Combining Fisher Vector and Convolutional Neural Networks for Image Retrieval

Giuseppe Amato, Fabrizio Falchi, Fausto Rabitti, and Lucia Vadicamo

ISTI-CNR, via G. Moruzzi 1, Pisa 56124, Italy
{firstname}.{lastname}@isti.cnr.it

Abstract. Fisher Vector (FV) and deep Convolutional Neural Network (CNN) are two popular approaches for extracting effective image representations. FV aggregates local information (e.g., SIFT) and have been state-of-the-art before the recent success of deep learning approaches. Recently, combination of FV and CNN has been investigated. However, only the aggregation of SIFT has been tested. In this work, we propose combining CNN and FV built upon binary local features, called BMM-FV. The results show that BMM-FV and CNN improve the latter retrieval performance with less computational effort with respect to the use of the traditional FV which relies on non-binary features.

1 Introduction

Convolutional neural networks (CNNs)[5] have attracted enormous interest within research community because of the state-of-the-art results achieved in several domains, like image classification, image retrieval, object recognition, and speech recognition, to cite some. Several works [8, 3, 2] have shown that the outputs of the intermediate layers of CNN can be effectively used as high-level image descriptors. These CNN features have topped the already high results achieved by other image descriptors, such as Fisher Vectors (FV)[7]. FV and CNN features capture different aspects of the image visual content and have different strengths. For example, CNN features achieve very high effectiveness but have limited level of rotation invariance. The FV, instead, is robust to rotations but seems to be more affected to small scale changes than CNN [2].

In order to leverage the positive aspects of both these methods, Chandrasekhar et al. [2] have proposed a fusion of FV and CNN features. The results shown that FV can help improving the already high effectiveness of CNN features. However, the cost of extracting SIFT can be considered too high with respect to the small increase in the quality of retrieval. ORB binary local features, which extraction is typically two order of magnitude faster than SIFT [9], have been used in computer vision whenever high efficiency is needed. In this work, we tested FV built upon ORB [9] binary local features in conjunction with CNN feature. Our results show that while FVs of binary local features are less effective than the ones obtained from SIFT, when used in combination with CNN feature their effectiveness is comparable. Thus, the proposed combination of CNN and BMM-FV results in a profitable solution for both efficiency and effectiveness.

2 Background on Image Representations

The state-of-the-art image representations are mainly based on the use of *local features*, which are mathematical representation of local structure of images. To date, the most used and cited local feature is the SIFT [6], which led to effectively find correct matches between images. However, SIFTs extraction is costly due to local image gradients computation. Recently, the cost for extracting, representing and comparing local visual descriptors has been dramatically reduced by the introduction of *binary local features*, such as ORB[9].

Even if binary local descriptors are more compact and faster than non-binary ones, each image is still represented by thousands of local descriptors making it difficult to scale up the search to large digital archives. Encoding techniques, such as *Fisher Vectors* (FVs) [7] allow to obtain a more compact image representation. The FV approch transforms an incoming set of local descriptors into a fixed-size vector representation, that describe how the sample of the descriptors deviates from a "probabilistic visual vocabulary" which usually is modeled by a Gaussian Mixture Model (GMM). In this work we want to encode local binary descriptors and so we used a Bernoulli Mixture Model (BMM) that describes binary outcomes better than GMM. The resulting image representation is referred to as *BMM-FV*.

Recently, a new class of image descriptor built upon *Deep Convolutional Neural Networks* (CNNs)[5] have been used as effective alternative to descriptors built upon local features. In particular, it has been proven that activations produced by an image within the intermediate layers of the CNN can be used as a high-level descriptor. To improve retrieval results and balance the lack of geometrical invariance of CNN, in [2] a fusion of FV and CNN features have been proposed.

3 Experiments

In the follow we evaluate the performance of the combination of BMM-FV and CNN features for image retrieval tasks.

Experimental Setup. The evaluation was performed on the public *INRIA Holidays* dataset [4] which is a benchmark for image retrieval. It contains $1,491$ images, 500 of them being used as queries. All the learning stages were performed off-line using the independent *Flickr60k* dataset [4]. The retrieval performance was measured by the mean average precision (mAP) with the query image removed from the ranking list.

The ORB descriptors were extracted by using OpenCV[1]. The BMM-FVs were computed on ORB binary features by using our Visual Information Retrieval library, that is publicly available on GitHub[2].

