Google DeepMind

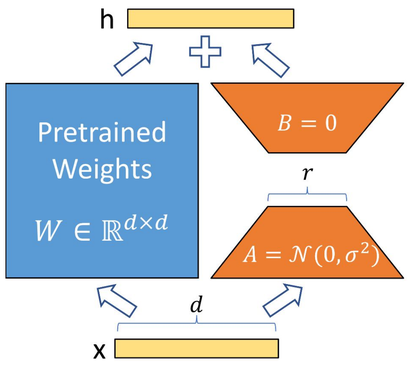
Low-Rank Adaptation for Multilingual Summarization: An Empirical Study

Chenxi Whitehouse,[1] Fantine Huot,[2] Jasmijn Bastings,[2] Mostafa Dehghani,[2] Chu-Cheng Lin,[3] Mirella Lapata[2]
[1] University of Cambridge [2] Google DeepMind [3] Google Cloud

Introduction

- LLMs are becoming increasingly powerful, but the growing size also makes training less practical
- Parameter-efficient fine-tuning (PEFT) approaches are desirable, especially for tasks requiring extensive memory, e.g., with long input
- We focus on LoRA (Low-rank Adaptation) and study a challenging task with long input:
  - Multilingual summarization, where LoRA is under-explored
  - Empirically evaluate LoRA vs Full Fine-tuning (FT) under different data availability scenarios

Low-Rank Adaptation



[Hu et al. 2021]

- Freezes the pre-trained model weights (W) and adds trainable low-rank matrices (A & B) into the Transformer architecture
- No extra cost or latency at inference time
  - Can merge LoRA with the frozen parameters.
- Marginal trainable parameters and less GPU
- Competitive performance vs Full Fine-tuning (on classification or monolingual generation tasks)

LoraHub [Huang et al. 2023]:

- Compose individually trained LoRA modules for cross-task generalization
- Available LoRA modules m_i are synthesized into module $\hat{m} = \sum_{i=1}^{N} w_i m_i$

LoRA for Multilingual Summarization (PaLM-2)

Multilingual Summarization is Complex:

- Models: fluently generate in many languages
- High/low resource: not all languages have (sufficient) data
- Long input and output

Datasets & Metrics:

Dataset	**XLSum**	**XWikis**
Source	BBC News	Wikipedia
Languages	44	5
Train/Val/Test Data	1.1M / 114K / 114K	1.4M / 40K / 35K
Input/Output Words	470 / 22	1043 / 64
- Compare summary **Relevance** (Rouge-L), **Faithfulness** (NLI), and **Conciseness** (Seahorse)

Different Data Regimes:

- High-data
- Low-data
- Cross-lingual Transfer (zero- and few-shot)

High-data Regime

- Train on all the data available for each language
- Full FT > LoRA on summary relevance (R-L). LoRA with higher ranks enhances summary relevance
- LoRA is superior on summary faithfulness (NLI) & conciseness (SH), lower ranks see better scores

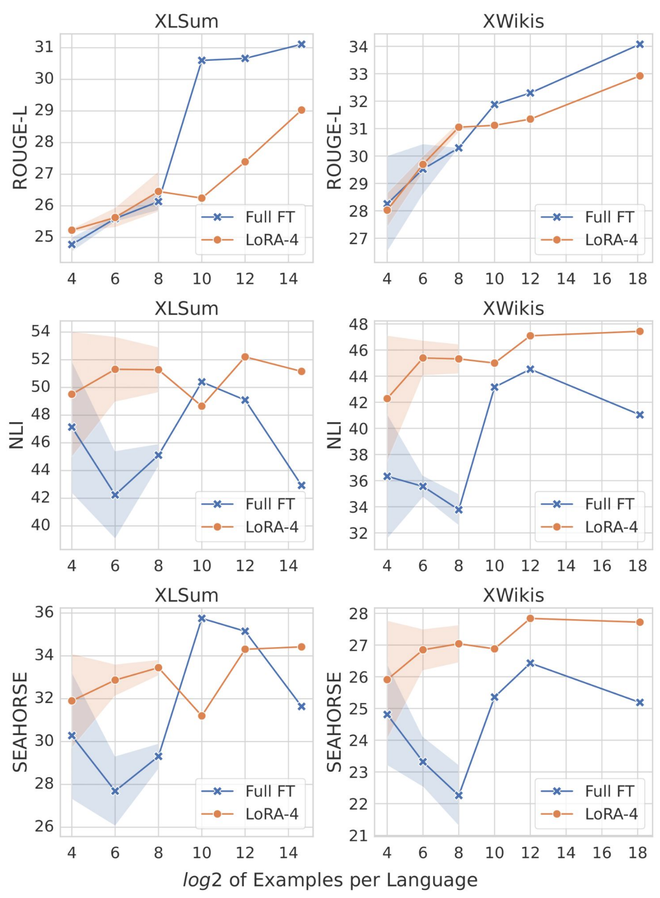
	XLSum			**XWikis**		
	R-L	NLI	SH	R-L	NLI	SH
Full-FT	**31.11**	42.93	31.64	**34.08**	41.04	25.19
LoRA-64	29.79	45.51	31.80	34.04	45.34	27.02
LoRA-16	29.77	48.48	33.25	33.80	46.10	27.42
LoRA-4	29.03	**51.16**	**34.42**	32.92	**47.43**	**27.72**

Average scores of all languages when training on all available data

Takeaway: For High-Data, FT is better for summary relevance, LoRA generates more factual and concise summaries.

