Branchformer: Parallel MLP-Attention Architectures to Capture Local and Global Context for Speech Recognition and Understanding

Yifan Peng, Siddharth Dalmia, Ian Lane, Shinji Watanabe

Carnegie Mellon University

{yifanpen, sdalmia, ianlane, swatanab}@andrew.cmu.edu

Introduction

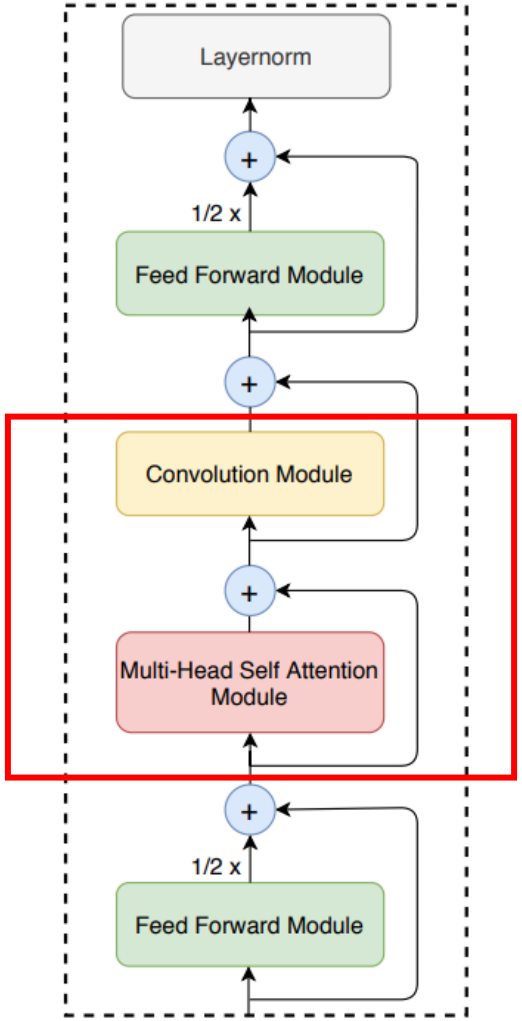
- Various types of neural networks have been applied to speech processing
  - Recurrent Neural Networks (RNNs)
  - Convolutional Neural Networks (CNNs)
  - Transformers with self-attention
  - Multi-Layer Perceptrons (MLPs)
- Different architectures have complementary capacities
- Convolution-augmented Transformer (Conformer) [1] has achieved state-of-the-art results in many speech processing tasks

[1] Gulati et al. Conformer: Convolution-augmented Transformer for Speech Recognition. In Proceedings of Interspeech, 2020.

Conformer Encoder

- Conformer combines self-attention and convolution sequentially
- It outperforms Transformer and convolution-based models

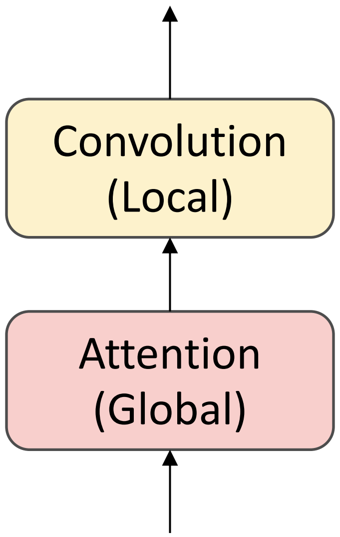
Sequential Combination



Gulati et al. Conformer: Convolution-augmented Transformer for Speech Recognition. In Proceedings of Interspeech, 2020.

