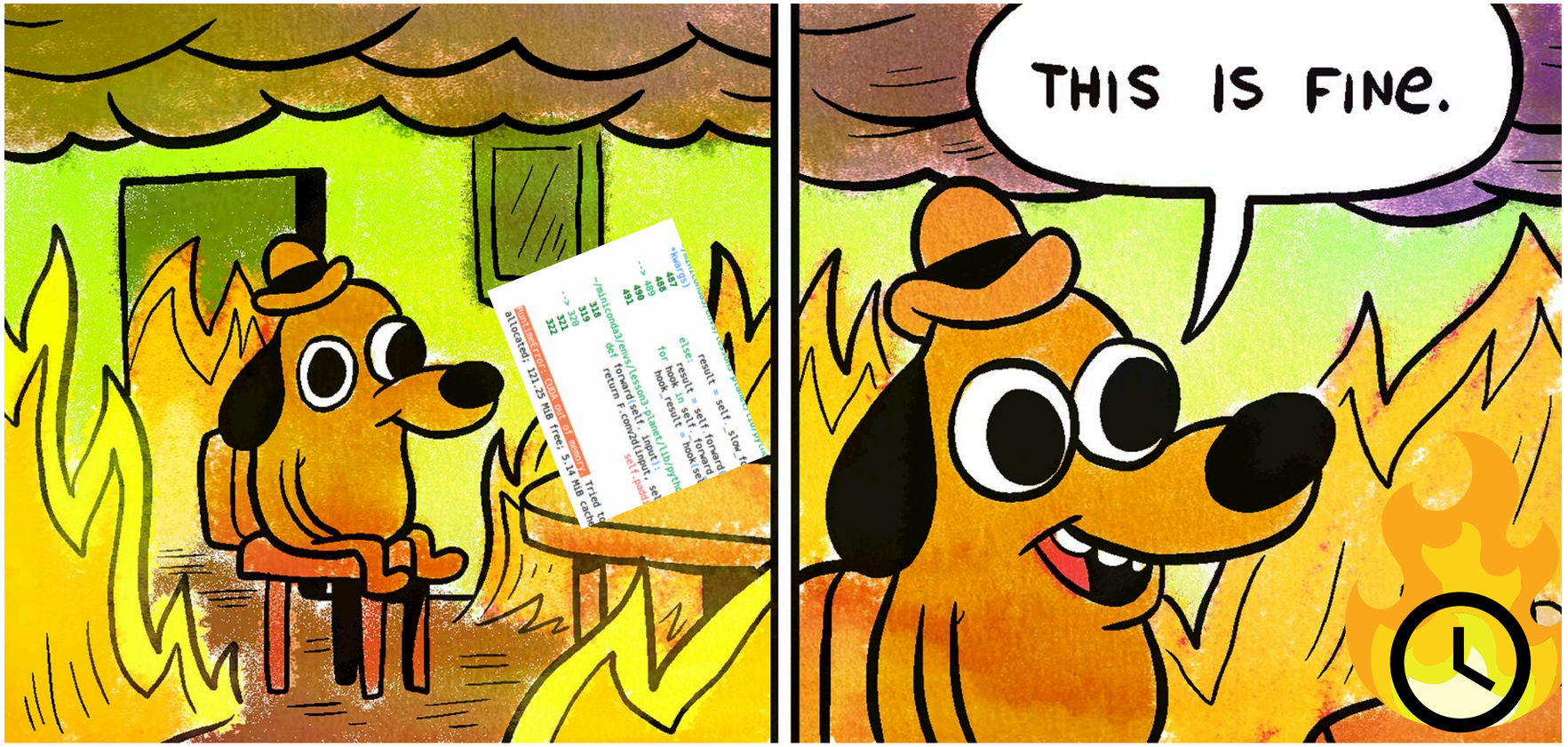
HOW TO READ PAPER AND IMPLEMENT CONCEPT

WHY READ???

WHY READ???



IDEA

HOW TO FIND PAPER?

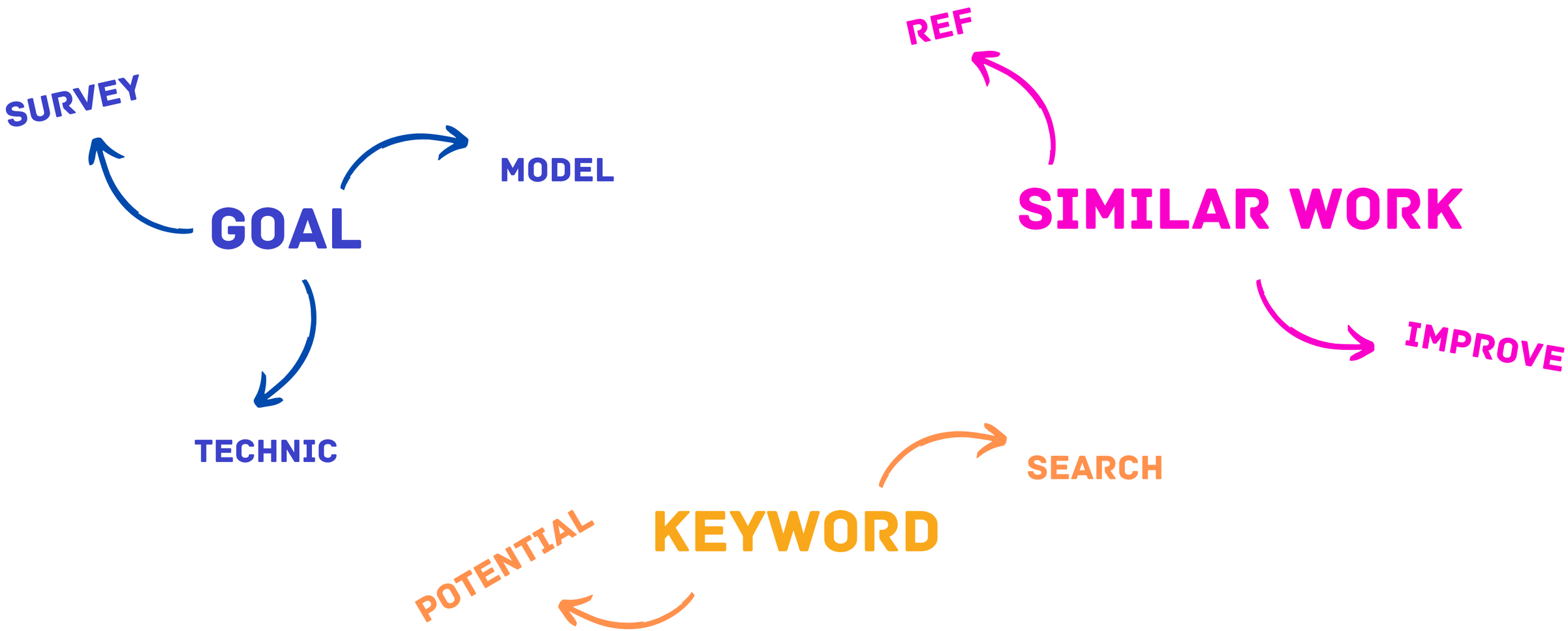
HOW TO FIND PAPER?

GOAL

SIMILAR WORK

KEYWORD

HOW TO FIND PAPER?



SOURCE / TOOLS

PAPER POOL

Google Scholar

Google Scholar provides a simple way to broadly search for scholarly literature. Search across a wide variety of disciplines and sources: articles, theses, books, abstracts and court opinions.

google.com

arXiv.org e-Print archive

Work on one of the world's most important websites and make an impact on open science.

arxiv.org

Semantic Scholar | AI-Powered Research Tool

Semantic Scholar uses groundbreaking AI and engineering to understand the semantics of scientific literature to help Scholars discover relevant research.

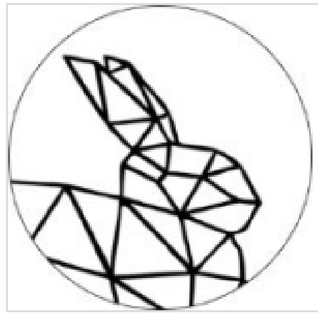
allenai.org

AI TOOLS FOR RESEARCH

Perplexity

Perplexity is a free AI-powered answer engine that provides accurate, trusted, and real-time answers to any question.

