INTISARI

PENGEMBANGAN MODEL *IMAGE CAPTIONING MULTI-VIEW* UNTUK *DATASET FASHION* INDONESIA DENGAN *MULTIMODAL TRANSFORMER*

Oleh

Jovinca Claudia Amarissa

21/479037/PA/20778

Dalam industri penjualan pakaian secara daring, deskripsi produk yang akurat penting untuk visibilitas dalam pencarian produk. Namun, model *captioning* konvensional masih menghadapi tantangan dalam menangkap detail pakaian lokal secara menyeluruh karena keterbatasan pada keragaman data pelatihan dan ketergantungan pada sudut pandang tunggal yang kurang representatif.

Penelitian ini mengembangkan model *image captioning multi-view* berbasis *multimodal Transformer* dengan memanfaatkan tiga citra dari sudut pandang berbeda pada setiap produk untuk menghasilkan *caption* dalam Bahasa Inggris. Pelatihan model dilaksanakan menggunakan *Cross-Entropy Loss* dan dilanjutkan *fine-tuning* menggunakan *Self-Critical Sequence Training* (*SCST*).

Hasil evaluasi menunjukkan bahwa pendekatan *multi-view* secara konsisten lebih unggul dibandingkan *single-view*. *Fine-tuning* dengan SCST meningkatkan performa pada *CIDEr* senilai 12,33%, dengan skor akhir terbaik di antara model lainnya, yaitu *BLEU-4* = 0,4070, *CIDEr* = 1,5697, dan *METEOR* = 0,3196. Evaluasi kualitatif turut menunjukkan keunggulan model dalam mendeteksi detail seperti motif, kombinasi warna, serta detail teknis secara akurat. Sehingga, integrasi *multi-view* dengan optimasi *SCST* terbukti mampu meningkatkan kualitas *caption* secara signifikan dan dapat menjadi solusi yang relevan dalam otomasi pembuatan deskripsi produk pada *platform e-commerce fashion* di Indonesia.

Kata Kunci: *Fashion* Lokal, Deskripsi Produk, *Image Captioning*, *Multi-view*, *Multimodal Transformer*, Pembelajaran Mesin Mendalam

ABSTRACT

DEVELOPMENT OF MULTI-VIEW IMAGE CAPTIONING MODEL FOR INDONESIAN FASHION DATASET WITH MULTIMODAL TRANSFORMER

By

Jovinca Claudia Amarissa

21/479037/PA/20778

In online fashion retail, accurate product descriptions are essential for search visibility. However, conventional captioning models still face challenges in comprehensively capturing the details of local clothing due to limited diversity in training data and reliance on a single, less representative viewpoint.

This study develops a multi-view image captioning model based on a multimodal *Transformer*, utilizing three images from different angles for each product to generate captions in English. The model was trained using *Cross-Entropy Loss* and further *fine-tuned with Self-Critical Sequence Training (SCST)*.

Evaluation results show that the multi-view approach consistently outperforms the single-view method. Fine-tuning with *SCST* improved performance on the *CIDEr* metric by 12.33%, resulting in the model achieving the best scores among all, with *BLEU*-4 = 0.4070, *CIDEr* = 1.5697, and *METEOR* = 0.3196. Qualitative evaluation also demonstrated the model's advantage in accurately identifying details such as patterns, color combinations, and technical features. Thus, the integration of multi-view input with *SCST* optimization significantly enhances caption quality and offers a relevant solution for automating product description generation on fashion e-commerce platforms in Indonesia.

Keywords: Local Fashion, Product Description, Image Captioning, Multi-view, Multimodal Transformer, Deep Learning