

Object Level Deep Feature Pooling for Compact Image Representation

Konda Reddy Mopuri and R. Venkatesh Babu
Video Analytics Lab, SERC,
Indian Institute of Science, Bangalore, India.

sercmkreddy@ssl.serc.iisc.in, venky@serc.iisc.ernet.in

Abstract

Convolutional Neural Network (CNN) features have been successfully employed in recent works as an image descriptor for various vision tasks. But the inability of the deep CNN features to exhibit invariance to geometric transformations and object compositions poses a great challenge for image search. In this work, we demonstrate the effectiveness of the objectness prior over the deep CNN features of image regions for obtaining an invariant image representation. The proposed approach represents the image as a vector of pooled CNN features describing the underlying objects. This representation provides robustness to spatial layout of the objects in the scene and achieves invariance to general geometric transformations, such as translation, rotation and scaling. The proposed approach also leads to a compact representation of the scene, making each image occupy a smaller memory footprint. Experiments show that the proposed representation achieves state of the art retrieval results on a set of challenging benchmark image datasets, while maintaining a compact representation.

1. Introduction

Semantic image search or content based image retrieval is one of the well studied problems in computer vision. Variations in appearance, scale and orientation changes and change in the view point pose a major challenge for image representation. In addition, the huge volume of image data over the Internet adds to the constraint that the representation should also be compact for efficient retrieval.

Typically, an image is represented as a set of local features for various computer vision applications. Bag of Words (BOW) model [31] is a well studied approach for content based image retrieval (CBIR). The robustness of the SIFT like local features [22] [3] [23] to various geometric transformations and applicability of different distance measures for similarity computation have led to a widespread adoption of this framework.

The BOW model is inspired from the familiar Bag of

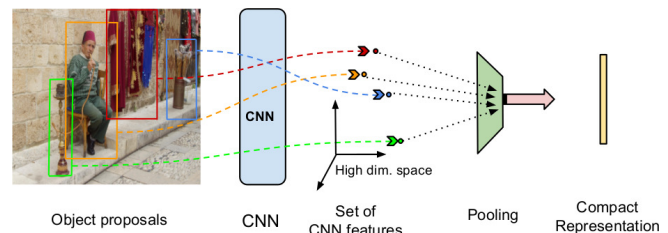


Figure 1. **The overview of the proposed system.** First stage obtains object proposals from the input image. Second stage extracts deep features from these regions. Finally, the last stage pools these features in order to obtain a compact representation.

Words approach for document retrieval. In this model, an image is represented as a histogram over a set of learned codewords after quantizing each of the image features to the nearest codeword. However, computing the BOW representation for an image is tedious and the computation time grows linearly with the size of the codebook. Although BOW model provides good accuracy for image search, it is not scalable for large image databases, as it is intensive both in memory and computations.

In order to have a more informative image representation, Peronin *et al.* proposed a generative model based approach called Fisher vector [27]. They estimate a parametric probability distribution over the feature space from a huge representative set of features, generally a gaussian mixture model (GMM). The features extracted from an image are assumed to be sampled independently from this probability distribution. Each feature in the sample is represented by the gradient of the probability distribution at that feature with respect to its parameters. Gradients corresponding to all the features with respect to a particular parameter are summed. The final image representation is the concatenation of the accumulated gradients. They achieve a fixed length vector from a varying set of features that can be used in various discriminative learning activities. This approach considers the 1st and 2nd order statistics of the local image features as well, whereas the BOW captures only the 0th order statistics, which is just their count.

Jégou *et al.* proposed Vector of Locally Aggregated Descriptors (VLAD) [16] as a non-probabilistic equivalent to Fisher vectors. VLAD is a simplified and special case of Fisher vector, in which the residuals belonging to each of the codewords are accumulated. The concatenated residuals represent the image, using which the search is carried out via simple distance measures like l_2 and l_1 . A number of improvements have been proposed to make VLAD a better representation by considering vocabulary adaptation and intra-normalization [1], residual normalization and local coordinate system [6], geometry information [33] and multiple vocabularies [14].

