Group Invariant Deep Representations for Image Instance Retrieval

Olivier Morère[*,1,2], Antoine Veillard[*,1], Jie Lin[2], Julie Petta[3], Vijay Chandrasekhar[2], Tomaso Poggio[4]

Abstract

Image instance retrieval pipelines are based on comparison of vectors known as global image descriptors between a query image and the database images. While CNN-based descriptors are generally known for good retrieval performance, they nevertheless present a number of drawbacks including the lack of robustness to common object transformations such as rotations compared with their interest point based Fisher Vector counterparts. In this paper, we propose a method for computing invariant global descriptors from CNNs. Our method implements a recently proposed mathematical theory for invariance in a sensory cortex modeled as a feedforward neural network. The resulting global descriptors can be made invariant to multiple arbitrary transformation groups while retaining good discriminativeness. Based on a thorough empirical evaluation using several publicly available datasets, we show that our method is able to significantly and consistently improve retrieval results every time a new type of invariance is incorporated.

Introduction

Image instance retrieval is the discovery of images from a database representing the same object or scene as the one depicted in a query image. The first step of a typical retrieval pipeline starts with the comparison of vectors representing the image contents known as *global image descriptors*. While CNN based descriptors are progressively replacing Fisher Vectors (FV) (Perronnin et al. 2010) as state-of-the-art descriptors for image instance retrieval (Babenko et al. 2014; Sharif Razavian et al. 2015), we have shown in our recent work thoroughly comparing both types of descriptors (Chandrasekhar et al. 2016) that the use of CNNs still presents a number of significant drawbacks compared with FVs. One of them is the lack of invariance to transformations of the input image such as rotations: the performance of CNN descriptors quickly degrade when the objects in the query and the database image are rotated differently.

In this paper, we propose a method to produce global image descriptors from CNNs which are both compact and robust to such transformations. Our method is inspired from a recent invariance theory (subsequently referred to as *i-theory*) for information processing in sensory cortex (Anselmi et al. 2013). After showing that CNNs are compatible with the *i-theory*, we propose a simple and practical way to apply the theory to the construction of global image descriptors which are robust to various types of transformations of the input image at the same time. Through a thorough empirical evaluation based on multiple publicly available datasets, we show that our method is able to significantly consistently improve retrieval results while keeping dimensionality low.

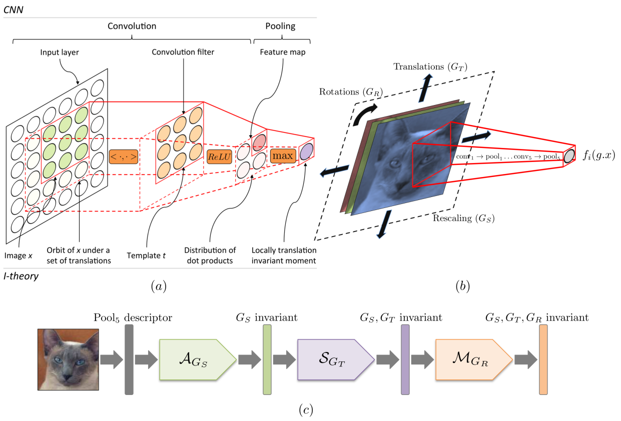

Figure 1: (a) A single convolution-pooling operation from a CNN schematized for a single input layer and single output neuron. (b) A specific succession of convolution and pooling operations learnt by the CNN (depicted in red) computes the *pool5* feature f_i for each feature map i from the RGB image data. A number of transformations g can be applied to the input x in order to vary the response $f_i(g.x)$. (c) Starting with raw *pool5* descriptors, it can be used to stack-up an arbitrary number of transformation group invariances while keeping the dimensionality under control.

The contributions of our work can be summarized as follows.

- A method based on *i-theory* for creating robust and compact global image descriptors from CNNs.
- The ability to iteratively incorporate different group in-

[**] O. Morère, A. Veillard, L. Jie, V. Chandrasekhar contributed equally.

[†1] Université Pierre et Marie Curie

[‡2] A*STAR Institute for Infocomm Research

[§3] CentraleSupélec

[¶4] CBMM, LCSL, IIT and MIT

variances, each new addition leading to consistent and significant improvements in retrieval results.

