ABSTRACT

Transformer-based models for image captioning offer high accuracy, but their large size and high computational cost often limit their use in resource-constrained applications. This research addresses this challenge by developing an image captioning model that uses a Vision Transformer (ViT) as the image encoder and a Distilled Generative Pre-Trained Transformer 2 (DistilGPT2) as the text decoder. The model was trained on the Flickr8k dataset and then optimized using a post-training dynamic quantization technique to improve its efficiency. A comparative evaluation of the original full-precision model and the 8-bit quantized model was conducted on CPU hardware, focusing on metrics of caption quality (ROUGE, BLEU), inference time, and model size. The results show that quantization provides a significant performance improvement. The model's size was reduced by 22.77% (from 912.63 MB to 704.80 MB), and its average inference speed increased by 28.48% (from 5.23 to 3.74 seconds). These efficiency gains were achieved with a negligible impact on accuracy, as the average ROUGE score remained unchanged and the average BLEU score decreased only marginally. The model demonstrated strong potential on individual images, achieving a peak BLEU score of 1.0 and a peak ROUGE. *While the model still faces challenges in generating detailed captions for complex scenes, this study successfully demonstrates that dynamic quantization offers a highly favorable trade-off, making large, accurate models practical for deployment in environments with limited computational resources without a significant compromise in their core performance.*

Keywords: *Image Captioning, Vision Transformer, DistilGPT2, Quantization*