Fisher Encoded Convolutional Bag-of-Windows for Efficient Image Retrieval and Social Image Tagging

Tiberio Uricchio* Marco Bertini* Lorenzo Seidenari Alberto Del Bimbo
Università di Firenze - MICC
Firenze
name.surname@unifi.it

Abstract

In this paper we present an efficient and accurate method to aggregate a set of Deep Convolutional Neural Network (CNN) responses, extracted from a set of image windows. CNN features are usually computed on the whole frame or with a dense multi scale approach. There is evidence that using multiple windows yields a better image representation nonetheless it is still not clear how windows should be sampled and how CNN responses should be aggregated. Instead of sampling the image densely in scale and space we show that selecting a few hundred windows is enough to obtain an effective image signature. We show how to use Fisher Vectors and PCA to obtain a short and highly descriptive signature that can be used effectively for image retrieval. We test our method on two relevant computer vision tasks: image retrieval and image tagging. We report state-of-the art results for both tasks on three standard datasets.

1. Introduction

In this paper we address the problem of efficient multimedia retrieval and automatic image annotation in the context of social media. In the first task we aim at obtaining a very compact and discriminative signature, that allows the creation of scalable image retrieval systems. The goal of the second task is to predict, for a given image, a finite set of tags from a given vocabulary, serving as a compact description of the image. A popular group of recent image annotation methods apply tag propagation using diversely defined image neighborhoods [4, 12, 19–21, 30]. These approaches have been successfully applied to the context of social and user generated media, that are typically annotated with tags that are likely to correlate with image content. However, this rich source of metadata is often hard to exploit both for the noise in labels and for the difficulty to find semantically meaningful visual features. Clearly a good image representation boosts the precision and recall of these techniques by providing a visually consistent neighborhood. In fact, many of these techniques apply a form of metric learning to make up for low quality image features. We point out that an essential requirement of these techniques is the ability to retrieve similar images to compose good image neighborhoods. Hence, excelling in image retrieval is likely to improve image tagging. A recent breakthrough in image representation has been achieved using Convolutional Neural Networks (CNN) with deep architectures. It has been shown that using a large corpus of images CNNs can learn compact and powerful image features. CNNs are typically applied to classification tasks and activations from the latest layers are used as features. These have been used by several approaches to extract generic features for image retrieval [11, 32]. While they show promising results, they leave several questions unaddressed. First, CNNs features are more semantically related to the global image and they hardly preserve local characteristics of objects. Second, existing approaches address CNNs invariance issues with extracting patches densely at multiple scales usually leading to a very onerous feature computation process.

Recently Wei *et al*. [31] have applied a multi-label variation of CNN extracting features from few hundred object proposals. We agree with their intuition and we believe that multiple image windows can be carefully selected in order to obtain a more comprehensive representation of image content. This is particularly relevant in the case of image tagging where more than one tag is sought. User tags may refer to the image as a whole but they are also likely to be associated with specific scene elements. Specifically, tags often refer to *things* (e.g. person, car, horse, etc.) and *stuff* (e.g. sky, sand, cloud, water, etc.) present in a scene.

In this paper we show a technique, derived from Fisher Vectors [26], to combine CNN features from multiple windows into a more discriminative representation for image retrieval and image tagging. Our representation improves upon the single global representation approach, obtaining state-of-the-art results with compact image signatures on

*Equal contribution and corresponding authors.

