Image Retrieval Under Fine-Grained and Long-Tailed Distribution

Shuai Chen, Fanman Meng(✉), Qingbo Wu, Yuxuan Liu, and Yaofeng Yang

University of Electronic Science and Technology of China,
Chengdu 611731, China
{s-chen,liuyuxuan1,202022011513}@std.uestc.edu.cn,
{fmmeng,qbwu}@uestc.edu.cn

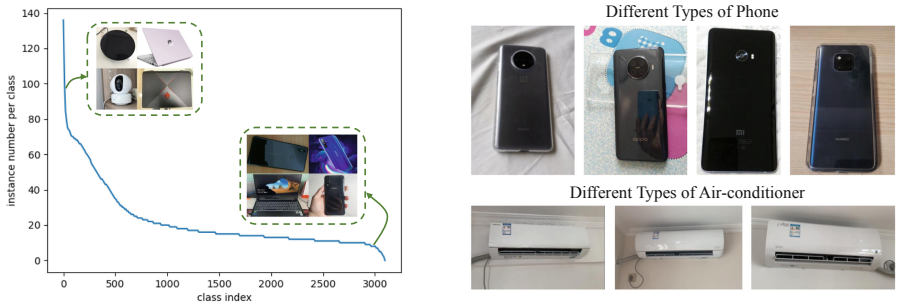
Abstract. Image retrieval is a promising task that aims to retrieve relevant images from a large-scale database based on user's requests. This paper is a solution to the Huawei 2020 digital device image retrieval competition. The core idea of the proposed solution is to employ metric-learning to perform fine-grained image retrieval. To be specific, to address the long-tailed distribution caused by the imbalance of samples in the dataset, an image retrieval tailored causal graph is first constructed, and a causal intervention is performed for counterfactual reasoning, which proves to be effective to alleviate the influence of long-tailed distribution. To solve the challenging fine-grained image retrieval issues, this paper proposes a novel global and local attention image retrieval framework, which simultaneously mines global and local features to obtain the most discriminative feature. In addition, an object detector is further developed to capture the object of interest, thereby an accurate representation of the foreground area can be acquired. Furthermore, some additional testing and model ensemble skills, such as re-ranking, fine-tuning on larger images, and multi-scale testing, are implemented to further boost the performance. Extensive experiments on the benchmark demonstrate the effectiveness of the proposed method.

Keywords: Fine-grained image retrieval · Long-tailed distribution · Attention mechanism · Causal intervention

1 Introduction

Image retrieval technology aims at retrieving user-interested images from a large-scale database according to the user's specific needs. Image retrieval is a fundamental task in the modern computer vision community and has a large number of application scenarios, such as e-commerce [8], fashion landmark detection [10,23], and so on. In the past few years, due to the rapid growth of multimedia technology and the continuous increase of social information, image retrieval has aroused growing attention and many effective image retrieval methods have been developed [3,7,19]. Although considerable progress has been made in image

G. Zhai et al. (Eds.): IFTC 2020, CCIS 1390, pp. 259–270, 2021.
https://doi.org/10.1007/978-981-16-1194-0_23



(a) The distribution of the dataset is long-tail distribution.

(b) The classes of dataset is fine-grained.

Fig. 1. Illustration of the long-tailed (a) and fine-grained (b) essence of the given dataset. In (a), some classes have hundreds of instances, while there are many classes whose instances are less than twenty. In (b), different brands of phones and air-conditioners belong to different classes, although they are very similar to each other.

retrieval, most existing image retrieval methods still face some challenges, such as long-tailed distribution and fine-grained classification problems, which will affect the practicability of image retrieval algorithms.

The long-tailed essence of the given dataset is shown in Fig. 1(a), some classes (termed as tail classes) only contain rarely few images, while their counterparts (termed as head classes) contain hundreds of images. This unbalanced distribution can inevitably lead to an inferior training process predominated by the head class, while deterioration of tail class is gained. The previous methods mainly adapt the re-sampling [11] or re-weighting [4] strategy to alleviate such an issue, however, these two strategies need future data distribution before training, thus cannot be applied to a variety of dynamic data streams. Motivated by [15,21], we establish a Structural Causal Model (SCM) [15], which considers the SGD momentum term as a confounder in long-tailed image retrieval, playing two roles simultaneously: benefiting the head class and misleading the tail class to the representation of the head class. The details of the utilized causal graph can be found in [21]. After establishing the causal graph, we perform backdoor adjustment [13] to obtain the real representation of tail class by wiping out the influence of the SGD momentum term. We refer readers to [15,21] for more details about SCM and it's applications in long-tailed image classification.