Perplexity AI



Research Rabbit

The most powerful discovery app ever built for researchers!

researchrabbitapp.com

SCISPACE

HTTPS://SCISPACE.COM/

IS IT GOOD?

- TIER REVIEW
- UP TO DATE
- BENCHMARK (DATASET)
- CODE/WEIGHT/ARCHITECTURE

FIGHT WITH PAPER

101

Conformer: Convolution-augmented Transformer for Speech Recognition

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs)....

arXiv.org

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.

{anmolgulati, jamesqin, chungchengc, nikip, ngyuzh, jiahuiyu, weihan, shibow, zhangzd, yonghui, rpang}@google.com

Abstract

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs). Transformer models are good at capturing content-based global interactions, while CNNs exploit local features effectively. In this work, we achieve the best of both worlds by studying how to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way. To this regard, we propose the convolution-augmented transformer for speech recognition, named *Conformer*. *Conformer* significantly outperforms the previous Transformer and CNN based models achieving state-of-the-art accuracies. On the widely used LibriSpeech benchmark, our model achieves WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/testother. We also observe competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

Index Terms: speech recognition, attention, convolutional neural networks, transformer, end-to-end

1. Introduction

End-to-end automatic speech recognition (ASR) systems based on neural networks have seen large improvements in recent years. Recurrent neural networks (RNNs) have been the de-facto choice for ASR [1, 2, 3, 4] as they can model the temporal dependencies in the audio sequences effectively [5]. Recently, the Transformer architecture based on self-attention [6, 7] has enjoyed widespread adoption for modeling sequences due to its ability to capture long distance interactions and the high training efficiency. Alternatively, convolutions have also been successful for ASR [8, 9, 10, 11, 12], which capture local context progressively via a local receptive field layer by layer.

However, models with self-attention or convolutions each has its limitations. While Transformers are good at modeling long-range global context, they are less capable to extract fine-grained local feature patterns. Convolution neural networks (CNNs), on the other hand, exploit local information and are used as the de-facto computational block in vision. They learn shared position-based kernels over a local window which maintain translation equivariance and are able to capture features like edges and shapes. One limitation of using local connectivity is that you need many more layers or parameters to capture global information. To combat this issue, contemporary work ContextNet [10] adopts the squeeze-and-excitation module [13] in each residual block to capture longer context. However, it is still limited in capturing dynamic global context as it only applies a global averaging over the entire sequence.

Recent works have shown that combining convolution and self-attention improves over using them individually [14]. Together, they are able to learn both position-wise local features, and use content-based global interactions. Concurrently, papers like [15, 16] have augmented self-attention with relative position based information that maintains equivariance. Wu et al. [17] proposed a multi-branch architecture with splitting the input into two branches: self-attention and convolution; and concatenating their outputs. Their work targeted mobile applications and showed improvements in machine translation tasks.

In this work, we study how to organically combine convolutions with self-attention in ASR models. We hypothesize that both global and local interactions are important for being parameter efficient. To achieve this, we propose a novel combination of self-attention and convolution will achieve the best of both worlds – self-attention learns the global interaction whilst the convolutions efficiently capture the relative-offset-based local correlations. Inspired by Wu et al. [17, 18], we introduce a novel combination of self-attention and convolution, sandwiched between a pair feed forward modules, as illustrated in Fig 1.

Our proposed model, named Conformer, achieves state-of-the-art results on LibriSpeech, outperforming the previous best published Transformer Transducer [7] by 15% relative improve-

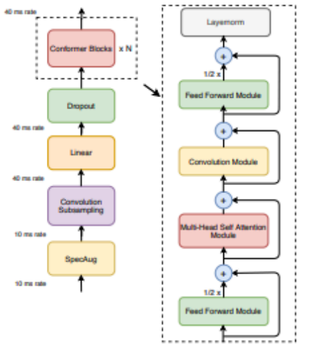


Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*


breath in

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.

{anmolgulati, jamesqin, chungchengc, nikip, ngyuzh, jiahuiyu, weihan, shibow, zhangzd, yonghui, rpang}@google.com

Abstract

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs). Transformer models are good at capturing content-based global interactions, while CNNs exploit local features effectively. In this work, we achieve the best of both worlds by studying how to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way. To this regard, we propose the convolution-augmented transformer for speech recognition, named *Conformer*. *Conformer* significantly outperforms the previous Transformer and CNN based models achieving state-of-the-art accuracies. On the widely used LibriSpeech benchmark, our model achieves WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/testother. We also observe competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

Index Terms: speech recognition, attention, convolutional neural networks, transformer, end-to-end

1. Introduction

End-to-end automatic speech recognition (ASR) systems based on neural networks have seen large improvements in recent years. Recurrent neural networks (RNNs) have been the de-facto choice for ASR [1, 2, 3, 4] as they can model the temporal dependencies in the audio sequences effectively [5]. Recently, the Transformer architecture based on self-attention [6, 7] has enjoyed widespread adoption for modeling sequences due to its ability to capture long distance interactions and the high training efficiency. Alternatively, convolutions have also been successful for ASR [8, 9, 10, 11, 12], which capture local context progressively via a local receptive field layer by layer.

However, models with self-attention or convolutions each has its limitations. While Transformers are good at modeling long-range global context, they are less capable to extract fine-grained local feature patterns. Convolution neural networks (CNNs), on the other hand, exploit local information and are used as the de-facto computational block in vision. They learn shared position-based kernels over a local window which maintain translation equivariance and are able to capture features like edges and shapes. One limitation of using local connectivity is that you need many more layers or parameters to capture global information. To combat this issue, contemporary work ContextNet [10] adopts the squeeze-and-excitation module [13] in each residual block to capture longer context. However, it is still limited in capturing dynamic global context as it only applies a global averaging over the entire sequence.

Recent works have shown that combining convolution and self-attention improves over using them individually [14]. Together, they are able to learn both position-wise local features, and use content-based global interactions. Concurrently, papers like [15, 16] have augmented self-attention with relative position based information that maintains equivariance. Wu et al. [17] proposed a multi-branch architecture with splitting the input into two branches: self-attention and convolution; and concatenating their outputs. Their work targeted mobile applications and showed improvements in machine translation tasks.

In this work, we study how to organically combine convolutions with self-attention in ASR models. We hypothesize that both global and local interactions are important for being parameter efficient. To achieve this, we propose a novel combination of self-attention and convolution will achieve the best of both worlds – self-attention learns the global interaction whilst the convolutions efficiently capture the relative-offset-based local correlations. Inspired by Wu et al. [17, 18], we introduce a novel combination of self-attention and convolution, sandwiched between a pair feed forward modules, as illustrated in Fig 1.

Our proposed model, named Conformer, achieves state-of-the-art results on LibriSpeech, outperforming the previous best published Transformer Transducer [7] by 15% relative improve-

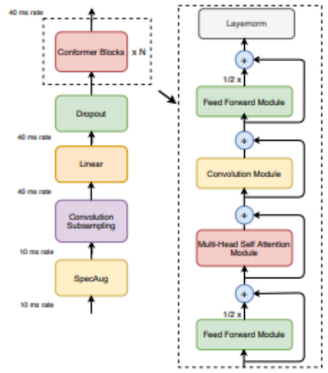


Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*

PAPER STRUCTURE

- ABSTRACT
- INTRODUCTION
- METHODS
- RESULTS
- CONCLUSION
- REFERENCES

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.

{anmolgulati, jamesqin, chungchengc, nikip, ngyuzh, jiahuiyu, weihan, shibow, zhangzd, yonghui, rpang}@google.com

Abstract

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs). Transformer models are good at capturing content-based global interactions, while CNNs exploit local features effectively. In this work, we achieve the best of both worlds by studying how to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way. To this regard, we propose the convolution-augmented transformer for speech recognition, named *Conformer*. *Conformer* significantly outperforms the previous Transformer and CNN based models achieving state-of-the-art accuracies. On the widely used LibriSpeech benchmark, our model achieves WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/testother. We also observe competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

Index Terms: speech recognition, attention, convolutional neural networks, transformer, end-to-end

1. Introduction

End-to-end automatic speech recognition (ASR) systems based on neural networks have seen large improvements in recent years. Recurrent neural networks (RNNs) have been the de-facto choice for ASR [1, 2, 3, 4] as they can model the temporal dependencies in the audio sequences effectively [5]. Recently, the Transformer architecture based on self-attention [6, 7] has enjoyed widespread adoption for modeling sequences due to its ability to capture long distance interactions and the high training efficiency. Alternatively, convolutions have also been successful for ASR [8, 9, 10, 11, 12], which capture local context progressively via a local receptive field layer by layer.

