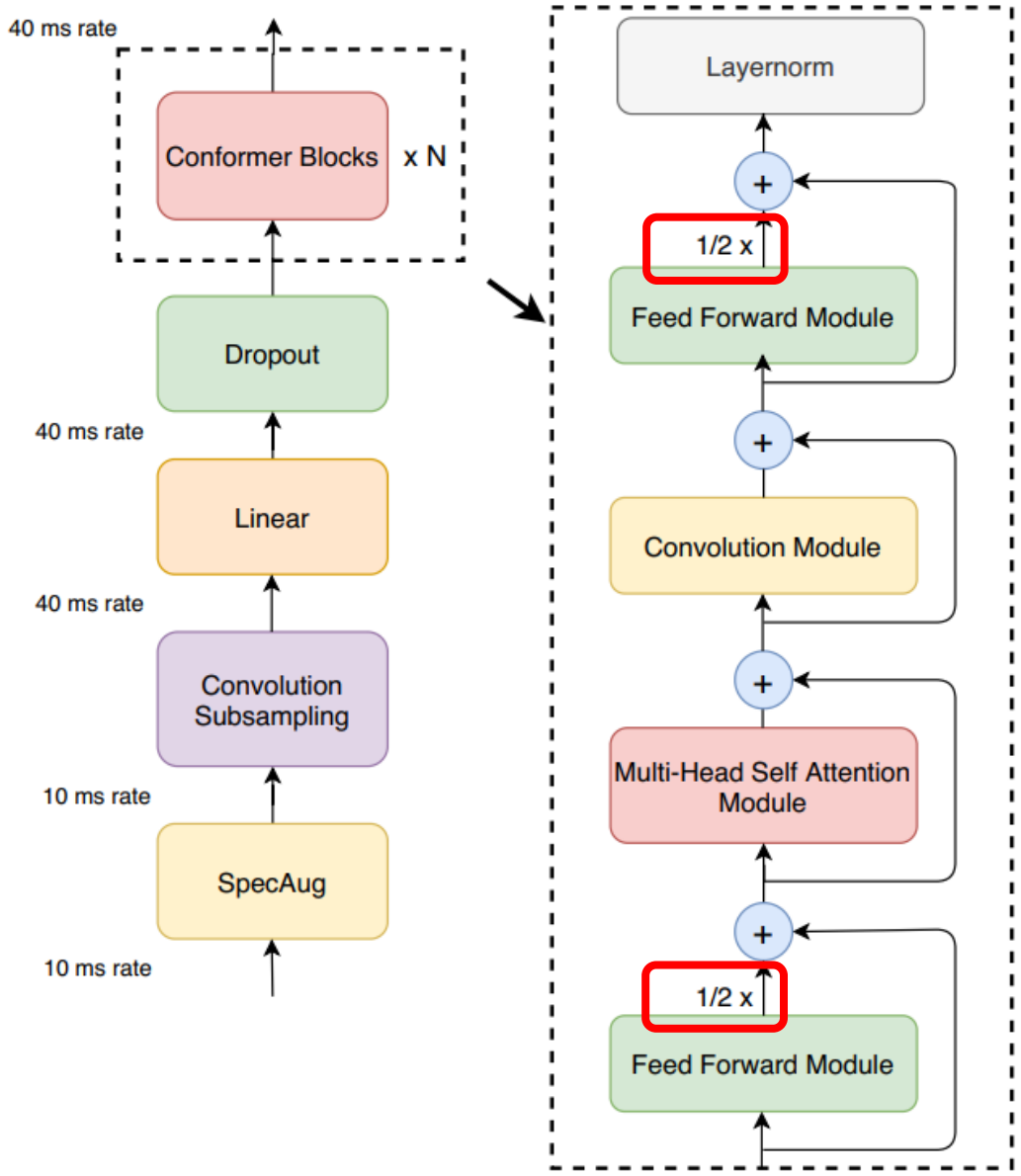


[1] Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han, "Lite transformer with long-short range attention," arXiv preprint arXiv:2004.11886, 2020.