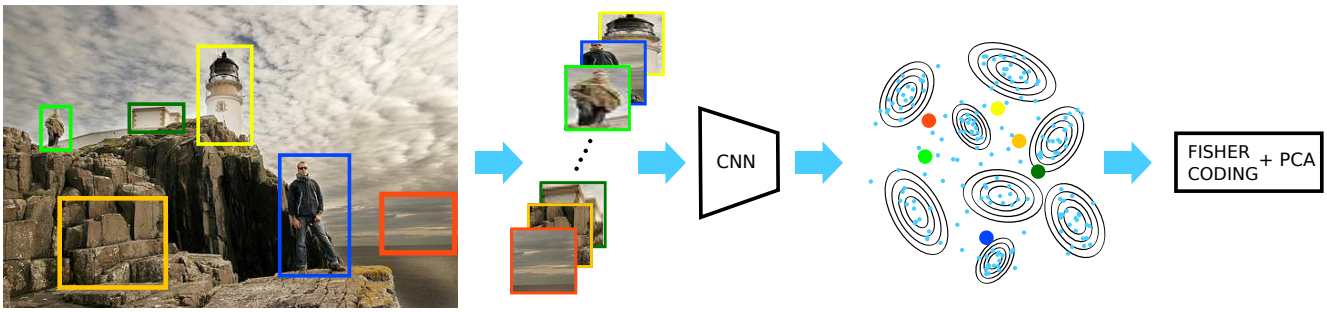

Figure 1. Full pipeline of the proposed method. Each image window is represented by the FC7 CNN activations. The final signature is obtained encoding activations (same color dots) with a Fisher Vector computed on a GMM dictionary (blue dots). PCA is further applied to image signature.

three popular public datasets.

2. Previous work

So far, the best performance in image retrieval has been obtained aggregating SIFT descriptors using Fisher Vectors [17, 26], VLAD [1, 17], or variations of these approaches e.g. pooling oriented local features [33]. A breakthrough in performance for computer vision algorithms has recently been obtained thanks to supervised image feature learning. Krizhevsky *et al*. revived supervised deep learning for computer vision proposing to solve large scale image classification problem using a deep CNN [18]. Following that, several architectures have been proposed in the last 3 years, all sharing a common principle: networks are usually built with a sequence of convolutional/max-pooling layers, followed by low-resolution fully-connected (FC) layers whose activations are fed to a soft-max classifier.

One interesting fact about CNNs is the ability to perform transfer learning. Indeed a very powerful image representation can be obtained by removing the soft-max classifier and keeping the activations of the last FC layer. This approach has been applied to many computer vision and multimedia retrieval tasks, with dramatic improvements over previously proposed techniques such as Fisher Vectors over local SIFT descriptors. Razavian *et al*. [24] made a comprehensive contribution on this matter testing CNN features for object and scene classification, attribute prediction and image retrieval. However, their spatial search approach in image retrieval has an unbearable computational cost: their method requires the extraction of CNN features for a large amount of image sub-windows and the computation of all pairwise distances between them. The approach has scalability issues, since it is quadratic in the number of windows.

Approaches close to ours have been proposed in [11, 22, 32]. Gong *et al*. [11], propose to aggregate CNN responses from multiple scales using VLAD, thus requiring a dense computation of multi-scale CNN responses. In contrast, we show how we can rely on the computation of CNN responses on a few hundreds of proposal windows. Ng *et al*. [22] have speeded up the approach of [11] applying the network only once to the input image and extracting features at each location of the convolutional feature map of each layer. Yoo *et al*. [32] propose to apply Fisher Vector encoding to dense multi-scale CNN activations. Compared to these methods our approach computes CNN activations on large parts of the image, which are likely to contain objects, rather than considering CNN activations of dense and small patches, that are more similar in spirit to SIFT descriptors. Another difference is that we introduce a simpler and effective multi-scale representation by concatenating the Fisher Vector with a global representation of image content, and reducing the overall descriptor size with PCA.

The identification of relevant patches in an image has been recently addressed in the object detection community, with the introduction of window proposal methods [9, 28]. Object proposals are cheap to compute and cover more than 90% of objects with few thousands windows of different scales and aspect ratios. This allows the application of expensive classifiers like [9] or kernelized bag-of-words classifiers [28] to perform object detection.

Regarding the task of social image tagging, our work is related to instance based tag assignment methods [20]. Makadia *et al*. [21], in their seminal work, showed that simple tag voting on nearest neighbor outperformed previous complex approaches. Li *et al*. [19] improved upon by adding a penalty on frequent tag votes. As low-level features are hardly semantically related, Guillaumin *et al*. [12] and Verma *et al*. [30] proposed to learn a weighted metric to improve on precision. Ballan *et al*. [4] proposed using KCCA to learn mid-level features to be used with previous nearest neighbors approaches.

