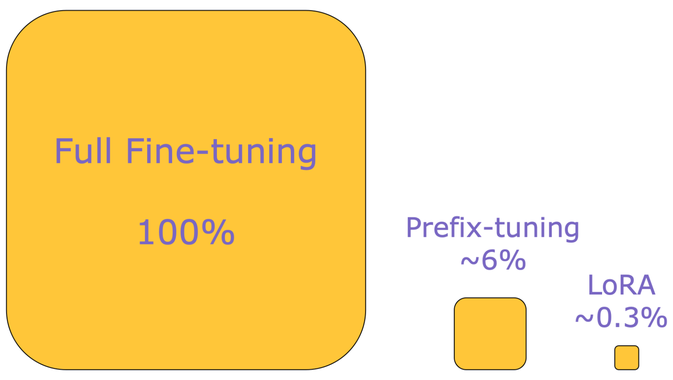
Leveraging Parameter-Efficient Fine-Tuning for Multilingual Abstractive Summarization

Jialun Shen [1] Yusong Wang* [1,2]

[1]Tokyo Institute of Technology [2]Guangdong Institute of Intelligence Science and Technology

Motivation

- Full fine-tuning of **Pre-trained Language Models (PLMs)** is **computationally expensive** and **resource-intensive**. As the demand for **multilingual NLP** grows, the need for efficient and scalable transfer learning methods becomes critical to make PLMs practical for real-world use in diverse linguistic settings.
- **Parameter-Efficient Fine-Tuning (PEFT)** methods like **Prefix-tuning** and **LoRA** have shown potential in reducing the number of parameters updated while retaining performance in monolingual tasks. However, **their effectiveness in multilingual tasks remains underexplored**.
- **Key Contribution**: We are the first to systematically evaluate PEFT methods in **multilingual abstractive summarization**, demonstrating clear **efficiency-performance trade-offs**. Our work sets new **benchmarks** for advancing **efficient multilingual NLP**.