All these approaches are built on top of the invariant SIFT-like local features. On the other hand, features like GIST [25] are proposed towards an image representation via a global image feature. Douze *et al.* [8] attempted to identify the cases in which a global description can reasonably be used in the context of object recognition and copy detection.

Recently, the deep features extracted from the Convolutional Neural networks (CNN) have been observed [7, 26] to show a better performance over the the state-of-the-art for important vision applications such as object detection, recognition and image classification [10]. This has inspired many researchers to explore deep CNNs in order to solve a variety of problems [2, 10, 12].

For obtaining an image representation using CNNs, the mean-subtracted image is forward propagated through a set of convolution layers. Each of these layers contains filters that convolve the outputs from previous layer followed by max-pooling of the resulting feature maps within a local neighborhood. After a series of filtering and sub-sampling layers, another series of fully connected layers process the feature maps and leads to a fixed size representation in the end. Similar to the multi layer perceptron (MLP), the output at each of the hidden units is passed through a non-linear activation function to induce the nonlinearity into the model.

Because of the local convolution operations, the representation preserves the spatial information to some extent. For example, Zeiler *et al.* [35] have shown that the activations after the max-pooling of the fifth layer can be reconstructed to a representation that looks similar to the input image. This makes the representation to be sensitive to the arrangement of the objects in the scene. Though max-pooling within each feature map contributes towards the invariance to small-scale deformations [21], transformations that are more global are not well-handled due to the preserved spatial information in the activations.

Intuitively, the fully connected layers, because of their complexity, are supposed to provide the highest level of visual abstraction. For the same reason, almost all the works represent the image content with the activations obtained from the fully connected layers. There is no concrete rea-

soning provided to comment on the invariance properties of the representations out of the fully connected layers. However, Gong *et al.*[12] have empirically shown that the final CNN representation is affected by the transformations such as scaling, rotation and translation. In their experiments, they report that this inability of the activations has been translated into direct a loss of classification accuracy and emphasize that the activations preserve the spatial information. As another evidence, Babenko *et al.* [2] show that the retrieval performance of the global CNN activations (called ‘neural codes’) increases considerably, when the rotation in the reference (database) images is nullified manually.

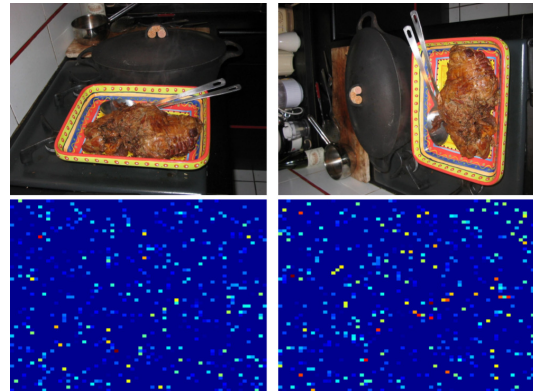


Figure 2. Sensitivity of the CNN features to global transformation. Top row: two images from the Holidays dataset, with the same objects but re-arranged. Bottom row: Corresponding l_2 normalized 4096D CNN features shown on 64×64 grid (Best viewed when zoomed-in).

We have considered two images of the same scene from the Holidays database. We have conducted an experiment to demonstrate the sensitivity of the CNN features to global transformation that results in a different arrangement of the underlying objects. We extracted the 4096 dimensional CNN feature for both the images after the first fully connected layer (fc6) in the Alexnet, as described in [19]. Figure 2 shows the two images and the corresponding deep features reshaped to a 64×64 matrix. We can clearly observe that the corresponding activations to be very different in both the representations, leading to an inferior similarity between the two images.

Gong *et al.*[12] proposed a simple framework called Multi-scale Orderless Pooling (MOP-CNN) towards an orderless representation via pooling the activations from a set of patches extracted at different scales. They resized the image to a fixed size and considered patches at three fixed scales in an exhaustive manner to extract CNN features from each of the patches. The resulting CNN features are VLAD pooled (aggregated) in order to achieve an image representation.

