PicSOM Experiments in TRECVID 2014

Workshop notebook paper – Revision: 1.16

Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Rao Muhammad Anwer, Jorma Laaksonen, Erkki Oja
Department of Information and Computer Science
Aalto University School of Science
P.O. Box 15400, FI-00076 Aalto, Finland

firstname.lastname@aalto.fi

Abstract

Our experiments in TRECVID 2014 include successful participation in the Semantic Indexing (SIN) task and unsuccessful participation in the Multimedia Event Detection (MED) and Multimedia Event Recounting (MER) tasks.

In semantic indexing, we participated in the main task only. We extended our last year's set of features with SIFT descriptors encoded with Fisher vectors and VLAD, and a total of 24 features based on convolutional neural network (CNN) activations. We also utilized hard negative mining to to acquire more relevant negative examples. We submitted the following four runs:

- 4 MUMINPAPPAN: Baseline run matching the best PicSOM SIN submission in TRECVID 2013
- 3 HATTIFNATTAR: Run based on CNN features only, also including hard negative mining
- 2 SNUSMUMRIKEN: Run with Fisher vector and VLAD features and the set of 24 CNN features included
- 1 MÅRRAN: Run combining all features and hard negative mining

The run 1 MÅRRAN obtained the highest MXIAP score of 0.2880.

In the Multimedia Event Detection and Recounting task we tried to participate in the MED14-EvalFull search task, but failed.

I. Introduction

In this notebook paper, we describe our experiments for the TRECVID 2014 evaluation [1]. We participated in the Semantic Indexing (SIN, Section II) and tried to participate in the Multimedia Event Detection and Multimedia Event Recounting (MED&MER, Section III) tasks. Overall conclusions are presented in Section IV.

II. Semantic indexing

Our submissions to the semantic indexing (SIN) task are based on fusing several supervised detectors trained for each concept, based on different shot-level image features. The basic system architecture is the same as we have used in previous editions of TRECVID [2]. As the concept-wise ground-truth for the supervised detectors we used the annotations gathered by the organized collaborative annotation effort [3]. All our runs were submitted to the *main* task and are either of *training type A* or *type D*. We did not participate in the *no annotation* condition nor the *localization* subtask.

A. Features and classifiers

In addition to the main keyframes provided in the master shot reference, we extracted additional frames from training data shots longer than two seconds and used all I-frames provided in the test data set.

1) Old global and BoV features: We used the six image features from our previous TRECVID submission: two global features (*Centrist* and *ScalableColor*) and four BoV-type features (*SIFT*, *ColorSIFT*, *SIFTds*, and *ColorSIFTds*). Non-linear SVM classifiers were used with the exponential χ^2 kernel for the BoV features and the RBF kernel for the global features. See [4], [2] for details.

2) Fisher vector and VLAD encoding: We extracted dense SIFT descriptors and encoded them using both Fisher vectors [5] and VLAD [6]. The codebooks were constructed using a 128-component GMM and k-means with 512 clusters, respectively. The corresponding classifiers were trained using linear SVMs.

3) CNN features: It has recently been observed that CNNs trained with one visual dataset can function as highly discriminative features even for considerably different data domains and tasks [7], [8], [9]. For our experiments, we extracted a total of 24 different CNN features from the images. The used CNNs were trained on ImageNet 2010 and 2012 training datasets, following as closely as possible the network structure parameters of Krizhevsky *et al* [10] and Zeiler & Fergus [11].

We use the activations of the first fully-connected layers of each network as our features, which results in 4096-dimensional feature vectors. Both a single center region or ten regions as suggested in [10] were extracted from the test images. In the case of ten regions, both average and maximum pooling of the region-wise features were used.

Furthermore, we use the reverse spatial pyramid pooling proposed in [8] with two scale levels. Our first level corresponds to the center region, and the second level consists of nine regions (3×3 grid) on the scale of two. The CNN activations of the regions are then pooled using average pooling, and the activations of the different scales are concatenated. The resulting spatial pyramid features are therefore 8192-dimensional. See [9] for more details.

As classifiers for the CNN features, we utilized linear SVMs

TABLE I
An overview of the submitted runs in the semantic indexing task. See text for details.

	features				hard neg.	
run id	glob.	BoV	FV	CNN	mining	MXIAP
4 Muminpappan	●	●				0.1951
3 Hattifnattar				●	●	0.2843
2 Snusmumriken		●	●	●		0.2722
1 Mårran		●	●	●	●	0.2880

with homogeneous kernel maps [12] of order $d = 2$ to approximate the intersection kernel.

B. Classifier fusion

Classifier outcomes were in the first stage fused over the features for each frame with arithmetic mean. In the second fusion stage over the frames of each shot we used the maximum value. This can be written as

$$r_i = \max_{j=1,\ldots,n_i} \frac{1}{N} \sum_{k=1}^{N} r_{i,j,k} \ , \tag{1}$$

where N is the number of used features, n_i is the number of frames in shot i and $r_{i,j,k}$ is the detection score for feature k in frame j of shot i.

