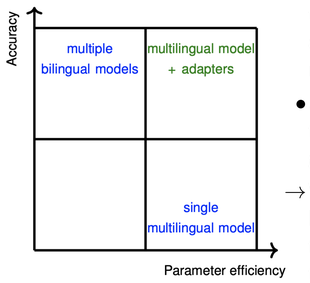
Lightweight Adapter Tuning for Multilingual Speech Translation

Hang Le[1] **Juan Pino**[2] **Changhan Wang**[2]
Jiatao Gu[2] **Didier Schwab**[1] **Laurent Besacier**[1,3]
[1]Univ. Grenoble Alpes, CNRS, LIG [2]Facebook AI [3]Naver Labs Europe

ACL-IJCNLP 2021

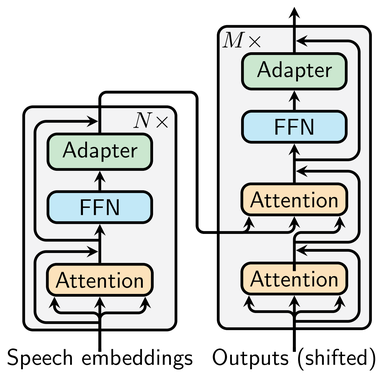
Context and Motivation



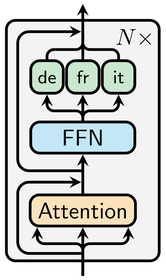
- **Adapter tuning** is one way to achieve models with **high accuracy** and **low maintenance complexity**.
- **Adapter tuning** has been so far **investigated** for **multilingual NMT**.

→ We introduce first comprehensive analysis of **adapters for multilingual speech-to-text translation** (ST).

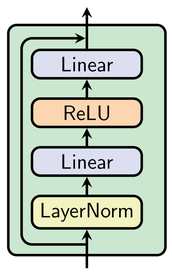
Adapters for multilingual ST



(a) Transformer with adapters at its FFN sub-layers



(b) Language-specific adapters



(c) A typical adapter cell

Adapter tuning for multilingual ST

(1) Pre-train a backbone model.
(2) Add adapters for each language pair.
(3) Finetune adapters on the corresponding bilingual data (the rest of the backbone is frozen).

Two scenarios to evaluate our adapters

(1) *refinement*, and (2) *transfer learning*.

Adapter layers

- Adapters are generally inserted between the layers of a pre-trained network and finetuned on the adaptation corpus.
- Adapter modules can be introduced into a Transformer in a *serial* or *parallel* fashion.

f: a component of the backbone model. g: an adapter layer. Instead of $\mathbf{y} = f(\mathbf{x})$, the new "adapted output" is given by:

$$\mathbf{y}_{\text{serial}} = g(f(\mathbf{x})) \qquad \mathbf{y}_{\text{parallel}} = f(\mathbf{x}) + g(\mathbf{x}).$$

Experimental setup

1. **Dataset:** MuST-C, from English to 8 target languages.
   (1) *MuST-C original:* from 385 hours (pt) to 504 hours (es).
   (2) *MuST-C imbalanced:* from 41 hours (de) to 504 hours (es).
2. **Model:** 12-layer encoder + 6-layer decoder.
   (1) $D = 256$ (small).
   (2) $D = 512$ (medium).
3. **Vocabulary:**
   - Bilingual models: 8K.
   - Multilingual models: 10K.
4. **Speech pre-processing:**
   - 80-d log mel filter-bank.
   - SpecAugment with Librispeech basic (LB) policy.

Refinement

Training settings: Finetune a (fully) pre-trained multilingual ST backbone on each language pair to boost performance and close potential gaps with bilingual models.

- **Partial fine-tuning:** Fine-tune only some components (e.g. encoder or decoder) on each language pair.
- **Full fine-tuning:** Fine-tune all the backbone on each pair.
- **Adapter tuning:** Add language-specific adapters to backbone and fine-tune them only.

Transfer learning

- **Training settings:** Initialize Transformer with pre-trained ASR encoder and mBART50 decoder, then fine-tune only some components on (multilingual) ST dataset.
- **Different configurations for comparison:**
  - With and without (language-specific) adapters.
  - Adapters added to decoder only, or both encoder & decoder.
  - Cross (encoder-decoder) attention is fine-tuned or frozen.