3. Proposed method

Our idea is to represent an image as a bag of windows, each one represented as CNN output activations. The final image signature is obtained using Fisher Encoding and reducing the final descriptor dimensionality using PCA, as shown in Figure 1. This powerful novel image signature

is used to boost performance in image retrieval and social image tagging.

3.1. Image representation

Patch Sampling We start by sampling a set of few hundred windows from each image to construct a bag-of-windows $\mathcal{X}$ as image representation. To perform the sampling we propose a content-based strategy and a random strategy.

Regarding the content-based strategy, we use the object proposal approach, namely EdgeBoxes, from Zitnick *et al*. [34] due to its computational efficiency and performance in terms of detection, recall and repeatability [13]. This method provides a ranked list of windows that typically contain instances of objects, disregarding areas with few edges. The second strategy is a simple random strategy where window coordinates are generated randomly.

We also consider the combination of the two strategies. This is motivated by the fact that some discriminative portions of images, often useful for retrieval, are not part of objects or *things* but rather are referred as *stuff*, i.e. part of larger textured regions like trees or mountains. In fact, we found in some experiments that employing a set of randomly sampled windows in addition to the EdgeBoxes may be beneficial.

CNN usually require, as it is in our case, a fixed size input patch. To this end we resize each window to 224×224 pixels disregarding the aspect ratio, as it is common practice in object detection [9]. We use the pre-trained CNN-S-128 CNN architecture from [5] in order to have a low dimensional representation (128D), comparable to that of SIFT. For each window, we extract the activation from the first fully connected layer (FC7).

Activation Aggregation To obtain a short signature for each image we perform an aggregation step. Given a set of patches $x \in \mathcal{X}$, we encode it using Fisher Encoding.

We first learn a Mixture of Gaussians codebook with diagonal covariances on a subset of the windows extracted at the previous step. Differently from [26] we do not apply PCA on the local window features. This is not needed, and actually slightly worsen the performance in our case, since our window representation has highly decorrelated features.

In Fig. 2 we show a comparison of the absolute values of correlation coefficients ρ among dimensions of CNN codes and SIFT descriptors extracted from the INRIA Holidays dataset. The ρ coefficients of the CNN codes are 1 only on the diagonal, while as a counter-example on SIFT descriptors extracted from the same dataset there are many directions with $|\rho| > .8$.

For each bag-of-windows we compute an Improved Fisher Vector (IFV) applying L2 and Power Normalization as in [26]. Finally, to compress the representation, we reduce the dimensionality of the IFVs using PCA.

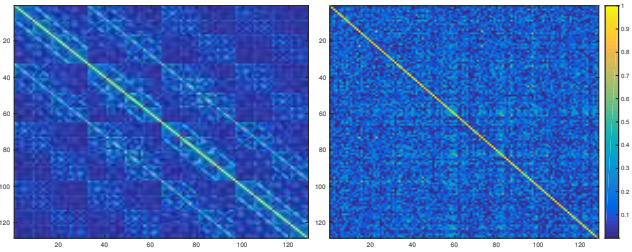

Figure 2. Correlation coefficients computed on a set of SIFT descriptors *(left)* and on a set of CNN features on image windows *(right)*.

Global-Local signature The PCA-compressed IFV signature provides a compact representation of the local sampled windows of an image. However, windows are aggregated independently, without considering their relation to the context. It can be further enhanced by integrating an explicit global scene representation. We propose a Global-Local (GL) signature made by the concatenation of the FC7 feature of the entire image with the generated PCA-compressed IFV signature. The FC7 feature has been proved to be very powerful [24] as we can also observe from our baseline experiments in Tab. 1, 2 and 3.

