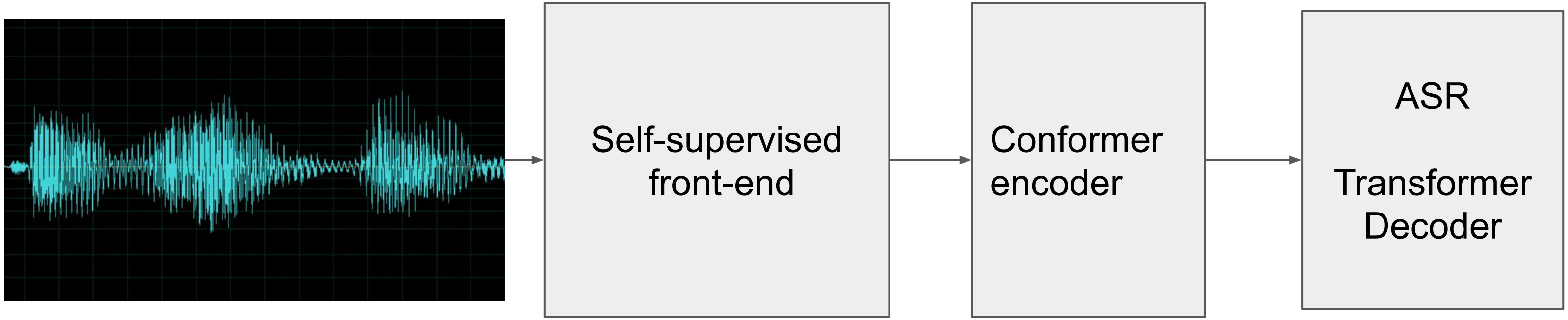
Watanabe's Audio and Voice Lab

Adapters in Speech Transformers

SSL in ESPnet : A Quick re-cap



Two ways we use upstream model features:

1. Freeze and take last layer
2. Weighted sum of hidden states

Watanabe's Audio and Voice Lab

SSL Fine-tuning challenges

1. The entire front-end + pre-encoder + encoder + post-encoder + decoder (A very big graph) difficult to optimize
2. Hyper-parameter search is costly
3. Batch-size difference between pre-training of these SSL models like Wav2vec2 and HuBERT and our fine-tuning creates instability in training

Should we just fine-tune last few layers ?

1. For task specific adaptation recent papers based on empirical analysis suggest re-initializing the last 2 layers
2. But higher pre-trained layers encode more phonemic information which is important for adapting to new languages , then how do fine-tune the higher layers in a parameter efficient manner

Watanabe's Audio and Voice Lab

Table 2: Validation WER on TD, LS, and SB of models pre-trained (PT) on various subsets of {TD, LS, SB}, and fine-tuned (FT) on TD-10h, LS-10h, or SB-10h.