Motivated by this approach, we construct an image representation on top of the CNN activations without binding

any spatial information. The proposed approach, constructs a novel object based invariant representation on top of the CNN descriptions extracted from an image. We describe the objects present in an image through the deep features and pool them into a representation, as to keep the effective visual content of the image.

Generating object proposals is one of the effective solutions in the recent times to achieve computational efficiency for object detection. Our proposed system employs the selective search object proposals scheme developed by Uijlings *et al.* [32] to extract regions from the given image. Selective search relies on a bottom-up grouping approach for segmentation that results in a hierarchical grouping. Hence, it naturally enables to generate object proposals at all scales.

After obtaining the object level deep features, constructing an image representation on top of them is another challenge. In order to have a holistic image representation that summarizes the entire visual content, we propose to keep the largest of the activations at each of the output units in the CNN. That means the representation is the max pooled output vector of all the object level CNN activations extracted from an image. We have experimentally observed that the proposed representation consistently outperforms other representations based on deep features (refer section 3).

Figure 1 shows the overview of the proposed system. The remainder of the paper is organized as follows: In section 2, we discuss each of the modules of the proposed retrieval system in detail. In section 3, we present the experimental setup using which we evaluate the invariance of the proposed image representation derived from the object level CNN activations and discuss the results. Section 4 summarizes our approach with a conclusion and possible expansions.

2. Proposed Approach

In this paper, we propose to use objectness prior on the Convolutional Neural Network (CNN) activations obtained from the image in order to represent the image compactly. Our proposed system as shown in Figure 1, is essentially a cascade of three modules. In this section we discuss these three modules in detail.

2.1. Region Proposals

The first stage of the proposed system is to obtain a set of object regions from the image. Recent research offers a variety of approaches [32] [4] [36] for generating a set of class independent region proposals on an image. The motivation for this module comes from the fact that tasks such as detection and recognition rely on the localization of the objects present in the image. Thus the approach is an object based representation learning built on the deep features.

The conventional approach to object detection has been an exhaustive search over the image region with a sliding window. The recent alternative framework of objectness proposals aims to propose a set of object bounding boxes with an objective of reducing the complexity in terms of the image locations needed to be further analyzed. Another attractive feature of the object proposals framework is that, the generated proposals are agnostic to the type of the objects being detected.

In the proposed system, we adapt the region proposals approach by Uijlings *et al.* [32]. This method combines the strength of both an exhaustive search and segmentation. This method relies on a bottom-up hierarchical segmentation approach, enabling itself naturally to generate regions at different scales. The object regions obtained are forwarded through a CNN for an efficient description.

Since the objects in the scene are crucial components to understand the underlying scene, we use object based representation for describing the scene. Further, in semantically similar scenes, these objects need not be present in the same spatial locations, making the object level description a necessity. Unlike MOP-CNN [12] which preserves the geometry information loosely because of the concatenation, the proposed approach does not encode any geometric information into the final representation. Wei *et al.* [34] proposed Hypotheses-CNN-Pooling (HCP), an object based multi-label image classification approach that aggregates the labels obtained at object level classification. They obtain objects present in the image as different hypotheses and classify them separately using a fine-tuned CNN. The final set of labels are obtained as the result of max-pooling the outputs corresponding to the hypotheses, emphasizing the importance of the objects in the scene.

2.2. Feature Extraction: CNN features

The objective of the CNN in the proposed system is to achieve a high level semantic description of the input image regions. For over more than a decade, the vision community has witnessed the dominance enjoyed by a variety of hand-crafted SIFT like [22, 3, 5] features for applications like object recognition and retrieval. But a number of recent works [7, 10, 12, 26] have claimed the supremacy of the features extracted from the Convolutional Neural Networks (CNN) and Deep learning over the conventional SIFT-like ones.