Fine-grained image retrieval is a challenging problem and needs further research. This issue mainly locates on the following two technical points. First, the category-level of fine-grained retrieval is much lower than traditional image retrieval. Take category 'cellphone' as an example, the traditional image retrieval only needs to identify 'cellphone' from other categories (such as 'air-conditioner' and 'watch'), but the fine-grained retrieval need to distinguish the specific kinds (such as 'Huawei' and 'iPhone') of 'cellphone' additionally, and even the 'Huawei cellphone' is also split into some tiny classes according to its specific type. Figure 1(b) illustrates some samples of 'cellphone' and 'air-conditioner'. Second, the fine-grained categories belonging to tail classes have a big difference

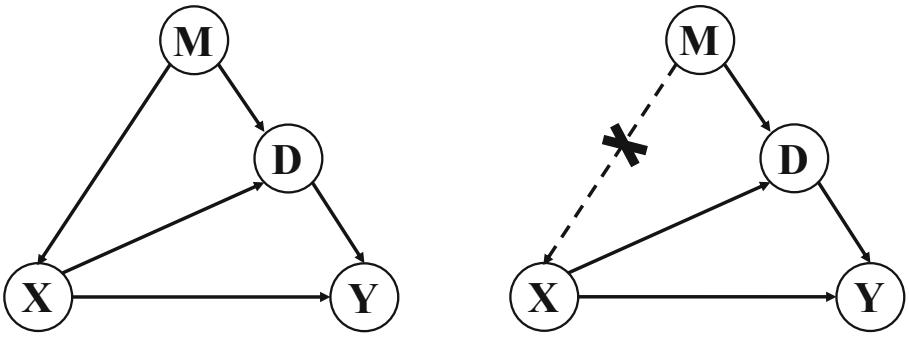
in posture, texture, color, and background, which brings great difficulty to fine-grained classification. Previous methods [1,2,5,12,16,18] propose a global structure in a single-forward pass to solve this task but the performance is limited. In our opinion, fine-grained image retrieval has a high demand for local details but such a global strategy cannot capture detailed and local information, thus it has poor performance. [2] proposes a local self-attention module combining with global structure and achieves great improvement. This module uses 1×1 convolution layers to convert middle feature maps into a coefficient matrix, and then the matrix is used to re-weight feature maps to filter information. Motivated by group convolution, we propose our structure: Global and Local Attention Network (GLANet) to solve such an issue. The key innovation is the multi-head attention module, and a fusion module is proposed to fuse the local features and global features, forming the final descriptor. In this way, each output has highlighted local channel-specific information, which can promote the network to get better recognition capability on fine-grained retrieval task.

Noted that the fusion module plays two roles in our method, fusing the global and local features in the GLANet, and simultaneously acting as a sampling method to perform the backdoor adjustment in the utilized causal graph. Moreover, a foreground detector and some additional strategies are implemented to further enhance performance. To summarize, the main contribution of the proposed method are four-fold:

1. Different from the existing methods, the proposed method utilizes an image retrieval tailored causal graph and a causal intervention strategy to perform counterfactual reasoning, capable of boosting performance under a long-tailed distribution setting.
2. To cope with the challenge of fine-grained retrieval, this paper proposes a global and local attention image retrieval framework, which jointly uses the global and local descriptors to fully exploit the most discriminative features.
3. Observing that the interesting object is usually located in the center of each picture, this paper develops an object detector for foreground detection, so that the model can focus on the object region and a precise representation of the interesting object can be obtained.
4. A series of additional strategies, namely re-ranking, fine-tuning on larger images, and model ensemble, are implemented to further improve the performance of the proposed model. Extensive experiments on the benchmark demonstrate the effectiveness of the proposed method.