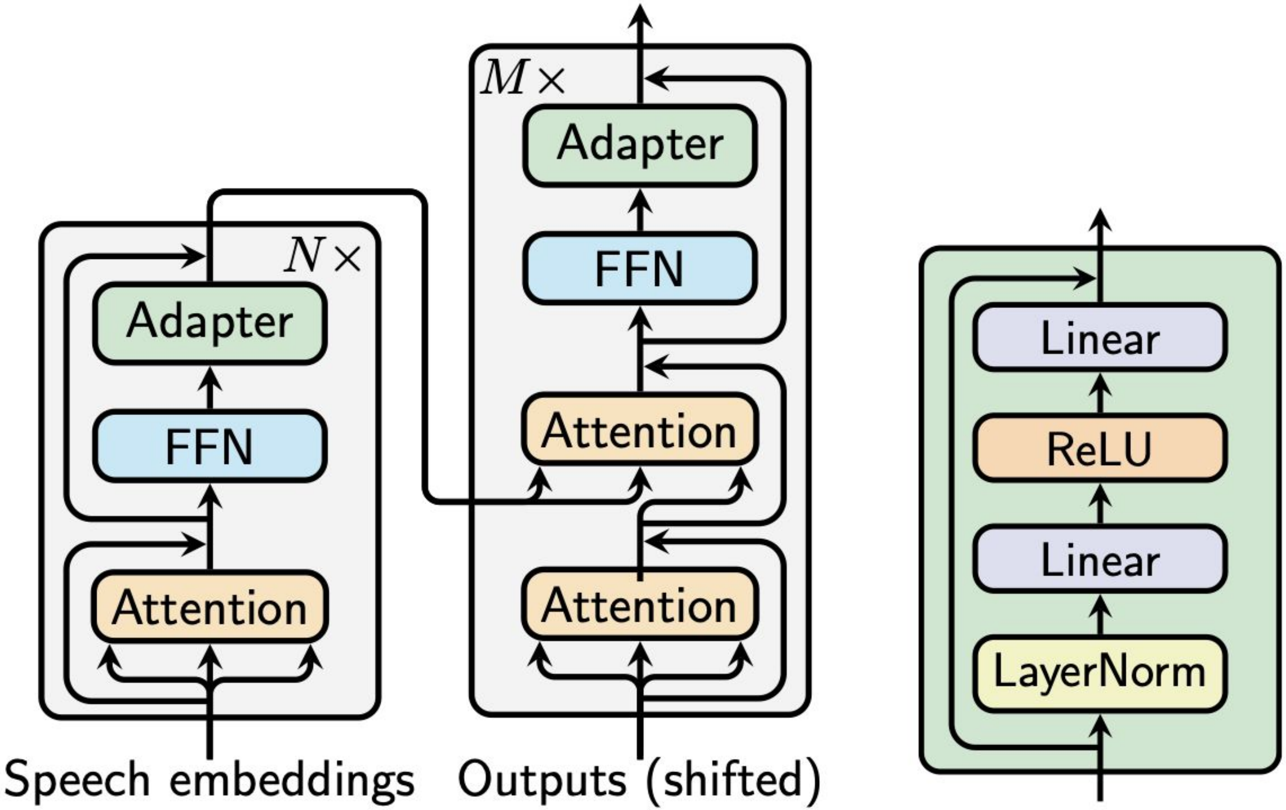
X	*TED-LIUM (TD) dev WER*					
	FT on TD-10h		FT on LS-10h		FT on SB-10h	
	PT on X	X+TD	PT on X	X+TD	PT on X	X+TD
None	diverge	9.93	diverge	10.99	diverge	11.32
SF	12.12	9.60	14.82	11.08	99.63	11.04
LS	9.81	8.59	12.92	8.91	13.08	10.39
SF+LS	9.13	8.91	10.61	9.67	12.25	10.75

X	*LibriSpeech (LS) dev-other WER*					
	FT on TD-10h		FT on LS-10h		FT on SB-10h	
	PT on X	X+LS	PT on X	X+LS	PT on X	X+LS
None	diverge	14.60	diverge	10.53	diverge	17.92
SF	28.91	14.30	20.36	10.44	94.38	15.53
TD	23.44	12.81	15.36	9.71	27.50	15.46
SF+TD	20.50	13.58	14.42	10.39	21.99	13.89

X	*Switchboard (SB) RT03 WER*					
	FT on TD-10h		FT on LS-10h		FT on SB-10h	
	PT on X	X+SF	PT on X	X+SF	PT on X	X+SF
None	diverge	18.90	diverge	19.30	diverge	10.80
TD	35.70	16.20	34.60	17.40	18.70	11.00
LS	33.60	17.80	36.50	16.10	18.20	11.00
TD+LS	29.70	17.40	28.90	16.90	15.60	10.80