02 **Conformer Encoder** | Overall architecture

- Subsampling layer + N x Conform Blocks

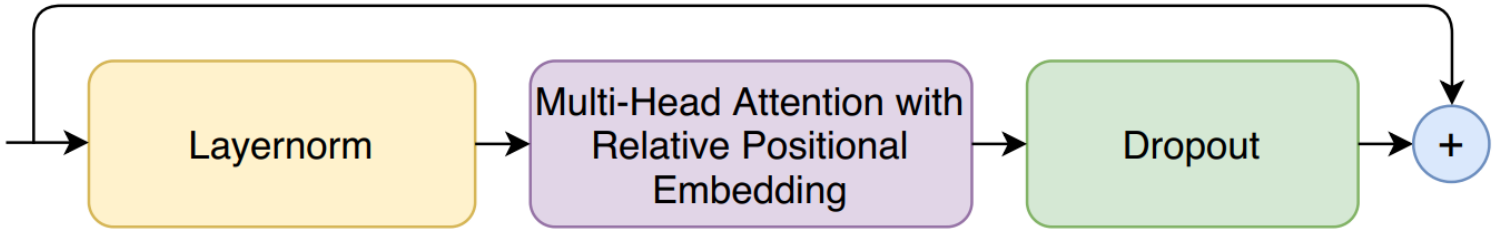
- Conformer Block of sandwich structure [2]
  - Macaron Feed Forward Module (first)
  - Multi-Head Self Attention Module
  - Convolution Module
  - Macaron Feed Forward Module (second)



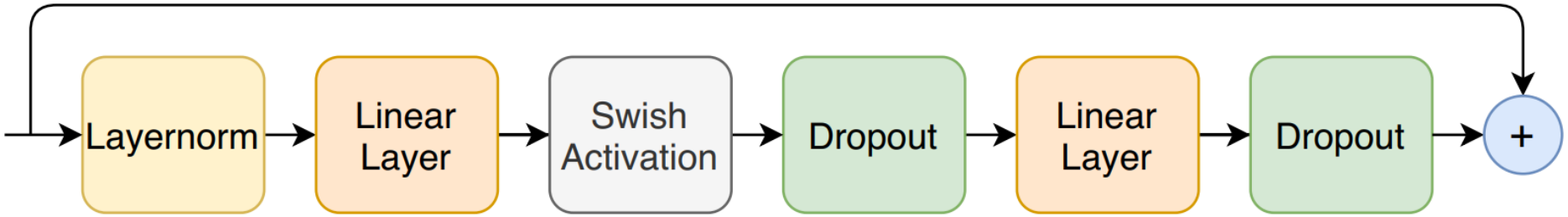
[2] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu, "Understanding and improving transformer from a multi-particle dynamic system point of view," arXiv preprint arXiv:1906.02762, 2019.

- Multi-Headed Self-Attention (MHSA) Module
  - Relative positional embedding [3] to improve robustness to varying utterance lengths
  - Pre-norm residual units [4] with dropout for regularization and training stabilization

$$x_{l+1} = x_l + F_l(\text{LayerNorm}(x_l))$$



- Feed Forward Module (FFN)



[3] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," 2019.
[4] T. Q. Nguyen and J. Salazar, "Transformers without tears: Improving the normalization of self-attention," arXiv preprint arXiv:1910.05895, 2019.

- Gating mechanism [5] consisting of a pointwise convolution and a gated linear unit (GLU)

$$\mathrm{GLU}(a, b) = a * \sigma(b)$$

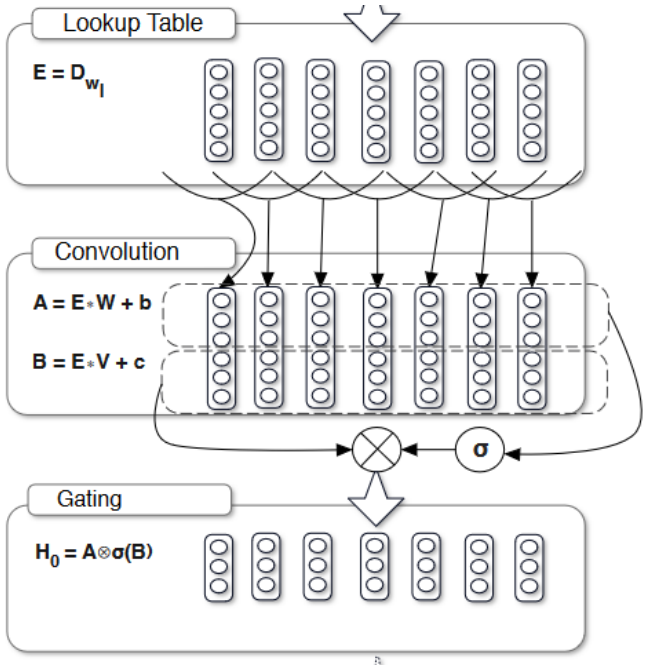
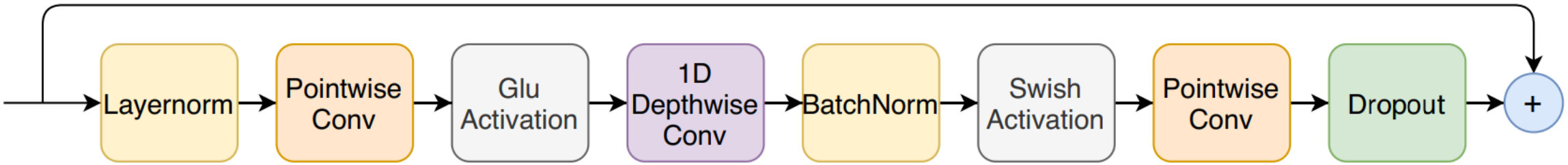


