Self-Supervised Composed Image Retrieval with Large Multi-Modal Model

No Author Given

No Institute Given

1 Introduction

2 Related Work

Composed Image Retrieval Composed Image Retrieval (CIR) has gained considerable attention due to its capability to facilitate sophisticated image search tasks by leveraging both visual and textual information. Early work [16] in CIR introduced a supervised learning framework utilizing annotated triplets consisting of a reference image, a text modifier, and a target image. This residual gating method effectively merges multimodal information by combining image and text features. Subsequent approaches [14] have integrated graph convolutional networks with existing composition methods, while others [7] have considered image style and content separately using distinct neural network modules. Advancements in cross-modal pretraining techniques [10, 18, 12, 9] have further enhanced supervised approaches in CIR. For instance, CLIP4CIR [3] employs CLIP [12] as the backbone, training an image-text combiner to merge the reference image and complementary text into a unified representation, which is then matched with the target image. BLIP4CIR+Bi [11] explores the complementary relationship in the mapping from ⟨target image, reversed modification text⟩ pairs to the reference image, while other works investigate hidden relations within image-text tuples from unique perspectives [5, 17]. The CASE method [8] utilizes large language models (LLMs) to generate similar triplets, augmenting existing datasets and extracting latent information from larger datasets. A significant recent advancement is the SPRC [1], which introduces Qformer, a feature of BLIP2 [9] that acts as a fusion encoder, thereby enhancing the capabilities of CIR networks. The sentence-level prompt technique from SPRC underscores the value of utilizing both explicit and implicit relationships in CIR tasks. Despite the successes of supervised CIR methods, their dependence on meticulously annotated triplet datasets poses significant challenges for large-scale data collection and labeling.

Zero-Shot Composed Image Retrieval To address these limitations, zero-shot composed image retrieval (ZSCIR) has been proposed. The ZSCIR approach circumvents the need for triplet-labeled data by [13] relying solely on image-text pairs, employing an MLP to convert CLIP visual features into single-word embeddings and using a cycle contrastive loss for training. SEARLE [2] enhances

this by leveraging large language models to generate additional descriptions, thereby improving the alignment between mapped image features and class semantics. Recent efforts [4] introduce a two-stage framework involving text inversion and distillation to map images into the text domain. Further advancements [15] in ZSCIR involve integrating external knowledge to effectively align visual features with textual semantics. The authors propose a diffusion-based CIR model that operates in the latent space of a frozen CLIP model. CompoDiff [6] uses a Transformer-based denoiser and is trained with classifier-free guidance, allowing it to accommodate diverse and complex conditions. In our research, we introduce a novel semi-supervised CIR method that overcomes the drawbacks of both supervised and zero-shot techniques. By seeking out reference and related target images from auxiliary data, we use a large language model-based Spoter to create text that describes the visual differences between them. Spoter's advanced language skills and model-agnostic approach produce pseudo-triplets that improve CIR model performance.

References

1. Bai, Y., Xu, X., Liu, Y., Khan, S., Khan, F., Zuo, W., Goh, R.S.M., Feng, C.M.: Sentence-level prompts benefit composed image retrieval (2023), https://arxiv.org/abs/2310.05473
2. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15338–15347 (October 2023)
3. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4959–4968 (June 2022)
4. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Composed image retrieval using contrastive learning and task-oriented clip-based features. ACM Trans. Multimedia Comput. Commun. Appl. **20**(3) (oct 2023). https://doi.org/10.1145/3617597, https://doi.org/10.1145/3617597
5. Chen, Y., Zhou, J., Peng, Y.: Spirit: Style-guided patch interaction for fashion image retrieval with text feedback. ACM Trans. Multimedia Comput. Commun. Appl. **20**(6) (mar 2024). https://doi.org/10.1145/3640345, https://doi.org/10.1145/3640345
6. Gu, G., Chun, S., Kim, W., Jun, H., Kang, Y., Yun, S.: Compodiff: Versatile composed image retrieval with latent diffusion (2024), https://arxiv.org/abs/2303.11916
7. Lee, S., Kim, D., Han, B.: Cosmo: Content-style modulation for image retrieval with text feedback. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 802–812 (2021). https://doi.org/10.1109/CVPR46437.2021.00086
8. Levy, M., Ben-Ari, R., Darshan, N., Lischinski, D.: Data roaming and quality assessment for composed image retrieval (2023), https://arxiv.org/abs/2303.09429
9. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023), https://arxiv.org/abs/2301.12597

10. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation (2022), https://arxiv.org/abs/2201.12086
11. Liu, Z., Sun, W., Hong, Y., Teney, D., Gould, S.: Bi-directional training for composed image retrieval via text prompt learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5753–5762 (January 2024)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
13. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval (2023), https://arxiv.org/abs/2302.03084
14. Shin, M., Cho, Y., Ko, B., Gu, G.: Rtic: Residual learning for text and image composition using graph convolutional network (2021), https://arxiv.org/abs/2104.03015
15. Suo, Y., Ma, F., Zhu, L., Yang, Y.: Knowledge-enhanced dual-stream zero-shot composed image retrieval (2024), https://arxiv.org/abs/2403.16005
16. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval - an empirical odyssey (2018), https://arxiv.org/abs/1812.07119
17. Wen, H., Zhang, X., Song, X., Wei, Y., Nie, L.: Target-guided composed image retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia. MM '23, ACM (Oct 2023). https://doi.org/10.1145/3581783.3611817, http://dx.doi.org/10.1145/3581783.3611817
18. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models (2022), https://arxiv.org/abs/2205.01917