Fine-tuning of CNN models for Instance Search with Pseudo-Relevance Feedback

Eva Mohedano, Kevin Mcguinness, Xavier Giró-i-Nieto and Noel E. O'Connor
Insight Centre for Data Analytica, Dublin, Ireland
Universitat Politecnica de Catalunya, Barcelona, Spain

CNN classification models trained on millions of labeled images have been proven to encode "general purpose" descriptors in their intermediate layers. These descriptors are useful for a diverse range of computer vision problems [1]. However, the target task of these models is substantially different to the instance search task. While classification is concerned with distinguishing between different classes, instance search is concerned with identifying concrete instances of a particular class.

In this work we propose an unsupervised approach to finetune a model for similarity learning [2]. For that, we combine two different search engines: one based on off-the-shelf CNN features, and another one on the popular SIFT features. As shown in the figure below, we observe that the information of pre-trained CNN representations and SIFT is in most of the cases complementary, which allows the generation of high quality rank lists. The fusion of the two rankings is used to generate training data for a particular dataset. A pseudo-relevance feedback strategy [3] is used for sampling images from rankings, considering the top images as positive examples of a particular instance and middle-low ranked images as negative examples.

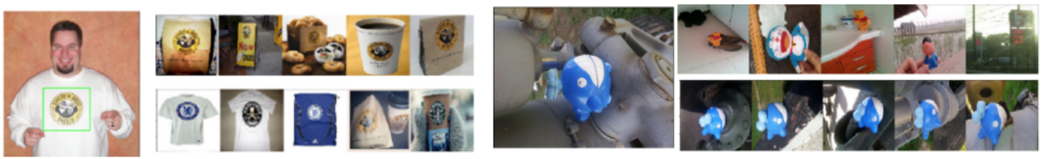

Figure 1: A logo instance query (left) and a whale toy instance query (right). On the right of each query, the top five retrieved images for a SIFT-based system (first row), and a CNN-based system (second row). SIFT appears to be superior in planar textured instances such as logos. CNN appear superior on textureless instances such as the whale toy, where color is one of the most discriminative features.

Preliminary results (table below) show that CNN representations fine-tuned with the proposed scheme significantly improve their performance within a particular domain at the cost of losing their generalization. Future work will explore large training datasets to improve the generalization of the fine-tuned model.

Table 1: Mean Average Precision (mAP) of off-the-shelf and fine-tuned models using Oxford images as training set.

	Oxford	Paris
Off-the-shelf CNN	0.480	0.698
Fine-tuned CNN	0.733	0.247

References

[1] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

[2] F. Radenović, G. Tolias, and O. Chum, "Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples," in *European Conference on Computer Vision*, pp. 3–20, Springer, 2016.

[3] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," *Image and Video Retrieval*, pp. 649–654, 2003.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.