SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

Scaling Vision-Language Models for Radiology Caption Prediction: A Cross-Scale Comparison of Parameter-Efficient Fine-Tuning Methods and Their Attention-Based Explanations

Mahmudul Hoque, Morgan State University; Raisa Nusrat Chowdhury; Md Mahmudur Rahman, PhD

Introduction/Background
The internal attention of a vision-language model is often read as showing which image regions support its output, yet this assumption is rarely tested. For radiology captioning, it is unknown which decoder layers track image content, whether some instead attend to fixed positions unrelated to the input, and how these behaviors shift with model scale or fine-tuning method.

Methods/Intervention
We fine-tuned LLaVA-OneVision models at three scales (0.5, 7, and 72 billion parameters) for radiology caption generation using two parameter-efficient methods: quantized low-rank adaptation (QLoRA) and its weight-decomposed variant (QDoRA). Training used a public multi-modality benchmark of 116,604 development and 15,249 test images. We extracted attention from every decoder layer during generation and computed four per-layer diagnostics: attention entropy, pairwise correlation of word attention, top-k patch overlap, and peak-to-mean ratio. These metrics distinguished layers attending to fixed positions regardless of the word from layers whose attention shifts per word and aligns with image regions.

Results/Outcome
Across all three scales, the diagnostics identified two layer types, fixed-position (register) layers and candidate grounding layers that localize image content. Their depth varied with scale. In the 0.5 billion model, the register occupied the final layer with grounding earlier; at 7 billion, the register appeared early with grounding in middle layers; and at 72 billion, the register appeared early with grounding in late non-terminal layers before final-layer register behavior. Only the smallest model placed its register at the network's end. QLoRA and QDoRA produced near-identical per-layer attention (Pearson correlation ≥0.97 across all metrics at 72 billion parameters), suggesting attention geometry is largely preserved across adaptation methods. QDoRA inference was about six times slower than QLoRA at the largest scale.

Conclusion
Attention-based interpretation of medical captioning models reveals a consistent division between register and grounding layers whose depth depends on model scale. Because the two methods yield equivalent attention, the more efficient QLoRA suffices for both captioning and explanation.

Statement of Impact
Localizing grounding layers provides region-level evidence for each caption phrase, supporting verification of model output. That efficient fine-tuning matches its costlier alternative can reduce the cost of building and deploying interpretable radiology captioning systems.

Keywords
medical image captioning; vision-language models; attention interpretability; visual grounding; parameter-efficient fine-tuning; radiology