CNNs are biologically inspired frameworks to emulate cortex-like mechanisms and the architectural depth of the brain for efficient scene understanding. The seminal works by LeCun *et al.* [20] has inspired many researchers to explore this framework for many vision tasks. A similar architecture proposed by Serre *et al.* [30] also tries to follow the organization of the visual cortex and builds an increasingly complex and invariant representation. Their architec-

ture alternates between template matching and max-pooling operations and is capable of learning from only a few set of training samples.

Compared to the standard feedforward networks with similar number of hidden layers, CNNs have fewer connections and parameters, making it easier to learn. In the proposed approach, we describe each of the extracted image regions with a fixed size feature vector obtained from a high capacity CNN. Our feature extraction is similar to that in [10]. We have used the Caffe [18] implementation of the CNN described by Krizhevsky *et al.* [19]. We pass the image region through five convolution layers, each of which is followed by a sub-sampling (or pooling) and a rectification. After all the convolution layers, the activations are passed through a couple of fully connected layers in order to extract a 4096-dimensional feature vector. More about the architecture details of the network can be found in [18] and [19].

2.3. Feature Pooling

At this stage, each image is essentially a set of object level CNN features. The extracted features from the CNN are max-pooled in order to obtain a compact representation for the input image.

Apart from the obvious aggregation to deliver a compact representation, the main motivation for the max-pooling stems from the aim to construct a representation that sums the visual content of the image. It is also observed by Girshick *et al.* [10] that, at the intermediate layers of the CNN (like pool 5 layer in [10]), each of the units responds with a high activation to a specific type of visual input. The specific image patches causing these high activations are observed to belong to a particular object class. It can be thought of as, each of the receptive fields in the network is probing the image region for the content of a specific semantic class. Thus, keeping the component-wise maximum of all the CNN activations belonging to the object proposals sums up the whole visual content of the image.

2.4. Query time Processing

During the query time, we apply selective search on the given query image to extract object proposals. These proposals are propagated forward through the CNN in order to describe them with a fixed length ($4096D$) representation. At this point, we rank the proposals using their objectness scores. We use a Jaccard intersection-over-union measure to reject the lower scoring proposals and retain the higher scoring ones.

Finally, we employ max-pooling to aggregate all the CNN activations in order to represent the visual content of the input image. We compare the pooled vector representation of the query with that of the database images to retrieve the similar images. Simple distance measures such as l_2 and

Hamming distance (for the binarized representations of the pooled output) are found suitable to compare the images.

3. Experiments and Results

In this section, we demonstrate the effectiveness of having an objectness prior on the Convolutional Neural Network (CNN) activations obtained from the object patches of an image in order to achieve an invariant image representation. We evaluate these invariant representations via a series of retrieval experiments performed on a set of benchmark datasets. Note that our primary aim is to have a compact representation for the images at both query and database ends over which we build an image retrieval system. We evaluate our approach over the following publicly available benchmark image databases. Sample images from the databases are shown in Fig. 3.

3.1. Datasets

INRIA Holidays[15]: This dataset contains 1491 photographs captured at different places that fall into 500 different classes. One query per class is considered for evaluating the proposed system through Mean Average Precision (mAP).

Oxford5K[28]: This database is a collection of 5062 images of 11 landmark buildings in the oxford campus that are downloaded from Flickr. 55 hold-out queries, uniformly distributed over the 11 landmarks evaluate the retrieval performance through Mean Average Precision (mAP).

Paris6K[29]: The *Paris* dataset consists of around 6400 high resolution images collected from Flickr by searching for 11 particular landmarks in the city of Paris. Similar to the *Oxford* database, 55 queries distributed over these 11 classes. Mean Average Precision (mAP) is considered as the metric for evaluation.

UKB[24]: *UKB* dataset contains 10200 images of 2550 different indoor objects. This can be considered as having 2550 classes each populated with four images. We have considered One query image per class and four times the precision at rank 4 ($4 \times precision@4$) as the evaluation metric for this database.