Low-Data Regime

- Randomly select 16, 64, 256, 1024, 4096 data per language and train together (balanced data)
- LoRA achieves overall better faithfulness (NLI) and conciseness (Seahorse) than Full FT
- For ROUGE-L, Full FT outperforms LoRA when provided > 1K training examples
- Low-data training on LoRA is more stable (Full FT more sensitive for the selection of checkpoints)



Takeaway: For Low-Data, LoRA is superior than FT across metrics and the learning curve is smoother and more stable.

Cross-lingual Transfer

Zero-shot transfer from English

- FT exhibits catastrophic forgetting

	XLSum			**XWikis**		
	Rouge-L	NLI	Seahorse	Rouge-L	NLI	Seahorse
Full-FT	5.20	4.49	6.88	17.51	35.95	22.43
LoRA-4	**21.13**	**39.07**	**23.08**	**23.86**	**45.54**	**25.96**

Performance of average non-English languages when training on English

Hausa

Target: Gwamnatin Najeriya ta ce 'yan kasar sun ga irin amfani da rufe iyakokin kasar ya yi a fannin tattalin arzikinta

Full FT: President Muhammadu Buhari has appointed his deputy, the BBC presenter and former minister, Shugaba Muhammadu Buhari, as the new chairman of the Presidential Council.

LoRA-4: Gwamnatin Nijeriya ta yi tsokacin da shawarar da zai rufe iyakokin kasar.

Indonesian

Target: Perempuan Vietnam yang dituding terlibat dalam pembunuhan Kim Jong-nam, saudara tiri dari pemimpin Korea Utara Kim Jong-un, telah dibebaskan.

Full FT: Kim Jong-nam, the wife of North Korean leader Kim Jong-un, has died in a fight with Malaysia Airlines flight MH17. Here are the key points of the ruling:

LoRA-4: Seorang wanita Vietnam yang didakwa sebagai bagian dari pembunuhan Kim Jong-nam, saudara tiri dari pemimpin Korea Utara, telah dibebaskan.

Zero-shot transfer from Multiple Languages

- Multilingual LoRA and Weight averaging of individual LoRA benefits different unseen languages
- Lower resource languages (Kirundi, Scottish, Somali, Yoruba) work best with individual LoRA training
- Similar languages may transfer better

SEEN \ UNSEEN	AZ	BN	JA	RN	KO	NE	GD	SO	TH	YO
AR	15.42	23.38	28.20	10.29	23.78	21.91	16.75	14.94	23.35	19.00
ZH	14.46	22.11	30.85	8.25	22.33	22.77	16.02	14.40	23.12	16.53
EN	15.12	22.24	28.91	8.90	23.09	23.43	15.54	18.30	22.23	20.85
HA	15.67	22.26	27.49	10.59	21.90	22.17	16.20	18.09	20.47	19.40
HI	13.60	22.71	28.81	9.75	21.31	24.96	18.13	12.90	22.54	19.30
ID	17.07	23.91	29.41	10.47	24.82	23.64	20.66	19.26	22.94	19.51
FA	10.66	22.15	27.59	10.19	20.77	20.68	16.26	15.86	22.28	17.62
PT	15.05	22.32	28.13	7.82	22.84	22.27	16.78	15.26	21.34	18.52
SW	17.10	22.69	28.67	11.87	24.37	24.84	18.18	18.74	21.42	19.49
TR	12.16	21.46	27.49	9.79	20.30	20.23	16.78	15.67	21.71	18.44
Full FT	15.89	5.97	22.61	13.17	8.45	21.72	17.92	12.15	13.17	13.75
LoRA-4	19.94	26.25	32.15	10.23	26.26	27.38	19.16	20.26	25.37	18.87
Avg. LoRA	18.22	23.05	29.71	16.25	25.03	24.57	22.67	21.51	23.42	22.96

ROUGE-L scores for 10 test languages on XLSum

Few-shot transfer from Multiple Languages

- Assume a handful target examples available (16, 64), compare LoRA continued learning (CL) and LoraHub
- LoRA continued learning superior performance
- A few examples significantly improves Full FT compared to the zero-shot results

Zero-shot				16-shot			
	R-L	NLI	SH		R-L	NLI	SH
Full FT	14.48	28.87	13.71	Full FT	22.31	30.15	18.79
LoRA	22.59	37.39	24.21	LoRA (CL)	**24.71**	**41.12**	**26.47**
Avg.LoRA	**22.74**	**49.14**	**32.44**	LoraHub	23.37	38.95	26.07

Zero-and 16-shot scores for average of 10 test languages on XLSum

Takeaway: For Cross-lingual Transfer, LoRA is a clear winner, FT suffers from catastrophic forgetting.

Conclusions

Main Takeaway: For multilingual summarization, LoRA is not only efficient, but overall a better option across different data availability and evaluation metrics!

- LoRA achieves **superior performance** vs FT:
  - Zero-shot and few-shot cross-lingual transfer
  - Low-data regime (< 1K examples)
  - Summary faithfulness and conciseness
  - In addition, LoRA continued learning outperforms LoraHub under few-shot settings
- LoRA achieves **on-par performance** vs Full FT in larger models (see paper)
- LoRA achieves **worse performance** vs Full FT:
  - Smaller models
  - High-data regime, particularly for summary relevance