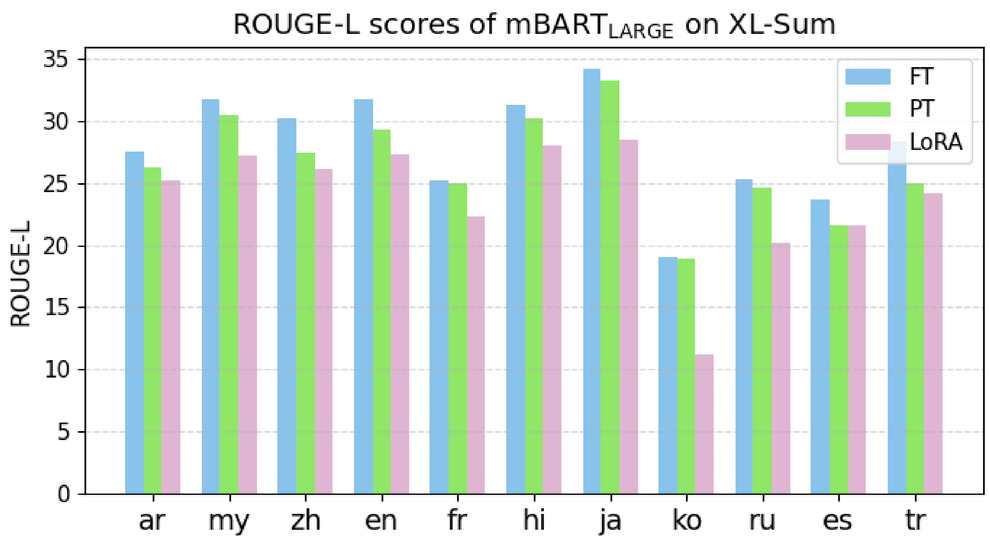
Experimental Setting

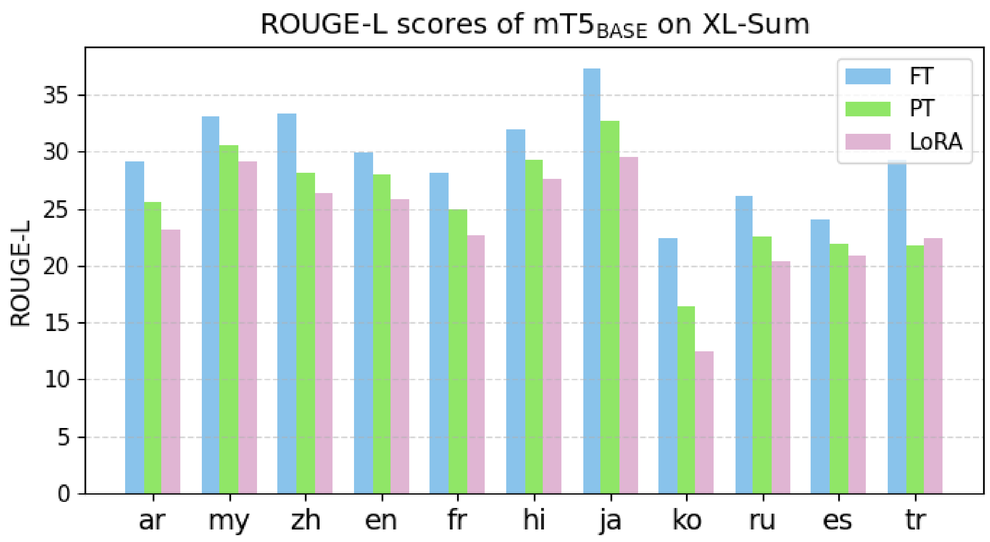
Dataset: XL-Sum, a large-scale multilingual abstractive summarization dataset containing news articles from BBC. We experimented on 11 languages containing both high- and low-resource ones: Arabic, Burmese, Chinese (Simplified), English, French, Hindi, Japanese, Korean, Russian, Spanish, and Turkish.

Lang.	ar	my	zh	en	fr	hi	ja	ko	ru	es	tr
#Train	37,519	4,569	37,360	306,522	8,697	70,777	7,113	4,407	62,243	38,110	27,176
#Dev	4,689	570	4,670	11,535	1,086	8,847	889	550	7,780	4,763	3,397
#Test	4,689	570	4,670	11,535	1,086	8,847	889	550	7,780	4,763	3,397
#Total	46,897	5,709	46,700	329,592	10,869	88,471	8,891	5,507	77,803	47,636	33,970

Models: $mBART_{LARGE}$ and **$mT5_{BASE}$**. We compare full fine-tuning (FT), prefix-tuning (PT) with prefix length 100, and LoRA with rank 16.