Conformer Encoder

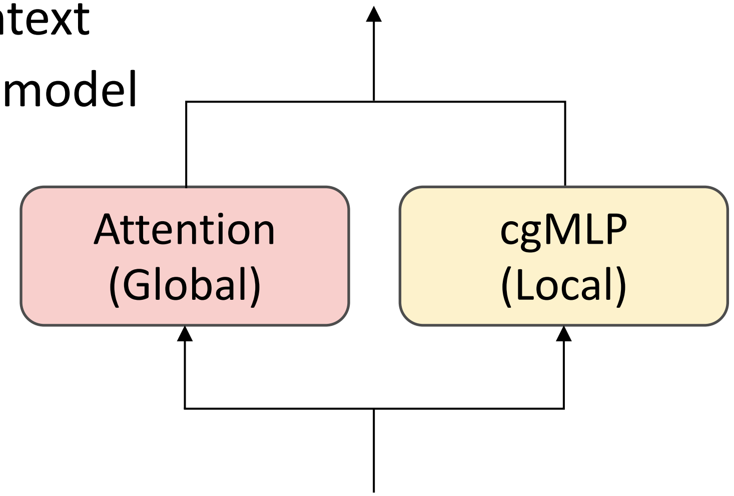
- Conformer combines self-attention and convolution sequentially
- It outperforms Transformer and convolution-based models
- Limitations
  - The static single-branch architecture is difficult to interpret and modify
  - The fixed, interleaving pattern of self-attention and convolution may not always be optimal
  - Self-attention has quadratic complexity w.r.t. the sequence length



Gulati et al. Conformer: Convolution-augmented Transformer for Speech Recognition. In Proceedings of Interspeech, 2020.

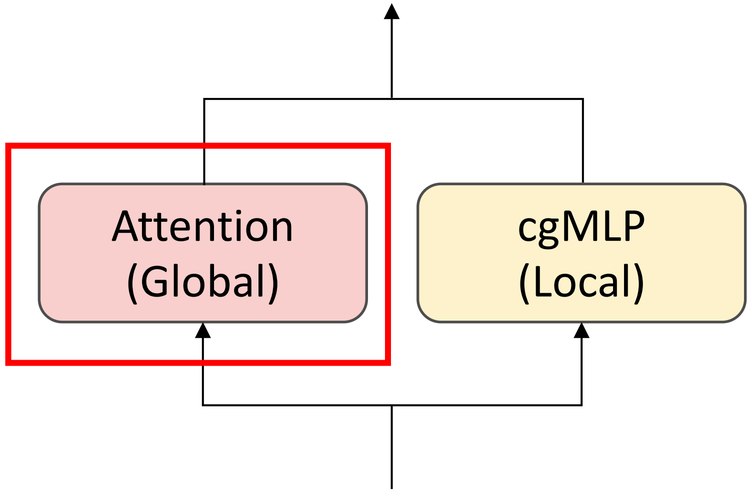
Proposed Model

- We propose a novel encoder alternative, *Branchformer*, with parallel branches for modeling various ranged dependencies
  - Effective in various speech recognition and understanding benchmarks
  - Stable to train for short utterances and limited data
  - Flexible to allow efficient attention variants
  - Interpretable to present interesting analysis on local and global context
  - Customizable to have different inference speeds in a single trained model
- Our code is released as part of ESPnet
  - https://github.com/espnet/espnet



Branchformer Encoder

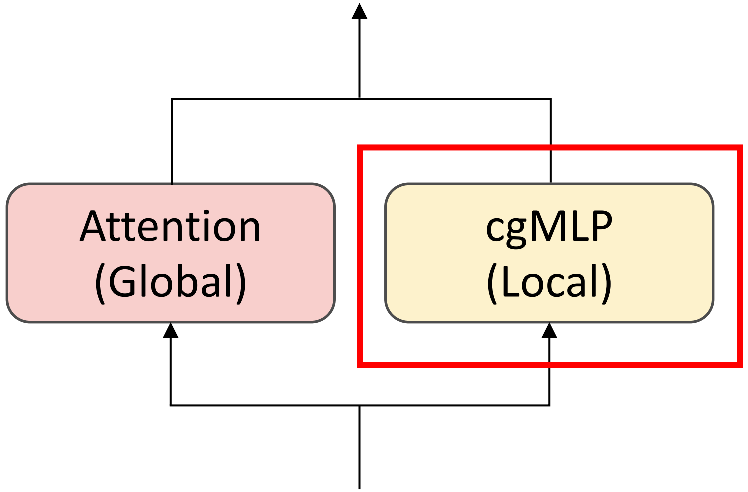
- Attention Branch for Global Context Modeling
  - Multi-headed self-attention
  - Efficient attention variants
    - E.g., Fastformer [1]