However, models with self-attention or convolutions each has its limitations. While Transformers are good at modeling long-range global context, they are less capable to extract fine-grained local feature patterns. Convolution neural networks (CNNs), on the other hand, exploit local information and are used as the de-facto computational block in vision. They learn shared position-based kernels over a local window which maintain translation equivariance and are able to capture features like edges and shapes. One limitation of using local connectivity is that you need many more layers or parameters to capture global information. To combat this issue, contemporary work ContextNet [10] adopts the squeeze-and-excitation module [13] in each residual block to capture longer context. However, it is still limited in capturing dynamic global context as it only applies a global averaging over the entire sequence.

Recent works have shown that combining convolution and self-attention improves over using them individually [14]. Together, they are able to learn both position-wise local features, and use content-based global interactions. Concurrently, papers like [15, 16] have augmented self-attention with relative position based information that maintains equivariance. Wu et al. [17] proposed a multi-branch architecture with splitting the input into two branches: self-attention and convolution; and concatenating their outputs. Their work targeted mobile applications and showed improvements in machine translation tasks.

In this work, we study how to organically combine convolutions with self-attention in ASR models. We hypothesize that both global and local interactions are important for being parameter efficient. To achieve this, we propose a novel combination of self-attention and convolution will achieve the best of both worlds – self-attention learns the global interaction whilst the convolutions efficiently capture the relative-offset-based local correlations. Inspired by Wu et al. [17, 18], we introduce a novel combination of self-attention and convolution, sandwiched between a pair feed forward modules, as illustrated in Fig 1.

Our proposed model, named Conformer, achieves state-of-the-art results on LibriSpeech, outperforming the previous best published Transformer Transducer [7] by 15% relative improve-

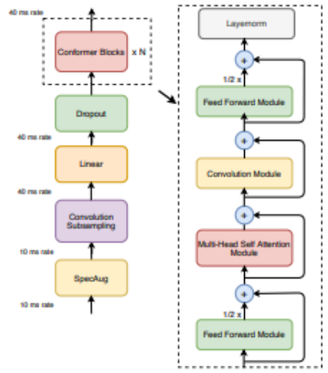


Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*

PAPER STRUCTURE

- **ABSTRACT**
- INTRODUCTION
- METHODS
- **RESULTS**
- **CONCLUSION**
- REFERENCES

Abstract

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs). Transformer models are good at capturing content-based global interactions, while CNNs exploit local features effectively. In this work, we achieve the best of both worlds by studying how to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way. To this regard, we propose the convolution-augmented transformer for speech recognition, named *Conformer*. *Conformer* significantly outperforms the previous Transformer and CNN based models achieving state-of-the-art accuracies. On the widely used LibriSpeech benchmark, our model achieves WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/testother. We also observe competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

Index Terms: speech recognition, attention, convolutional neural networks, transformer, end-to-end

arXiv:2005.08100v1 [eess.AS] 16 May 2020

Abstract

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs). Transformer models are good at capturing content-based global interactions, while CNNs exploit local features effectively. In this work, we achieve the best of both worlds by studying how to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way. To this regard, we propose the convolution-augmented transformer for speech recognition, named *Conformer*. *Conformer* significantly outperforms the previous Transformer and CNN based models achieving state-of-the-art accuracies. On the widely used LibriSpeech benchmark, our model achieves WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/testother. We also observe competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

Index Terms: speech recognition, attention, convolutional neural networks, transformer, end-to-end

TRANSFORMER

CNN

LIBRISPEECH

ASR

WER

CONFORMER

arXiv:2005.08100v1 [eess.AS] 16 May 2020

Abstract

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs). Transformer models are good at capturing content-based global interactions, while CNNs exploit local features effectively. In this work, we achieve the best of both worlds by studying how to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way. To this regard, we propose the convolution-augmented transformer for speech recognition, named *Conformer*. *Conformer* significantly outperforms the previous Transformer and CNN based models achieving state-of-the-art accuracies. On the widely used LibriSpeech benchmark, our model achieves WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/testother. We also observe competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

Index Terms: speech recognition, attention, convolutional neural networks, transformer, end-to-end

TASK: ASR

“CONFORMER”

TRANSFORMER

+

CNN

BENCH DATASET: LIBRISPEECH

METRIC: WORD ERROR RATE (WER)

KEYWORD

Table 1: *Model hyper-parameters for Conformer S, M, and L models, found via sweeping different combinations and choosing the best performing models within the parameter limits.*

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

Table 2: *Comparison of Conformer with recent published models. Our model shows improvements consistently over various model parameter size constraints. At 10.3M parameters, our model is 0.7% better on testother when compared to contemporary work, ContextNet(S) [10]. At 30.7M model parameters our model already significantly outperforms the previous published state of the art results of Transformer Transducer [7] with 139M parameters.*

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**

4. Conclusion

In this work, we introduced Conformer, an architecture that integrates components from CNNs and Transformers for end-to-end speech recognition. We studied the importance of each component, and demonstrated that the inclusion of convolution modules is critical to the performance of the Conformer model. The model exhibits better accuracy with fewer parameters than previous work on the LibriSpeech dataset, and achieves a new state-of-the-art performance at 1.9%/3.9% for test/testother.

I LIKED IT

HOW TO IMPLEMENT?

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.

{anmolgulati, jamesqin, chungchengc, nikip, ngyuzh, jiahuiyu, weihan, shibow, zhangzd, yonghui, rpang}@google.com

Abstract

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs). Transformer models are good at capturing content-based global interactions, while CNNs exploit local features effectively. In this work, we achieve the best of both worlds by studying how to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way. To this regard, we propose the convolution-augmented transformer for speech recognition, named *Conformer*. *Conformer* significantly outperforms the previous Transformer and CNN based models achieving state-of-the-art accuracies. On the widely used LibriSpeech benchmark, our model achieves WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/testother. We also observe competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

Index Terms: speech recognition, attention, convolutional neural networks, transformer, end-to-end

1. Introduction

End-to-end automatic speech recognition (ASR) systems based on neural networks have seen large improvements in recent years. Recurrent neural networks (RNNs) have been the de-facto choice for ASR [1, 2, 3, 4] as they can model the temporal dependencies in the audio sequences effectively [5]. Recently, the Transformer architecture based on self-attention [6, 7] has enjoyed widespread adoption for modeling sequences due to its ability to capture long distance interactions and the high training efficiency. Alternatively, convolutions have also been successful for ASR [8, 9, 10, 11, 12], which capture local context progressively via a local receptive field layer by layer.

However, models with self-attention or convolutions each has its limitations. While Transformers are good at modeling long-range global context, they are less capable to extract fine-grained local feature patterns. Convolution neural networks (CNNs), on the other hand, exploit local information and are used as the de-facto computational block in vision. They learn shared position-based kernels over a local window which maintain translation equivariance and are able to capture features like edges and shapes. One limitation of using local connectivity is that you need many more layers or parameters to capture global information. To combat this issue, contemporary work ContextNet [10] adopts the squeeze-and-excitation module [13] in each residual block to capture longer context. However, it is still limited in capturing dynamic global context as it only applies a global averaging over the entire sequence.

Recent works have shown that combining convolution and self-attention improves over using them individually [14]. Together, they are able to learn both position-wise local features, and use content-based global interactions. Concurrently, papers like [15, 16] have augmented self-attention with relative position based information that maintains equivariance. Wu et al. [17] proposed a multi-branch architecture with splitting the input into two branches: self-attention and convolution; and concatenating their outputs. Their work targeted mobile applications and showed improvements in machine translation tasks.

In this work, we study how to organically combine convolutions with self-attention in ASR models. We hypothesize that both global and local interactions are important for being parameter efficient. To achieve this, we propose a novel combination of self-attention and convolution will achieve the best of both worlds – self-attention learns the global interaction whilst the convolutions efficiently capture the relative-offset-based local correlations. Inspired by Wu et al. [17, 18], we introduce a novel combination of self-attention and convolution, sandwiched between a pair feed forward modules, as illustrated in Fig 1.