The score values for the shots were obtained in the same manner for each run as the maximum over the frame-wise scores resulting from the arithmetic mean over all features.

C. Mining hard negatives

A concept-wise, two-class classifier generally produced false positives on negative examples that were similar to the positive examples according to the used feature space. Therefore, to acquire more relevant negative examples, we performed n rounds of hard negative mining [13] and sampled 10 000 negative examples on each round. The final classifier for a given feature was obtained by fusing the classifier trained with the original, randomly sampled negatives and the n classifiers using mined relevant negatives.

In preliminary experiments, we observed that a single round of mining hard negatives already brought the greatest improvement. We therefore used the value $n = 1$ in the following experiments.

D. Submitted runs

This section describes our submitted semantic indexing runs. Table I shows an overview, where the four columns in the middle refer to the used features: global non-BoV features, BoV features, Fisher vectors + VLAD, and CNN features. The next column indicates whether hard negative mining was used, and the rightmost column lists the corresponding mean extended inferred average precision (MXIAP) [14] values. Figure 1 illustrates the concept-wise XIAP results of the runs.

The run 4 Muminpappan is intended to match the best PicSOM submission in TRECVID 2013, denoted as `PicSOM_M_1`, i.e. to use the same features, classifiers, and method of fusion [2].

In the run 2 Snusmumriken, the Fisher vector and VLAD features and the set of 24 CNN features were included. The global image features were discarded.

The run 3 Hattifnattar uses only the CNN features, together with hard negative mining.

The run 1 Mårran combined the characteristics of 2 Snusmumriken and 3 Hattifnattar, that is, all SIFT-based and CNN features with hard negative mining.

Except 4 Muminpappan, all our submitted runs were of training type D, due to the use of non-IACC non-TRECVID training data (ImageNet 2010 and 2012 training datasets to train the CNNs for feature extraction). 4 Muminpappan was of training type A as it only used IACC training data.

The most striking observation on the results is the notable increase of performance compared to our last year's submissions. This is mostly due to the extended set of features, in particular the CNN activation features. By comparing 4 Muminpappan with 2 Snusmumriken, we observe a 40% increase on MXIAP induced by the different feature sets.

Second, the mining of hard negatives further improved the results, as can be observed by comparing 2 Snusmumriken and 1 Mårran, the latter including the mining step and obtaining the highest MXIAP among our runs, 0.2880 (a 6% increase). The solid performance of the CNN features can furthermore be observed from the run 3 Hattifnattar, which contains only the CNN features but still almost reaches the MXIAP value of 1 Mårran.

III. Multimedia Event Detection & Recounting

In Multimedia Event Detection (MED) and Recounting (MER) tasks, we tried to participate in the MED14-EvalFull search task. However, we underestimated the time required for the metadata generation phase and were not able to proceed to the event query generation stage. Additionally we ran out of available disk space and drained out our computational resources before the metadata was fully generated.

IV. Conclusions

Concerning the SIN task results, it seems that the utilization of CNN features and hard negative mining raised us to the second position in the result ranking. We indeed expected improvements in terms of both the absolute MXIAP values and the relative placement among the participants. The outcome, however, gave us a happy surprise.

Concerning the MED&MER tasks, we plan to participate again next year, but with a less ambitious selection of different features in our system. In that way, we will hopefully be able to stay within the limits of the time allocations and computational resources and to submit at least some kind of results.

Acknowledgments

This work has been funded by the grants 255745 and 251170 of the Academy of Finland and *Data to Intelligence (D2I)* SHOK projects. The calculations were performed using computer resources within the Aalto University School of Science "Science-IT" project.