- A set of highly competitive global descriptors for image instance retrieval compared with other state-of-the-art compact descriptors at similar bitrates.
- A low risk of overfitting: few parameters and many reasonable settings which can generalize well across all datasets.
- A thorough empirical study based on several publicly available datasets.

Invariant Global Image Descriptors

I-theory in an Nutshell

Many common classes of image transformations such as rotations, translations and scale changes can be modeled by the action of a group G over the set E of images. Let $x \in E$ and a group G of transformations acting over E with group action $G \times E \rightarrow E$ denoted with a dot (.). The orbit of x by G is the subset of E defined as $O_x = \{g.x \in E | g \in G\}$. It can be easily shown that O_x is globally invariant to the action of any element of G and thus any descriptor computed directly from O_x would be globally invariant to G.

The *i-theory* predicts that an invariant descriptor for a given object $x \in E$ is computed in relation with a predefined template $t \in E$ from the distribution of the dot products $D_{x,t} = \{< g.x, t > \in \mathbb{R} | g \in G\} = \{< x, g.t > \in \mathbb{R} | g \in G\}$ over the orbit. One may note that the transformation can be applied either on the image or the template indifferently. The proposed invariant descriptor extracted from the pipeline should be a histogram representation of the distribution with a specific bin configuration. Such a representation is mathematically proven to have proper invariance and selectivity properties provided that the group is compact or at least locally compact (Anselmi et al. 2013).

In practice, while a compact group (e.g. rotations) or locally-compact group (e.g. translations, scale changes) is required for the theory to be mathematically provable, the authors of (Anselmi et al. 2013) suggest that the theory extends well (with approximate invariance) to non-locally compact groups and even to continuous non-group transformations (e.g. out-of-plane rotations, elastic deformations) provided that proper class-specific templates can be provided. Additionally, the histograms can also be effectively replaced by statistical moments (e.g. mean, min, max, standard deviation, etc.).

CNNs are i-theory Compliant Networks

All popular CNN architectures designed for image classification share a common building block: a succession of convolution-pooling operations designed to model increasingly high-level visual representations of the data. As shown in detail on Figure 1 (a), the succession of convolution and pooling operations in a typical CNN is in fact a way to incorporate local translation invariance strictly compliant with the framework proposed by the *i-theory*. The network architecture provides the robustness such as predicted by the invariance theory while training via back propagation ensures a proper choice of templates. In general, multiple convolution-pooling steps are applied resulting in increased robustness and higher level templates. Note that the iterative composition of local translation invariance approximately translates into robustness to local elastic distortions for the features at the *pool5* layer.

In this study, instead of the popular first fully-connected layer (*fc6*) which is on average the best single CNN layer to use as a global out-of-the-box descriptor for image retrieval (Chandrasekhar et al. 2016), we decide to use the locally invariant *pool5* as a starting representation for our own global descriptors and further enhance their robustness to selected transformation groups in a way inspired from *i-theory*.

Transformation Invariant CNN Descriptors

For every feature map i of the *pool5* layer ($0 \leq i < 512$ in the case of the presently used *VGG OxfordNet*), we denote $f_i(x)$ the corresponding feature obtained from the RGB image data through a succession of convolution-pooling operations. Note that the transformation f_i is non-linear and thus not strictly a mathematical dot product with a template but can still be viewed as an inner product. In this study, we propose to further improve the invariance of *pool5* CNN descriptors by incorporating global invariance to several transformation groups. The specific transformation groups considered in this study are translations G_T, rotations G_R and scale changes G_S. As shown on Figure 1 (b), transformations g are applied on the input image x varying the output of the *pool5* feature $f_i(g.x)$ accordingly.

The invariant statistical moments computed from the distributions $\{f_i(g.x)|g \in G\}$ with $G \in \{G_T, G_R, G_S\}$ are averages, maxima and standard deviations, respectively:

$$\mathcal{A}_{G,i}(x) = \frac{1}{\int_G dg} \int_G f_i(g.x) dg \tag{1}$$

$$\mathcal{M}_{G,i}(x) = \max_G \left(f_i(g.x) \right) \tag{2}$$

$$\mathcal{S}_{G,i}(x) = \frac{1}{\int_G dg} \sqrt{\int_G f_i(g.x)^2 dg - (\int_G f_i(g.x) dg)^2} \tag{3}$$

with corresponding global image descriptors obtained by simply concatenating the moments for the individual features:

$$\mathcal{A}_G(x) = (\mathcal{A}_{G,i}(x))_{0 \leq i < 512} \tag{4}$$

$$\mathcal{M}_G(x) = (\mathcal{M}_{G,i}(x))_{0 \leq i < 512} \tag{5}$$

$$\mathcal{S}_G(x) = (\mathcal{S}_{G,i}(x))_{0 \leq i < 512} \tag{6}$$

In principle, G is always measurable and of finite measure as required since G_T and G_S must be restricted to compact subsets due to image border effects.