2 Method

In this section, we will elaborate the strategies adopted to deal with long-tailed distribution, namely the causal graph and casual intervention firstly. Then we describe the detailed architecture of the proposed GLANet, addressing the fine-grained image retrieval task. Finally, we present the core idea of the foreground detector.



(a) Casual Graph (b) Casual Intervention

Fig. 2. Illustration of the causal graph and the causal intervention process. M, D, X, Y denote momentum term, projection on head class, image feature, image label respectively. By cutting off the $M \longrightarrow X$ edge, the true $X \longrightarrow Y$ is obtained.

2.1 Causal Graph for Long-Tailed Distribution

Causal Graph Establishment. To deal with the long-tailed distribution of the given dataset, a causal graph is established, shown in Fig. 2, containing four data variables: the SGD momentum M, the image feature X, projection on the head direction D, and the image label Y. The directed edges among these variables indicate how each variable influences each other, where the root is the cause and the goal is the effect. The edges are explained as follows:

$M \longrightarrow X$. This edge denotes the image feature X form a specific image, with the influence of optimizer momentum M.

$M \longrightarrow D \longleftarrow X$. These two edges illustrate that the momentum M makes X deviating from its real direction.

$X \longrightarrow Y \longleftarrow D$. The final model prediction Y is determined by two variables, where $X \longrightarrow Y$ is the real effect, and also what we want, while $Y \longleftarrow D$ can be taken as a median that X influence Y indirectly.

Causal Intervention. To obtain the direct representation of head class and tail class, known as the Total Direct Effect (TDE) [14,22]. A de-confound training step is needed, in other words, the classifier of the image retrieval model should be normalized:

$$Y_i = \frac{\tau}{M} \sum_{m=1}^{M} \frac{\left(w_i^m\right)^T x^m}{\left(\left\|w_i^m\right\| + \gamma\right) \left\|x^m\right\|} \tag{1}$$

where τ, γ is two hyper-parameters controlling the classifiers, M is the multi-head number, Y_i is the logits of i^{th} class, and w_i^m is the weight of m^{th} classifier without bias.

To accomplish such a de-confound training, a moving average feature $\overline{x}_i$ should be maintained across training iterations, and the direction of such a

moving average feature is considered as the inclination direction $\hat{d}$ of training dynamics on the head class.

$$\overline{x}_i = \alpha \overline{x}_{i-1} + (1-\alpha) x_i \tag{2}$$

$$\hat{d} = \overline{x} / \|\overline{x}\| \tag{3}$$

where α is the momentum term, and x_i is the image feature under current model.

During the validation and test step, a counterfactual TDE inference step is performed, that is, part of the class logits, having the excessive tendency towards the head class, should be taken out:

$$TDE\left(Y_i\right) = \frac{\tau}{M} \sum_{m=1}^{M} \left(\frac{\left(w_i^m\right)^T x^m}{\left(\left\|w_i^m\right\| + \gamma\right) \left\|x^m\right\|} - \alpha \cdot \frac{\cos\left(x^m, \hat{d}^m\right) \cdot \left(w_i^m\right)^T \hat{d}^m}{\left\|w_i^m\right\| + \gamma} \right) \tag{4}$$

After obtaining the TDE logits of a specific image, the image retrieval process can be built on this representation.

2.2 Global and Local Attention Network

To extract the efficient representation of images, we propose the GLANet, which is shown in Fig. 3. GLANet mainly contains three components: a local feature extraction module, a global feature extraction module, and a feature fusion module. Furthermore, when extracting local features, we introduce an attention mechanism to obtain local features at diverse spatial locations.

