INTISARI

Multi-scale Visual Attention: Pendekatan *Low-Resource Image Captioning* Bahasa Indonesia Berbasis Transformer dengan Representasi Citra Multi-skala, *Attention* Spasial, dan *Beam Search*

Oleh
Krisna Bayu Dharma Putra
21/482071/PA/21017

Image captioning adalah suatu teknik dalam bidang pengolahan citra dan pemrosesan bahasa alami yang memungkinkan sistem komputer untuk memberikan deskripsi atau caption secara otomatis terhadap suatu citra. *Image captioning* telah dilakukan untuk berbagai bahasa, termasuk bahasa Indonesia. Namun, kebanyakan *image captioning* berbahasa Indonesia belum maksimal. Hal ini disebabkan oleh kurangnya dataset yang tersedia.

Penelitian ini berfokus untuk mengimplementasikan *deep learning* transformers untuk *image captioning* berbahasa Indonesia dengan dataset yang sedikit. Modifikasi dilakukan pada transformers dengan mengganti *attention* pada enkodernya dengan kombinasi *multi-scale feature extraction* serta *spatial-awareness attention*. Selanjutnya, inferensi juga akan diperkuat dengan implementasi pencarian *beam search*.

Hasil penelitian menunjukkan skor evaluaasi rata-rata BLEU, METEOR, dan CIDER sebesar 0,482, 0,336, dan 0,383 secara berturut-turut. Hasil ini merupakan peningkatan sebesar 17,5 % dan 50,7% pada skor evaluasi BLEU dan CIDER serta 19,1% untuk METEOR jika dibandingkan dengan transformers biasa. Penambahan metode inferensi *beam search* juga berhasil meningkatkan kemampuan model dalam melakukan inferensi. *Beam* dengan skor evaluasi terbaik diraih oleh beam dengan n=7. Hasil evaluasi beam ini berhasil mendapatkan skor rata-rata BLEU, METEOR, dan CIDER sebesar 0,520, 0,320, dan 0,540 secara berturut-turut. Hasil ini merupakan sebuah peningkatan sebesar 7,8% untuk evaluasi BLEU serta 40,9% untuk evaluasi CIDER jika dibandingkan dengan metode *greedy search* pada model yang sama.

Kata Kunci: *Image Captioning*, Pembelajaran Mendalam, Transformers

ABSTRACT

Multi-Scale Visual Attention: A Transformer-Based Approach for Low-Resource Indonesian Image Captioning Using Multi-Scale Image Representation, Spatial Attention, and Beam Search

By
Krisna Bayu Dharma Putra
21/482071/PA/21017

Image captioning is a technique in the field of image processing and natural language processing that enables computer systems to automatically generate a description or caption for an image. Image captioning has been implemented in various languages, including Indonesian. However, most Indonesian-language image captioning systems are still suboptimal. This is mainly due to the limited availability of datasets.

This research focuses on implementing deep learning transformers for Indonesian-language image captioning using a small dataset. Modifications were made to the transformer architecture by replacing the encoder's attention mechanism with a combination of multi-scale feature extraction and spatial-awareness attention. Furthermore, inference performance was enhanced through the implementation of beam search.

The research results show average evaluation scores of BLEU, METEOR, and CIDEr at 0.482, 0.336, and 0.383, respectively. These results represent an improvement of 17.5% and 50.7% in BLEU and CIDER evaluation scores, and an 19.1% increase for METEOR, compared to the standard transformer model. The addition of the beam search inference method also successfully improved the model’s inference capability. The best evaluation score was achieved with a beam size of n=7. The evaluation results for this beam configuration yielded average BLEU, METEOR, and CIDEr scores of 0.520, 0.320, and 0.540, respectively. These results indicate an improvement of 7.8% in BLEU and 40.9% in CIDER compared to the greedy search method on the same model.

Keywords : Image Captioning, Deep Learning, Transformers