3.2. Compact Representation and Retrieval

The proposed approach extracts the possible object patches as suggested by the selective search [32] approach. Because of the objectness prior associated with each of these patches, our approach is more accurate and reliable in aggregating the effective visual content of the image. From each image, we extract on an average 2000 object proposals. These patches are described by the 4096 dimensional features obtained using a trained CNN. Through max-pooling, we retain the component-wise maxima of all these features. The resulting representation is still a 4096 dimensional vector that sums up the visual information present in the image.

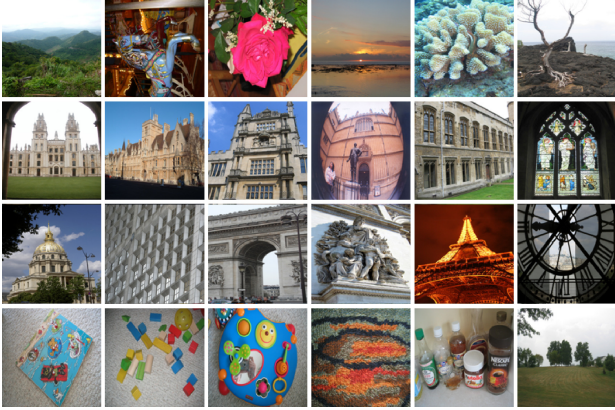


Figure 3. Sample database images. First row shows images from the *Holidays*, second from the *Oxford5K*, third from the *Paris6K* and the fourth from the *UKB* dataset.

We use the Principal Component Analysis (PCA) to reduce the dimensionality of the representation. The retrieval is conducted using these lower dimensional representations on all the databases and the results are presented in Tables 1, 2, 3 and 4. We compare the performance of the proposed approach with the existing methods for various sizes of the compact codes.

3.3. Binarization

A great deal of compression can be achieved via binary coding. We adopt the recent work, Iterative Quantization (ITQ) by Gong *et al.* [11] to obtain binary representation for our pooled representation while preserving the similarity. The method proposes a simple iterative optimization strategy to rotate a set of mean centered data so that the quantization error is minimized to assign each point to vertices of a unit binary cube, in the projected space of arbitrary dimension.

The results presented in Figure 5 show that the binary representation retains the retrieval performance in spite of the very compact representation.

3.4. Number of Proposals

In this subsection, we investigate the retrieval performance as a function of the number of extracted region proposals per image. The proposed method, using the selective search [32] approach provides around 2000 regions per image as potential object locations. We rank the regions based on their scores. Jaccard IoU measure is employed to neglect the regions with lower scores.

The plot in Fig. 6 shows the mAP considering only the top N proposals for three datasets. It can be observed that with as minimum as 100 proposals itself, the proposed approach achieves very competitive results on all the

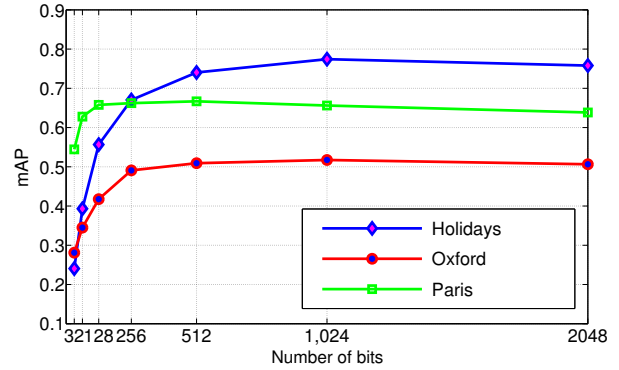


Figure 5. Retrieval performance on various datasets for different number of bits in the binary representation obtained via ITQ.

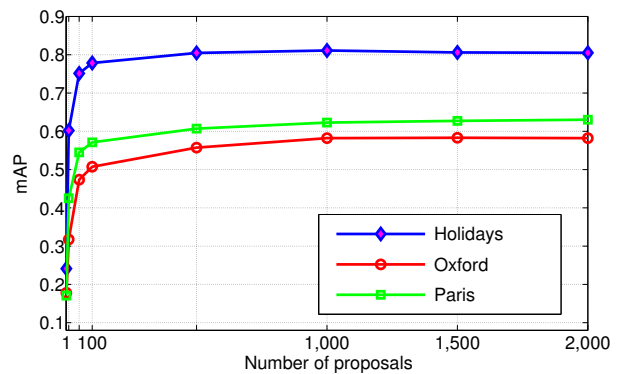


Figure 6. Retrieval performance on various databases considering different number of object proposals per image.

databases. The approach can be computationally very efficient by considering only these top scored CNN activations.