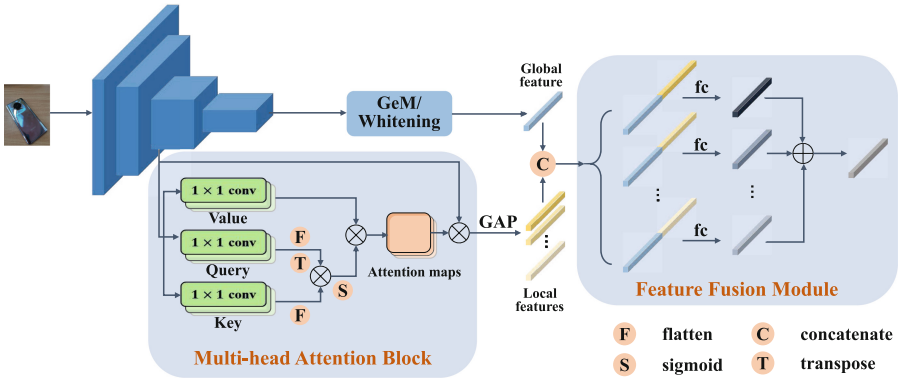


Fig. 3. Overview of our method. It contains three components: local feature extraction module, global feature extraction module and feature fusion module.

Global Feature Extraction Module. For the extraction of global features, we implement it based on the deep features to obtain global and high-level semantic information of images. Specifically, the deep features of size $C_D \times H_D \times W_D$ is obtained from the output of the last layer firstly. Then, we apply Generalized Mean Pooling (GeM) [16] to aggregate deep features into a global feature. Afterward, whitening is applied to remove redundant information, which utilizes a fully connected layer F with learning bias b to get the final global features v_G.

$$v_G = F \times \left(\frac{1}{H_D W_D} \sum_{h,w} d_{h,w}^p \right)^{1/p} + b, \tag{5}$$

where p denotes the generalized mean power parameter, $d_{h,w}$ is the feature vector of size C_D at location h, w.

Local Feature Extraction Module. As for local features with low-level information, we apply our proposed local feature extraction module to obtain them from the features of the penultimate layer. In addition, to obtain richer local features from diverse locations, a Multi-Head Attention Block is constructed during the process of local features generation.

Given an input feature of size $C_M \times H_M \times W_M$. Firstly, we apply three 1×1 convolutions to the features to get query, value, and key. Then tensor $q \in N_M \times C_M$ is obtained through flattening and transposing the query, and tensor $k \in C_M \times N_M$ is obtained through flattening the key, where $N_M = H_M \times W_M$. Then We multiply q and k together and use softmax to get a relationship matrix S of size $N_M \times N_M$

$$S = softmax(q^\top k). \tag{6}$$

Then we multiply the relation matrix S and value to get the attention map of size $C_M \times H_M \times W_M$. We implement the attention operation M times to get M attention maps focusing on different locations and then multiply the attention map with the original feature to get the weighted feature

$$F_{Li} = S \cdot F_i, i \in [1, M] \tag{7}$$

where M represents the number of attention maps, F_{Li} denotes the i^{th} local feature. Then, we apply Global Average Pool (GAP) to the obtained local features to obtain a feature vector v_{Li} of size $C_M \times 1$.

$$v_{Li} = f_{GAP}(F_{Li}) \tag{8}$$

Feature Fusion Module. New feature vectors $f_C = \{\boldsymbol{f}_{C1}, \ldots, \boldsymbol{f}_{CM}\}$ are obtained by concatenating the global feature vector and each local feature vector. We reduce the dimension of the newly obtained feature vectors through the

fully connected layer and then perform the average operation on the obtained results. In this way, we get the feature vector v_{out} that we finally output.

$$v_{\text{out}} = \frac{1}{M} \sum_{i=1}^{M} f_{FC}\left(f_{Ci}\right) \tag{9}$$

Subsequently, we adapt the Sphere loss [9] for global and local features learning. Specifically, the Sphere loss can be described as follows:

$$L_{\text{ang}} = \frac{1}{N} \sum_{i} -\log\left(\frac{e^{\|\boldsymbol{v}_{out}\|\psi\left(\theta_{y_i,i}\right)}}{e^{\|\boldsymbol{v}_{out}\|\psi\left(\theta_{y_i,i}\right)} + \sum_{j \neq y_i} e^{\|\boldsymbol{v}_{out}\| \cos\left(\theta_{j,i}\right)}}\right) \tag{10}$$

where $\psi\left(\theta_{y_i,i}\right)$ denotes angular margin, which can be formulated as follows:

$$\psi\left(\theta_{y_i,i}\right) = (-1)^k \cos\left(m\theta_{y_i,i}\right) - 2k \tag{11}$$

where $\theta_{y_i,i} \in \left[\frac{k\pi}{m}, \frac{(k+1)\pi}{m}\right]$, and m is used to control the size of angular margin.

