Large-scale Image Retrieval using Neural Net Descriptors

David Novak
Masaryk University
Brno, Czech Republic
david.novak@fi.muni.cz

Michal Batko
Masaryk University
Brno, Czech Republic
batko@fi.muni.cz

Pavel Zezula
Masaryk University
Brno, Czech Republic
zezula@fi.muni.cz

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

metric indexing; deep convolutional neural network; content-based image retrieval; k-NN search

1. ONLINE IMAGE RETRIEVAL SYSTEM

One of current big challenges in computer science is development of data management and retrieval techniques that would keep pace with the evolution of contemporary data and with the growing expectations on data processing. Various digital images became a common part of both public and enterprise data collections and there is a natural requirement that the retrieval should consider more the actual visual content of the image data. In our demonstration, we aim at the task of retrieving images that are visually and semantically similar to a given example image; the system should be able to online evaluate k nearest neighbor queries within a collection containing tens of millions of images. The applicability of such a system would be, for instance, on stock photography sites, in e-shops searching in product photos, or in collections from a constrained Web image search.

A successful content-based image retrieval system must stand on two pillars: *effective* image-processing technique to achieve high-quality retrieval, and *efficient* search techniques to make the system work in real time and on a large scale. Our demonstration uses the cutting edge approach of deep convolutional neural networks [4] to obtain powerful visual features that carry a certain semantic footprint of the image content. These descriptors compared by a distance function seem to very well correspond to the human perception of general visual similarity. Our main contribution is the search engine that can organize large volumes of these complex descriptors so that the similarity queries can be evaluated efficiently. The engine exploits our recent distance-based index PPP-Codes [5] which organizes 320 GB of neural network descriptors extracted from a collection of 20 million images; given a query image, the memory part of the index is able to identify a relatively small candidate set of descriptors that is loaded from the disk and refined to obtain the final result of the similarity query.

SIGIR'15, August 09-13, 2015, Santiago, Chile.
ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767868.

Figure 1: Examples of visual queries in the demo.[1]

One of the main objectives of our demonstration is to share the unique user experience of online search powered by neural networks on a large collection of high-quality images. Figure 1 shows examples from the demo, which is available at `http://disa.fi.muni.cz/demos/profiset-decaf/`.

2. SYSTEM ARCHITECTURE

Let us describe in detail the components and overall architecture of the demonstrated system.

2.1 Visual Search with Deep Neural Networks

Recently, the successful application of deep convolutional neural networks revolutionized the area of image and video recognition. Our system exploits the breakthrough image classifier by Krizhevsky et al. [4] that gained significant attention by winning the 2012 ImageNet challenge, defeating other approaches by a significant margin. This neural network was trained on about 1.2M images classified into 1000 categories. Moreover, it was soon observed that intermediate outputs of hidden layers of the network can be used as *features* for assessment of general mutual similarity of various types of images – even without any need of retraining the neural network [2, 4].

[1]The copyright of the images belongs to their authors; they are used for research purposes according to *Profiset usage agreement* at `http://disa.fi.muni.cz/profiset/`

Specifically, we use the DeCAF$_7$ feature produced by the last hidden layer of the neural network model provided by Caffe[2] [3], which has been trained by the procedure described in the original paper [4]. We have extracted these features from a collection of 20 million images (see Section 2.3 for details). The extraction process consists of image normalization to 256×256 pixels, generation of ten overlapping patches of 224×224, and a forward feed of these patches through the network [4]. The resulting DeCAF$_7$ feature of an image is the average of the last-but-one layer outputs for the ten image patches. Such extraction takes about 100 ms on our GPU (30 days for the 20M collection). One DeCAF$_7$ feature is a 4096-dimensional vector taking 16 KB on the disk; the whole 20M dataset has thus some 320 GB of uncompressed data. Given a query image, the same extraction procedure generates feature $q \in$ DeCAF$_7$, where DeCAF$_7$ denotes the 4096-dimensional feature space; this is schematically depicted in the top part of Figure 2.

2.2 Distance Indexing for Similarity Search

In our system, we adopt a broad similarity model based on mutual object distances, specifically, $\mathbb{D}$ is a *domain of data* and δ is a total *distance* function $\delta : \mathbb{D} \times \mathbb{D} \longrightarrow \mathbb{R}_0^+$; we assume that this space $(\mathbb{D}, \delta)$ satisfies metric postulates of *identity*, *symmetry* and *triangle inequality* [6]. In our specific case, domain $\mathbb{D}$ is the 4096-dimensional DeCAF$_7$ space and δ is Euclidean distance, as proposed in the image recognition papers [4, 2]. The actual set of 20M features is denoted as $\mathbb{X} \subseteq \mathbb{D}$; the search is modeled by the *nearest neighbors query* k-NN(q) returning the k objects from $\mathbb{X}$ with the smallest distances to given $q \in \mathbb{D}$: $\delta(q, x)$, $x \in \mathbb{X}$.

