Global-Local Similarity for Efficient Fine-Grained Image Recognition with Vision Transformers

- Cover Letter

Edwin Arkel Rios, Min-Chun Hu, and Bo-Cheng Lai

—————————— ◆ ——————————

# Cover Letter

Dear Editors:

Please find enclosed a manuscript entitled: " Global-Local Similarity for Efficient Fine-Grained Image Recognition with Vision Transformers" which we would like to submit for publication as a Regular Paper in ACM Transactions on Intelligent Systems and Technology.

Fine-grained image recognition (FGIR) is a widely studied area with various applications such as automatic biodiversity monitoring and multimedia analysis. FGIR is challenging due to small inter-class differences and large intra-class variations. To tackle these challenges, state-of-the-art work in this area has mostly focused on utilizing attention to select discriminative regions that encapsulate subtle differences between classes. However, employing attention effectively can incur a significant computational cost. Therefore, in this work we propose a computationally light alternative to select discriminative regions with vision transformers (ViTs). Our method crops an image based on the cosine similarity between the global representation of an image as given by the ViT’s CLS token, which is typically used for classification, and the local representation of individual image patches. These crops are then resized and forwarded through a ViT encoder. Finally, high-level features of the original and cropped representations are further refined together in order to make more robust predictions. To validate our method, we conduct a thorough analysis of fine-grained recognition models by comparing models across 10 datasets spanning a wide spectrum of tasks. Our proposed model achieves the highest accuracy in 8 of these datasets, and on average, reduces the relative classification error by 10.23% compared to the baseline ViT. Moreover, our model achieves these results with 9.26x less VRAM and a 2.59x higher inference throughput than the best performing model in the remaining 2 datasets.

We believe our findings and results would appeal to the readership of ACM Transactions on Intelligent Systems and Technology. All authors have approved the manuscript and agree with its submission to ACM Transactions on Intelligent Systems and Technology. We look forward to hearing from you at your earliest convenience.

Sincerely,

Edwin Arkel Rios

Ph.D. Candidate

Institute of Electronics

National Yang Ming Chiao Tung University

Hsinchu, Taiwan