An interesting aspect of the *i-theory* is the possibility in practice to chain multiple types of group invariances one after the other. In this study, we construct descriptors invariant to several transformation groups by successively applying the method to different transformation groups as shown on Figure 1 (c). For instance, following scale invariance with average by translation invariance with standard deviation for feature i would correspond to:

$$\max_{g_t \in G_T} \left(\frac{1}{\int_{g_s \in G_S} dg_s} \int_{g_s \in G_S} f_i(g_t g_s.x) dg_s \right) \tag{7}$$

One may note that the operations are sometimes commutable (e.g. $\mathcal{A}_G$ and $\mathcal{A}_{G'}$) and sometimes not (e.g. $\mathcal{A}_G$ and $\mathcal{M}_{G'}$) depending on the specific combination of moments.

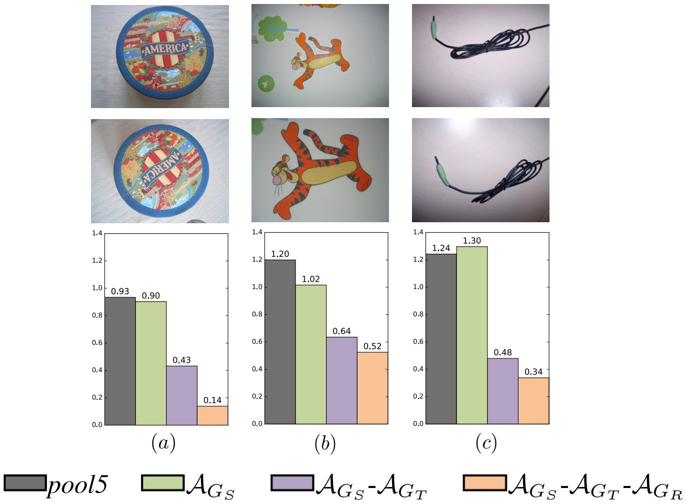

Figure 2: Distances for 3 matching pairs from *UKBench*. For each pair, 4 pairwise distances (L_2-normalized) are computed corresponding to the following descriptors: *pool5*, $\mathcal{A}_{G_S}$, $\mathcal{A}_{G_S}$-$\mathcal{A}_{G_T}$ and $\mathcal{A}_{G_S}$-$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$. Adding scale invariance makes the most difference on (b), translation invariance on (c), and rotation on (a) which is consistent with the scenarii suggested by the images.

Image Instance Retrieval Results

We evaluate our invariant image descriptors in the context of image instance retrieval. As our starting representation, we use the *pool5* layer from the 16 layers *VGG OxfordNet* with a total dimensionality of 25088 organized in 512 feature maps of size 7×7. The step size for rotations is 10 degrees yielding 36 rotated images per orbit. For scale changes, 10 different center crops geometrically spanning from 100% to 50% of the total image have been taken. For translations, the entire feature map is used for every feature, resulting in an orbit size of $7 \times 7 = 49$. We evaluate the performances of the descriptors against four popular data sets: *Holidays*, *Oxford buildings (Oxbuild)*, *UKBench (UKB)* and *Graphics*.

Figure 2 provides an insight on how adding different types of invariance with our proposed method will affect the matching distance on different image pairs of matching objects. With the incorporation of each new transformation group, we notice that the relative reduction in matching distance is the most significant with the image pair which is the most affected by the transformation group.

Transformations, Order and Moments

Our first set of results summarized in Table 1 study the effects of incorporating various transformation groups and using different moments. Table 1 also provides results for all possible combinations of transformation groups for average pooling (order does not matter as averages commute) and for the single best performer which is $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ (order matters).