[1] Wu et al. Fastformer: Additive attention can be all you need. arXiv preprint arXiv:2108.09084, 2021

Branchformer Encoder

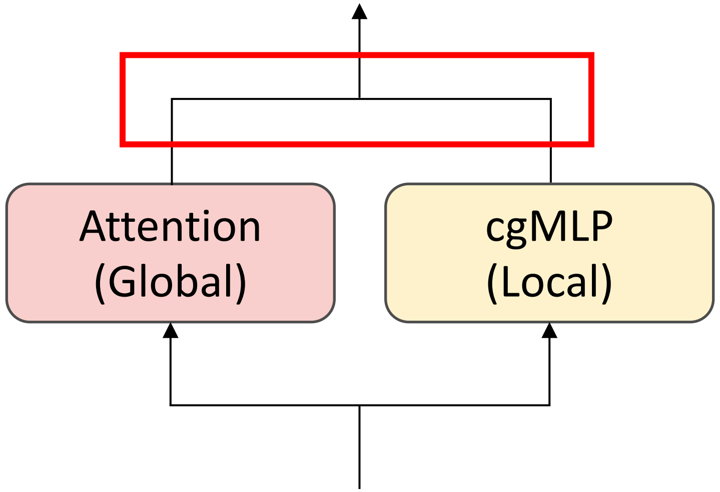
- MLP Branch for Local Context Modeling
  - MLP with convolutional gating (cgMLP) [1]



[1] Sakuma et al. MLP-based architecture with variable length input for automatic speech recognition, 2022. URL https://openreview.net/forum?id=RA-zVvZLYIy