3.2. Image retrieval

The first task we address with our novel image representation is image retrieval. To retrieve images means that given an image as query we want to rank a dataset of images in order to assign high ranks to images with the same content of the query. We perform this task in a very straightforward manner. Given a query image I and a dataset of images Y_i, we consider their respective sets of image window features $\mathcal{I}$ and $\mathcal{Y}_i$ and signatures $\phi(\mathcal{I})$ and $\phi(\mathcal{Y}_i)$. For each query I we rank images by cosine distances:

$$d(I, Y_i) = 1 - \frac{\phi(I)^T \cdot \phi(\mathcal{Y}_i)}{\|\phi(I)\| \|\phi(\mathcal{Y}_i)\|}$$

3.3. Social image tagging

In this task we aim at annotating social images, using other social images as training data. A collection of social images, e.g. obtained from Flickr, can be modeled as a set of tuples $\mathcal{T}_i = \langle Y, \mathcal{W} \rangle$ where Y is an image, $\mathcal{W}$ is a set of tags provided by the users and the vocabulary $\mathcal{V}$ is the set of all the tags of $\mathcal{W}$. These tags are typically ambiguous, imprecise, and tend to follow user preferences [27]. This is a different setup from that of using images from datasets annotated by experts.

When performing image annotation we would like to predict tags for an untagged image I. This problem is usu-

ally solved with voting algorithms based on nearest neighbor search [3, 12, 19], because of their scalability and relatively good performance [20]. We use the ranking described in Sect. 3.2 to obtain the first K neighbors, and use the following three different algorithms.

NN voting The simplest voting algorithm is nearest neighbor tag voting, which is close to the method first proposed by Makadia *et al*. [21]. We count the tag occurrences of images in the neighborhood and rank tags per image using their frequencies.

Tag Relevance With NN voting we assume that the more frequently the tag occurs in the neighbor set, the more relevant it might be for the image. However tags occurring frequently in the whole training set are not necessary relevant for all the images. So to moderate this effect, Li *et al*. [19] proposed a tag relevance measure that takes into account both the tags distributions of the neighbor set and of the entire training set.

TagProp Guillaumin *et al*. [12] have proposed TagProp, a method that learns a weighted nearest neighbor model. Weights can be learned based on distance or rank. Moreover, to compensate for varying frequencies of tags, a tag-specific sigmoid is used to boost the probability for rare tags and decrease that of frequent ones. Sigmoids and metric parameters can be learned by maximizing the log-likelihood of tag predictions.

4. Experiments

Datasets For the image retrieval task we use the popular INRIA Holidays dataset [16]. The dataset is composed by 1,491 images in total. We measure average precision (AP) for 500 queries and 991 corresponding relevant images.

We test image tagging on the MIRFLICKR-25K and NUS-WIDE datasets. The MIRFLICKR-25K dataset [15] is composed of 25,000 images from Flickr with 1,386 user tags that occur in at least 20 images, and is split in 12,500 for training and 12,500 for testing, with exactly the same partition as [4, 12]. In addition ground truth annotations for 18 tags are provided on the whole set. The NUS-WIDE dataset [7] is composed of 269,648 images from Flickr with 355,913 user tags, and is split in training and testing sets of 161,789 and 107,859 images, respectively. Ground truth is available for 81 tags. Since there is no common experimental setup for NUS-WIDE, we have adopted the same setup of [10], i.e. following the train/test splits of the dataset, ignoring the small subset of images that are not annotated by any tag and using only the ground-truth tags. The resultant train and test sets have a respective total of 125,449 and 83,898 images. Since it is feasible to evaluate tagging performance only on ground truth tags, the experiments are performed with the user tags provided in the ground truth annotations, as in [29].

Baselines The natural baseline for our method is the extraction of a single CNN code per image. We refer to this baseline as CNN-Image. We warp the whole image to 224×224 and use the FC7 output as image signature. We develop two other baselines to see if the use of an aggregated signature is relevant to keep the expressiveness of the many windows extracted or if sampling multiple CNN responses is enough to boost retrieval and annotation performance. The first one is obtained by averaging the output of all the CNN features of the bag-of-windows, we refer to it as AVG-Pooling. The latter is computed with a max pooling operation over the CNN activations, which we denote as MAX-Pooling.

Experimental results: retrieval We first evaluate the parameters affecting retrieval performance on INRIA Holidays, evaluated in terms of mean average precision (MAP). In a set of preliminary experiments we found that the final PCA step slightly improves results but not significantly. This step is indeed mostly relevant to compress the image signature. The size of the GMM codebook is instead extremely relevant for performance.