Table 1: Retrieval results (mAP) for different sequences of transformation groups and moments.

SEQUENCE	DIMS	DATASET			
		Oxbd	Holi	UKB	Graphics
pool5	25088	0.427	0.707	3.105	0.315
fc6	4096	0.461	0.782	3.494	0.312
$\mathcal{A}_{G_S}$	25088	0.430	0.716	3.122	0.394
$\mathcal{A}_{G_T}$	512	0.477	0.800	3.564	0.322
$\mathcal{A}_{G_R}$	25088	0.462	0.779	3.718	0.500
$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$	512	0.418	0.796	3.725	0.417
$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_S}$	512	0.537	0.811	3.605	0.430
$\mathcal{A}_{G_R}$-$\mathcal{A}_{G_S}$	25088	0.494	0.815	3.752	0.552
$\mathcal{A}_{G_S}$-$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$	512	0.484	0.833	3.819	0.509
$\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$	512	**0.592**	**0.838**	**3.842**	**0.589**
$\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$	512 **bits**	0.523	0.787	3.741	0.552

Results are computed with the mean average precision (mAP) metric. 4×Recall@4 results are provided for UKBench, as is standard practice. G_T, G_R, G_S denote the groups of translations, rotations and scale changes respectively. Best results are achieved by choosing specific moments. $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ corresponds to the best average performer for which a binarized version is also given. *fc6* and *pool5* are provided as a baseline.

First, we point out the effectiveness of *pool5*. Although it performs notably worse than *fc6* as-is, a simple average pooling over the space of translations $\mathcal{A}_{G_T}$ makes it both better and 8 times more compact than *fc6*. As shown in Figure 3, accuracy increases with the number of transformation groups involved. On average, single transformation schemes perform 21% better compared to *pool5*, 2-transformations schemes perform 34% better, and the 3-transformations scheme performs 41% better. Second, selecting statistical moments different than averages can further improve the retrieval results. In Figure 4, we observe that $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ performs roughly 17% better (average results over all datasets) than $\mathcal{A}_{G_S}$-$\mathcal{A}_{G_T}$-$\mathcal{A}_{G_R}$. Notably, the best combination corresponds to an increase in the orders of the moments: $\mathcal{A}$ being a first-order moment, $\mathcal{S}$ second order and $\mathcal{M}$ of infinite order. A different way of stating this fact is that a more invariant representation requires a higher order of pooling. Overall, $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ improves results over *pool5* by 29% (Oxbuild) to 86% (UKBench) with large discrepancies according to the dataset. Better improvements with UKBench can be explained with the presence of many rotations in the dataset (smaller objects taken under different angles) while Oxbuild consisting mainly of upright buildings is not significantly helped by incorporating rotation invariance.

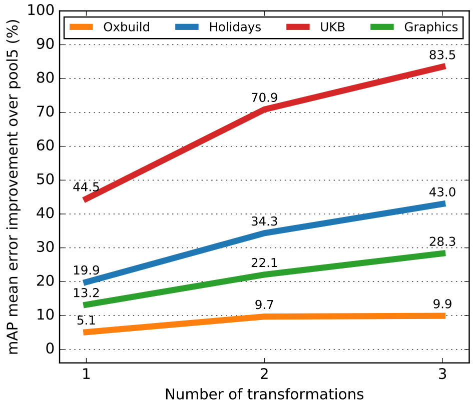

Figure 3: Results from Table 1 for the 7 strategies using averages only (rows 3 to 9) expressed in terms of improvement in mAP over *pool5*, and aggregated by number of invariance groups. On all 4 datasets, results clearly improve with the amount of groups considered.

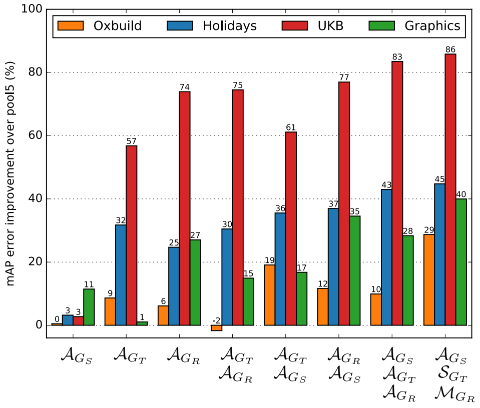

Figure 4: Results from Table 1 expressed in terms of improvement in mAP over *pool5*. Most strategies yield significant improvements over *pool5* on most datasets.