Watanabe's Audio and Voice Lab

https://arxiv.org/pdf/2104.01027.pdf

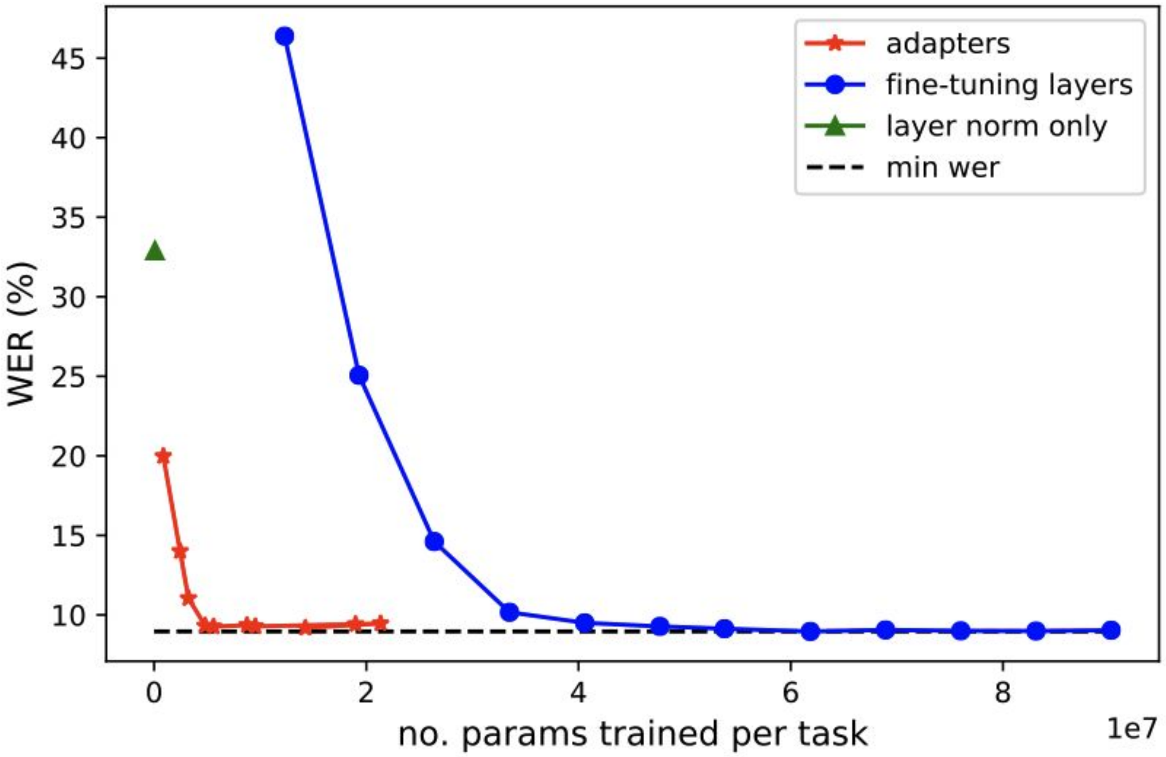


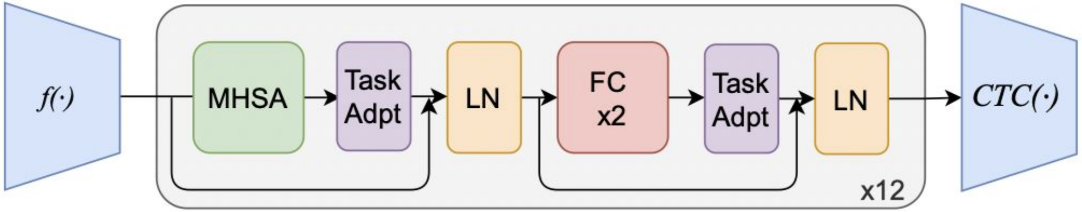
(a) Transformer with adapters. (b) An adapter cell.

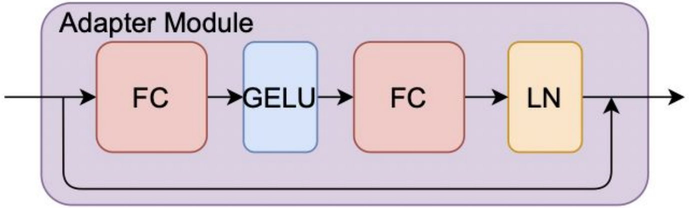
EFFICIENT ADAPTER TRANSFER OF SELF-SUPERVISED SPEECH MODELS FOR AUTOMATIC SPEECH RECOGNITION

Bethan Thomas[†] *Samuel Kessler*[*‡] *Salah Karout*[†]

[†]Huawei R&D UK [‡]University of Oxford


adapters
fine-tuning layers
layer norm only
min wer
WER (%)
no. params trained per task
1e7
45
40
35
30
25
20
15
10
0
2
4
6
8