Increasing the number of Gaussians allows to model the distribution of CNN activations more precisely, as it has been observed also for SIFT features [26], where increasing the number of Gaussians improves the performance. To see how the codebook size affects retrieval performance we fixed the final PCA dimension to 512 which we found improving performance across codebook sizes.

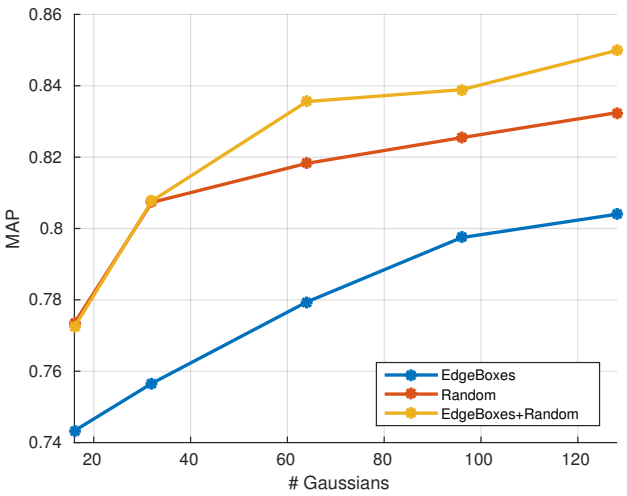

Figure 3. Mean average precision of our proposed approaches varying the number of Gaussians on Holidays dataset.

We sample the top 400 ranked EdgeBoxes and 400 Ran-

dom windows for a total of 800 windows in each image. Our method is efficient since it does not require to compute window correspondences exhaustively. Finally, we represent images with a very short 512D signature that scales in terms of space and time complexity.

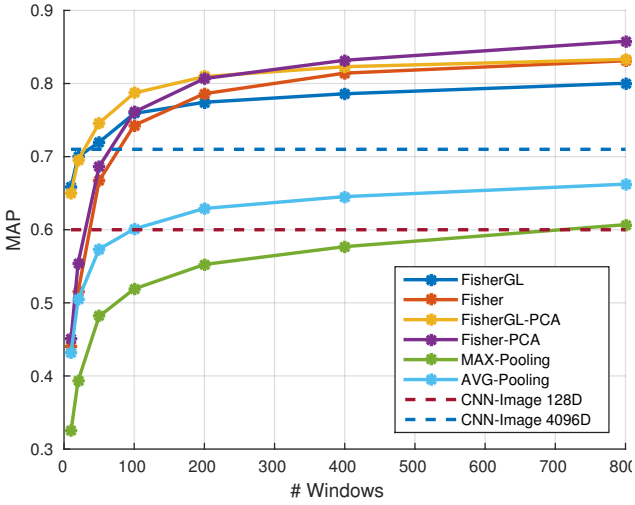

Figure 4. Mean average precision of our proposed approaches varying the number of EdgeBoxes + Random windows.

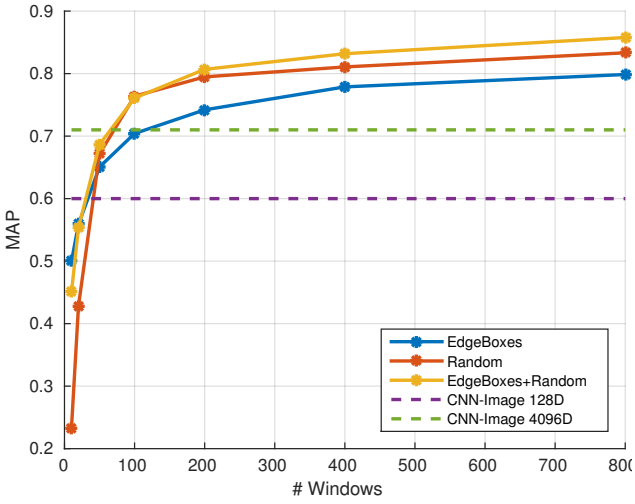

Figure 5. Mean average precision of our proposed approaches varying the number of windows, using Fisher-PCA coding.

