

ABSTRAK

Perkembangan teknologi kecerdasan buatan (*Artificial Intelligence*) saat ini telah mendorong integrasi sistem multimodal yang mampu mengidentifikasi dan menjembatani domain visual serta domain tekstual. Salah satu tantangan utama dalam ranah *visual storytelling* adalah bagaimana mentransformasikan karakteristik fitur citra mentah (*raw image*) menjadi rangkaian narasi cerita pendek berbahasa Indonesia yang runtut, estetik, dan kaya akan konteks semantik. Penelitian ini bertujuan untuk merancang, mengimplementasikan, dan menguji sistem pembangkit cerita pendek otomatis berbasis citra digital dengan menggabungkan arsitektur *Convolutional Neural Network* (CNN) dan model teks berbasis *Transformer*. Metodologi hibrida yang diterapkan memanfaatkan arsitektur VGG16 yang dimodifikasi secara spesifik sebagai *Visual Feature Extractor (Encoder)*. Modifikasi dilakukan dengan memotong jaringan pada lapisan *Fully Connected* kedua (fc2) untuk memetakan informasi spasial citra masukan berukuran 224×224 piksel menjadi sebuah vektor fitur numerik padat berdimensi 1×4096 . Vektor fitur tersebut kemudian diteruskan ke model *Bootstrapping Language-Image Pre-training* (BLIP) yang bertindak sebagai *Multimodal Text Decoder* untuk membangkitkan teks deskripsi dasar (*image captioning*) dalam bahasa Inggris. Luaran teks deskripsi tersebut selanjutnya dialirkan ke komponen *Google Translator API* untuk pengalihan bahasa ke dalam bahasa Indonesia baku. Pada tahap akhir, aturan logika pemrosesan bahasa alami (*Rule-Based Natural Language Processing*) dieksekusi untuk mengekspansi kalimat terjemahan menjadi narasi cerita pendek yang utuh dengan target panjang teks minimal 80+ kata. Seluruh alokasi komputasi ini diintegrasikan ke dalam antarmuka interaktif berbasis web memanfaatkan pustaka Gradio. Hasil pengujian komputasional secara empiris pada berbagai kategori citra (pemandangan alam, fauna, dan aktivitas manusia) menunjukkan tingkat keberhasilan sebesar 100% dalam mengenali objek utama serta membangkitkan narasi cerita yang selaras secara kontekstual. Pengukuran kinerja sistem mencatatkan rata-rata waktu respons *end-to-end* (*inference time*) sebesar ± 11.5 detik, yang terdiri dari fase ekstraksi fitur VGG16 (± 4.2 detik), generasi teks BLIP (± 4.5 detik), serta proses translasi dan ekspansi narasi (± 3.8 detik). Penelitian ini berhasil memberikan kontribusi dalam menghadirkan solusi multimodal yang responsif, efisien, dan interaktif untuk otomatisasi penulisan kreatif berbasis data visual.

Kata Kunci: *Convolutional Neural Network*, VGG16, *Transformer*, BLIP, *Visual Storytelling*, Cerita Pendek Otomatis, Multimodal.