2.3 Foreground Detector

Observing that the images of the given dataset have the following characteristics: each image in the dataset has only one object. Moreover, small objects occupy a certain proportion of the dataset. Based on the above observations, the proposed method adopts an object detector [17] to firstly extract the object of interest from the image, and subsequently extract features based on the detection result. By adopting the above measures, the interference of some useless information, such as the background area, can be eliminated to some extent, thus an accurate representation of the foreground area can be acquired.

3 Experiment

In this section, the dataset and experiment setup, comparison with previous methods, and the tricks and model ensemble strategy will be introduced.

3.1 Dataset and Experiment Setup

Dataset. The track B of the 2020 Digix Global AI Challenge is a digital device image retrieval competition, and the given dataset contains three parts: the training part, the test-A part, and test-B part. The training part contains 3094 classes of common digital devices and 68811 images in total. As shown in Fig. 1, the training data is long-tailed and fine-grained, making the competition especially challenging. The test-A part contains a query set of 9600 images and a gallery set of 49804 images. The organization of test-B part is the same as test-A part, but with different images inside. Note that only the class label of the training part is available, no other information except the images is available in test-A and test-B part. To validate the trained models, the original training part is split into a new training part and validation part.

Basic Configuration. The model framework is shown in Fig. 3, the ImageNet pre-trained ResNet-101 [6] is used as the backbone in our baseline method, while some larger models are also used, such as EfficientNet [20] and ResNext [24]. And the dim of GLANet is set to 512 because larger dim leads to over-fitting, the multi-head number is set to 10 for the balance of total parameters and efficiency. The sphere-loss and a common cross-entropy loss without weighting are used. For image augmentation, several strategies are adapted, such as horizontally flip, brightness, saturation, and contrast change, center-crop. The baseline model takes the 512 × 512 training images as input. The SGD optimizer is selected for training, where learning rate, momentum, weight decay are set to 5e−5, 0.9, 1e−5, a cosine annealing adjustment is performed. The batch-size is set to 16. The total epoch is set to 200. For the validation set and test set, the baseline model also takes the 512 × 512 training images as input. All the experiments are based on the PyTorch framework with 4 Nvidia Titan XP GPUs.

Evaluation Metric. The top-1 accuracy and the mean average precision of top-10 (mAP@10) are used, the final average score is formed as:

$$Score = 0.5Acc + 0.5mAP@10 \tag{12}$$

3.2 Comparison with Previous Methods

To accomplish the comparison, we select some current image retrieval method, such as R-MAC [5], GeM [16], GeM-AP [18], GEM-SOLAR [12], GEM-SmoothAP [1]. As shown in Table 1, we compare the proposed method with some existing methods on the given image retrieval dataset. As can be seen, compared with these methods, our method can achieve the best retrieval performance in both the validation set and the test-A set, with a 70.52 top-1 accuracy, 67.12 mAP@10, and an average score of 68.82 in the validation set, and an average score of 65.50 in test-A set.

To demonstrate the effectiveness of the proposed solution, some images from the query set and their corresponding retrieved images from the gallery set are illustrated in Fig. 4.

Table 1. Comparison with other image retrieval method

Methods	Val set			Test-A
	Top-1 acc	mAP@10	Avg	Avg
R-MAC [5]	67.16	63.12	65.14	62.10
GeM [16]	68.55	64.19	66.37	63.12
GeM-AP [18]	69.94	65.74	67.84	64.14
GEM-SOLAR [12]	69.62	66.42	68.02	64.73
GEM-SmoothAP [1]	70.30	66.84	68.57	65.28
Ours	70.52	67.12	68.82	65.50

