To Adapt or to Fine-tune: A Case Study on Abstractive Summarization

Zheng Zhao Pinzhen Chen

School of Informatics, University of Edinburgh

{zheng.zhao, pinzhen.chen}@ed.ac.uk

Main Contributions

In this work, we carry out multifaceted investigations on fine-tuning and adapters for summarization tasks with varying complexity:

1. **languages involved:** monolingual, cross-lingual, and multilingual;
2. **data availability:** high, medium, low, and scarce resources;
3. **knowledge being transferred:** languages, domains, and tasks.

In our experiments, we find that:

1. fine-tuning a pre-trained language model is superior to using adapters;
2. the performance gap positively correlates with the amount of training data used;
3. adapters exceed fine-tuning under extremely low-resource conditions.

Methodology

Our aim is to study two fine-tuning variants for summarization under several settings using a PLM: the ***fine-tuning*** paradigm, and the ***adapter*** strategy.

mBART-FT initializes a mBART model from a pre-trained checkpoint, then trains and updates the whole model on a summarization dataset. We provide a cross-lingual demonstration for our model in Figure 1.

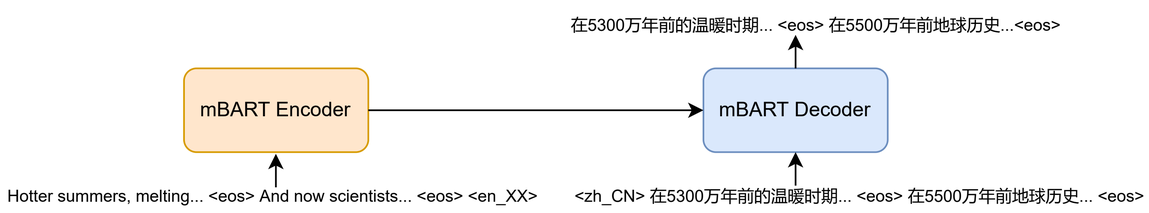
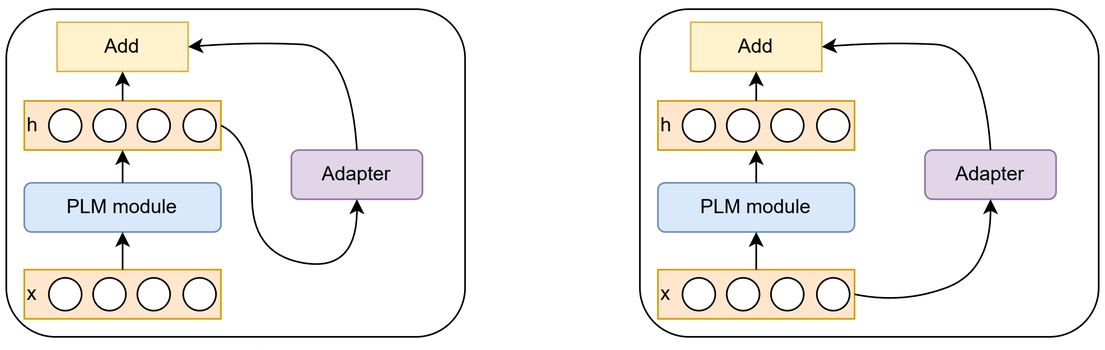


Figure 1. An illustration of mBART-FT for cross-lingual summarization from English to Chinese.

mBART-Adapt also initializes a mBART model from a pre-trained checkpoint, with adapter modules then inserted into the model. We experiment with two adapter variants: sequential and parallel, illustrated in Figure 2.



(a) A sequential adapter (b) A parallel adapter

Figure 2. An illustration of adapter variants.

Results for Domain Experiments

We conduct experiments on domain adaptation, which is typically tackled using the same pre-training then fine-tuning paradigm.

In our setting, we adapt CNN/Daily Mail to XL-Sum, both in English, with various data sizes. In addition to the XL-Sum dataset, which is in the news domain, we also experimented with adapting to the BookSum dataset, a collection of narratives from the literature domain. We provide results in Table 1.

Domain	Data Size		BART-FT			BART-Adapt		
			R1	R2	RL	R1	R2	RL
XL-Sum	original	306.5k	**34.48**	**14.73**	**28.93**	32.94	13.46	27.60
	medium	30.65k	**30.63**	**11.38**	**25.31**	30.15	11.10	25.05
	small	3065	27.27	**8.91**	**22.27**	**27.32**	8.79	22.20
	tiny	307	24.10	**6.52**	19.38	**24.29**	6.41	**19.50**
	micro	31	19.69	4.26	15.73	**20.74**	**4.65**	**16.45**
BookSum	111.6k		**20.27**	**4.01**	15.50	20.22	3.95	**15.57**

Table 1. Results for domain adaptation from CNN/Daily Mail to XL-Sum/BookSum.