Branchformer Encoder

- Merging Two Branches
  - Concatenation
  - Weighted average



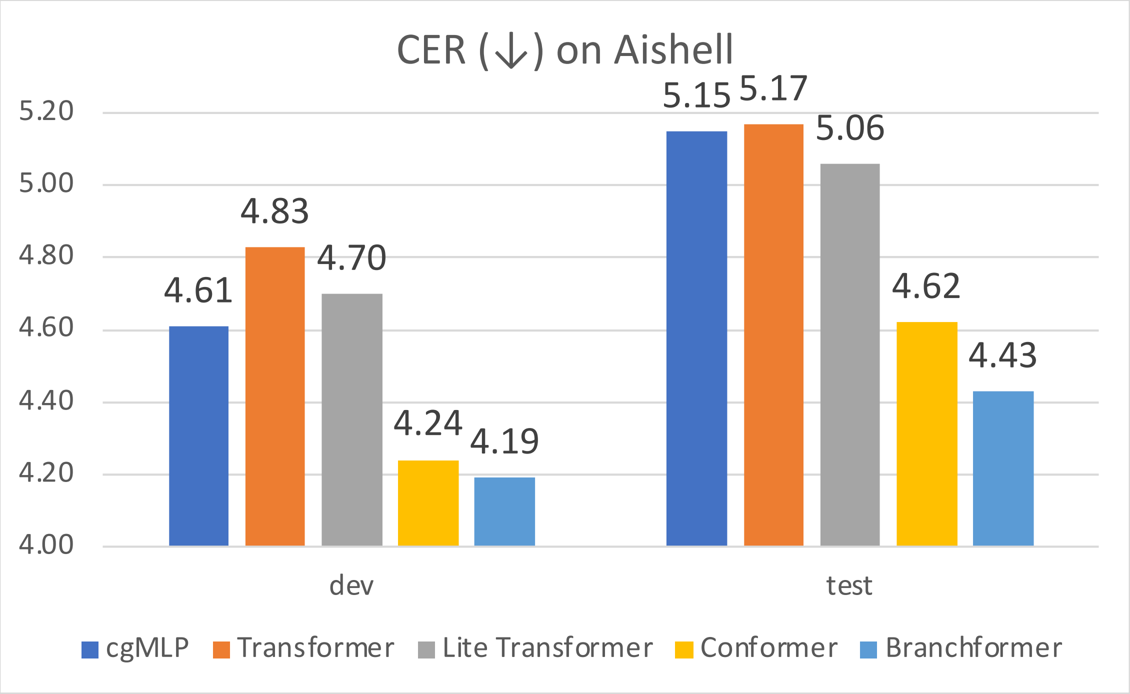
Tasks and Datasets

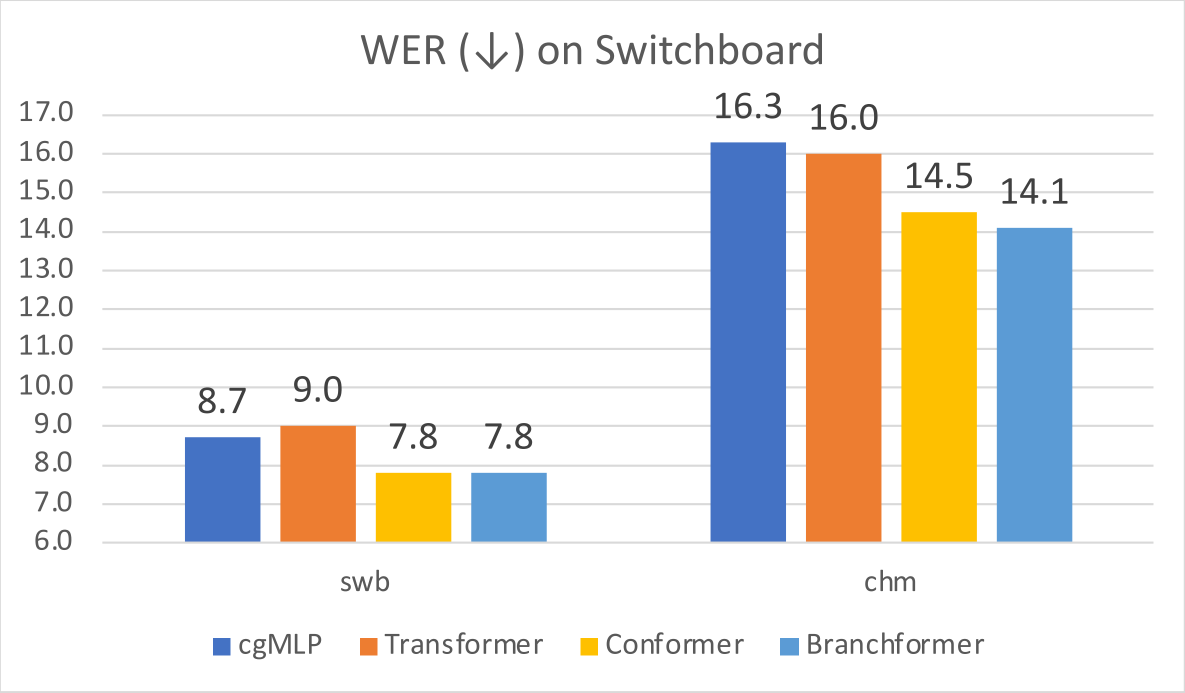
- Automatic Speech Recognition (ASR)
  - Aishell: 170 hours of Mandarin speech data
  - Switchboard (SWBD): 300 hours of English telephone conversations
  - LibriSpeech: 960 hours of English read audiobooks
- Spoken Language Understanding (SLU)
  - SLURP: intent classification and entity prediction
  - Speech Commands: limited-vocabulary speech recognition

Our Branchformer is a general encoder model which can be utilized in other sequence modeling tasks. We also tested the efficacy on machine translation. Preliminary results are shown in Appendix G of our paper.

Main Results: Aishell and SWBD

- We adopt the standard self-attention and concatenation-based merging
- Branchformer outperforms cgMLP and Transformer baselines by a large margin. It matches with or outperforms our reproduced Conformer.

