Image Search Engine by Deep Neural Networks

Yao, Y.; Zhang, Q.; HU, Y.; Meo, C.; Wang, Y.; Nanetti, Andrea ; Dauwels, J.H.G.

Publication date
2022

Document Version
Final published version

Published in
42nd WIC Symposium on Information Theory and Signal Processing in the Benelux (SITB 2022)

Image Search Engine by Deep Neural Networks

Yuanyuan Yao[1*], Qi Zhang[1*], Yanan Hu[1*], Cristian Meo[1], Yanbo Wang[1], Andrea Nanetti[2], Justin Dauwels[1*]

***Abstract*—We typically search for images by keywords, e.g., when looking for images of apples, we would enter the word "apple" as query. However, there are limitations. For example, if users input keywords in a specific language, then they may miss results labeled in other languages. Moreover, users may have an image of the object they want to obtain more information about, e.g., a landmark, but they may not know the name of it. In such scenario, word-based search is not adequate, while image-based search would be ideally suited. These needs drive us to develop a purely content-based image search engine, meaning that users can search images with an image as query. Motivated by this use case with numerous applications, in this paper we propose and validate an image query based search engine. The image processing pipeline contains the following modules (Fig. 1): feature extraction, approximate nearest-neighbour search and re-ranking.**

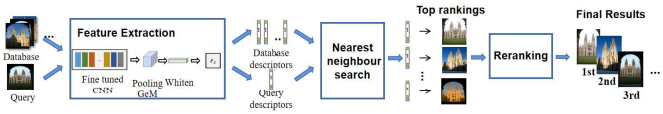

Fig. 1. Proposed pipeline. First, features are extracted from the query image. Similarly, such features are also extracted from all images in the database. These features are extracted only once, and are then stored for processing the image-based queries. The feature extractor comprises a CNN, followed by an aggregation layer, a whitening layer and an L2 normalization layer, which can be trained end-to-end with various loss functions. The extracted features are then forwarded to an approximate nearest-neighbour search module to produce the initial ranking results, which are next refined by a re-ranking module, generating the final results.

Feature extraction is the cornerstone of an image retrieval system. In this paper, ResNet101 is selected as the backbone of the extractor. In order to obtain more discriminative and compact representations of the target image, GeM Pooling [1] is employed as aggregation methods. After processing by the GeM pooling layer, the feature maps are reduced to one global descriptor. Besides the extraction procedure design, the optimisation of the loss function also needs to be considered. Most work in image retrieval considers pairwise (e.g., contrastive) or tuplewise (e.g., triplet-based, n-tuple-based) loss functions. Both methods train the model by learning the corresponding distances between the positive and negative and the target. However, these methods show limitations in the cluster distribution when processing hard positive images. In this paper, we applied the second-order similarity loss [2] combined with above loss functions to optimize clusters and explore better model learning mechanisms.

Once we obtain the feature vectors, the image retrieval problem becomes a nearest neighbour search problem: finding the relevant images is then finding the database vectors that are close to the query vector. A trivial way is linear scan, which has linear search complexity and may lead to unacceptable time costs when the database is very large. Therefore, approximate nearest-neighbour (ANN) search methods are proposed to achieve sub-linear complexity, which mainly fall into two categories: compression-based and graph/tree-based. Compression-based methods aim to encode the vector into a much more compact representation, and graph/tree-based methods enable us to calculate and compare distances between only a small portion of the database vectors and the query vector. However, these two types of approaches are usually not discussed and compared together. Also, the possibility of combining them together has not been fully studied. In this paper, we apply methods in both categories, e.g., product quantization [3] and hierarchical navigable small world [4], and explore how to get the best of both worlds.

After the ANN search, the engine outputs the preliminary results of top-K images. However, these results are not robust when there are illumination and viewpoint changes in images. Therefore, we need to implement reranking. The reranking can be divided into two types: global and local feature based reranking. The global feature represents an image with a single feature vector. The global feature based reranking uses the global features from the previous feature extraction to obtain more representative features and implements reranking. By contrast, the local feature represents an image with a multidimensional feature matrix. The local feature based reranking extracts local features of images, calculates the geometric similarity and implements reranking. The global feature based reranking is faster, while the local feature based reranking can provide pixel-to-pixel similarity analysis. However, previous researchers spend little attention to reranking. Therefore, we propose and apply accurate and efficient global and local feature based reranking (Diffusion and SIFT) [5], [6] at a small speed cost.

The most important contribution of this paper is to provide a fine-grained instance-level search engine that can be applied in real-world applications. The system is highly modular and therefore flexible, allowing for easier adaptation to requirements, striking a balance between speed and accuracy. Extensive tests on various datasets have shown that our pipeline achieves state-of-the-art results across the public benchmarks with acceptable time costs.

*These authors contributed equally to the work
[1]Department of Microelectronics, Delft University of Technology
[2]School of Art, Design and Media, Nanyang Technological University
*Email: J.H.G.Dauwels@tudelft.nl

REFERENCES

[1] Radenović, Filip and Tolias, Giorgos and Chum, Ondřej, "Fine-tuning CNN image retrieval with no human annotation," IEEE transactions on pattern analysis and machine intelligence, vol. 41, 2018, pp.1655–1668.
[2] Ng, Tony and Balntas, Vassileios and Tian, Yurun and Mikolajczyk, Krystian, "SOLAR: second-order loss and attention for image retrieval," European Conference on Computer Vision, 2020, pp.253–270.
[3] Jegou, Herve and Douze, Matthijs and Schmid, Cordelia, "Product quantization for nearest neighbor search," IEEE transactions on pattern analysis and machine intelligence, vol. 33, 2010, pp.117–128.
[4] Malkov, Yu A and Yashunin, Dmitry A, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," IEEE transactions on pattern analysis and machine intelligence, vol. 42, 2018, pp.824–836.
[5] F. Yang, R. Hinami, Y. Matsui, S. Ly, and S. Satoh, "Efficient image retrieval via decoupling diffusion into online and offline processing," 2018.
[6] D. Lowe, "Object recognition from local scale-invariant features," in Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 vol.2, 1999.