Perolehan Citra Fashion dari Street Domain ke Shop Domain Menggunakan Self-Supervised Learning dan Structural Matching = Fashion Image Retrieval from Street Domain to Shop Domain Using Self-Supervised Learning and Structural Matching

Arief Pratama, author

Abstrak

Sebagai salah satu industri terbesar di dunia, pemasaran fashion pada platform ecommerce menarik jutaan pengguna setiap harinya. Salah satu fitur yang penting untuk dimiliki platform ecommerce adalah kemampuan mencari produk fashion menggunakan foto pengguna sebagai query. Hasil pencarian yang akurat akan memberikan manfaat bagi pengguna dan bagi pelaku bisnis. Persoalan ini sangat menantang karena adanya perbedaan domain antara citra query yang diunggah pengguna dengan citra galeri produk yang menjadi target pencarian. Perolehan citra lintas domain dapat diselesaikan dengan metode konvensional seperti pemelajaran metrik menggunakan dataset berlabel. Namun metode ini tidaklah feasible dalam jangka panjang mengingat selalu bertambahnya inovasi di bidang fashion sehingga dibutuhkan anotasi terhadap citra yang berkesinambungan agar model tetap relevan. Pada penelitian ini diusulkan penggunaan self-supervised learning untuk meningkatkan kebermanfaatan data tanpa label dan mengurangi ketergantungan terhadap data berlabel. Pelatihan dengan metode ini menghasilkan sebuah encoder CNN dengan arsitektur ResNet-50, yang dilatih dengan sekumpulan citra tidak berlabel, agar mampu menghasilkan fitur umum dari citra. Model ini kemudian di-finetune dengan data berlabel agar mampu melakukan downstream task, yaitu perolehan citra lintas domain. Untuk meningkatkan hasil perolehan, dilakukan structural matching menggunakan Wasserstein distance (optimal transport) terhadap fitur spasial luaran encoder CNN pada saat inference dan finetuning. Selain itu, structural matching juga dapat menjelaskan bagian mana dari citra yang berkontribusi atas keseluruhan kesamaan atau jarak. Hasil menunjukkan bahwa kinerja encoder yang dilatih dengan self-supervised learning secara kuantitatif masih belum melampaui kinerja encoder baseline ImageNet, dengan perbedaan 1-2% dari sisi akurasi dan mAP menggunakan Triplet Loss, dan 6-10% dengan InfoNCE. Structural matching secara umum dapat meningkatkan hasil perolehan pada encoder yang dilatih dengan self-supervised learning. Hasil kualitatif menunjukkan bahwa semua varian model mampu mencari citra yang mirip dengan query, baik dari sisi kategori, warna, bentuk, dan motif.

.....Being one of the largest industries in the world, fashion marketing on ecommerce platforms attracts millions of users every day. One of the essential features for an ecommerce platform is the ability to retrieve fashion items using user photos as queries. Good search results will yield benefits for users and for businesses. This problem is challenging due to the domain differences of the query images uploaded by the users and of product gallery images as retrieval targets. Cross-domain image retrieval can be accomplished by conventional methods such as metric learning using labeled datasets. However, this method is not feasible in the long term since innovations in this sector are fast such that continuous image annotations are required for the model to stay relevant. In this study, we propose to use self-supervised learning to increase usefulness of unlabeled data and to reduce dependency on labeled data. Training with this method produces a CNN encoder with ResNet-50 architecture, trained on a collection of unlabeled images, to infer generic

features of images. The model is then finetuned with labeled data so that it can perform the downstream task, which is cross-domain image retrieval. To improve retrieval results, we performed structural matching by calculating Wasserstein distance (optimal transport) using spatial features inferred from CNN encoder during inference and finetuning. In addition, structural matching can also explain which parts of two images contribute to overall similarity or distance. Results show that an encoder trained with self-supervision quantitatively has not yet outperformed off-the-shelf ImageNet encoder baseline, with a difference in terms of accuracy and mAP of 1-2% for Triplet Loss, and 6-10% for InfoNCE. Generally, structural matching can improve retrieval results for self-supervised encoders. Qualitative results show that all model variants are able to retrieve images similar to the query, in terms of categories, colors, shapes, and patterns.