Our proposed model, named Conformer, achieves state-of-the-art results on LibriSpeech, outperforming the previous best published Transformer Transducer [7] by 15% relative improve-

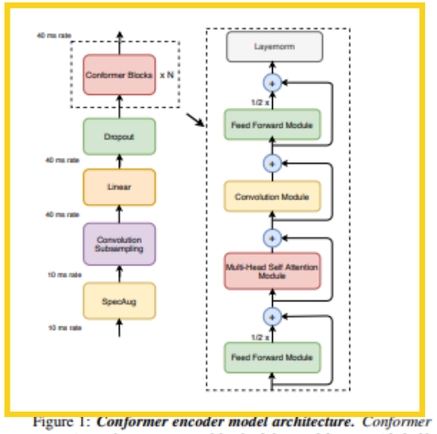
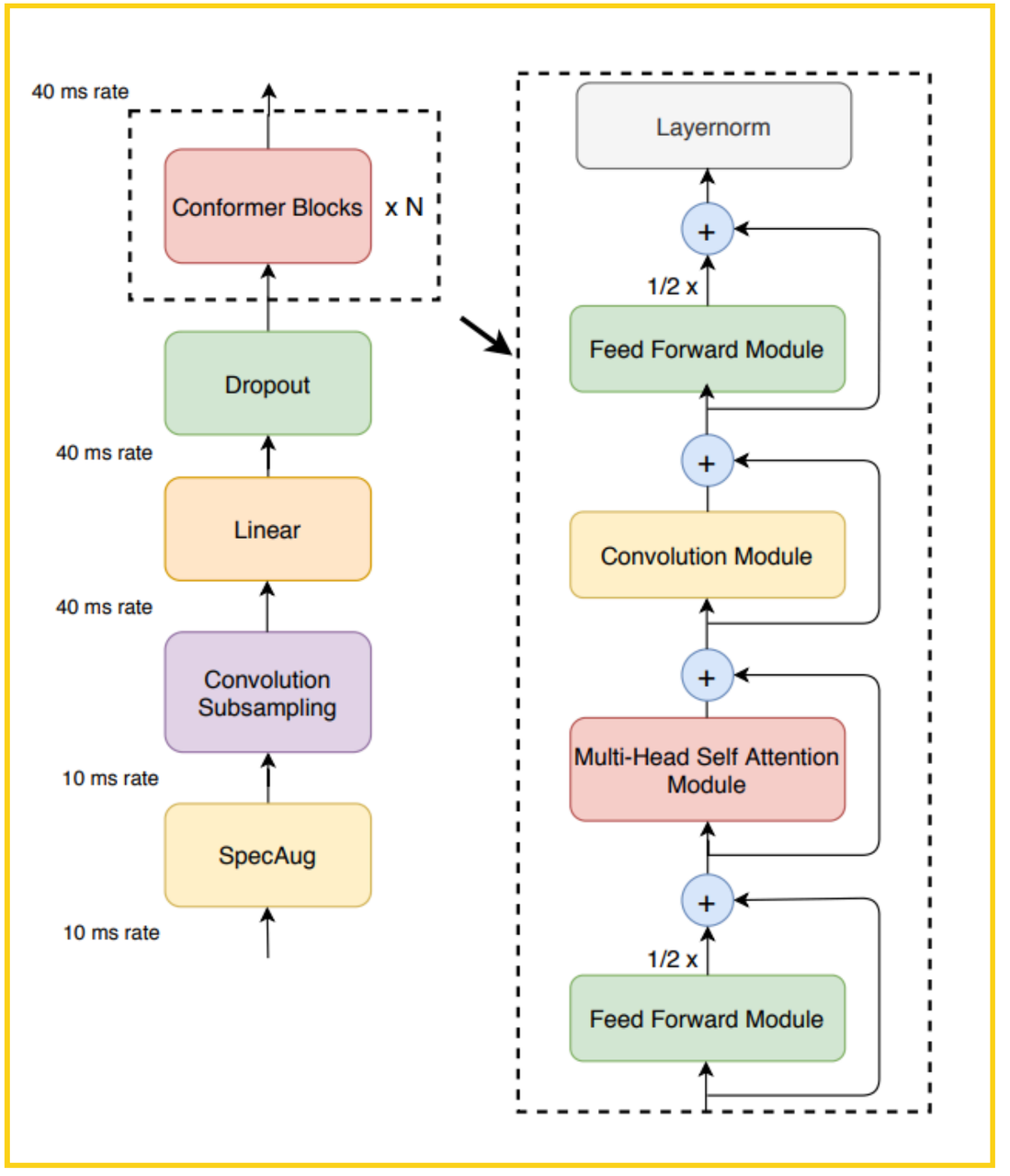


Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*



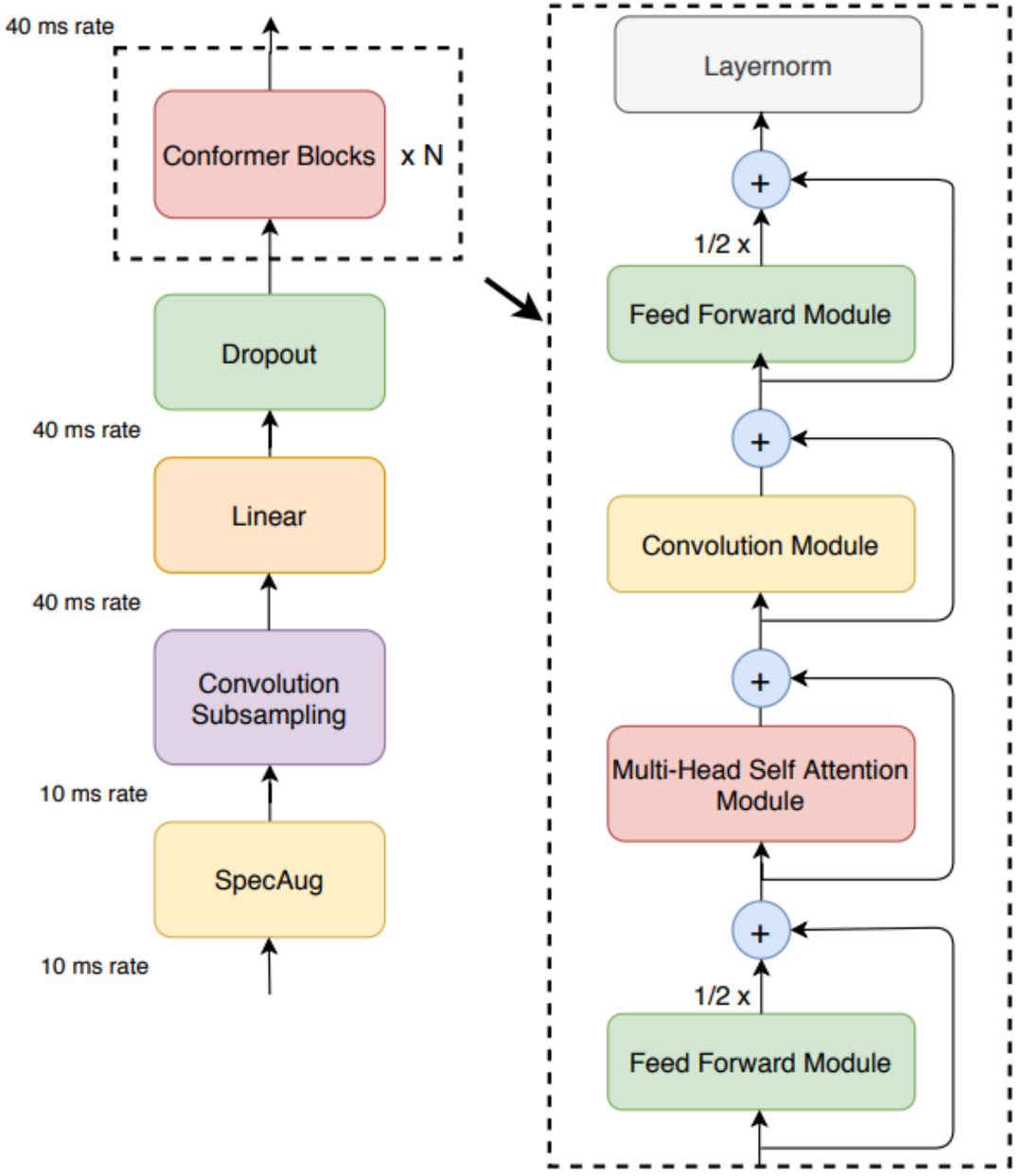
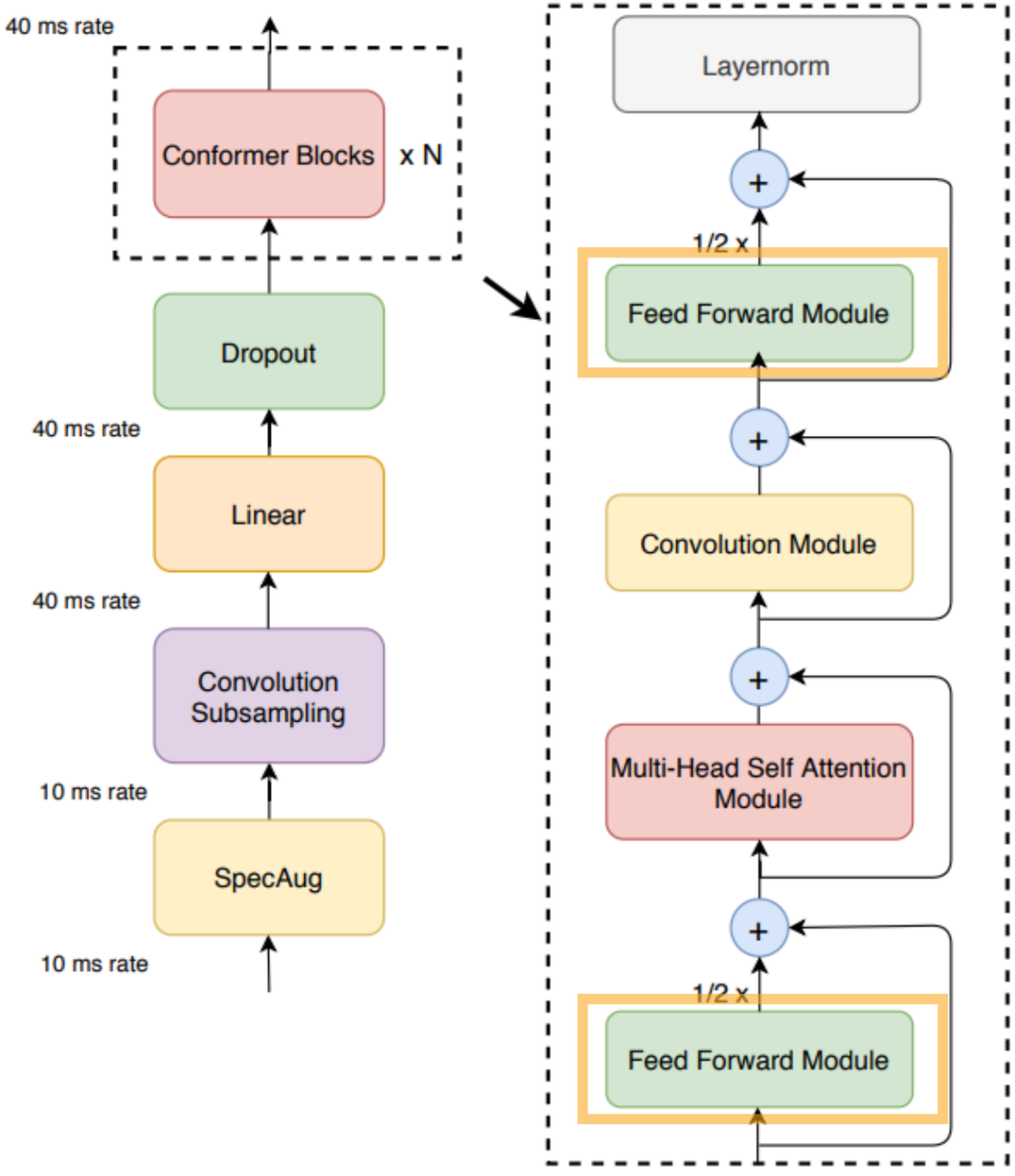


Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*



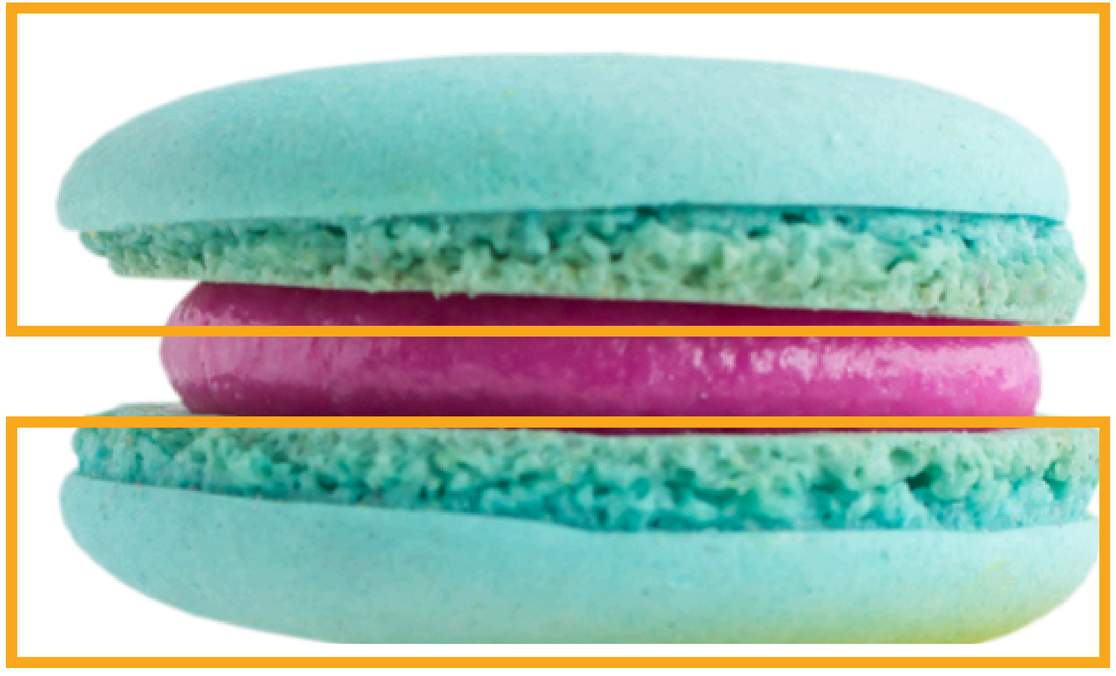
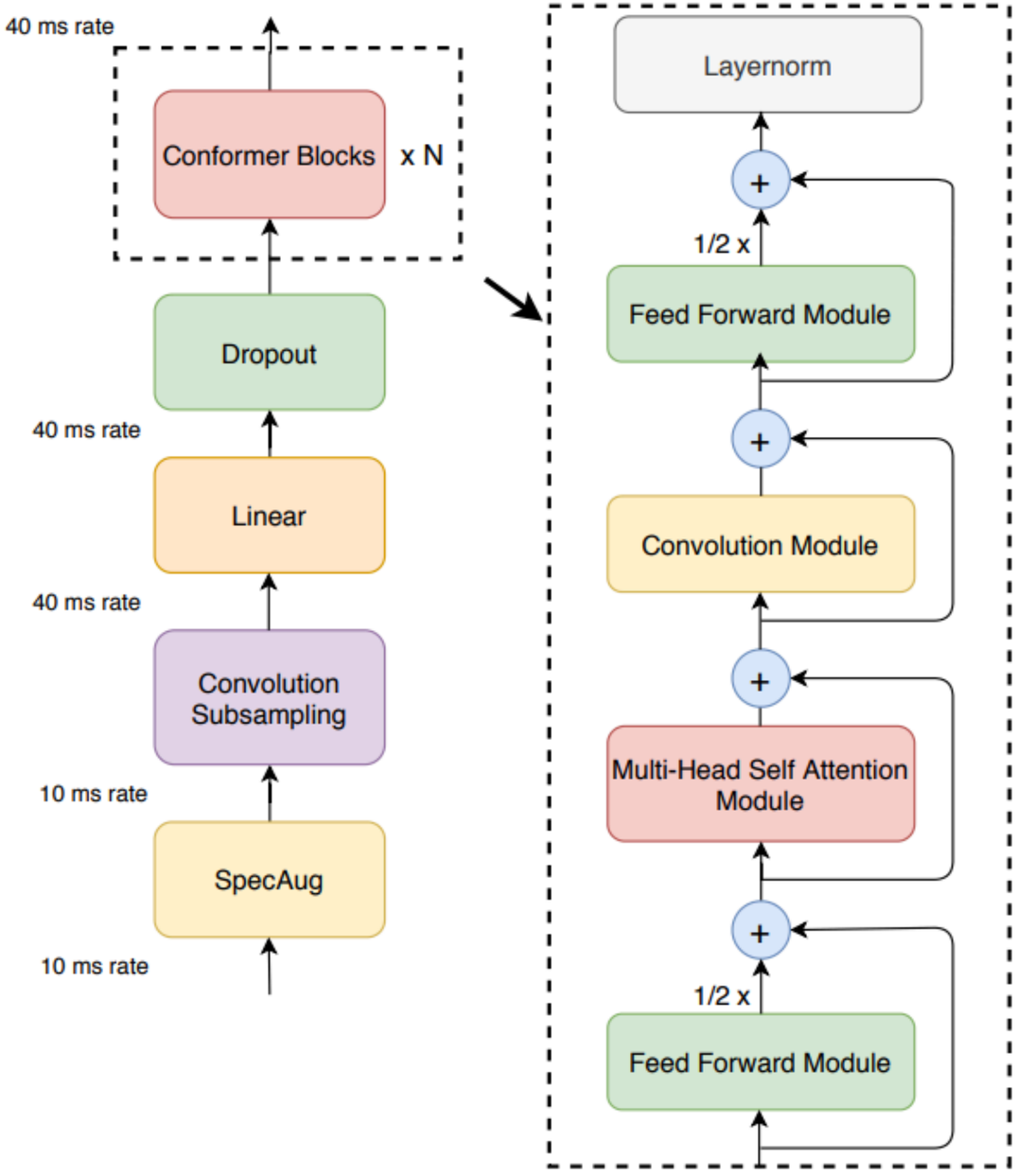


Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*



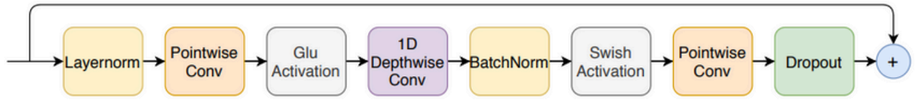


Figure 2: ***Convolution module.*** *The convolution module contains a pointwise convolution with an expansion factor of 2 projecting the number of channels with a GLU activation layer, followed by a 1-D Depthwise convolution. The 1-D depthwise conv is followed by a Batchnorm and then a swish activation layer.*

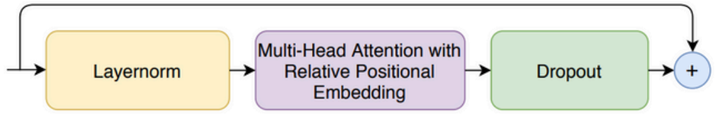


Figure 3: ***Multi-Headed self-attention module.*** *We use multi-headed self-attention with relative positional embedding in a pre-norm residual unit.*

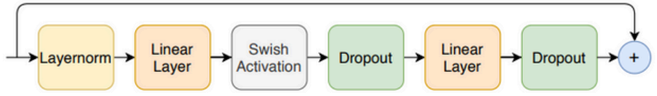


Figure 4: ***Feed forward module.*** *The first linear layer uses an expansion factor of 4 and the second linear layer projects it back to the model dimension. We use swish activation and a pre-norm residual units in feed forward module.*

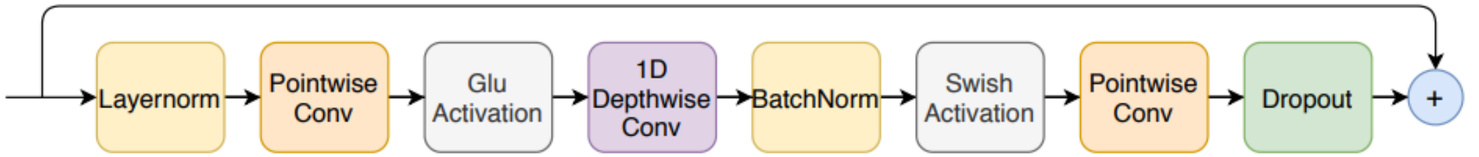


Figure 2: ***Convolution module.*** *The convolution module contains a pointwise convolution with an expansion factor of 2 projecting the number of channels with a GLU activation layer, followed by a 1-D Depthwise convolution. The 1-D depthwise conv is followed by a Batchnorm and then a swish activation layer.*

```python
# --- 3. Convolution Module ---
# Conv1D + GLU → DepthwiseConv → BN → Swish → Conv1D
# local acoustic pattern
class ConvolutionModule(nn.Module):
    def __init__(self, d_model, kernel_size=15, dropout=0.1):
        super().__init__()
        self.norm = nn.LayerNorm(d_model)
        self.pw1 = nn.Conv1d(d_model, 2 * d_model, 1)  # Pointwise conv
        self.glu = nn.GLU(dim=1)
        self.dw = nn.Conv1d(d_model, d_model, kernel_size,
                            padding=kernel_size // 2, groups=d_model)  # Depthwise
        self.bn = nn.BatchNorm1d(d_model)
        self.act = nn.SiLU()  # Swish
        self.pw2 = nn.Conv1d(d_model, d_model, 1)
        self.dropout = nn.Dropout(dropout)

    def forward(self, x):
        x = self.norm(x).transpose(1, 2)  # [B, D, T]
        x = self.glu(self.pw1(x))        # gating
        x = self.act(self.bn(self.dw(x)))
        x = self.dropout(self.pw2(x))
        return x.transpose(1, 2)  # back to [B, T, D]
```

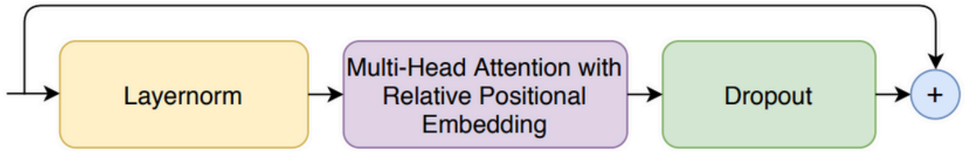


Figure 3: ***Multi-Headed self-attention module.*** *We use multi-headed self-attention with relative positional embedding in a pre-norm residual unit.*

```python
# --- 2. Multi-Head Self-Attention (MHSA) ---
# global context + relative positional encoding
class SelfAttentionModule(nn.Module):
    def __init__(self, d_model, n_heads, dropout=0.1):
        super().__init__()
        self.attn = nn.MultiheadAttention(embed_dim=d_model, num_heads=n_heads,
                                          dropout=dropout, batch_first=True)

    def forward(self, x):
        return self.attn(x, x, x)[0]  # x: [B, T, D]
```

