A BASELINE FOR VISUAL INSTANCE RETRIEVAL WITH DEEP CONVOLUTIONAL NETWORKS

Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, Stefan Carlsson
Computer Vision and Active Perception Lab (CVAP),
Royal Institute of Technology (KTH), 100 44 Stockholm, Sweden

1 EXTENDED ABSTRACT

This work presents simple pipelines for visual image retrieval exploiting image representations based on convolutional networks (ConvNets), and demonstrates that ConvNet image representations outperform other state-of-the-art image representations on six standard image retrieval datasets. ConvNet based image features have increasingly permeated the field of computer vision and are replacing hand-crafted features in many established application domains. Much recent work has illuminated the field on how to design and train ConvNets to maximize performance (Simonyan et al., 2013; Girshick et al., 2014; Chatfield et al.; Zeiler & Fergus, 2014; Azizpour et al., 2014) and also how to exploit learned ConvNet representations to solve visual recognition tasks (Oquab et al., 2014; Donahue et al., 2014; Fischer et al., 2014; Razavian et al., 2014). We have built on these findings to tackle visual image retrieval.

Beside the performance, another issue for visual instance retrieval is the dimensionality and memory requirements of the image representation. Usually two separate categories are considered, for which we report the results. These are the *small* footprint representations encoding each image with less than 1kbytes and the *medium* footprint representations which have dimensionality between 10k and 100k. The small regime is required when the number of images is huge and memory is a bottleneck, while the medium regime is more useful when the number of images is less than 50k. In our pipeline for the *small* we extract the features for 576×576 images, and for *medium* we use those features combined with the spatial search method described in (Azizpour et al., 2014). Furthermore, inspired by the recent work of Chatfield et al. (2014), we report the results also for a *tiny* representation (Torralba et al., 2008; Jégou et al., 2012; Arandjelović & Zisserman, 2013; Jégou & Zisserman, 2014). We define a *tiny* image representation as one that takes 32bytes or less to store and is learnt independently of the test dataset. Such a compressed representation would allow large scale searches to be completed on mobile phones (Panda et al., 2013) or on the cloud (Quack et al., 2004).

2 RESULTS SUMMARY

To evaluate our model, we used two networks. First, one which we refer to as AlexNet Krizhevsky et al. (2012) is the publicly available network implemented by `caffe`. The second network which we call OxfordNet has the same structure as (Simonyan & Zisserman, 2015) except the network that we trained has 256 kernels at the final convolutional layers as opposed to the Oxford paper with 512 kernels. Among available alternatives we used last convolutional layers response followed by a max pooling operation as the basic representation for small regime retrieval. For tiny representation, we quantized the basic representation and for medium regime, we followed Azizpour et al. (2014) but optimized the parameters. The details of our pipeline is presented in table 1 (we extract square patches of L different sizes in the search).

	Spat. search	**ConvNet rep**	**Rep. Post-processing**					**Matching**
Rep.	L	max-pool grid sz	Norm.	PCA	Quant.	Pow. norm.	Norm.	Dist. met.
Medium	4	2×2	L_2	yes	no	no	L_2	L_2
Small	1	1×1	none	no	4 bits	$\alpha = 2$	L_1	city-block
Tiny	1	1×1	none	no	1 bit	no	none	cosine

Table 1: **Parameter settings for the representations of different memory footprint studied in this work.** The parameters are listed from left to right in the order in which they are applied especially in regard to the post-processing. The heading PCA refers to the process of PCA whitening and PCA dimensionality reduction. Norm. refers to the normalization that is applied to the representation extracted from 576×576 image patches.

We report results on six standard image retrieval datasets: **Oxford5k buildings** (Philbin et al., 2007), **Oxford105k buildings** (Philbin et al., 2007), **Paris6k buildings** (Philbin et al., 2008), **Sculptures6k** (Arandjelović & Zisserman, 2011), **Holidays** (Jégou et al., 2008) and **UKbench** (Nistér & Stewénius, 2006). In table 2 we report results for our small and tiny memory representations and compare our results to s.o.a. methods with 128 or more dimensions. In table 3 we report results for our medium sized representations and compare their performance to that of other s.o.a. image representations. From examining the numbers one can see that our simple image retrieval pipeline based on ConvNet representations outperforms hand-crafted image representations, such as VLAD and IFV, which require several iterations of learning from specialized training data. The ConvNet representations we use have been extracted from generically trained ConvNets, and the only recourse we make to specialized training data is in the PCA whitening.

	Rep. size	**Dataset**					
Method	# dim(size)	Oxford5k	Paris6k	Sculp6k[1]	Holidays	UKB	Ox105k
AlexNet	256(128B)	45.2	58.8	23.1	**74.2**	**88.6**	41.0
OxfordNet	256(128B)	**53.3**	**67.0**	**37.7**	71.6	84.2	**48.9**
AlexNet	256(32B)	33.5	38.0	16.3	61.1	78.7	27.1
OxfordNet	256(32B)	43.6	54.9	26.1	57.8	69.3	38.0
VLAD+CSurf (Xioufis et al., 2014)	128	29.3	-	-	73.8	83.0	-
mVLAD+Surf (Xioufis et al., 2014)	128	38.7	-	-	71.8	87.5	-
T-embedding (Jégou & Zisserman, 2014)	128	43.3	-	-	61.7	85.0	35.3
T-embedding (Jégou & Zisserman, 2014)	256	47.2	-	-	65.7	86.3	40.8

Table 2: **Performance of small & tiny memory footprint regimes.**

	Dataset				
Method	Oxford5k	Paris6k	Sculp6k[1]	Holidays	UKB
AlexNet	77.4	82.5	50.9	**89.7**	**95.0**
OxfordNet	**84.4**	**85.3**	**67.4**	88.1	93.1
BoB (Arandjelović & Zisserman, 2011)	-	-	45.4	-	-
CVLAD (Zhao et al., 2013)	51.4	-	-	82.7	90.5
PR-proj (Simonyan et al., 2014)	82.5	81.0	-	-	-
ASMK+MA (Tolias et al., 2013)	83.8	80.5	-	88.0	-

Table 3: **Performance of medium memory footprint regimes**.