Table 2: Retrieval performance (mAP) comparing our method to other state-of-the-art methods.

METHOD	REP. SIZE	DATASET		
	#dim (bits)	Oxbd	Holi	UKB
Our results	512 (512)	0.523	**0.787**	**0.958**
(Sharif Razavian et al. 2015)	256 (1024)	**0.533**	0.716	0.842
(Sharif Razavian et al. 2015)	256 (256)	0.436	0.578	0.693
(Jégou and Zisserman 2014)	256 (2048)	0.472	0.657	0.863
(Jégou and Zisserman 2014)	128 (1024)	0.433	0.617	0.850
(Spyromitros-Xioufis et al. 2014)	128 (1024)	0.293	0.738	0.830
(Spyromitros-Xioufis et al. 2014)	128 (1024)	0.387	0.718	0.875

Only methods within a comparable range of bitrates are selected.

As shown in Table 1, a simple binarization strategy (thresholding at dataset mean) applied to our best performing descriptor $\mathcal{A}_{G_S}$-$\mathcal{S}_{G_T}$-$\mathcal{M}_{G_R}$ degrades retrieval performance only very marginally and is in fact sufficient to produce a hash that compares favourably with other state-of-the-art methods. Compared to *fc6* feature, the hash performs 26% better while being 256 times smaller. As mentioned, the invariant hash also performs well compared to other state-of-the-art. In Table 2, we compare our invariant hash against other approaches designed to produce compact representations with comparable bit sizes, 512 being considered fairly compact by current standards (Sharif Razavian et al. 2015). In addition, note that the hashing scheme used in this experiment is very naive and better hashes (better retrieval results and/or higher compression rates) could most certainly be obtained by applying better techniques based on stacked RBMs, like in (Lin et al. 2015).

Conclusion

We proposed a novel method based on *i-theory* for creating robust and compact global image descriptors from CNNs for image instance retrieval. Through a thorough empirical study, we show that the incorporation of every new group invariance property following the method leads to consistent and significant improvements in retrieval results. Our method has a number of parameters (sequence of the invariances and choice of statistical moments) but experiments show that many default and reasonable settings produce results which can generalise well across all datasets meaning that the risk of overfitting is low. Our method produces a set of compact binary descriptors that compare favourably with other state-of-the-art compact descriptors at similar bitrates.

References

Anselmi, F.; Leibo, J. Z.; Rosasco, L.; Mutch, J.; Tacchetti, A.; and Poggio, T. 2013. Unsupervised learning of invariant representations in hierarchical architectures. *arXiv:1311.4158*.

Babenko, A.; Slesarev, A.; Chigorin, A.; and Lempitsky, V. 2014. Neural Codes for Image Retrieval. In *Proceedings of European Conference on Computer Vision (ECCV)*.

Chandrasekhar, V.; Jie, L.; Morere, O.; Goh, H.; and Veillard, A. 2016. A practical guide to cnns and fisher vectors for image instance retrieval. *Signal Processing*.

Jégou, H., and Zisserman, A. 2014. Triangulation embedding and democratic aggregation for image search. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 3310–3317. IEEE.

Lin, J.; Morère, O.; Chandrasekhar, V.; Veillard, A.; and Goh, H. 2015. Deephash: Getting regularization, depth and fine-tuning right. *CoRR* abs/1501.04711.

Perronnin, F.; Liu, Y.; Sanchez, J.; and Poirier, H. 2010. Large-scale Image Retrieval with Compressed Fisher Vectors. In *Computer Vision and Pattern Recognition (CVPR)*, 3384–3391.

Sharif Razavian, A.; Sullivan, J.; Maki, A.; and Carlsson, S. 2015. A baseline for visual instance retrieval with deep convolutional networks. In *International Conference on Learning Representations, May 7-9, 2015, San Diego, CA*. ICLR.

Spyromitros-Xioufis, E.; Papadopoulos, S.; Kompatsiaris, I. Y.; Tsoumakas, G.; and Vlahavas, I. 2014. A comprehensive study over vlad and product quantization in large-scale image retrieval. *Multimedia, IEEE Transactions on* 16(6):1713–1728.