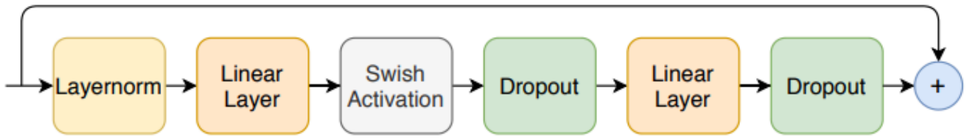
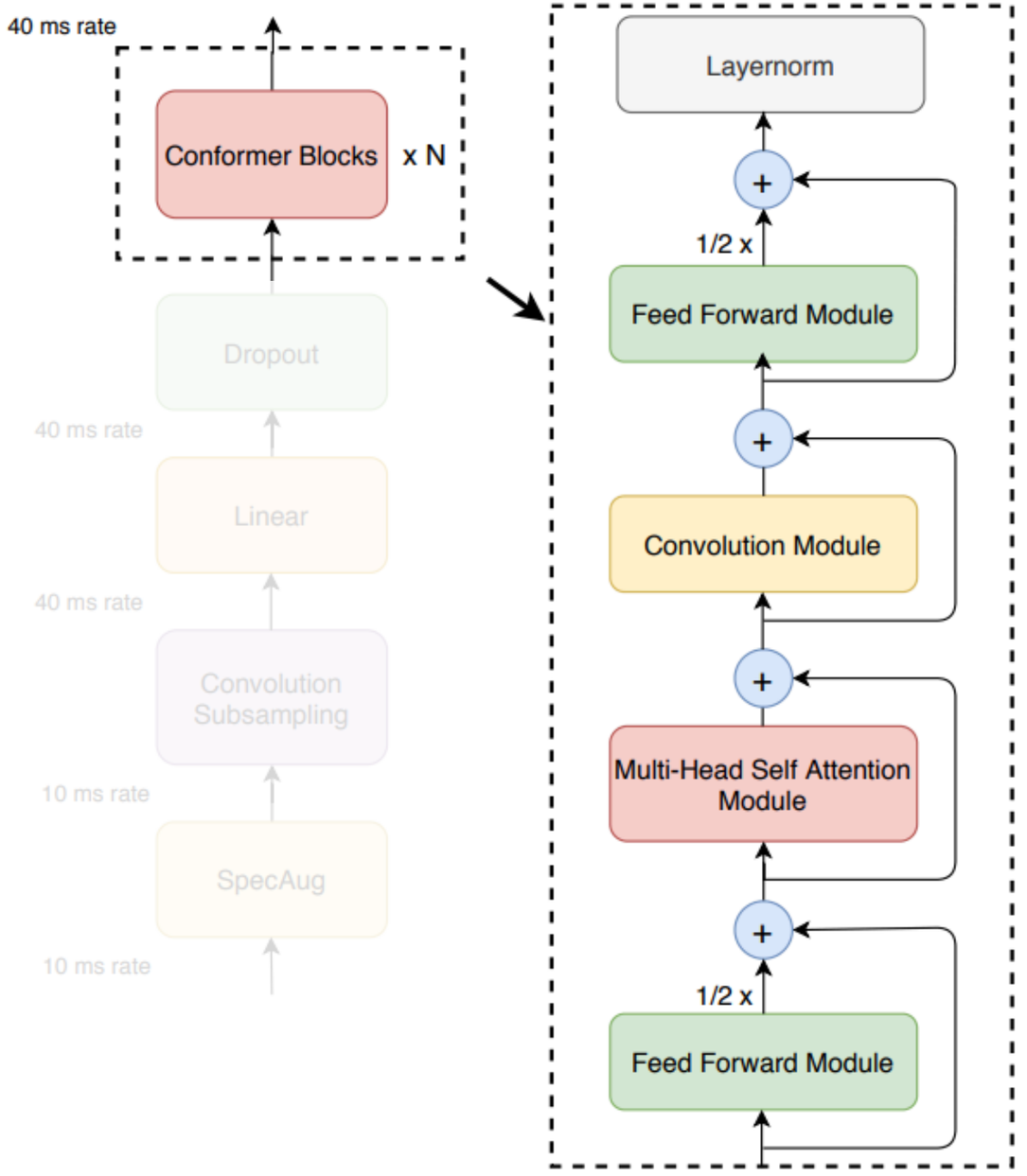


Figure 4: ***Feed forward module.*** *The first linear layer uses an expansion factor of 4 and the second linear layer projects it back to the model dimension. We use swish activation and a pre-norm residual units in feed forward module.*

```python
# --- 1. FeedForward Module (Macaron-style) ---
# FFN = Linear → Activation(Swish) → Linear → Dropout
class FeedForwardModule(nn.Module):
    def __init__(self, d_model, d_ff, dropout=0.1):
        super().__init__()
        self.linear1 = nn.Linear(d_model, d_ff)
        self.activation = nn.SiLU()  # Swish = x * sigmoid(x)
        self.linear2 = nn.Linear(d_ff, d_model)
        self.dropout = nn.Dropout(dropout)

    def forward(self, x):
        return self.dropout(self.linear2(self.activation(self.linear1(x))))
```




```python
# --- 4. Conformer Block ---
# FFN(1) + MHSA + Conv + FFN(2) + Final LayerNorm
class ConformerBlock(nn.Module):
    def __init__(self, d_model, d_ff, n_heads, kernel_size, dropout=0.1):
        super().__init__()
        self.ffn1 = FeedForwardModule(d_model, d_ff, dropout)
        self.mhsa = SelfAttentionModule(d_model, n_heads, dropout)
        self.conv = ConvolutionModule(d_model, kernel_size, dropout)
        self.ffn2 = FeedForwardModule(d_model, d_ff, dropout)
        self.final_norm = nn.LayerNorm(d_model)

    def forward(self, x):
        x = x + 0.5 * self.ffn1(x)  # Macaron FFN #1
        x = x + self.mhsa(x)        # MHSA
        x = x + self.conv(x)        # Conv block
        x = x + 0.5 * self.ffn2(x)  # Macaron FFN #2
        return self.final_norm(x)
```

HOW TO IMPLEMENT?

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.

{anmolgulati, jamesqin, chungchengc, nikip, ngyuzh, jiahuiyu, weihan, shibow, zhangzd, yonghui, rpang}@google.com

Abstract

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs). Transformer models are good at capturing content-based global interactions, while CNNs exploit local features effectively. In this work, we achieve the best of both worlds by studying how to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way. To this regard, we propose the convolution-augmented transformer for speech recognition, named *Conformer*. *Conformer* significantly outperforms the previous Transformer and CNN based models achieving state-of-the-art accuracies. On the widely used LibriSpeech benchmark, our model achieves WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/testother. We also observe competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

Index Terms: speech recognition, attention, convolutional neural networks, transformer, end-to-end

1. Introduction

End-to-end automatic speech recognition (ASR) systems based on neural networks have seen large improvements in recent years. Recurrent neural networks (RNNs) have been the de-facto choice for ASR [1, 2, 3, 4] as they can model the temporal dependencies in the audio sequences effectively [5]. Recently, the Transformer architecture based on self-attention [6, 7] has enjoyed widespread adoption for modeling sequences due to its ability to capture long distance interactions and the high training efficiency. Alternatively, convolutions have also been successful for ASR [8, 9, 10, 11, 12], which capture local context progressively via a local receptive field layer by layer.

However, models with self-attention or convolutions each has its limitations. While Transformers are good at modeling long-range global context, they are less capable to extract fine-grained local feature patterns. Convolution neural networks (CNNs), on the other hand, exploit local information and are used as the de-facto computational block in vision. They learn shared position-based kernels over a local window which maintain translation equivariance and are able to capture features like edges and shapes. One limitation of using local connectivity is that you need many more layers or parameters to capture global information. To combat this issue, contemporary work ContextNet [10] adopts the squeeze-and-excitation module [13] in each residual block to capture longer context. However, it is still limited in capturing dynamic global context as it only applies a global averaging over the entire sequence.

Recent works have shown that combining convolution and self-attention improves over using them individually [14]. Together, they are able to learn both position-wise local features, and use content-based global interactions. Concurrently, papers like [15, 16] have augmented self-attention with relative position based information that maintains equivariance. Wu et al. [17] proposed a multi-branch architecture with splitting the input into two branches: self-attention and convolution; and concatenating their outputs. Their work targeted mobile applications and showed improvements in machine translation tasks.

In this work, we study how to organically combine convolutions with self-attention in ASR models. We hypothesize that both global and local interactions are important for being parameter efficient. To achieve this, we propose a novel combination of self-attention and convolution will achieve the best of both worlds – self-attention learns the global interaction whilst the convolutions efficiently capture the relative-offset-based local correlations. Inspired by Wu et al. [17, 18], we introduce a novel combination of self-attention and convolution, sandwiched between a pair feed forward modules, as illustrated in Fig 1.

Our proposed model, named Conformer, achieves state-of-the-art results on LibriSpeech, outperforming the previous best published Transformer Transducer [7] by 15% relative improve-

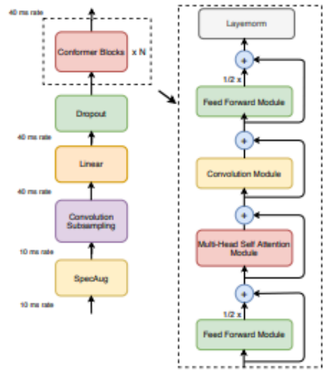


Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*

- CODE
- WEIGHT
- ARCHITECTURE

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.

