INTISARI

PERBANDINGAN MODEL *TRANSFORMER* (VIT, SWIN, DAN DEIT) UNTUK DETEKSI CITRA AI GENERATIF

Oleh

Gian Luky Saputra

21/474481/PA/20487

Teknologi kecerdasan buatan atau *Artificial Intelligence* (AI) telah mengalami perkembangan pesat dan memberikan dampak signifikan dalam berbagai sektor. Salah satu inovasi AI yang menonjol adalah model generatif, yang digunakan dalam berbagai platform seperti DALL-E, Stable Diffusion, dan MidJourney. Platform ini memungkinkan pengguna menghasilkan citra sintetis berdasarkan perintah (*prompt*) yang diberikan. Namun, kemudahan dalam menciptakan citra sintetis ini juga menimbulkan tantangan dalam menjaga keaslian dan integritas konten berbasis visual. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan model klasifikasi guna membedakan citra asli dan citra yang dihasilkan oleh AI menggunakan model berbasis *transformer*. Tiga arsitektur utama yang digunakan adalah *Vision Transformer* (ViT), *Swin Transformer*, dan *Data-Efficient Image Transformer* (DeiT). Hasil penelitian menunjukkan bahwa model berbasis *transformer* sangat efektif dalam mendeteksi citra hasil AI generatif. *Swin Transformer* memiliki median *balanced accuracy* tertinggi sebesar 98,78%, namun dengan variabilitas performa yang lebih besar dibandingkan dua model lainnya. DeiT memiliki variabilitas paling rendah sehingga lebih stabil, dengan median *balanced accuracy* sebesar 97,35%. Sementara itu, ViT memiliki sebaran nilai *balanced accuracy* dalam rentang 70–100% dengan median sebesar 97,85%. Model dengan *learning rate* lebih kecil (0,00001 dan 0,0001) serta *batch size* yang lebih besar (64) cenderung memberikan performa yang lebih baik, meskipun membutuhkan waktu pelatihan yang lebih lama. Dari ketiga arsitektur, model terbaik yang diperoleh adalah *Swin Transformer* varian *base* dengan *batch size* 32 dan *learning rate* 0,0001, menghasilkan performa *balanced accuracy* tertinggi sebesar 99,52%.

Kata Kunci: klasifikasi citra, AI generatif, *Vision Transformer* (ViT), *Swin Transformer*, *Data-Efficient Image Transformer* (DeiT)

ABSTRACT

COMPARISON OF TRANSFORMER MODELS (VIT, SWIN, AND DEIT) FOR GENERATIVE AI IMAGE DETECTION

By

Gian Luky Saputra

21/474481/PA/20487

Artificial Intelligence (AI) has rapidly advanced, significantly impacting various sectors. One of the most notable AI innovations is generative models, which are utilized in platforms such as DALL-E, Stable Diffusion, and MidJourney. These platforms enable users to generate synthetic images based on given prompts. However, the ease of generating synthetic images also presents challenges in maintaining the authenticity and integrity of visual content. Therefore, this study aims to develop a classification model to distinguish between real and AI-generated images using transformer-based architectures. The three primary architectures employed in this research are Vision Transformer (ViT), Swin Transformer, and Data-Efficient Image Transformer (DeiT). The results demonstrate that transformer-based models are highly effective in detecting AI-generated images. Swin Transformer achieved the highest median balanced accuracy of 98,78%, although it exhibited greater performance variability than the other two architectures. In contrast, DeiT displayed the lowest variability, making it the most stable model, with a median balanced accuracy of 97,35%. Meanwhile, ViT exhibited a broader distribution of balanced accuracy, ranging from 70% to 100%, with a median of 97,85%. Models trained with a lower learning rate (0,00001 and 0,0001) and a larger batch size (64) tended to achieve better performance, albeit at the cost of longer training times. Among the three architectures, the best-performing model was Swin Transformer (base variant) with a batch size of 32 and a learning rate of 0,0001, achieving the highest balanced accuracy of 99,52%.

Keywords: image classification, generative AI, Vision Transformer (ViT), Swin Transformer, Data-Efficient Image Transformer (DeiT)