Illustration of gating mechanism

- Swish activation function [6]

$$\mathrm{Swish}(x) = x * \sigma(x)$$



[5] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017, pp. 933–941.
[6] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," arXiv preprint arXiv:1710.05941, 2017.

- 970-hour LibriSpeech dataset and 800M word token text-only corpus for language model
- Decoder: a single-layer LSTM
- Model hyper-parameters:

Model	Conformer (S)	Conformer (M)	Conformer (L)
Num Params (M)	10.3	30.7	118.8
Encoder Layers	16	16	17
Encoder Dim	144	256	512
Attention Heads	4	4	8
Conv Kernel Size	32	32	32
Decoder Layers	1	1	1
Decoder Dim	320	640	640

- Language model (LM): 3-layer LSTM with width 4096

- Word error rate (WER) results on LibriSpeech test-clean / test-other

Method	#Params (M)	WER Without LM		WER With LM	
		testclean	testother	testclean	testother
Hybrid					
Transformer [33]	-	-	-	2.26	4.85
CTC					
QuartzNet [9]	19	3.90	11.28	2.69	7.25
LAS					
Transformer [34]	270	2.89	6.98	2.33	5.17
Transformer [19]	-	2.2	5.6	2.6	5.7
LSTM	360	2.6	6.0	2.2	5.2
Transducer					
Transformer [7]	139	2.4	5.6	2.0	4.6
ContextNet(S) [10]	10.8	2.9	7.0	2.3	5.5
ContextNet(M) [10]	31.4	2.4	5.4	**2.0**	4.5
ContextNet(L) [10]	112.7	**2.1**	4.6	**1.9**	4.1
Conformer (Ours)					
Conformer(S)	10.3	**2.7**	**6.3**	**2.1**	**5.0**
Conformer(M)	30.7	**2.3**	**5.0**	**2.0**	**4.3**
Conformer(L)	118.8	**2.1**	**4.3**	**1.9**	**3.9**

- Disentangling Conformer towards a vanilla Transformer by
  1. Replacing Swish with ReLU
  2. Removing convolution blocks
  3. Replacing FFN pairs with a single FFN
  4. Removing relative positional encoding from self-attention layer

Model Architecture	dev clean	dev other	test clean	test other
Conformer Model	1.9	4.4	2.1	4.3
– SWISH + ReLU	1.9	4.4	2.0	4.5
– **Convolution Block**	2.1	4.8	2.1	4.9
– Macaron FFN	2.1	5.1	2.1	5.0
– Relative Pos. Emb.	2.3	5.8	2.4	5.6

- Study the effects of convolution & MHSA combination by
  1. Replacing depthwise convolution with lightweight convolution [7]
  2. Placing the convolution module before the MHSA module
  3. Split the input into parallel modules and concatenate the output

Model Architecture	dev clean	dev other
Conformer	1.9	4.4
– Depthwise conv + Lightweight convolution	2.0	4.8
Convolution block before MHSA	1.9	4.5
Parallel MHSA and Convolution	2.0	4.9

[7] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin, and M. Auli, “Pay less attention with lightweight and dynamic convolutions,” arXiv preprint arXiv:1901.10430, 2019.

- Impact of changing the Conformer block to use a single FFN or full-step residuals.