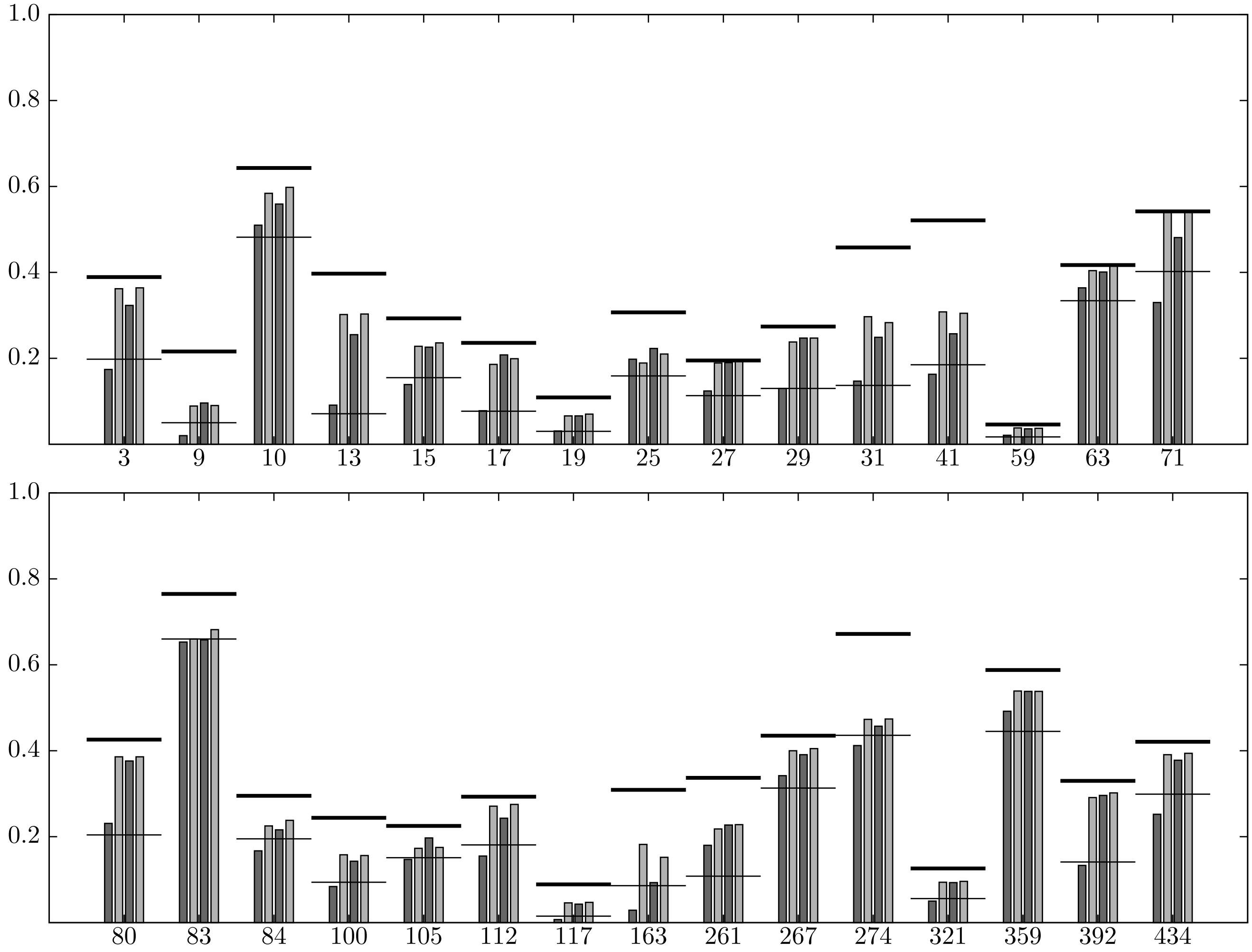

Fig. 1. The concept-wise XIAP results of our submitted runs for each evaluated concept in the semantic indexing task. The order of the runs is as in Table I, i.e. 4 MUMINPAPPAN, …, 1 MÅRRAN. The median and maximum values over all submissions are illustrated as horizontal lines.

REFERENCES

[1] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, and Georges Quénot. Trecvid 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2014*. NIST, USA, 2014.

[2] Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Erkki Oja, Ehsan Amid, Kalle Palomäki, Annamaria Mesaros, and Mikko Kurimo. PicSOM experiments in TRECVID 2013. In *Proceedings of the TRECVID 2013 Workshop*, Gaithersburg, MD, USA, November 2013.

[3] Stéphane Ayache and Georges Quénot. Video corpus annotation using active learning. In *Proceedings of 30th European Conference on Information Retrieval (ECIR'08)*, pages 187–198, Glasgow, UK, March-April 2008.

[4] Markus Koskela, Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen, and Philip Prentis. PicSOM experiments in TRECVID 2007. In *Proceedings of the TRECVID 2007 Workshop*, Gaithersburg, MD, USA, November 2007.

[5] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1 –8, June 2007.

[6] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, June 2010.

[7] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML 2014*, 2014.

[8] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. arXiv.org:1403.1840, March 2014.

[9] Markus Koskela and Jorma Laaksonen. Convolutional network features for scene recognition. In *Proceedings of the 22nd International Conference on Multimedia*, Orlando, Florida, November 2014.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoff Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[11] Matthew Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. arXiv:1311.2901, November 2013.

[12] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2010)*, 2010.

[13] Xirong Li, Cees G. M. Snoek, Marcel Worring, Dennis C. Koelma, and Arnold W. M. Smeulders. Bootstrapping visual categorization with relevant negatives. *IEEE Transactions on Multimedia*, 15(4):933–945, June 2013.

[14] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08)*, pages 603–610, 2008.