IMPLEMENTASI MODEL *VISUAL QUESTION ANSWERING* MENGGUNAKAN *VISION TRANSFORMER* DAN *EFFICIENTNET-V2* DENGAN BERT

Ahmed Nizhan Haikal

ABSTRAK

Visual Question Answering (VQA) merupakan tugas *multimodal* yang memadukan visi komputer dan pemrosesan bahasa alami untuk menjawab pertanyaan berdasarkan citra. Penelitian ini mengeksplorasi pengaruh arsitektur visual (*EfficientNet V2* dan *Vision Transformer*/ViT) terhadap performa VQA, dengan BERT digunakan sebagai *backbone* teks untuk pemahaman konteks semantik *bidirectional*. Penelitian ini menggunakan *dataset* DT-VQA dan menerapkan strategi *thresholding* untuk mengatasi ketidakseimbangan label. Hasil menunjukkan bahwa model ViT-BERT dengan *fine-tuning* mencapai performa terbaik di antara varian ViT-BERT, mencatat akurasi 53,19% dan ANLS 61,95% pada *threshold* 75. Performa terbaik dicatat oleh model *EfficientNet V2*-BERT dengan *transfer learning*, dengan akurasi sebesar 56,3% dan ANLS 62,9%. Temuan ini menggarisbawahi pentingnya penyesuaian arsitektur dan strategi pelatihan untuk optimasi performa pada karakteristik data.

Kata kunci: BERT, *EfficientNet V2,* Pemrosesan *Multimodal , Visual Question Answering*, *Vision Transformer*

IMPLEMENTASI MODEL *VISUAL QUESTION ANSWERING* MENGGUNAKAN *VISION TRANSFORMER* DAN *EFFICIENTNET-V2* DENGAN BERT

Ahmed Nizhan Haikal

ABSTRACT

Visual Question Answering (VQA) is a multimodal task that integrates computer vision and natural language processing to answer image-based questions. This research explores the impact of visual architectures (EfficientNet V2 and Vision Transformer/ViT) on VQA performance, utilizing BERT as the text backbone for bidirectional semantic context understanding. We employed the DT-VQA dataset and implemented a thresholding strategy to address label imbalance. Results indicate that the ViT-BERT model with fine-tuning achieved the best performance among ViT-BERT variants, recording an accuracy of 53.19% and an ANLS of 61.95% at a threshold of 75. Overall, the highest performance was achieved by the EfficientNet V2-BERT model with transfer learning, with an accuracy of 56.3% and an ANLS of 62.9%. These findings underscore the importance of tailoring architectural choices and training strategies for optimal performance given specific data characteristics.

Keywords: BERT, *EfficientNet V2, Multimodal Processing, Visual Question Answering*, *Vision Transformer*