$f(\cdot)$
MHSA
Task Adpt
LN
FC x2
Task Adpt
LN
x12
$CTC(\cdot)$


Adapter Module
FC
GELU
FC
LN

Combining task and language knowledge

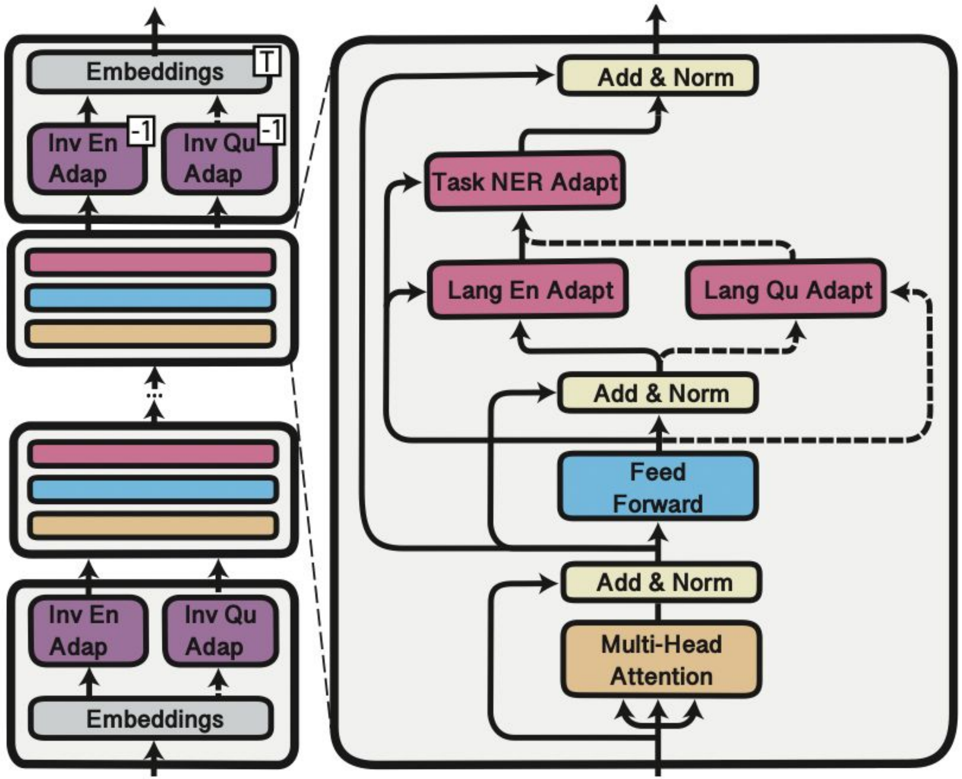
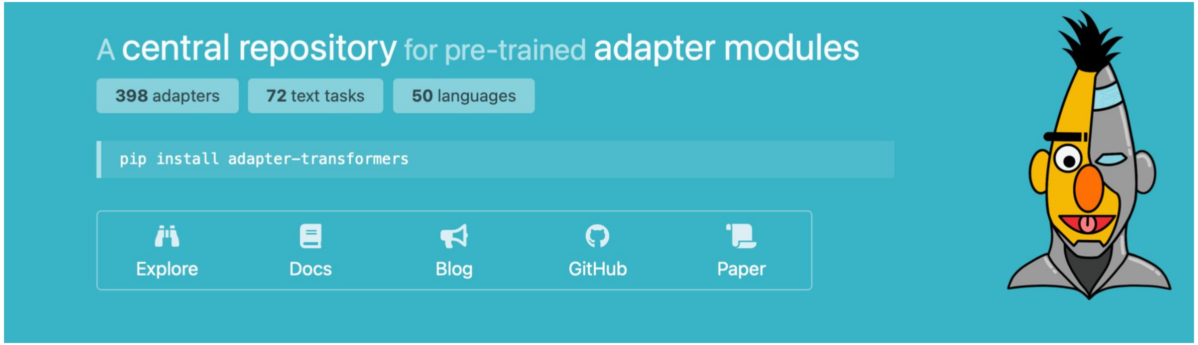


Image source: MAD-X paper https://arxiv.org/pdf/2005.00052.pdf




```
model = AutoModelForSequenceClassification.from_pretrained("bert-base-
model.load_adapter("sentiment/sst-2@ukp")
model.set_active_adapters("sst-2")
```

```
tokenizer = AutoTokenizer.from_pretrained("bert-base-uncased")
tokens = tokenizer.tokenize("AdapterHub is awesome!")
input_tensor = torch.tensor([
    tokenizer.convert_tokens_to_ids(tokens)
])
outputs = model(input_tensor)
```

Watanabe's Audio and Voice Lab

Related work

1. Lightweight Adapter Tuning for Multilingual Speech Translation
   https://arxiv.org/pdf/2106.01463.pdf (Must-C Dataset)

   https://github.com/formiel/fairseq/blob/master/examples/speech_to_text/docs/adapters.md

2. Exploiting Adapters for Cross-lingual Low-resource Speech Recognition
   https://arxiv.org/pdf/2105.11905.pdf (Source code not available,but experiments conducted using ESPnet

Watanabe's Audio and Voice Lab

Possible benefits

1. Task + language adaptation
2. Easy to share pre-trained SSL models for different type of tasks due to smaller memory footprint
3. Support for extreme low resource-setting

Thank you for your attention!

Watanabe's Audio and Voice Lab