

ABSTRACT

The massive increase in digital information production has led to the phenomenon of information overload, making it difficult for humans to absorb information efficiently, thus requiring reliable automatic text summarization technology. This study aims to analyze the performance and computational efficiency of three language model training methods, namely Full Fine-Tuning (FFT), Low-Rank Adaptation (LoRA), and Knowledge Distillation with LoRA (KD-LoRA), applied to the IndoT5 model for abstractive Indonesian news summarization. The three methods were applied to the IndoSum dataset with a 64% training, 16% validation, and 20% testing split, using an input token length of 512 and an output token length of 150. The evaluation results show that KD-LoRA achieves the best performance across all quality metrics, with a ROUGE-1 score of 65.04, a BERTScore F1 score of 84.61, and a Novelty 1-gram score of 10.26%, outperforming FFT by 0.69 points in ROUGE-1 and 0.20 points in BERTScore F1, as well as outperforming LoRA by 0.88 points in ROUGE-1 and 0.26 points in BERTScore F1. In terms of efficiency, KD-LoRA trains only 1.24% of the total parameters, representing a reduction of 98.77% compared to FFT and 61.10% compared to LoRA. It also produces inference FLOPs of 9.71G, approximately 76.95% lower than FFT and LoRA, a checkpoint size of only 13.60 MB, representing a reduction of 98.56% compared to FFT and 93.77% compared to LoRA, a shorter average training time per epoch of 1.82 minutes compared to FFT at 2.25 minutes and LoRA at 2.37 minutes, lower peak VRAM usage by approximately 47.40% compared to FFT and 43.34% compared to LoRA, and faster inference time by approximately 22.18% compared to FFT and 23.22% compared to LoRA. These results indicate that KD-LoRA is the most balanced method in terms of summarization quality and computational efficiency for abstractive Indonesian text summarization.

Keywords : abstractive text summarization, IndoT5, Full Fine-Tuning, LoRA, KD-LoRA, IndoSum