In Figure 3 we evaluate the performance of the proposed approach with a varying number of Gaussians and with different window sampling strategy. We can see how using EdgeBoxes alone for retrieval is not sufficient. Adding random windows increases the performance also for a small amount of Gaussians (32). In Figure 4 we report MAP values obtained using different numbers of EdgeBoxes and random windows, with different encodings. The combination with the global signature (FisherGL and FisherGL-PCA) does not improve the MAP for large codebooks but instead allows to get very high results even for small codebooks. Fisher Vectors always outperform max and average pooling. In Figure 5 we evaluate the performance of Fisher Vector + PCA coding with varying number of windows, either from EdgeBoxes, random, or EdgeBoxes + random sampling. As for Fig.4 it can be observed that FV + PCA outperforms the use of single global CNN descriptors, when using more than 100 windows. Considering the random windows step we report the average of five runs.

Method	Features	Codebook	Dim.	MAP
Fisher-PCA	FC7-CNN	128	512	**85.8**
Fisher-GL-PCA	FC7-CNN	128	635	83.3
Fisher-GL	FC7-CNN	128	32,889	81.2
Fisher	FC7-CNN	128	32,768	80.3
AVG-Pooling	FC7-CNN	–	128	66.2
MAX-Pooling	FC7-CNN	–	128	60.1
CNN-Image 128D	FC7-CNN	–	128	60.0
CNN-Image 4096D	FC7-CNN	–	4,096	71.0
Spatial Pooling [25]	CONV-CNN	–	256	74.2
CNNaug-ss [24]	FC7-CNN	–	4,096	84.3
VLAD+PCA [22]	CONV-CNN	100	128	83.6
Neural codes [2]	FC7-CNN	–	128	78.9
VLAD+PCA [11]	FC7-CNN	100	2,048	80.2
VLAD+PCA [11]	FC7-CNN	100	512	74.2
Perronin [23]	FC7-SIFT+LCS	1,024	4,096	84.7
Fisher [26]	SIFT	4,096	524,288	70.0
Zhao [33]	SIFT	32	32,768	68.8
Delhumeau [8]	SIFT	64	8,192	65.8
Arandjelovic [1]	SIFT	256	32,536	65.3
Fisher [17]	SIFT	256	16,384	62.5
Fisher [17]	SIFT	64	4,096	59.5
VLAD [17]	SIFT	256	16,384	58.7
VLAD [17]	SIFT	64	4,096	55.6

Table 1. Image retrieval results on INRIA Holidays compared with state-of-the-art approaches.

Finally, we compare our method with other global methods aggregating local features in Table 1 and some recent methods that use either convolutional or fully connected layers of CNNs [2, 11, 22, 24, 25]. We can clearly see that although the 128D CNN is competitive with some smaller size representations based on SIFT features [17] the 4096D outperforms all the approaches based on engineered features. Average pooling of 128D activations outperforms the single image 128D representation indicating that more information is contained in multiple windows. Adoption of Improved Fisher Vector coding improves over the majority of the other methods based on CNN features except [22,24]. Finally we can see how applying the Fisher encoding and PCA outperforms all other methods, including [22,24], with a very small signature (512D).

Experimental results: tagging In this set of experiments we show how our novel representation improves performance on image tagging. We report results as Mean Average Precision (MAP) and Mean image Average Precision

(MiAP) in Tab. 2 and Tab. 3. MAP measures the quality of image ranking and can be affected by the performance on rare tags, while MiAP measures the quality of tag ranking and is biased toward frequent tags [20]. In each experiment we fix the number of nearest-neighbor K to 1,000 as suggested by the authors [19]. For TagProp we employ the best combination reported (distance + sigmoids) [29]. To speedup computations on these larger datasets we have used a GMM codebook of only 32 elements and halved the number of windows (200 EdgeBoxes and 200 Random for a total of 400) with respect to the experimental setup used for retrieval. We reduce the dimension of the final IVF to 512 dimensions as in the previous case using PCA.