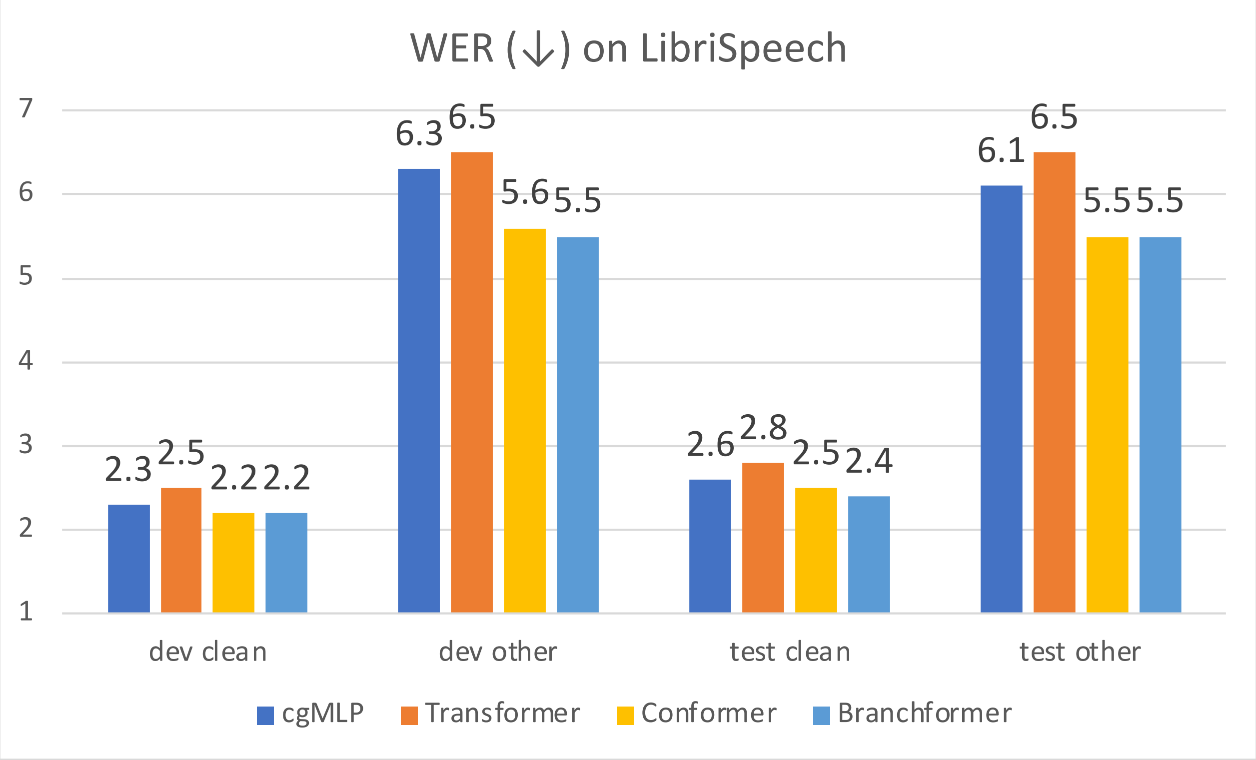


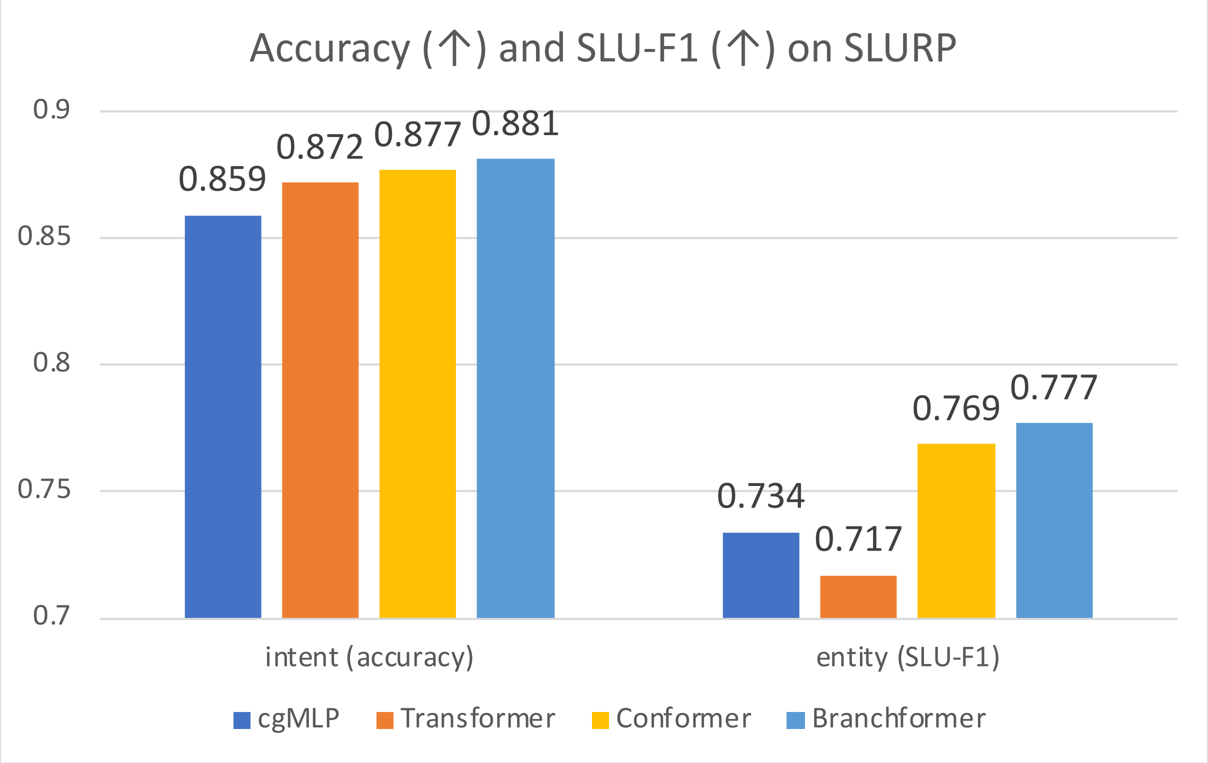


Carnegie Mellon University

Main Results: LibriSpeech and SLURP

- Branchformer outperforms cgMLP and Transformer baselines by a large margin. It matches with or outperforms our reproduced Conformer.





Carnegie Mellon University