3.5. Fine Tuning

Apart from using the pre-trained model of the Alexnet CNN [19], we have also fine tuned it with a very small amount of training data from the dataset under investigation. For each dataset, we have presented around 100,000 patches populated equally across all the classes, as the training data for fine tuning. The final fully connected layer (*fc8*) is adjusted to have as many units as the number of classes in the corresponding dataset. The learning rate is kept very low for the initial layers compared to *fc8*. The results for these experiments are presented in Figs. 7, 8, 9 and 10. Fine tuning of the model is observed to improve the representation for most of the databases.



Figure 4. **Retrieval results on Holidays.** First image in each row (separated by green line) is the query image and the subsequent images are the corresponding ranked images. The red boxes show images that are irrelevant and ranked higher than the relevant images. Note that irrelevant images ranked after the relevant images are not shown in red boxes. Note that queries 1, 3, 4 and 5 are picked to demonstrate the invariance to rotation achieved by the proposed approach. Note that, the query image is removed from the ranked list while computing the mAP.

Table 1. Retrieval results on the *Holidays* dataset. Best performances in each column are shown in bold. (∇ indicates result obtained with manual geometric alignment and retraining the CNN with similar database.)

METHOD	Dimension									
	32	64	128	256	512	1024	2048	4096	8064	$\geq 10K$
VLAD [16]	48.4	52.3	55.7	-	59.8	-	62.1	55.6		
Fisher Vector[27]	48.6	52	56.5	-	61	-	62.6	59.5		
VLAD +adapt+ innorm [1]	-	-	62.5	-	-	-	-	-	-	64.6
Fisher+color [13]	-	-	-	-	-	-	-	77.4		
Multivoc-VLAD [14]	-	-	61.4	-	-	-	-	-		
Triangulation Embedding [17]	-	-	61.7	-	-	72.0	-	-	77.1	
Sparse-coded Features [9]	-	-	0.727	-	-	-	-	-	-	76.7
Neural Codes [2]	68.3	72.9	78.9 ∇	74.9	74.9	-	-	79.3 ∇		
MOP-CNN [12]	-	-	-	-	-	-	80.2	78.9		
gVLAD [33]	-	-	77.9	-	-	-	-	81.2		
Proposed	73.96	80.67	85.09	87.77	88.46	86.58	85.94	85.94		

4. Conclusion

In this paper, we have demonstrated the effectiveness of the objectness prior on the Convolutional Neural Network (CNN) activations of image regions, to impart invariance to the common geometric transformations such as translation, scaling and rotation. The object patches are extracted at multiple scales in an efficient non-exhaustive manner. The

activations obtained for the extracted patches are aggregated in an order less manner to obtain an invariant image representation. Without incorporating any spatial information, our representation exhibits the state of the art performance for the image retrieval application with compact codes of dimensions less than 4096. The binary representation with a memory footprint as small as $2Kbits$ per image also per-

Table 2. Retrieval results on the *Oxford5K* dataset. Best performances in each column are shown in bold. (* indicates result obtained with retraining the CNN with similar database.)

METHOD	Dimension								
	32	64	128	256	512	1024	2048	4096	8064
VLAD [16]	-	-	28.7	-	-	-	-	-	-
Fisher Vector[27]	-	-	30.1	-	-	-	-	-	-
Neural Codes [2]	39.0	42.1	43.3	43.5	43.5	-	-	54.5*	
VLAD +adapt+ innorm [1]	-	-	44.8	-	-	-	-	55.5	
gVLAD [33]	-	-	60	-	-	-	-	62.6	
Triangulation Embedding [17]	-	-	43.3	-	-	56.0	57.1	62.4	67.6
Proposed	40.1	48.02	56.24	59.78	60.71	59.42	58.92	58.20	

Table 3. Retrieval results on the *Paris6K* dataset. Best performances in each column are shown in bold. (* indicates result obtained with retraining the CNN with similar database.)