Our PPP-Codes index uses multiple *recursive Voronoi partitioning* of the metric space; in this way, each object $x \in \mathbb{X}$ is mapped onto its $PPP(x)$ code, which carries information about location of x in these Voronoi diagrams [5]. A memory index is then created on $PPP(x)$ codes, $\forall x \in \mathbb{X}$; each object x is identified by its ID. Given a query k-NN(q), $q \in \mathbb{D}$, this index determines a set $C(q) \subseteq \mathbb{X}$ of candidate objects, or rather object IDs. Objects from $C(q)$ are read from an ID-object disk store and are refined by evaluation of $\delta(q, x)$, $\forall x \in C(q)$ in order to obtain final answer for k-NN(q); this process is sketched in the lower part of Figure 2.

The result of this search process is an approximation of the precise k-NN(q) answer. The precision of the answer depends on the size of the candidate set and so does the response time, because the search costs mainly consist of (1) CPU costs of $C(q)$ generation, (2) I/O costs of reading $C(q)$ objects from the disk and (3) CPU costs of $C(q)$ refinement. Because we work with complex and bulky features, the I/O costs are important, even though we compress the features down to 1/3 of their original size. The PPP-Codes index has about 1 GB in memory and it is designed to provide a very accurate candidate set; for instance, refining 5000 candidate objects results in 80% average recall for 10-NN queries with average response time around 400 ms.

2.3 Dataset and Demo Description

Our demonstration uses a unique collection Profiset[3] consisting of 20 million high quality images provided for research purposes by a stock photography company [1]. The search engine is running on a commodity machine with 8 GB of memory, 8 CPU cores and an SSD disk with the ID-object store maintaining the set of the DeCAF$_7$ features. The actual images to be displayed are stored separately.

[2] `http://caffe.berkeleyvision.org`
[3] `http://disa.fi.muni.cz/profiset/`

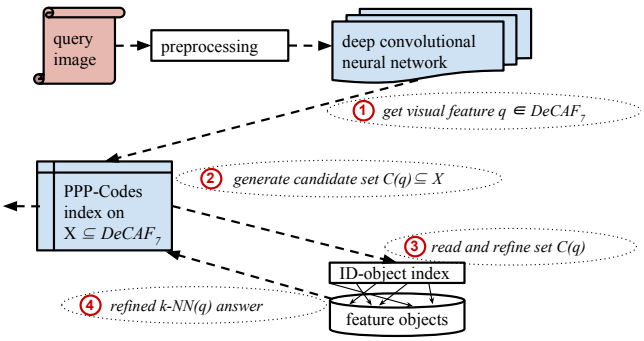

Figure 2: Schema of the demonstration system.

The application provides several actions. The default one displays a random selection of images, which demonstrates the diversity of the collection. Each of the images can serve as a query image for the k-NN visual search as described above; in order to initiate such query, the DeCAF$_7$ feature is retrieved from the ID-object store. Also, an external image can be uploaded and serve as the query; in this case, the DeCAF$_7$ feature is first extracted by the neural network. The Profiset collection contains also keywords for each image and our application provides also text search realized via Lucene index. These keywords are not combined anyhow with the DeCAF features in order to demonstrate the pure strength of the deep convolutional neural networks.

Acknowledgments

This work was supported by the Czech Research Foundation project P103/12/G084.

3. REFERENCES

[1] P. Budikova, M. Batko, and P. Zezula. Evaluation Platform for Content-based Image Retrieval Systems. In *International Conference on Theory and Practice of Digital Libraries*, LNCS, pages 130–142. Springer, 2011.

[2] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proceedings of the International Conference on Machine Learning*, pages 647–655, 2014.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1106–1114, 2012.

[5] D. Novak and P. Zezula. Rank Aggregation of Candidate Sets for Efficient Similarity Search. In *Database and Expert Systems Applications: 25th International Conference, DEXA 2014. Proceedings, Part II*, vol. 8645 of *LNCS*, pages 42–58. Springer, 2014.

[6] P. Zezula, G. Amato, V. Dohnal, and M. Batko. *Similarity Search: The Metric Space Approach.* Springer, 2006.