Model Scalability

- Branchformer achieves the best performance at the three scales

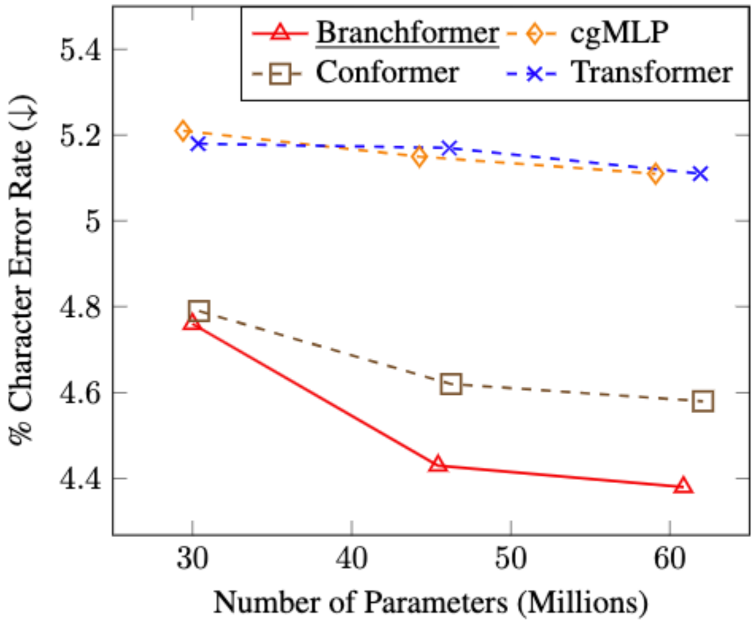


Figure 1. Character Error Rate (%) vs. Model Size. Our Branchformer outperforms previously proposed Conformer, cgMLP and Transformer at all scales on the benchmark Aishell ASR task.