{anmolgulati, jamesqin, chungchengc, nikip, ngyuzh, jiahuiyu, weihan, shibow, zhangzd, yonghui, rpang}@google.com

Abstract

Recently Transformer and Convolution neural network (CNN) based models have shown promising results in Automatic Speech Recognition (ASR), outperforming Recurrent neural networks (RNNs). Transformer models are good at capturing content-based global interactions, while CNNs exploit local features effectively. In this work, we achieve the best of both worlds by studying how to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way. To this regard, we propose the convolution-augmented transformer for speech recognition, named *Conformer*. *Conformer* significantly outperforms the previous Transformer and CNN based models achieving state-of-the-art accuracies. On the widely used LibriSpeech benchmark, our model achieves WER of 2.1%/4.3% without using a language model and 1.9%/3.9% with an external language model on test/testother. We also observe competitive performance of 2.7%/6.3% with a small model of only 10M parameters.

Index Terms: speech recognition, attention, convolutional neural networks, transformer, end-to-end

1. Introduction

End-to-end automatic speech recognition (ASR) systems based on neural networks have seen large improvements in recent years. Recurrent neural networks (RNNs) have been the de-facto choice for ASR [1, 2, 3, 4] as they can model the temporal dependencies in the audio sequences effectively [5]. Recently, the Transformer architecture based on self-attention [6, 7] has enjoyed widespread adoption for modeling sequences due to its ability to capture long distance interactions and the high training efficiency. Alternatively, convolutions have also been successful for ASR [8, 9, 10, 11, 12], which capture local context progressively via a local receptive field layer by layer.

However, models with self-attention or convolutions each has its limitations. While Transformers are good at modeling long-range global context, they are less capable to extract fine-grained local feature patterns. Convolution neural networks (CNNs), on the other hand, exploit local information and are used as the de-facto computational block in vision. They learn shared position-based kernels over a local window which maintain translation equivariance and are able to capture features like edges and shapes. One limitation of using local connectivity is that you need many more layers or parameters to capture global information. To combat this issue, contemporary work ContextNet [10] adopts the squeeze-and-excitation module [13] in each residual block to capture longer context. However, it is still limited in capturing dynamic global context as it only applies a global averaging over the entire sequence.

Recent works have shown that combining convolution and self-attention improves over using them individually [14]. Together, they are able to learn both position-wise local features, and use content-based global interactions. Concurrently, papers like [15, 16] have augmented self-attention with relative position based information that maintains equivariance. Wu et al. [17] proposed a multi-branch architecture with splitting the input into two branches: self-attention and convolution; and concatenating their outputs. Their work targeted mobile applications and showed improvements in machine translation tasks.

In this work, we study how to organically combine convolutions with self-attention in ASR models. We hypothesize that both global and local interactions are important for being parameter efficient. To achieve this, we propose a novel combination of self-attention and convolution will achieve the best of both worlds – self-attention learns the global interaction whilst the convolutions efficiently capture the relative-offset-based local correlations. Inspired by Wu et al. [17, 18], we introduce a novel combination of self-attention and convolution, sandwiched between a pair feed forward modules, as illustrated in Fig 1.

Our proposed model, named Conformer, achieves state-of-the-art results on LibriSpeech, outperforming the previous best published Transformer Transducer [7] by 15% relative improve-

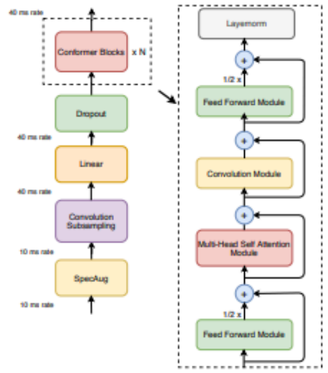
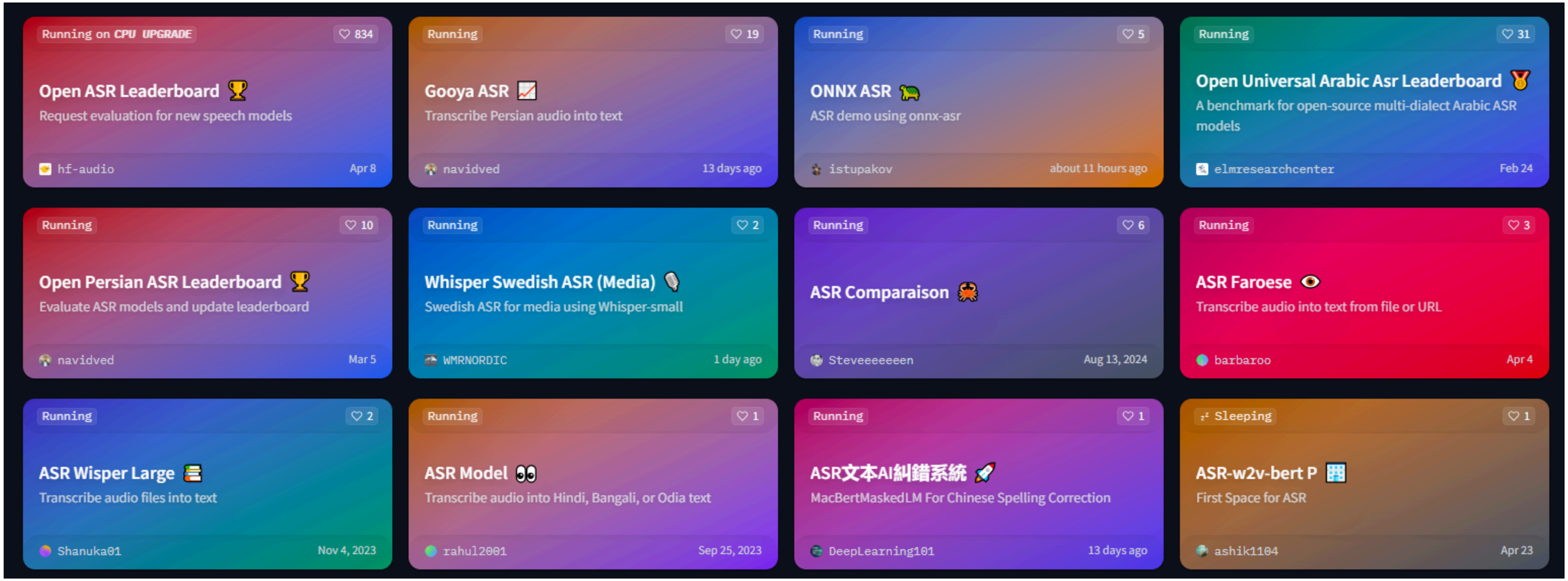


Figure 1: ***Conformer encoder model architecture.*** *Conformer comprises of two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules. This is followed by a post layernorm.*

- CODE
- WEIGHT
- **ARCHITECTURE**

PRE-TRAINED MODEL

PRE-TRAINED MODEL



PRE-TRAINED MODEL

Leaderboard | Metrics | Request a model here! | About

model	Average WER ↓	RTFx ↑	License	AMI	Earnings22	Gigaspeech	LS Clean	LS Other	SPGISpeec
nvidia/parakeet-tdt-0.6b-v2	6.05	3386.02	Open	11.16	11.15	9.74	1.69	3.19	2.17
microsoft/Phi-4-multimodal-instruct	6.14	62.12	Open	11.45	10.5	9.77	1.67	3.82	3.11
nvidia/canary-1b-flash	6.35	1045.75	Open	13.11	12.77	9.85	1.48	2.87	1.95
nvidia/canary-1b	6.5	235.34	Open	13.9	12.19	10.12	1.48	2.93	2.06
nyrahealth/CrisperWhisper	6.67	84.05	Open	8.71	12.89	10.24	1.82	4	2.7
elevenlabs/scribe_v1	6.88	NA	Proprietary	14.43	12.14	9.66	1.79	3.31	3.3
nvidia/parakeet-tdt-1.1b	7.01	2390.61	Open	15.87	14.49	9.52	1.4	2.6	3.16
assemblyai/assembly_best	7.03	NA	Proprietary	15.64	13.54	9.5	1.74	3.11	1.81
revai/fusion	7.12	NA	Proprietary	10.93	12.09	9.41	2.88	6.23	4.05
nvidia/canary-180m-flash	7.12	1233.58	Open	14.86	12.33	10.51	1.73	4.35	2.26
nvidia/parakeet-rnnt-1.1b	7.12	2053.15	Open	17.01	13.94	9.89	1.45	2.5	2.93

“Now it’s your turn to find a paper”