[1] `http://opencv.org/`
[2] `https://github.com/ffalchi/it.cnr.isti.vir`

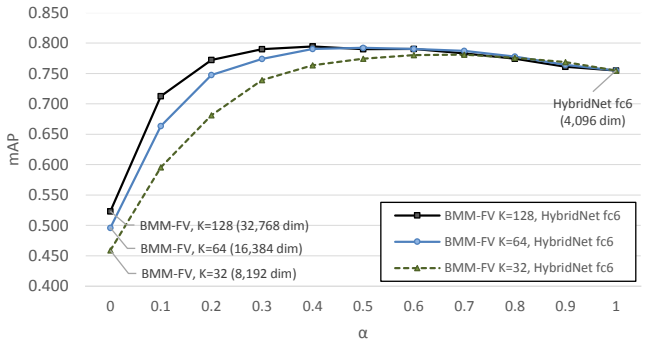

Fig. 1. Retrieval performance of the combinations of BMM-FV and HybridNet *fc6* for various number K of Bernoulli mixtures. The BMM-FVs were computed using ORB binary features. α is the parameter used in the combination: $\alpha = 0$ corresponds to use only FV, while $\alpha = 1$ correspond to use only the HybridNet feature.

The CNN features were computed using the pre-trained HybridNet [1], public available on the Caffe Model Zoo[3]. We used Caffe to extract the output of the first fully-connected layer (*fc6*). The resulting $4,096$ dimensional descriptors were L_2 normalized.

Combination of FV and CNN Features. We represented each image by a couple (c, f), where c and f were respectively the CNN descriptor and the BMM-FV of the image. Then, we evaluated the distance d between two couples (c_1, f_1) and (c_2, f_2) as the convex combination between the L_2 distances of the CNN descriptors and the BMM-FV descriptors, i.e.

$$d\big((c_1, f_1), (c_2, f_2)\big) = \alpha \, \|c_1 - c_2\|_2 + (1 - \alpha) \, \|f_1 - f_2\|_2 \tag{1}$$

with $0 \leq \alpha \leq 1$. Choosing $\alpha = 0$ corresponds to use only FV approach, while $\alpha = 1$ correspond to use only CNN features.

Results. In [2] it has been shown that combining the HybridNet *fc6* with FV representation led to obtain a relative improvement of **4.9%** mAP respect to the use of the CNN feature alone. Specifically, they used a FV computed on 64-dimensional PCA-reduced SIFTs, using $K = 256$ mixtures of Gaussians, which results in a $32,768$ dimensional vector.

In this work we propose to combine the CNN feature with the less expensive BMM-FV built on ORB binary features. In figure 1 we plot the mAP obtained for three different number K of Bernoulli mixtures used in the BMM-FV representation, namely $K = 32, 64, 128$. It is worth noting that all the three BMM-FVs led to improve the performance when combined with the HybridNet *fc6* and that exists an optimal α to be used in the convex combination (equation (1)). When using $K = 64$ the optimal α was obtained around 0.5, which correspond to give

[3] `https://github.com/BVLC/caffe/wiki/Model-Zoo`

the same importance to both FV and CNN feature. In this case the achieved mAP was **79.2%** which correspond to a relative improvement of **4.9%** respect to the use of the CNN feature alone (whose mAP was **75.5%**). The combination with BMM-FV for $K = 128$ achieves the best effectiveness (mAP of **79.5%**) for $\alpha = 0.4$. However, since the cost for computing and storing FV increase with the number K of Bernoulli, the improvement obtained using $K = 128$ respect to that of $K = 64$ doesn't worth the extra cost of using a bigger value of K. Moreover for $K = 64$ we obtain the same relative improvement of [2] using a less expensive FV representation that takes advantage from both the use of binary local features and smaller number K of mixtures.

4 Conclusion

The retrieval performance of CNN features is improved when using information provided by other image representations. In particular, the combination of CNN and FV features has been proved to be effective. However, state-of-the-art FV is generally computed using non-binary feature, as SIFT, which extraction is time-consuming. This paper shows that the more efficient BMM-FV, built upon ORB features, can be profitable use to this scope. In fact, the relative improvement in the retrieval performance obtained using the BMM-FV is similar to that obtained using the more expensive FV built upon SIFT.

References

1. Learning deep features for scene recognition using places database. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 27, pp. 487–495. Curran Associates, Inc. (2014)
2. Chandrasekhar, V., Lin, J., Morère, O., Goh, H., Veillard, A.: A practical guide to cnns and fisher vectors for image instance retrieval. CoRR abs/1508.02496 (2015)
3. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. CoRR abs/1310.1531 (2013)
4. Jégou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: European Conference on Computer Vision. LNCS, vol. I, pp. 304–317. Springer (oct 2008)
5. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (may 2015)
6. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
7. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. pp. 1–8 (June 2007)
8. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on. pp. 512–519. IEEE (2014)
9. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 2564–2571 (Nov 2011)