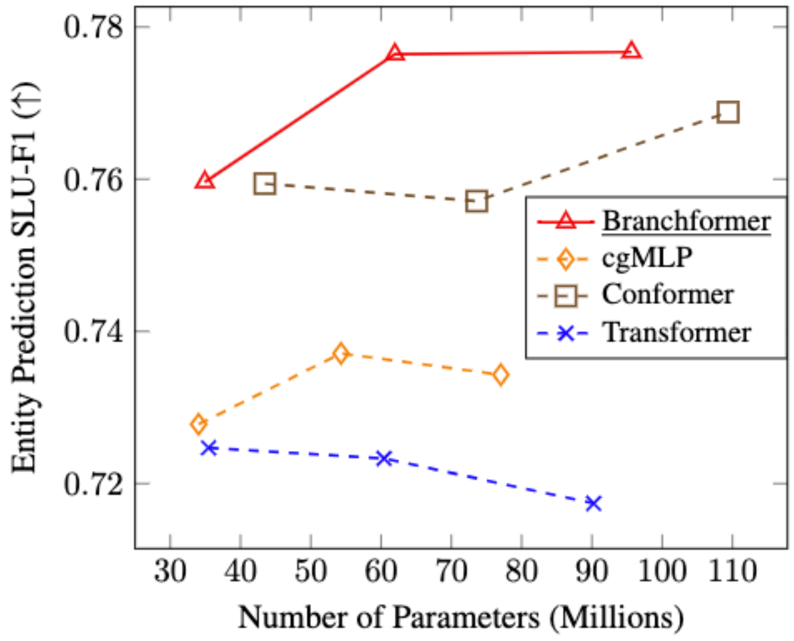


Figure 3. SLURP Entity Prediction SLU-F1 vs. Model Size. Our Branchformer outperforms the previously proposed Conformer, cgMLP and Transformer models at all scales for the SLURP benchmark Spoken Language Understanding task.

Carnegie Mellon University

Training Stability

- We have found that Branchformer is more stable to train than Conformer on short utterances and limited data.

Table 5. Accuracy performance of Branchformer vs. other architectures on Google Speech Commands (35 commands). Training the vanilla Conformer model is unstable on this dataset, but Branchformer achieves similar performance as other models.

Method	Params (M)	Accuracy (↑)	
		dev	test
SpeechBrain (Ravanelli et al., 2021)			
TDNN (+ xvector)	-	-	0.974
ESPnet (Arora et al., 2022)			
Conformer (w/o BatchNorm)	-	**0.974**	**0.975**
Our Baselines (reproduced based on ESPnet)			
cgMLP	30.7	0.966	0.966
Transformer	42.9	**0.973**	**0.974**
Conformer (w/ BatchNorm)	43.0	diverged	
Our Proposed Model			
Branchformer	41.8	**0.973**	0.973

Carnegie Mellon University

Results of Efficient Attention

- Standard self-attention → more efficient attentions such as Fastformer
- The complexity becomes lower, and the performance is still competitive

Table 6. Comparison of the Fastformer-based model with others on Aishell (% CER) and Switchboard 300h (% WER). Fastformer has linear complexity w.r.t. the sequence length T, while self-attention has quadratic complexity. K denotes the convolution kernel size.

Method	Complexity	Aishell		SWBD 300h	
		dev	test	swb	chm
cgMLP	$O(TK)$	4.61	5.15	8.7	16.3
Transformer	$O(T^2)$	4.83	5.17	9.0	16.0
Conformer	$O(T^2)$	4.24	4.62	**7.8**	14.5
Branchformer					
w/ self-attention	$O(T^2)$	**4.19**	**4.43**	**7.8**	**14.1**
w/ Fastformer	$O(TK)$	4.22	4.58	7.9	14.5

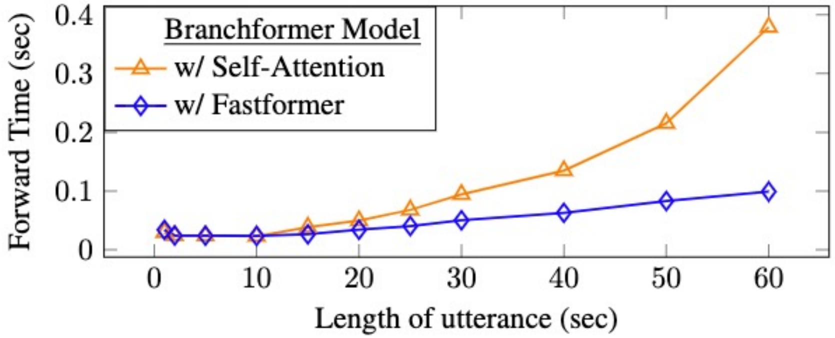


Figure 4. Encoder forward time vs. input audio length using different attention mechanisms for modeling global dependencies in Branchformer. Branchformer w/ Fastformer achieves linear scaling in forward time with different utterance lengths.