Results

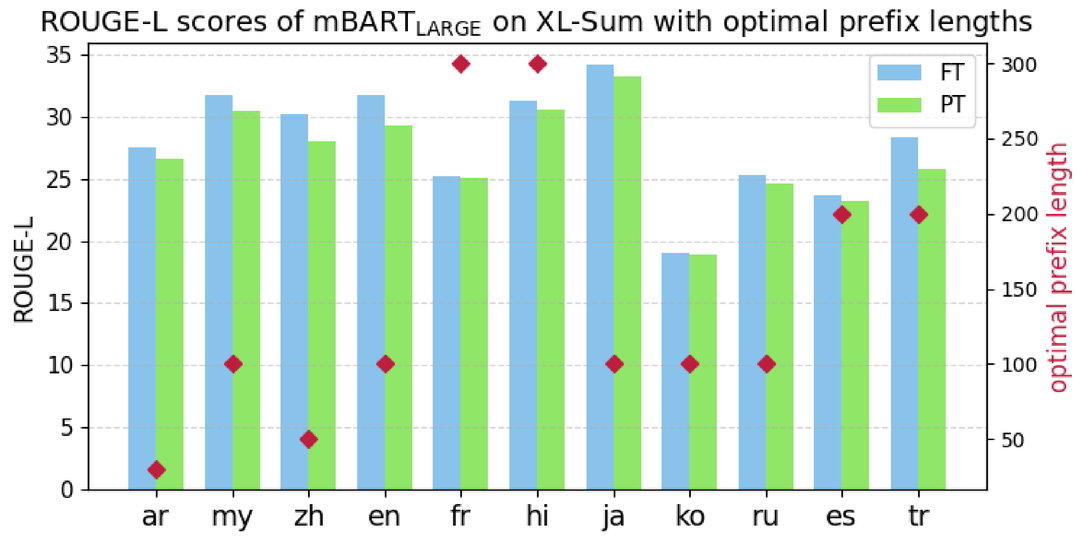




Further Investigation into Prefix-tuning

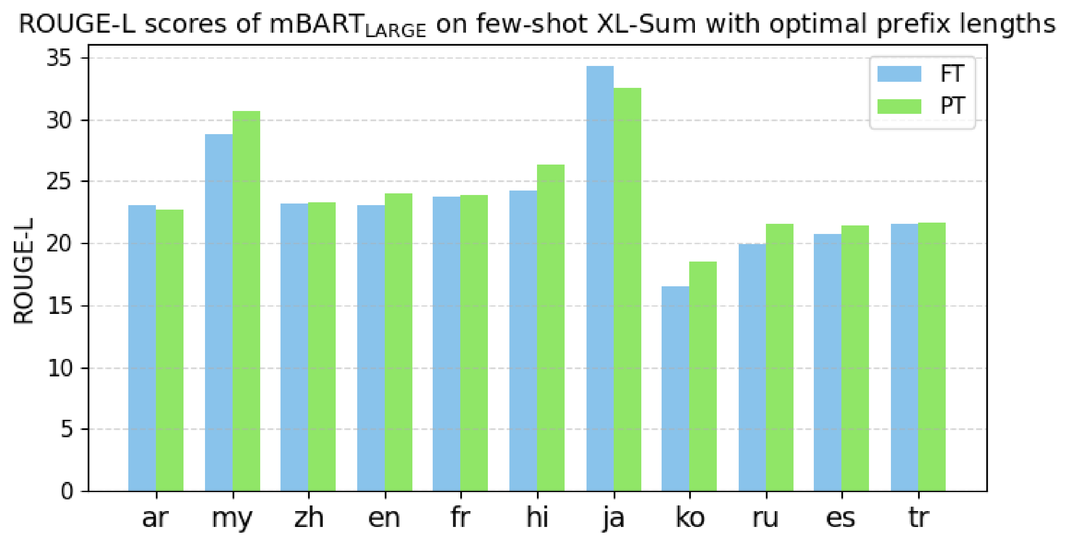
Impact of Prefix Length

Performance optimization requires careful tuning of prefix length, which varies significantly by language. A one-size-fits-all approach to prefix length can be ineffective for multilingual abstractive summarization.



Few-shot Performance

In few-shot scenarios with limited data available, prefix-tuning offers competitive and sometimes superior performance compared to full fine-tuning. This makes prefix-tuning an ideal choice for resource-constrained settings.



Conclusion

1. While PEFT methods can significantly reduce computational costs and memory usage, they exhibit a performance drop across languages when compared to full fine-tuning.
2. We present the first comprehensive evaluation of PEFT methods for multilingual abstractive summarization, providing key insights into the balance between efficiency and performance and establishing benchmarks for future research.
3. We include a detailed investigation into prefix-tuning, shedding light on its effectiveness under few-shot condition and providing valuable insights for optimizing its performance in multilingual settings.

References

[1] T. Hasan, A. Bhattacharjee, M. S. Islam, K. Mubasshir, Y.-F. Li, Y.-B. Kang, M. S. Rahman, and R. Shahriyar. XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, Aug. 2021. Association for Computational Linguistics.

[2] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

[3] X. L. Li and P. Liang. Prefix-Tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online, Aug. 2021. Association for Computational Linguistics.

[4] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[5] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.