Results for Language Experiments

We provide results on high-recourse cross-lingual summarization on NCLS in Table 2. Table 3 lists results on medium and low-recourse cross-lingual summarization on WikiLingua. We also provide results of both monolingual (Table 4) and multilingual (Table 5) summarization on XL-Sum.

Lang.	mBART-FT			mBART-Adapt		
	R1	R2	RL	R1	R2	RL
zh→en	**46.46**	**30.18**	**42.26**	41.41	22.73	36.56
en→zh	**45.22**	**22.49**	**34.38**	40.74	16.83	29.27

Table 2. High-resource, NCLS.

Lang.	mBART-FT			mBART-Adapt		
	R1	R2	RL	R1	R2	RL
en→ar	**25.85**	**7.35**	**21.01**	24.68	7.26	20.40
en→vi	**33.63**	**15.17**	**26.65**	30.98	13.94	24.59
en→ja	**35.70**	**12.34**	**28.34**	34.06	11.43	27.08
ja→en	**35.24**	**12.38**	**28.09**	33.14	11.54	26.46

Table 3. Medium and low-resource, WikiLingua.

Lang.	Monolingual					
	mBART-FT			mBART-Adapt		
	R1	R2	RL	R1	R2	RL
gu	**20.23**	**6.43**	**17.67**	19.20	5.95	16.96
fr	**33.29**	**13.68**	**25.13**	32.37	13.02	24.73
ne	**24.06**	**9.05**	**21.62**	23.31	8.36	21.01
ko	**19.73**	9.12	**18.07**	19.05	**9.24**	17.73
si	**25.59**	12.25	**21.92**	24.99	**12.30**	21.44

Table 4. Results for low-resource monolingual summarization on XL-Sum.

Lang.	Multilingual					
	mBART-FT			mBART-Adapt		
	R1	R2	RL	R1	R2	RL
gu	**20.18**	**6.96**	**18.09**	20.12	6.82	17.99
fr	**33.53**	**14.37**	**26.11**	33.44	14.01	25.63
ne	**24.70**	**9.52**	**22.23**	23.26	8.55	20.94
ko	17.73	**8.76**	16.27	**18.82**	8.12	**17.23**
si	**26.95**	**13.51**	**22.36**	25.68	12.69	21.80

Table 5. Results for low-resource multilingual summarization on XL-Sum.

Results for Task Transfer

In addition to experiments with the fine-tuning paradigm on the subject of language and domain adaption, we also experiment with adapting a news summarization model to dialogue summarization. We report the experiment results in Table 6.

Task	Data Size	Model	R1	R2	RL
DialogSum	12.5k	BART-FT	**47.40**	**24.66**	**39.03**
		BART-Adapt	47.24	24.57	38.56
SAMSum	14.7k	BART-FT	**49.52**	**24.91**	40.64
		BART-Adapt	49.38	24.69	**40.99**

Table 6. Results for task adaption from CNN/Daily Mail to DialogSum and SAMSum.

Effect of Data Availability on Performance

We observe that the amount of training data affects the performance gap between the two fine-tuning and adapters. We plot the percentage change in ROUGE performance (between those of fine-tuning and those of adapters) against the training size (log-scale) in Figure 3.

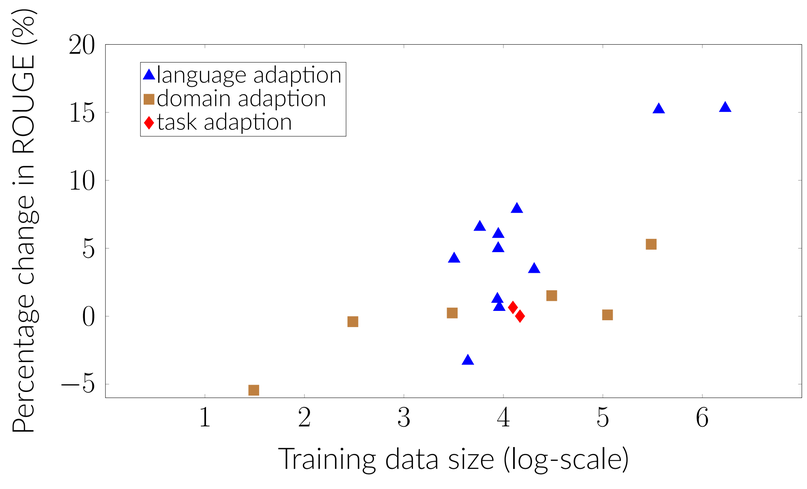


Figure 3. The effect of the training data size on ROUGE difference between the fine-tuning and adapter strategy.

Acknowledgement

Zheng Zhao is supported by the UKRI Centre for Doctoral Training in Natural Language Processing (grant EP/S022481/1). Pinzhen Chen is supported by a donation to Kenneth Heafield. This work reflects the view of the authors, and not necessarily that of the funders.