Experimental results for Refinement

	Dict	D	Adapter d	Adapter ENC	Adapter DEC	Finetune ENC	Finetune DEC	# params (M) trainable/total	de	es	fr	it	nl	pt	ro	ru	avg
MuST-C original training data (hours)									408	504	492	465	442	385	432	489	
1	mono		-	-	-	-	-	8×31.1/8×31.1	22.16	30.42	27.92	22.92	24.10	27.19	21.51	14.36	23.82
2	multi		-	-	-	-	-	32.1/32.1	22.37	30.40	27.49	22.79	24.42	27.32	20.78	14.54	23.76
3	multi	256	128	-	✓	-	-	8×0.4/35.3	22.45	30.85	27.71	23.06	24.57	27.52	20.93	14.57	23.96
4	multi	256	128	✓	✓	-	-	8×1.2/41.7	22.84*	31.25*	28.29*	23.27*	24.98*	28.16*	21.36*	14.71	24.36
5	multi		-	-	-	-	✓	8×14.6/8×32.1	23.49	31.29	28.40	23.63	25.51	28.71	21.73	15.22	24.75
6	multi		-	-	-	✓	✓	8×32.1/8×32.1	23.13*	31.39*	28.67*	23.80*	25.52*	29.03*	22.25*	15.44*	24.90
7	mono		-	-	-	-	-	8×74.3/8×74.3	21.93	30.46	27.90	22.64	23.98	25.98	20.50	14.01	23.42
8	multi		-	-	-	-	-	76.3/76.3	23.98	32.47	29.24	24.97	26.20	29.81	22.74	15.30	25.59
9	multi		256	-	✓	-	-	8×1.6/89.1	24.38	32.78	29.69	24.72	26.25	29.93	22.63	15.40	25.72
10	multi		256	✓	✓	-	-	8×4.8/114.7	24.61	32.94	29.67	**25.12**	26.16	**30.53**	22.66	15.31	25.88
11	multi	512	-	-	-	-	✓	8×35.5/8×36.3	**24.67**	**33.12**	**30.11**	25.05	**26.33**	29.85	**23.04**	**15.61**	**25.97**
12	multi	512	-	-	-	✓	✓	8×76.3/8×76.3	24.54*	32.95*	29.96*	25.01	26.31	30.04	22.66	15.54*	25.88
MuST-C Imbalanced training data (hours)									41	504	492	232	89	38	86	245	
1	multi	256	-	-	-	-	-	32.1/32.1	15.99	30.51	28.17	21.80	20.27	22.47	17.38	13.18	21.22
2	multi	256	128	✓	✓	-	-	8×1.2/41.7	17.02	30.71	28.42	22.37	21.01	23.74	18.55	13.52	21.92
3	multi	256	-	-	-	✓	✓	8×32.1/8×32.1	16.93	30.86	28.34	22.42	20.86	23.44	18.49	13.63	21.87
4	multi	512	-	-	-	-	-	76.3/76.3	17.05	31.92	29.06	22.91	21.64	24.15	19.18	14.09	22.50
5	multi	512	256	✓	✓	-	-	8×4.8/114.7	17.46	**31.94**	29.09	**23.11**	21.76	**24.96**	**19.50**	14.10	**22.74**
6	multi	512	-	-	-	✓	✓	8×76.3/8×76.3	**17.49**	31.67	**29.27**	22.97	**21.80**	24.80	19.43	**14.17**	22.70

- Both adapter tuning and fine-tuning yield improvements over the multilingual baseline.
- Adding adapters to the encoder improve the overall performance.
- On MuST-C original, fine-tuning slightly outperforms adapter-tuning.
- On MuST-C Imbalanced, adapter-tuning achieved the best performance, especially for the low-resource languages (pt, de, ro nl).

Experimental results for Transfer learning

	Adapter d	Adapter ENC	Adapter DEC	Finetune xattn	# params (M) trainable/total	de	es	fr	it	nl	pt	ro	ru	avg
1	-	-	-	-	8×31.1/8×31.1	22.16	30.42	27.92	22.92	24.10	27.19	21.51	14.36	23.82
2	-	-	-	✓	38 / 486	18.41	25.42	23.46	18.44	20.87	20.55	17.19	11.79	19.52
3	512	-	✓	-	101 / 587	0.94	0.65	0.93	0.76	0.95	0.89	0.52	0.93	0.82
4	512	-	✓	✓	139 / 587	21.98	29.47	27.05	22.89	24.06	26.34	21.0	14.35	23.39
5	512	✓	✓	-	152 / 638	11.04	18.62	16.10	12.37	13.18	14.29	10.62	6.95	12.90
6	512	✓	✓	✓	190 / 638	22.62	30.85	28.23	23.09	24.43	26.56	22.13	14.92	24.10

- Fine-tuning cross-attention is crucial to transfer to multilingual ST.
- Adding adapters to the backbone decoder or to both encoder and decoder further boosts performance.

→ Adapters is able to connect off-the-shelf models in a modular fashion.