An Efficient and Lightweight Medical Vision–Language Framework for Synthetic Data Generation and Radiological Image Captioning Using LoRA and Diffusion Models

Hsiao-Ju Lin
Department of Orthopedic Surgery, Taoyuan General Hospital, Ministry of Health and Welfare, Taoyuan, Taiwan

Shi-Jinn Horng
Department of Computer Science and Information Engineering, Asia University, Taiwan
Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan

Pin-Siang Huang
Agaruda Research, Agaruda, Taipei, Taiwan

Abstract

The scarcity and imbalance of annotated medical imaging datasets remain major obstacles to reliable medical AI development. Although diffusion models have enabled practical synthetic data generation, their integration into medical vision–language tasks under limited computational resources remains insufficiently explored. In addition, general-domain vision–language models, such as BLIP and InstructBLIP, often struggle in radiological applications due to complex visual patterns, specialized clinical terminology, and substantial GPU memory requirements. In this paper, we propose an efficient and lightweight medical vision–language framework for synthetic data generation and radiological image captioning using Low-Rank Adaptation (LoRA) and Stable Diffusion v1.5. Focusing on clinically significant conditions, including pneumonia, cardiomegaly, and brain tumors, the proposed framework efficiently fine-tunes BLIP and diffusion models to enable robust captioning and synthetic image generation on commodity GPUs with ≤12 GB VRAM. To evaluate the effectiveness of synthetic data, we conducted cardiomegaly-versus-normal classification experiments under two settings: real data only and real data combined with synthetic data. Results show that precision, recall, and F1-score improved from 0.80 to 0.89 after incorporating synthetic data. For radiological image captioning, we evaluated BLIP, BLIP-2, and InstructBLIP, representing lightweight to complex architectures. Qualitative assessment demonstrated that BLIP-2 generated more clinically grounded and specific descriptions, particularly for cardiomegaly and multifocal pneumonia. These findings demonstrate the feasibility of scalable and resource-efficient medical vision–language modeling for practical clinical applications.

Index Terms

Stable Diffusion, VLM, Image Caption, LoRA

Acknowledgement: This work was supported by National Science and Technology Council under contract number 114-2221-E-468 -014 -