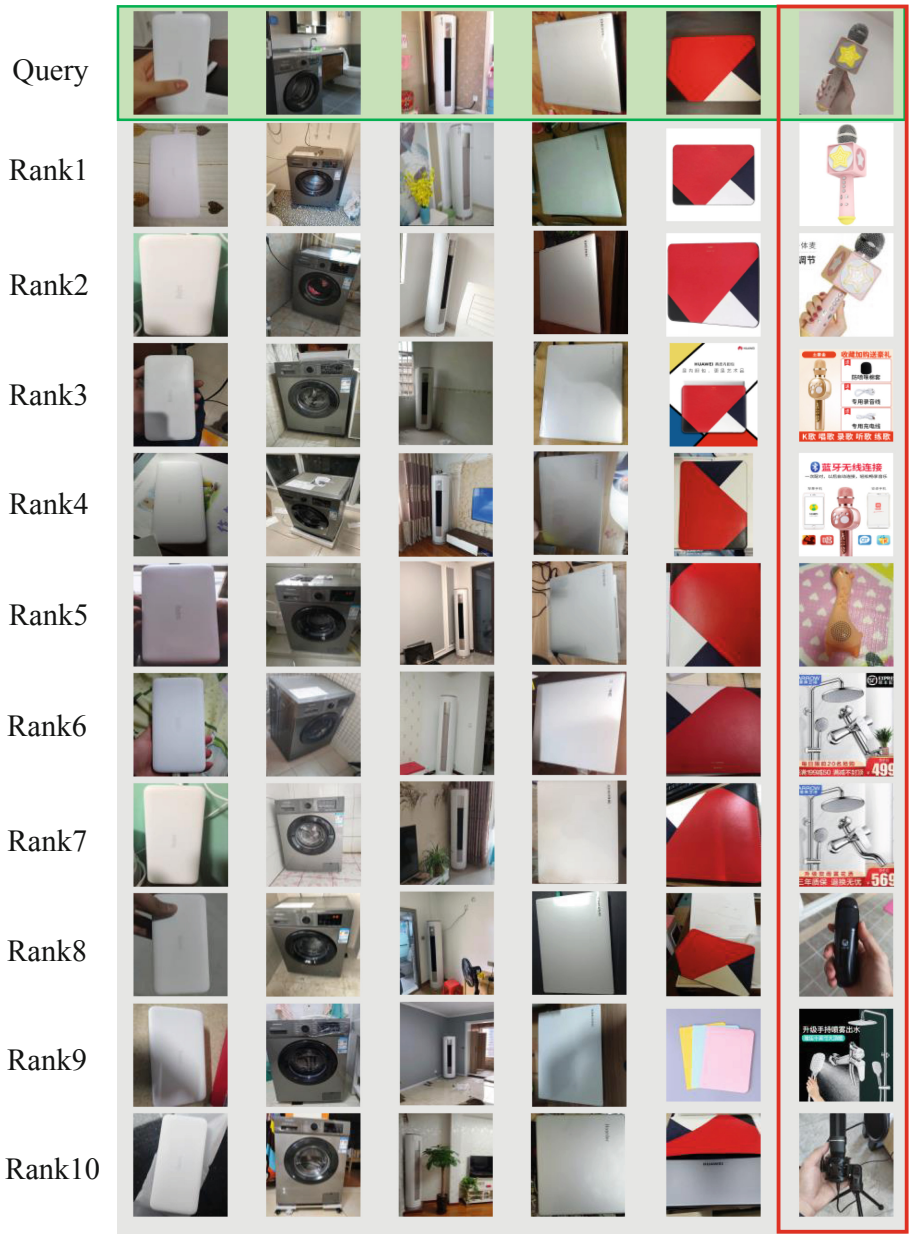


Fig. 4. The visualization results of our method. For each query image, the top 10 retrieval results are shown ranked up-down. The first 5 columns indicate some good cases, while the last column shows the failure case.

Table 2. Some additional tricks are adapted to further boost the performance.