METHOD	Dimension								
	32	64	128	256	512	1024	2048	4096	
VLAD [16]	-	-	28.7	-	-	-	-	-	
Fisher Vector[27]	-	-	30.1	-	-	-	-	-	
Neural Codes [2]	39.0	42.1	43.3	43.5	43.5	-	-	54.5 *	
gVLAD [33]	-	-	59.2	-	-	-	-	63.1	
Proposed	65.38	71.47	70.39	68.43	66.23	64.11	62.84	63.20	

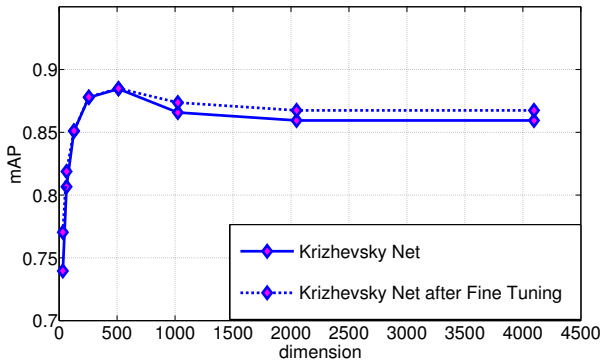


Figure 7. Retrieval performance on Holidays dataset with the object level deep features obtained after fine tuning the net.

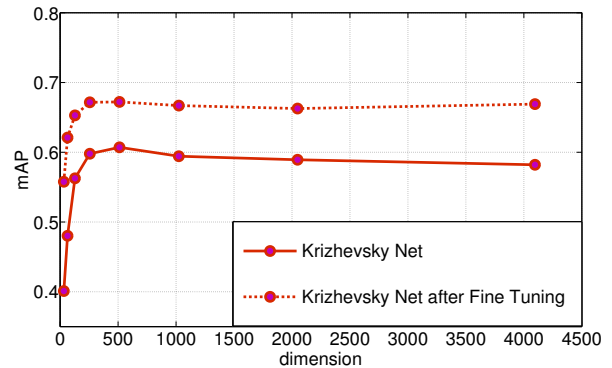


Figure 8. Retrieval performance on Oxford5K dataset with the object level deep features obtained after fine tuning the net.

forms on par with the state of the art. Our proposed method achieves these results with only 100 – 500 region proposals per image, making it computationally non-intensive. The database specific fine tuning of the learned model is observed to improve the representation using minimal training data.

5. Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the K40 GPU used for this

research.

Table 4. Retrieval results on the UKB dataset. Best performances in each column are shown in bold. (* indicates result obtained with retraining the CNN with similar database.)

METHOD	Dimension									
	32	64	128	256	512	1024	2048	4096	8064	$\geq 10K$
VLAD [16]	-	-	3.35	-	-	-	-	-	-	-
Fisher Vector [27]	-	-	3.33	-	-	-	-	-	-	-
Triangulation Embedding [17]	-	-	3.4	3.45	3.49	3.51	-	-	3.53	-
Neural Codes [2]	3.3*	3.53*	3.55*	3.56*	3.56*	-	-	3.56*	-	-
Sparse-Coded Features [9]	3.52	3.62	3.67	-	-	-	-	-	-	3.76
Proposed	3.4	3.61	3.71	3.77	3.81	3.84	3.84	3.84	-	-

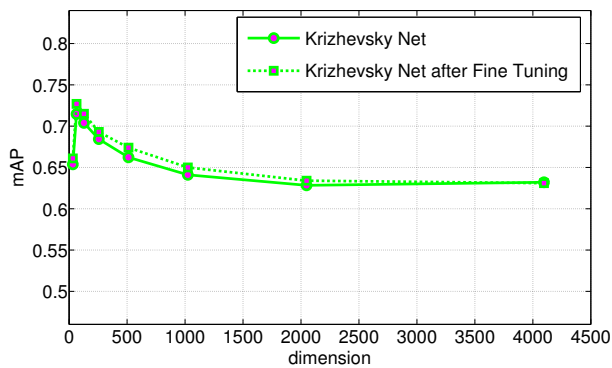


Figure 9. Retrieval performance on Paris6K dataset with the object level deep features obtained after fine tuning the net.

