Executive Summary

Efficient and Discriminative Image Feature Extraction for Multi-Domain Image Retrieval

Morris Florek

January 30, 2024

The prevalence of image capturing devices has led to the growth of digital image collections and the need for advanced image retrieval systems. Content-based image retrieval (CBIR) finds semantically similar images from a large database given a query image [15]. CBIR has many applications in various domains [20, 30, 28, 9], but current methods are often limited by their domain-specificity [17, 3] and encounter difficulties with out-of-domain images and lack of generalization. To address the cost and inconvenience associated with multiple per-domain models in a unified image retrieval system [6], this study delves into the realm of multi-domain visual-semantic feature extraction. The objective is to efficiently develop and train a multi-domain image encoder capable of extracting discriminative image features tailored for fine-grained image retrieval. Therefore, this study presents the following contributions:

- Demonstrating close SOTA results on the 2022 Google Universal Image Embedding Challenge (GUIEC) [1] while using significantly less computational resources for inference and training by solely fine-tuning the last projection layer (i.e. linear probing).
- Curation of a streamlined multi-domain training dataset allowing for resource-efficient training. The dataset will be publicly accessible.
- Ablations on the efficacy of various visual foundation models and margin-based softmax losses for the utilization of multi-domain feature extraction.

The multi-domain feature extraction model was conceptualized based on the leading methodologies [22, 7] observed in the GUIEC [1]. The challenge served as a catalyst for exploring innovative ideas and methodologies to train universal image representations that can efficiently retrieve images across diverse domains. Figure 1 depicts the model architecture and experimental framework employed in the study. To accommodate computational constraints, we confined the fine-tuning process to the last projection layer, which required us to freeze

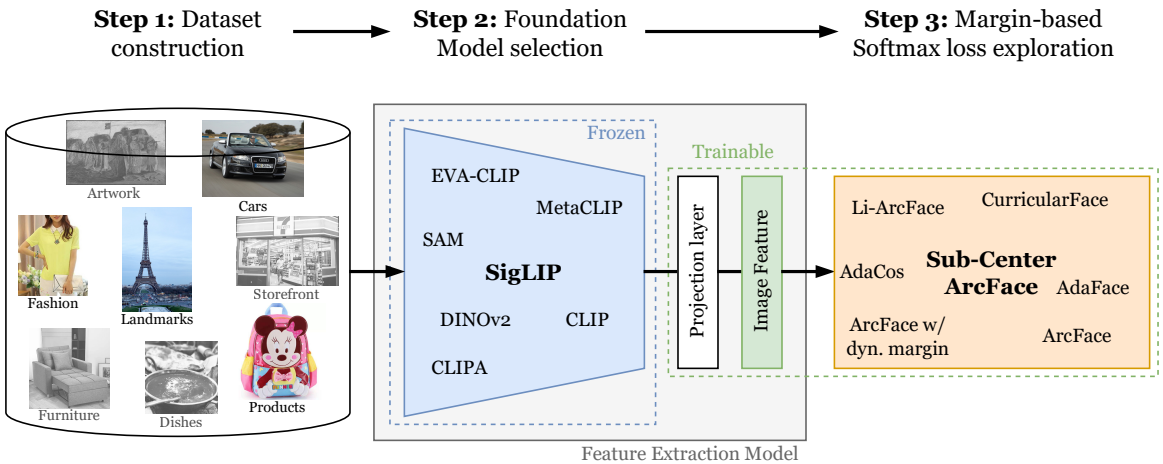

Figure 1: Overview of the model architecture and the experimental framework. Showing the trainable and non-trainable parts of the model and as well as the considered foundation models and margin-based softmax losses. Highlighting the selected domains of the curated training dataset, the utilized foundation model and the margin-based softmax loss to achieve close SOTA result on the GUIEC [1].

the entire backbone. However, the overall goal was to adeptly develop and fine-tune a model within resource constraints while maintaining its competitiveness against top-tier approaches in the GUIEC.