Notes on comparisons of our work with prior arts in the field:
1) Our method extract square patches form an image and therefore includes background context in the representation while other methods only use the features extracted from the bounding box. This can be fixed by extracting rectangular patches from an image and the results vary for $\pm 2\%$. (Oxford5k performance drops to 82.6 while Paris6k performance increases to 87.5.)
2) Our pipelines are the first pipelines that work for both textured-less items (e.g. sculptures) and highly-textured items (e.g. buildings) using exactly the same settings.
3) Previous methods are often specialized and learn their parameters on similar datasets and could then suffer from domain shift. On the other hand, our pipeline does not rely on the bias of the dataset but it can still be specialized to a high degree (fine-tuning the OxfordNet with landmark dataset Babenko et al. (2014) increases the performance on Oxford5k up to 85.3).

In sum, the work shows that ConvNet image representations outperform other s.o.a. image representations for visual image retrieval if one selects the appropriate responses from a generic deep ConvNet. Our result should only be viewed as a baseline and by no means we claim that our method is optimal yet. Even the simple additions such as concatinating different architecture representations gives a boost in performance (e.g. 87.2 for Oxford5k).

Acknowledgment. We would like to thank NVIDIA Co. for the generous donation of K40 GPUs.

[1]Results for the sculpture dataset are reported using image size of $227 \times 277 \times 3$.

REFERENCES

Arandjelović, Relja and Zisserman, Andrew. Smooth object retrieval using a bag of boundaries. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

Arandjelović, Relja and Zisserman, Andrew. All about VLAD. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

Azizpour, Hossein, Razavian, Ali Sharif, Sullivan, Josephine, Maki, Atsuto, and Carlsson, Stefan. From generic to specific deep representations for visual recognition. *arXiv:1406.5774v1 [cs.CV]*, June 2014.

Babenko, Artem, Slesarev, Anton, Chigorin, Alexander, and Lempitsky, Victor S. Neural codes for image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Chatfield, Ken, Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Chatfield, Ken, Simonyan, Karen, and Zisserman, Andrew. Efficient on-the-fly category retrieval using convnets and gpus. *arXiv:1407.4764 [cs.CV]*, 2014.

Donahue, Jeff, Jia, Yangqing, Vinyals, Oriol, Hoffman, Judy, Zhang, Ning, Tzeng, Eric, and Darrell, Trevor. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Fischer, Philipp, Dosovitskiy, Alexey, and Brox, Thomas. Descriptor matching with convolutional neural networks: a comparison to SIFT. *arXiv:1405.5769 [cs.CV]*, 2014.

Girshick, Ross B., Iandola, Forrest N., Darrell, Trevor, and Malik, Jitendra. Deformable part models are convolutional neural networks. *arXiv:1409.5403 [cs.CV]*, 2014.

Jégou, Hervé and Zisserman, Andrew. Triangulation embedding and democratic aggregation for image search. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Jégou, Hervé, Douze, Matthijs, and Schmid, Cordelia. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.

Jégou, Hervé, Perronnin, Florent, Douze, Matthijs, Sánchez, Jorge, Pérez, Patrick, and Schmid, Cordelia. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Nistér, David and Stewénius, Henrik. Scalable recognition with a vocabulary tree. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

Oquab, Maxime, Bottou, Léon, Laptev, Ivan, and Sivic, Josef. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Panda, Jayaguru, Brown, Michael S., and Jawahar, C. V. Offline mobile instance retrieval with a small memory footprint. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013.

Philbin, James, Chum, Ondrej, Isard, Michael, Sivic, Josef, and Zisserman, Andrew. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

Philbin, James, Chum, Ondrej, Isard, Michael, Sivic, Josef, and Zisserman, Andrew. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

Quack, Till, Mönich, Ullrich J., Thiele, Lars, and Manjunath, B. S. Cortina: a system for large-scale, content-based web image retrieval. In *Proceedings of the International Conference on Multimedia (ACM Multimedia)*, 2004.

Razavian, Ali Sharif, Azizpour, Hossein, Sullivan, Josephine, and Carlsson, Stefan. CNN features off-the-shelf: an astounding baseline for recognition. *arxiv:1008.2849v3 [cs.CV]*, May 2014.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034 [cs.CV]*, 2013.

Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1573–1585, 2014.

Tolias, Giorgos, Avrithis, Yannis S., and Jégou, Hervé. To aggregate or not to aggregate: Selective match kernels for image search. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2013.

Torralba, Antonio, Fergus, Robert, and Weiss, Yair. Small codes and large image databases for recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

Xioufis, Eleftherios Spyromitros, Papadopoulos, Symeon, Kompatsiaris, Yiannis, Tsoumakas, Grigorios, and Vlahavas, Ioannis P. A comprehensive study over VLAD and product quantization in large-scale image retrieval. *IEEE Transactions on Multimedia*, 16(6):1713–1728, 2014.

Zeiler, Matthew D. and Fergus, Rob. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Zhao, Wan-Lei, Jégou, Hervé, Gravier, Guillaume, et al. Oriented pooling for dense and non-dense rotation-invariant features. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2013.