INTISARI

ADAPTASI LAYER-RESIDUAL CO-ATTENTION NETWORK (LRCN) UNTUK *MEDICAL VISUAL QUESTION ANSWERING*

Oleh

Rhafael Chandra
22/498550/PA/21528

Medical Visual Question Answering (MedVQA) menuntut pemodelan yang menjaga detail visual *fine-grained*. Pada arsitektur *deep co-attention*, peningkatan kedalaman dapat memicu *information dispersion*, yaitu bobot *attention* makin menyebar pada lapisan dalam. Penelitian ini mengimplementasikan *Layer-Residual Co-Attention Network* (LRCN) untuk MedVQA dengan ViT-B/32 dan BioBERT, lalu mengevaluasi *Layer-Residual Mechanism* (LRM) melalui studi *ablation* (LRM ON/OFF, kedalaman L, dan ukuran model *tiny*/*small*/*base*).

Eksperimen pada SLAKE dan VQA-RAD menunjukkan bahwa reliabilitas evaluasi klasifikasi bergantung pada cakupan jawaban uji: SLAKE hampir sepenuhnya *learnable* (99.81%), sedangkan VQA-RAD terbatas (74.50% overall; 39.11% open-ended) sehingga *closed accuracy* lebih stabil pada VQA-RAD. Pada SLAKE, pemisahan performa LRM ON vs OFF makin jelas pada L menengah–dalam dan paling kuat pada model berkapasitas lebih besar, yang mengindikasikan LRM membantu menahan penurunan performa akibat kedalaman. Analisis *layer-wise* pada SLAKE *base* $L = 12$ juga mendukung indikasi dispersi: tanpa LRM, entropi cenderung lebih tinggi dan Top-5 mass menurun menuju lapisan akhir pada *visual self-attention*, sedangkan dengan LRM *attention* lebih terkonsentrasi pada lapisan dalam; pada *guided-attention* terlihat konsentrasi yang lebih kuat di lapisan akhir sebagai efek tidak langsung dari perubahan representasi visual. Pada VQA-RAD, $\Delta = \text{ON} - \text{OFF}$ berfluktuasi di sekitar nol sehingga akurasi tidak memberikan estimasi efek LRM yang stabil karena dibatasi OOV dan skala data.

Kata kunci: *Medical Visual Question Answering*, *Co-Attention*, *Information Dispersion*, *Layer-Residual Mechanism*, Vision Transformer, BioBERT

ABSTRACT

ADAPTATION OF THE LAYER-RESIDUAL CO-ATTENTION NETWORK (LRCN) FOR MEDICAL VISUAL QUESTION ANSWERING

By

Rhafael Chandra
22/498550/PA/21528

Medical Visual Question Answering (MedVQA) requires retaining fine-grained visual evidence. In deep co-attention architectures, increasing depth may induce *information dispersion*, where attention becomes increasingly diffuse in later layers. This work implements a MedVQA-adapted *Layer-Residual Co-Attention Network* (LRCN) with ViT-B/32 and BioBERT, and evaluates the *Layer-Residual Mechanism* (LRM) through an ablation over LRM ON/OFF, co-attention depth (L), and model size (*tiny*/*small*/*base*).

Experiments on SLAKE and VQA-RAD highlight that classification reliability depends on test-answer coverage: SLAKE is nearly fully learnable (99.81%), while VQA-RAD is constrained (74.50% overall; 39.11% open-ended), making closed-ended accuracy the more reliable comparison on VQA-RAD. On SLAKE, the ON–OFF separation becomes clearer at medium-to-deep L and is strongest for larger models, suggesting LRM helps limit depth-related performance degradation. A layer-wise analysis on SLAKE *base* $L = 12$ supports the dispersion hypothesis: without LRM, entropy remains higher and Top-5 mass declines toward deeper layers in visual self-attention, whereas with LRM attention stays more concentrated in late layers; guided-attention shows a similar late-layer concentration as an indirect effect of altered visual representations. On VQA-RAD, $\Delta = \text{ON} - \text{OFF}$ fluctuates around zero, so accuracy alone cannot provide a stable estimate of LRM due to OOV and data-scale constraints.

Keywords: Medical Visual Question Answering, Co-Attention, Information Dispersion, Layer-Residual Mechanism, Vision Transformer, BioBERT