The experimental study involved curating a customized training dataset, identifying a robust foundational model to serve as a backbone, and exploring various margin-based softmax losses. The datasets considered for inclusion were derived from those utilized in the GUIEC [1], encompassing 15 distinct datasets across 8 diverse domains. The curation process entailed the meticulous selection and incorporation of datasets that substantially contributed to the model's performance. To optimize the dataset size, rare classes were discarded, and limits were imposed on the maximum number of samples per class. This resulted in a curated multi-domain training dataset comprising four domains from four different datasets: Products-10k [2], Google Landmark v2 [24], DeepFashion [16], and a custom version of Stanford Cars [12] with enhanced class granularity. Owing to the inability to fine-tune the backbone owing to hardware capacities, several SOTA foundation models [21, 14, 23, 25, 27, 18, 11] (see Figure 1) were identified for assessment. Following zero-shot and linear probing evaluations, the SigLIP [27] model demonstrated superior performance, outperforming all other candidates. Subsequently, the feature extraction model was linear probed using a diverse set of margin-based softmax losses [5, 4, 13, 8, 10, 29] to enhance inter-class discrimination, as shown in Figure 1. Among all investigated methods, linear probing with Sub-Center ArcFace [4] proved to be the most effective approach.

In summary, the study emphasizes the significance of meticulous selection and adequate amount of training data, the importance of a robust visual-semantic foundation model, and the appropriate choice of the loss function. The incorporation of these observations led to comparable performance on the GUIEC [1], despite limited computational resources. Leveraging the SigLIP [27] model as the backbone and fine-tuning the projection layer on the curated dataset using Sub-Center ArcFace [4] resulted in an excellent mMP@5 score of 0.722 on the GUIEC evaluation dataset. Notably, this approach, while employing a smaller model (based on the number of model parameters) and without end-to-end fine-tuning, trailed the GUIEC leaderboard by only 0.8 percentage points (see Table 1). Furthermore, it outperformed the highest-ranked methodology with similar computational prerequisite (5th place [19]), achieving a significant 3.6 percentage point improvement.

Rank (GUIEC)	**Method**	**# Total params**	**# Trainable params**	**mMP@5**
1st [22]	Fine-Tuning	661M	661M	0.730
2nd [7]	Fine-Tuning	667M	667M	0.711
5th [19]	Linear Probing	633M	1.1M	0.686
10th[1]	Linear Probing	1,045M	22.0M	0.675
Own Approach	Linear Probing	431M	2.3M	0.722

Table 1: Performance and model size comparison of different utilized training methods (end-to-end fine-tuning or linear probing) on GUIEC [1] evaluation dataset. It improves the total model parameters at inference by 32% compared to the leanest approach (5th place), reduces the number of trainable parameters by 289x compared to the fine-tuning approaches (1st and 2nd place), and achieves a performance close to SOTA, surpassing 2nd place and just behind 1st place.

Future Direction

Further exploration of the proposed feature extraction model is directed toward its application to a novel large-scale multi-domain dataset known as UnED [26]. This dataset was released towards the end of the thesis editing period and was therefore not considered in the study. Similar to the curated training dataset of this research, UnED incorporates images from several publicly available datasets spanning eight different domains. Introduced by the same team that organized the GUIEC [1], UnED represents a notable advancement in the field of universal image embedding. In contrast to the GUIEC evaluation dataset, UnED includes fewer domains (8 compared to 11). While the GUIEC required the curation of a custom dataset for training, UnED provides a comprehensive framework for both training and evaluation, allowing for a more pronounced focus on model development. Therefore, we plan to submit the results presented in this study, together with the upcoming research on the UnED dataset, to the **student track** of the **2024 GCPR** conference.

[1] GUIEC 10th place solution

References

[1] Araujo, A., Cao, B., bbl, B., Chen, F., Maggie, Lipovský, M., Seyedhosseini, M., Dogan, P., Dane, S., and Cukierski, W. Google Universal Image Embedding, 2022.

[2] Bai, Y., Chen, Y., Yu, W., Wang, L., and Zhang, W. Products-10K: A Large-scale Product Recognition Dataset, Aug. 2020. arXiv:2008.10545 [cs].

[3] Cao, B., Araujo, A., and Sim, J. Unifying Deep Local and Global Features for Image Search. In *Computer Vision – ECCV 2020* (Cham, 2020), A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 726–743.

[4] Deng, J., Guo, J., Liu, T., Gong, M., and Zafeiriou, S. Sub-center ArcFace: Boosting Face Recognition by Large-Scale Noisy Web Faces. In *Computer Vision – ECCV 2020* (Cham, 2020), A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Lecture Notes in Computer Science, Springer International Publishing, pp. 741–757.

[5] Deng, J., Guo, J., Xue, N., and Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019), pp. 4685–4694. ISSN: 2575-7075.