Model Architecture	dev clean	dev other	test clean	test other
Conformer	1.9	4.4	2.1	4.3
Single FFN	1.9	4.5	2.1	4.5
Full step residuals	1.9	4.5	2.1	4.5

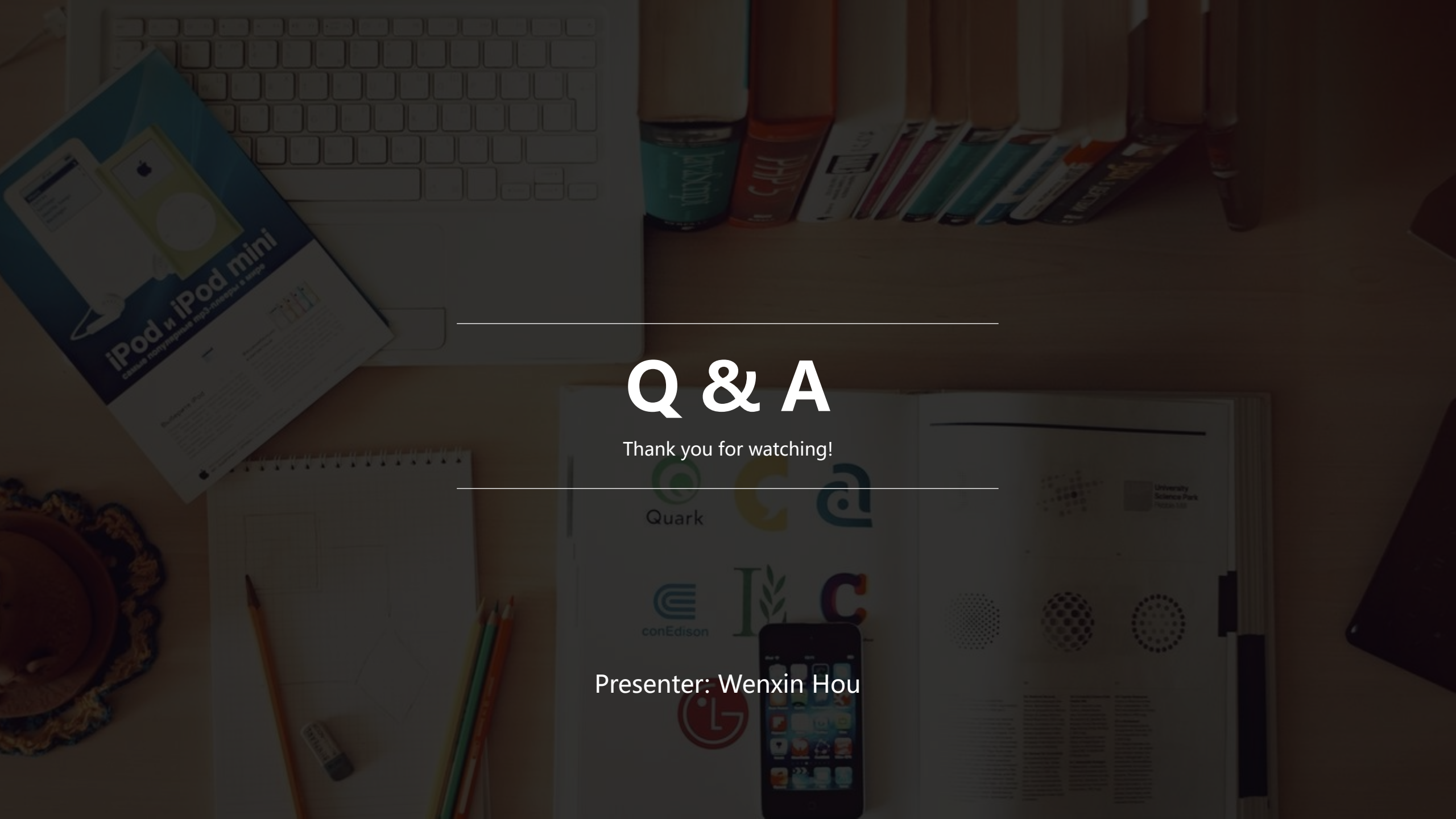
- **Left**: effect of varying the number of attention heads from 4 to 32
- **Right**: effect of kernel sizes in the depthwise convolution

Attention Heads	Dim per Head	dev clean	dev other	test clean	test other
4	128	1.9	4.6	2.0	4.5
8	64	1.9	4.4	2.1	4.3
16	32	2.0	4.3	2.2	4.4
32	16	1.9	4.4	2.1	4.5

Kernel size	dev clean	dev other	test clean	test other
3	1.88	4.41	1.99	4.39
7	1.88	4.30	2.02	4.44
17	1.87	4.31	2.04	4.38
32	1.83	4.30	2.03	4.29
65	1.89	4.47	1.98	4.46

Contribution

- Introduce Conformer integrating CNNs and Transformers for end-to-end speech recognition
- Demonstrate the effectiveness of including convolution module
- Achieve state-of-the-art WER at 1.9%/3.9% on LibriSpeech test-clean/test-other



Q & A

Thank you for watching!

Presenter: Wenxin Hou