INTISARI

ARSITEKTUR HIBRIDA CONVOLUTIONAL NEURAL NETWORK DAN VISION TRANSFORMER UNTUK REKOGNISI JUMLAH SUARA SAH PADA FORMULIR MODEL C HASIL PEMILIHAN UMUM 2024

Oleh

Muhammad Akbar Hamid

23/530003/PPA/06706

Formulir Model C Hasil merupakan dokumen pencatatan hasil penghitungan suara di TPS yang perlu ditranskripsikan ke format digital untuk mendukung rekapitulasi dan pengarsipan. Namun, citra formulir pada kondisi riil memiliki variasi kualitas yang berbeda-beda sehingga dapat membatasi kemampuan generalisasi jika model dilatih pada data terkontrol. Pada pengembangan model rekognisi karakter citra, CNN memiliki kelebihan mengekstraksi fitur lokal, tetapi terbatas dalam menangkap relasi global. Sebaliknya, ViT mampu memodelkan konteks global, namun berpotensi kurang kuat dalam mempertahankan detail lokal.

Oleh karena itu, penelitian ini mengembangkan arsitektur hibrida CNN-ViT yang menggabungkan ekstraksi fitur lokal CNN dan pemodelan konteks global ViT untuk rekognisi karakter numerik “0-9” dan simbolik “X” pada Jumlah Suara Sah di Formulir Model C Hasil PEMILU 2024. Kesimpulan penelitian menunjukkan arsitektur hibrida CNN-ViT dapat mengatasi kelemahan masing-masing metode CNN dan ViT dengan hasil kinerja akurasi tertinggi pada data uji yaitu 0,999879 dengan *loss* 0,000880, melampaui CNN dengan akurasi 0,999818 dan *loss* 0,005208, serta ViT dengan akurasi 0,996424 dan *loss* 0,013718, sekaligus mempertahankan kinerja per kelas yang mendekati sempurna pada hampir seluruh kelas. Dari sisi efisiensi, CNN-ViT berada pada posisi kompromi dengan *overhead* inferensi yang masih wajar dibandingkan ViT. Pada penerapan model yang dilatih pada dataset terkontrol ke ROI Formulir Model C Hasil, ditemukan adanya *domain shift* yang menyebabkan penurunan performa.

Kata Kunci: Jumlah Suara Sah, Formulir Model C Hasil, PEMILU 2024, Arsitektur Hibrida CNN-ViT.

ABSTRACT

HYBRID ARCHITECTURE OF CONVOLUTIONAL NEURAL NETWORK AND VISION TRANSFORMER FOR RECOGNIZING VALID VOTES ON THE FORMULIR MODEL C HASIL OF THE 2024 GENERAL ELECTION

By

Muhammad Akbar Hamid

23/530003/PPA/06706

The Model C Result Form is a document for recording vote-count results at polling stations that needs to be transcribed into a digital format to support recapitulation and archiving. However, real-world form images have varying quality, which can limit generalization ability if a model is trained on controlled data. In developing image character recognition models, CNNs have the advantage of extracting local features but are limited in capturing global relations. In contrast, ViT can model global context but may be less strong in preserving local details.

Therefore, this study develops a hybrid CNN-ViT architecture that combines CNN local feature extraction and ViT global context modeling for recognizing numeric characters "0-9" and the symbolic character "X" in the Valid Vote Count field on the 2024 General Election Model C Result Form. The study concludes that the hybrid CNN-ViT architecture can address the weaknesses of CNN and ViT, achieving the highest performance with a test accuracy of 0.999879 and a test loss of 0.000880, outperforming CNN with an accuracy of 0.999818 and a loss of 0.005208, and ViT with an accuracy of 0.996424 and a loss of 0.013718, while maintaining near-perfect per-class performance for almost all classes. In terms of efficiency, CNN-ViT represents a compromise with inference overhead that remains reasonable compared to ViT. When applying a model trained on a controlled dataset to the ROI of the Model C Result Form, a domain shift was found that caused a decrease in performance.

Keywords: Valid Vote Count, Model C Result Form, PEMILU 2024, Hybrid CNN-ViT Architecture.