Layerwise Analysis of Local/Global Branches

- We use weighted average to merge two branches. The learned weights represent the importance of local and global context in different layers.
- In most layers, one branch is dominant
- Initial layers: interleaved attention and cgMLP
- Intermediate layers: multiple attention
- Final layers: multiple cgMLP
- More results and discussions are in Section 4.6 and Appendix D of our paper

Aishell (24 layers)

Layer	Attention	cgMLP
0	0.538	0.462
1	0.962	0.038
2	0.912	0.088
3	0.078	0.922
4	0.063	0.937
5	0.000	1.000
6	0.957	0.043
7	0.979	0.021
8	0.978	0.022
9	0.059	0.941
10	0.989	0.011
11	0.985	0.015
12	0.990	0.010
13	0.994	0.006
14	0.174	0.826
15	0.992	0.008
16	0.985	0.015
17	0.981	0.019
18	0.912	0.088
19	0.090	0.910
20	0.087	0.913
21	0.069	0.931
22	0.089	0.911
23	0.095	0.905

Aishell (36 layers)

Layer	Attention	cgMLP
0	0.000	1.000
1	0.092	0.908
2	0.075	0.925
3	0.924	0.076
4	0.077	0.923
5	0.060	0.940
6	0.049	0.951
7	0.978	0.022
8	0.935	0.065
9	0.011	0.989
10	0.082	0.918
11	0.991	0.009
12	0.789	0.211
13	0.031	0.969
14	0.990	0.010
15	0.986	0.014
16	0.988	0.012
17	0.988	0.012
18	0.989	0.011
19	0.991	0.009
20	0.993	0.007
21	0.992	0.008
22	0.988	0.012
23	0.993	0.007
24	0.991	0.009
25	0.992	0.008
26	0.980	0.020
27	0.034	0.966
28	0.978	0.022
29	0.060	0.940
30	0.958	0.042
31	0.067	0.933
32	0.065	0.935
33	0.072	0.928
34	0.071	0.929
35	0.042	0.958

Switchboard (24 layers)

Layer	Attention	cgMLP
0	0.213	0.787
1	0.884	0.116
2	0.038	0.962
3	0.077	0.923
4	0.968	0.032
5	0.937	0.063
6	0.052	0.948
7	0.047	0.953
8	0.043	0.957
9	0.000	1.000
10	0.955	0.045
11	0.978	0.022
12	0.974	0.026
13	0.043	0.957
14	0.980	0.020
15	0.967	0.033
16	0.970	0.030
17	0.958	0.042
18	0.044	0.956
19	0.044	0.956
20	0.958	0.042
21	0.069	0.931
22	0.085	0.915
23	0.052	0.948

Model Pruning Using Branch Dropout

- A single Branchformer model can have two different inference speeds
  - During training, the attention branch is dropped at random
  - During inference, the model can work in two modes
    - Mode 1: both branches are employed, which is more accurate but slower
    - Mode 2: only the cgMLP branch is utilized, which has lower complexity
  - This approach does not require fine-tuning or re-training

Conclusion

- We propose Branchformer, a novel encoder architecture with parallel branches for modeling global and local context in speech processing
- Branchformer outperforms Transformer and cgMLP by a large margin in various ASR and SLU benchmarks. It is also comparable with or superior to Conformer.
- Branchformer is stable to train, flexible to allow efficient attentions and interpretable to present interesting design analysis
- With branch dropout, Branchformer can have two inference speeds within a single model

Poster Session

- Location: Hall E #127
- Time: Tue 19 Jul 6:30 p.m. EDT — 8:30 p.m. EDT