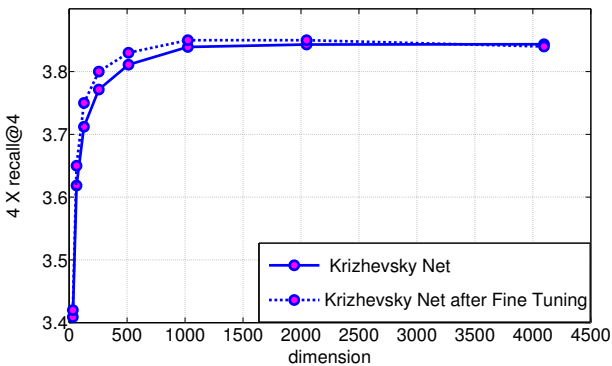


Figure 10. Retrieval performance on UKB dataset with the object level deep features obtained after fine tuning the net.

References

- [1] R. Arandjelović and A. Zisserman. All about VLAD. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 2, 6, 7
- [2] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky. Neural codes for image retrieval. In *ECCV- European Conference Computer Vision*, 2014. 2, 6, 7, 8
- [3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346–359, June 2008. 1, 3
- [4] M. M. Cheng, Z. Zhang, W. Y. Lin, and P. H. S. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *IEEE conference on Computer Vision and Pattern Recognition*, 2014. 3
- [5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *In CVPR*, pages 886–893, 2005. 3
- [6] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez. Revisiting the VLAD image representation. In *ACM Multimedia*, Oct. 2013. 2
- [7] 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 *International Conference in Machine Learning (ICML)*, 2014. 2, 3
- [8] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR*, 2009. 2
- [9] T. Ge, Q. Ke, and J. Sun. Sparse-coded features for image retrieval. In *British Machine Vision Conference*, 2013. 6, 8
- [10] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3, 4
- [11] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 5
- [12] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV- European Conference Computer Vision*, 2014. 2, 3, 6
- [13] A. Gordo, J. Rodriguez-Serrano, F. Perronnin, and E. Valveny. Leveraging category-level labels for instance-level image retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, 2012. 6
- [14] H. Jégou and O. Chum. Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening. In *ECCV - European Conference on Computer Vision*, 2012. 2, 6

- [15] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of European Conference on Computer Vision*. 2008. 4
- [16] H. Jégou, F. Perronnin, M. Douze, and C. Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012. 2, 6, 7, 8
- [17] H. Jégou and A. Zisserman. Triangulation embedding and democratic aggregation for image search. In *CVPR*, 2014. 6, 7, 8
- [18] 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
- [19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 2012. 2, 4, 5
- [20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. 3
- [21] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *International Conference on Machine Learning*, pages 609–616, 2009. 2
- [22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2), 2004. 1, 3
- [23] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proceedings of the British Machine Vision Conference*, 2002. 1
- [24] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168, June 2006. 4
- [25] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001. 2
- [26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 2, 3
- [27] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1–8, 2007. 1, 6, 7, 8
- [28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 4
- [29] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 4
- [30] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:411–426, 2007. 3
- [31] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1470–1477, 2003. 1
- [32] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013. 3, 4, 5
- [33] Z. Wang, W. Di, A. Bhardwaj, V. Jagadeesh, and R. P. Ramathu. Geometric VLAD for large scale image search. *CoRR*, 2014. 2, 6, 7
- [34] 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. 3
- [35] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, 2013. 2
- [36] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 3