The use of the Global-Local (GL) component of the descriptor, which accounts for scales variations providing an holistic representation of the image content, improves the results. This is reasonable because nearest neighbors approaches applied to social image datasets typically work better with descriptors that deal with the gist of the image (e.g. global descriptors or low-dimensionality BoF descriptors, as those provided by the authors of NUS-WIDE dataset [7]) rather than its details (e.g. performing spatial verification of matching local features). In this case the single image approach outperforms [29]. This means that CNN features are indeed a strong representation for image annotation. In this case average pooling is not improving over the single image approach. Finally we can see how adding the Global-Local part of the descriptor boosts MAP and MiAP for all voting methods; compressing the descriptor with PCA does not reduce the performance despite the high reduction in dimensionality. It has to be noted that TagProp always outperforms the simpler NN Voting and TagRel methods, exploiting better the improved visual neighborhood obtained with the proposed method. This is visible when comparing the performance obtained with the single CNN-Image descriptor w.r.t. that of Fisher-GL-PCA.

Method	NN Voting		TagRel		TagProp	
	MAP	MiAP	MAP	MiAP	MAP	MiAP
Fisher-GL-PCA	**51.4**	**48.6**	47.6	51.4	**58.0**	54.8
Fisher-GL	50.9	48.0	**48.4**	**51.5**	57.9	**54.9**
Fisher-PCA	46.1	44.9	43.7	48.2	51.6	50.9
Fisher	46.2	45.2	44.0	48.2	51.6	50.8
MAX-Pooling	40.7	45.6	41.5	47.1	47.6	49.2
AVG-Pooling	40.2	45.0	40.5	46.6	45.9	48.6
CNN-Image	48.3	46.6	46.0	50.1	55.7	53.7

Table 2. Image annotation results on MIRFLICKR-25K compared with the state-of-the-art (200 EdgeBoxes + 200 random windows).

Tab. 4 compares the best performance of the proposed method with the original TagProp method, on MIRFLICKR-25K, showing in particular a good improvement in terms of MiAP. Tab. 5 compares the best performance of the proposed method on NUS-WIDE with two other approaches that have a similar experimental setup, showing a very good performance.

Method	NN Voting		TagRel		TagProp	
	MAP	MiAP	MAP	MiAP	MAP	MiAP
Fisher-GL-PCA	26.7	**43.4**	**27.7**	**40.1**	**39.7**	**50.9**
Fisher-GL	**26.8**	**43.4**	27.6	**40.1**	39.7	50.8
Fisher-PCA	21.7	40.4	24.1	37.0	35.9	48.0
Fisher	21.3	40.3	23.6	36.6	35.5	47.4
MAX-Pooling	18.8	37.8	22.1	34.9	29.1	45.0
AVG-Pooling	19.9	40.2	22.4	37.1	29.8	45.9
CNN-Image	24.4	42.0	25.3	38.7	31.9	48.2

Table 3. Image annotation results on NUS-WIDE compared with the state-of-the-art (200 EdgeBoxes + 200 random windows).

Method	Features	MAP	MiAP
Fisher-GL-PCA + TagProp	FC7-CNN	**58.0**	**54.8**
Guillaumin [29]	local+global features[1]	38.4	47.3

Table 4. Image annotation results on MIRFLICKR-25K: comparison of the proposed method with other approaches.

Method	Features	MAP
Fisher-GL-PCA + TagProp	FC-7-CNN	**39.7**
Hash SISO [14]	NUS-WIDE [2]	25.5
LSMP [6]	NUS-WIDE	18.5

Table 5. Image annotation results on NUS-WIDE: comparison of the proposed method with other approaches.

5. Conclusion

In this paper we have shown the importance of extracting CNN activations from multiple windows. We investigated two different window sampling strategies and found out that the best performance is obtained by their combination. This confirm the intuition that image information is not fully captured by object proposals alone. In fact, adding randomly sampled windows improves our image representation.

We have shown that Fisher Vectors can be effectively used to aggregate low-dimensional CNN responses improving over more simplistic max and average pooling approaches. Finally applying PCA on the Fisher Vector representation allows to reduce the computational footprint of our method. Our method is computationally efficient since it relies on few hundred windows and has a low memory footprint representing each image with just 2.5Kb of data.

We tested our representation on two tasks, image retrieval and image tagging on three publicly available datasets collected from social networks showing state-of-the art results.