Tricks				Val set			Test-A
Larger-Size	Multi-Scale	Re-ranking	Ensemble	Top-1 acc	mAP@10	Avg	Avg
				70.52	67.12	68.82	65.50
√				71.3	68.14	69.72	66.34
√	√			72.81	71.55	72.18	69.36
√	√	√		74.6	72.68	73.64	71.3
√	√	√	√	76.8	74.16	75.48	73.32

3.3 Tricks and Model Ensemble Strategy

Some tricks are also adapted to boost the performance on the test-A part. First, a progressive enlarging of training images is utilized. 512 × 512 image is used at the beginning, and alternating to 786 × 768 and 1024 × 1024 sequentially, when converged to the last input size. Second, multi-scale testing is performed. At each testing, we collect the outputs under different scales, varying among $[512, 768, 840, 1024]$, and a concatenation step is performed before searching the most similar samples. Third, some re-ranking skills are performed, the descriptor of a query image is the average of the top-3 retrieved images and itself, and a second search is performed based on this new descriptor. Forth, the model ensemble strategy is used to further boost performance, descriptors from different backbones, including ResNet101, Efficient-7, and Efficient-8, different methods, such as GEM-SmoothAP [1] and GEM-SOLAR [12]. The above tricks make that the proposed solution obtains an average score of 73.32 on test-A data. Details can be seen in Table 2.

4 Conclusion

In this paper, we propose a novel global and local attention image retrieval network based on metric learning to solve the long-tailed distribution and fine-grained image retrieval. To tackle the long-tailed distribution issues, we leverage an image retrieval tailored causal graph and a causal intervention strategy to perform counterfactual reasoning. By jointly constructing the global and local descriptors, we propose a GLANet to extract the efficient representation of images, which is capable of solving the challenging fine-grained image retrieval task. To further improve performance, we utilize an object detector to detect the foreground area, which enables the network to obtain an accurate representation of the object of interest. Finally, we also implement a series of effective strategies to enhance the retrieval capability of the proposed model. Extensive experiments on the benchmark demonstrate the effectiveness of the proposed method.

References

1. Brown, A., Xie, W., Kalogeiton, V., Zisserman, A.: Smooth-AP: smoothing the path towards large-scale image retrieval. arXiv preprint arXiv:2007.12163 (2020)

2. Cao, B., Araujo, A., Sim, J.: Unifying deep local and global features for image search. arXiv (2020)
3. Chen, B., Deng, W.: Hybrid-attention based decoupled metric learning for zero-shot image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2750–2759 (2019)
4. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9268–9277 (2019)
5. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: learning global representations for image search. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 241–257. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_15
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
7. Jang, Y.K., Cho, N.I.: Generalized product quantization network for semi-supervised image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3420–3429 (2020)
8. Lang, Y., He, Y., Yang, F., Dong, J., Xue, H.: Which is plagiarism: fashion image retrieval based on regional representation for design protection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2595–2604 (2020)
9. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: SphereFace: deep hypersphere embedding for face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 212–220 (2017)
10. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1096–1104 (2016)
11. Mahajan, D., et al.: Exploring the limits of weakly supervised pretraining. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 185–201. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_12
12. Ng, T., Balntas, V., Tian, Y., Mikolajczyk, K.: Solar: second-order loss and attention for image retrieval. arXiv preprint arXiv:2001.08972 (2020)
13. Pearl, J.: Causal diagrams for empirical research. Biometrika **82**(4), 669–688 (1995)
14. Pearl, J.: Direct and indirect effects. arXiv preprint arXiv:1301.2300 (2013)
15. Pearl, J., Glymour, M., Jewell, N.P.: Causal Inference in Statistics: A Primer. Wiley, Hoboken (2016)
16. Radenović, F., Tolias, G., Chum, O.: Fine-tuning CNN image retrieval with no human annotation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1655–1668 (2018)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
18. Revaud, J., Almazán, J., Rezende, R.S., de Souza, C.R.: Learning with average precision: training image retrieval with a listwise loss. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5107–5116 (2019)
19. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1979–1988 (2019)

20. Tan, M., Le, Q.V.: EfficientNet: rethinking model scaling for convolutional neural networks. arXiv preprint arXiv:1905.11946 (2019)
21. Tang, K., Huang, J., Zhang, H.: Long-tailed classification by keeping the good and removing the bad momentum causal effect. In: NeurIPS (2020)
22. VanderWeele, T.J.: A three-way decomposition of a total effect into direct, indirect, and interactive effects. Epidemiology (Cambridge, Mass.) **24**(2), 224 (2013)
23. Wang, W., Xu, Y., Shen, J., Zhu, S.C.: Attentive fashion grammar network for fashion landmark detection and clothing category classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4271–4280 (2018)
24. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)