INTISARI

Image Captioning Bahasa Indonesia Berbasis Transformer dengan Learnable Sinusoidal Positional Encoding

Oleh

Muhammad Arsya Putra

20/462186/PA/20158

Image captioning adalah suatu tugas dalam bidang pemrosesan gambar dan bahasa alami yang bertujuan untuk menghasilkan deskripsi teks yang menjelaskan konten sebuah gambar. Tugas ini melibatkan pemahaman visual dari gambar dan kemampuan untuk menggabungkannya dengan bahasa alami dalam sebuah kalimat deskripsi yang relevan. *Image captioning* terinspirasi dari metode machine translation, mengadopsi kerangka kerja seq2seq dengan dua komponen utama: encoder dan decoder.

Dalam penelitian ini dibangun model *image captioning* berbasis *pretrained* CNN (InceptionV3) dan Transformer dengan menggunakan dataset berbahasa Indonesia yaitu Flickr8k Bahasa. Transformer sudah terbukti memiliki performa yang lebih baik dibanding model rekurensi terdahulu seperti RNN atau LSTM. Transformer yang tidak mengolah datanya secara rekurensi membutuhkan suatu *positional encoding* untuk mengetahui letak sekuens yang diolahnya. Pada penelitian ini dilakukan pengujian yang membandingkan *fix sinusoidal positional encoding* dari arsitektur asli transformer dengan *learnable sinusoidal positional encoding*.

Pada eksperimen dilakukan hyperparameter tuning untuk mencari nilai optimal dimensi *embedding*, dimensi *feedforward,* jumlah *head*, dan *beam width* pada proses inferensi. Kemudian evaluasi dilakukan dengan melakukan perhitungan skor BLEU.

Kata kunci: Pendeskripsian Gambar, *Transformer*, Inception V3, Citra Komputer.

ABSTRACT

Transformer Based Image Captioning In Bahasa Indonesia With Learnable Sinusoidal Positional Encoding

By

Muhammad Arsya Putra

20/462186/PA/20158

Image captioning is a task in the field of image processing and natural language that aims to produce a text description that explains the content of an image. This task involves visual understanding of the image and the ability to combine it with natural language in a relevant description sentence. Image captioning is inspired by the machine translation method, adopting a seq2seq framework with two main components: the encoder and the decoder.

In this study, we built an image captioning model based on CNN (InceptionV3) and Transformers using Indonesian-language datasets namely Flickr8k Bahasa. Transformers have proven to have better performance than previous recursion models such as RNN or LSTM. Transformers, that do not process their data recursively, need a positional encoding to know the sequence placement they are processing. The study compared the fix sinusoidal positional encoding of the original transformer architecture with the learnable sinusoid positional Encoding.

In the experiments performed hyperparameter tuning to find optimum values of embedding dimensions, feedforward dimension, number of head, and beam width on the inference process. Then the evaluation is done by doing the calculation of the BLEU score.

Keyword: Image Captioning, Transformer, Inception V3, Computer Vision.