[6] Feng, Y., Peng, F., Zhang, X., Zhu, W., Zhang, S., Zhou, H., Li, Z., Duerig, T., Chang, S.-F., and Luo, J. Unifying Specialist Image Embedding into Universal Image Embedding, Mar. 2020. arXiv:2003.03701 [cs].

[7] Huang, X., and Li, Q. 2nd Place Solution to Google Universal Image Embedding, Oct. 2022. arXiv:2210.08735 [cs].

[8] Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., and Huang, F. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. pp. 5901–5910.

[9] Jain, A. K., Klare, B., and Park, U. Face Matching and Retrieval in Forensics Applications. *IEEE MultiMedia 19*, 01 (Jan. 2012), 20, 28–20, 28. Publisher: IEEE Computer Society.

[10] Kim, M., Jain, A. K., and Liu, X. AdaFace: Quality Adaptive Margin for Face Recognition. pp. 18750–18759.

[11] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment Anything, Apr. 2023. arXiv:2304.02643 [cs].

[12] Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3D Object Representations for Fine-Grained Categorization. pp. 554–561.

[13] Li, X., Wang, F., Hu, Q., and Leng, C. AirFace: Lightweight and Efficient Model for Face Recognition. pp. 0–0.

[14] Li, X., Wang, Z., and Xie, C. An Inverse Scaling Law for CLIP Training, May 2023. arXiv:2305.07017 [cs].

[15] Li, X., Yang, J., and Ma, J. Recent developments of content-based image retrieval (CBIR). *Neurocomputing 452* (Sept. 2021), 675–689.

[16] Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. DeepFashion: Powering Robust Clothes Recognition and Retrieval With Rich Annotations. pp. 1096–1104.

[17] Noh, H., Araujo, A., Sim, J., Weyand, T., and Han, B. Large-Scale Image Retrieval with Attentive Deep Local Features, Feb. 2018. arXiv:1612.06321 [cs].

[18] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. DINOv2: Learning Robust Visual Features without Supervision, Apr. 2023. arXiv:2304.07193 [cs].

[19] Ota, N., Yokoi, S., and Yamaoka, S. 5th Place Solution to Kaggle Google Universal Image Embedding Competition, Oct. 2022. arXiv:2210.09495 [cs].

[20] Qayyum, A., Anwar, S. M., Awais, M., and Majid, M. Medical image retrieval using deep convolutional neural network. *Neurocomputing 266* (Nov. 2017), 8–20.

[21] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning* (July 2021), PMLR, pp. 8748–8763. ISSN: 2640-3498.

[22] Shao, S., and Cui, Q. 1st Place Solution in Google Universal Images Embedding, Oct. 2022. arXiv:2210.08473 [cs].

[23] Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. EVA-CLIP: Improved Training Techniques for CLIP at Scale, Mar. 2023. arXiv:2303.15389 [cs].

[24] Weyand, T., Araujo, A., Cao, B., and Sim, J. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. pp. 2575–2584.

[25] Xu, H., Xie, S., Tan, X. E., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. Demystifying CLIP Data, Oct. 2023. arXiv:2309.16671 [cs].

[26] Ypsilantis, N.-A., Chen, K., Cao, B., Lipovský, M., Dogan-Schönberger, P., Makosa, G., Bluntschli, B., Seyedhosseini, M., Chum, O., and Araujo, A. Towards Universal Image Embeddings: A Large-Scale Dataset and Challenge for Generic Image Representations. pp. 11290–11301.

[27] Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid Loss for Language Image Pre-Training. pp. 11975–11986.

[28] Zhang, X., Wang, S., Li, Z., and Ma, S. Landmark Image Retrieval by Jointing Feature Refinement and Multimodal Classifier Learning. *IEEE Transactions on Cybernetics 48*, 6 (June 2018), 1682–1695. Conference Name: IEEE Transactions on Cybernetics.

[29] Zhang, X., Zhao, R., Qiao, Y., Wang, X., and Li, H. AdaCos: Adaptively Scaling Cosine Logits for Effectively Learning Deep Face Representations. IEEE Computer Society, pp. 10815–10824.

[30] Zhang, Y., Pan, P., Zheng, Y., Zhao, K., Zhang, Y., Ren, X., and Jin, R. Visual Search at Alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY, USA, July 2018), KDD '18, Association for Computing Machinery, pp. 993–1001.