INTISARI

Vision Foundation Models (VFM) telah mengalami kemajuan pesat, beralih dari arsitektur yang spesifik untuk tugas tertentu ke model yang lebih umum dan fleksibel seperti CLIP, yang memanfaatkan pra-pelatihan kontrasif pada kumpulan data gambar-teks besar. Pendekatan multimodal ini sangat penting untuk pembelajaran *zero-shot*, memungkinkan model untuk menggeneralisasi ke kategori yang belum terlihat dengan mengintegrasikan informasi visual dan tekstual. Meskipun kemajuan ini terus berkembang, tantangan yang signifikan tetap ada: kurangnya pengujian standar untuk VFM dalam *zero-shot image classification*, yang sering kali mengarah pada hasil penelitian yang tidak konsisten dan kesulitan dalam perbandingan model yang objektif.

Selain itu, penyelidikan dampak variasi metode prompting terhadap kinerja model CLIP dalam tugas *zero-shot image classification* untuk melihat signifikansi pengaruh *prompting* terhadap performa model. Proses yang dilakukan dengan mengevaluasi beberapa varian model CLIP, termasuk CoNvNeXt base, SIGLIP, dan PE Core Base, di empat dataset yang berbeda: CIFAR100-C, Animals with Attributes 2 (AwA2), Oxford Pets, dan MIT Indoor67. Tiga strategi prompting yang berbeda diterapkan: label kelas "[kelas]", prompt berbasis template (misalnya, "foto [kelas]"), dan *Meta-Prompting for Visual Recognition* (MPVR) yang menggunakan LLM untuk menghasilkan *prompt* yang lebih deskriptif. Kinerja model dievaluasi terutama menggunakan metrik *top-1 accuracy* dan *top-5 accuracy*.

Hasil eksperimen menunjukkan bahwa model CLIP Perception Encoder (PE) Core Base/16 secara konsisten mencapai kinerja terbaik di semua eksperimen, dengan rata-rata akurasi top-1 sebesar 83.77% dan akurasi top-5 sebesar 96.30%. Model PE mengungguli *baseline model* ResNet-50 yang memiliki rerata metrik akurasi *top-1* sebesar 75.84% dan akurasi top-5 sebesar 91.09%. Meskipun metode prompting MPVR yang lebih canggih umumnya meningkatkan kinerja, terutama untuk dataset dengan atribut spesifik seperti Oxford Pets, analisis Two-Way ANOVA mengungkapkan tidak adanya dampak yang signifikan secara statistik dari variasi prompting atau arsitektur model terhadap akurasi top-1 atau top-5. Ini menunjukkan bahwa meskipun prompting dapat memberikan konteks yang lebih kaya, efeknya terhadap kinerja keseluruhan dalam tugas klasifikasi gambar zero-shot untuk model yang diuji tidak signifikan secara statistik, terutama jika dibandingkan dengan ketahanan yang melekat pada model itu sendiri.

Kata kunci : *Vision Foundation Models*, *Zero-Shot Image Classification*, CLIP, *Prompting Methods*, *Benchmarking*

ABSTRACT

Vision Foundation Models (VFMs) have seen rapid advancements, shifting from task-specific architectures to more generalized and flexible models like CLIP, which leverages contrastive pre-training on large image-text datasets. This multimodal approach is crucial for zero-shot learning, enabling models to generalize to unseen categories by integrating visual and textual information. Despite these advancements, a significant challenge remains: the lack of standardized benchmarking for VFMs in zero-shot image classification, which often leads to inconsistent research results and difficulties in objective model comparison.

In addition, the investigation examines the impact of various prompting methods on the performance of CLIP models in zero-shot image classification tasks to assess the significance of prompting's influence on model performance. The research evaluates several CLIP model variants, including CoNvNeXt base, SIGLIP, and PE Core Base, across four diverse datasets: CIFAR100-Corrupted, Animals with Attributes 2 (AwA2), Oxford Pets, and MIT Indoor67. Three distinct prompting strategies are employed: a simple class name, a template-based prompt (e.g., "a photo of a [class]"), and Meta-Prompting for Visual Recognition (MPVR), which utilizes Large Language Models (LLMs) to generate more descriptive prompts. Model performance is primarily assessed using top-1 and top-5 accuracy metrics.

The experimental results show that the Perception Encoder (PE) Core Base/16 model consistently outperformed the baseline model, ResNet-50, across all evaluated datasets. The PE Core Base/16 achieved an average top-1 accuracy of 83.77% and a top-5 accuracy of 96.30%, while the ResNet-50 baseline model yielded a top-1 accuracy of 75.84% and a top-5 accuracy of 91.09%. This performance gap highlights the superior capability of the PE Core model in zero-shot image classification tasks. Although the advanced Meta-Prompting for Visual Recognition (MPVR) method generally provided a performance boost, particularly for attribute-specific datasets like Oxford Pets, a Two-Way ANOVA analysis revealed no statistically significant effect of prompting variations or model architecture on either top-1 or top-5 accuracy. This suggests that, despite the higher performance of PE Core, the prompting methods, including MPVR, had a minimal impact on model performance when compared to the inherent strength of the PE Core model itself.

Keywords: *Vision Foundation Models*, *Zero-Shot Image Classification*, CLIP, *Prompting Methods*, *Benchmarking*