[1] GIST, colour histograms (RGB, LAB, HSV), SIFT + hue local descriptors BoW

[2] 225-D block- wise color moments, 128-D wavelet texture and 75-D edge direction histogram

Acknowledgments. This work is partially supported by "THE SOCIAL MUSEUM AND SMART TOURISM", MIUR project no. CTN01_00034_23154_SMST. Tiberio Uricchio is supported by Telecom Italia PhD grant funds (Italy).

References

[1] R. Arandjelovic and A. Zisserman. All about VLAD. In *Proc. of CVPR*, 2013. 2, 5

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *Proc. of ECCV*, 2014. 5

[3] L. Ballan, M. Bertini, T. Uricchio, and A. Del Bimbo. Data-driven approaches for social image and video tagging. *Multimedia Tools and Applications*, 74(4):1443–1468, 2014. 4

[4] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo. A cross-media model for automatic image annotation. In *Proc. of ACM ICMR*, 2014. 1, 2, 4

[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. of BMVC*, 2014. 3

[6] X. Chen, Y. Mu, S. Yan, and T.-S. Chua. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *Proc. of ACM MM*, 2010. 6

[7] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A real-world web image database from National University of Singapore. In *Proc. of ACM CIVR*, 2009. 4, 6

[8] J. Delhumeau, P.-H. Gosselin, H. Jegou, and P. Perez. Revisiting the VLAD image representation. In *Proc. of ACM MM*, 2013. 5

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of CVPR*, 2014. 2, 3

[10] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe. Deep convolutional ranking for multilabel image annotation. *CoRR*, abs/1312.4894, 2013. 4

[11] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *Proc. of ECCV*, 2014. 1, 2, 5

[12] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proc. of ICCV*, 2009. 1, 2, 4

[13] J. Hosang, R. Benenson, and B. Schiele. How good are detection proposals, really? In *Proc. of BMVC*, 2014. 3

[14] J. Huang, H. Liu, J. Shen, and S. Yan. Towards efficient sparse coding for scalable image annotation. In *Proc. of ACM MM*, 2013. 6

[15] M. Huiskes, B. Thomee, and M. Lew. New trends and ideas in visual concept detection: the MIR Flickr retrieval evaluation initiative. In *Proc. of ACM MIR*, 2010. 4

[16] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. of ECCV*, 2008. 4

[17] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012. 2, 5

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of NIPS*, 2012. 2

[19] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009. 1, 2, 4, 6

[20] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval. *arXiv preprint arXiv:1503.08248*, 2015. 1, 2, 4, 6

[21] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proc. of ECCV*, 2008. 1, 2, 4

[22] J. Y.-H. Ng, F. Yang, and L. S. Davis. Exploiting local features from deep networks for image retrieval. *arXiv preprint arXiv:1504.05133*, 2015. 2, 5

[23] F. Perronnin and D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *Proc. of CVPR*, 2015. 5

[24] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for visual recognition. In *Proc. of CVPR Workshop of DeepVision*, 2014. 2, 3, 5

[25] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson. Visual instance retrieval with deep convolutional networks. In *Proc. of ICLR Workshops*, 2015. 5

[26] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the Fisher vector: Theory and practice. *International Journal of Computer Vision*, 105(3):222–245, 2013. 1, 2, 3, 4, 5

[27] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proc. of WWW*, 2008. 3

[28] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013. 2

[29] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid. Image annotation with tagprop on the MIRFLICKR set. In *Proc. of ACM MIR*, 2010. 4, 6

[30] Y. Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *Proc. of ECCV*, 2012. 1, 2

[31] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. CNN: single-label to multi-label. *CoRR*, abs/1406.5726, 2014. 1

[32] D. Yoo, S. Park, J.-Y. Lee, and I. Kweon. Multi-scale pyramid pooling for deep convolutional representation. In *Proc. of CVPR Workshops*, pages 71–80, 2015. 1, 2

[33] W. Zhao, H. Jegou, and G. Gravier. Oriented pooling for dense and non-dense rotation-invariant features. In *Proc. of BMVC*, 2013. 2, 5

[34] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *Proc. of ECCV*, 2014. 3