CEA LIST@imageCLEF 2013: Scalable Concept Image Annotation

Hervé Le Borgne, Adrian Popescu, and Amel Znaidia

CEA, LIST,
Vision & Content Engineering
Gif-sur-Yvettes, France
firstname.lastname@cea.fr

Abstract. We report the participation of the CEA LIST to the Scalable Concept Image Annotation Subtask of ImageCLEF 2013. The full system is based on both textual and visual similarity to each concept, that are merged by late fusion. Each image is visually represented with a bag of visterm, computed from a dense grid of SIFT every 3 pixels, that a locally soft coded and max pooled on a codebook of size 1024 and spatially extended with a pyramid $1 \times 1 + 3 \times 1 + 2 \times 2$, resulting into a vector of size 8192. The visual neighbors of a query are given by the L_1 distance to the images of the learning database. The similarity of a query to one of the 95/116 concepts to identify is the sum of the similarity of each neighbor to the concept. The decision is set at 1 for all concepts above one standard deviation from the average similarity to all concepts for the considered query.

The similarity between a training image and a concept is computed from an intermediate vectorial representation of its tags. We tested several spaces of representation for the tags, including wikipedia concepts sorted according to their popularity or characterized according FlickR data. As well, the size of space was pruned at several values, from 5000 to $200,000$. The tag representation are max-pooled to make the training image vector, such that the resulting vector contain the maximal similarity to each concept of the intermediate space. The 96/116 concepts to identify are represented in the same space and their similarity to the training image is the cosine between the intermediate space representation.

As well, we ranked all training images to each concept to identify in order to learn visual classifiers (linear SVM). We tested several strategies to select positive and negative examples, including the visual coherency, but the simplest strategies was finally the most efficient. It consisted in setting the 100 most similar images as positive and the 500 least similar as negative.

Finally, a simple weighted late fusion of the visual and textual similarity scores appeared to be more efficient than more sophisticated strategies, resulting to 0.4 MAP on the development query and 0.34 on the testing ones.

1 Introduction

The Scalable Concept Image Annotation Subtask of ImageCLEF 2013[1] is described in detail in [7]. The system we propose rely on both visual and textual cues. We conducted many preliminary experiments in order to iteratively improve the provided baseline system (see section 1.1). These experiments dealt with visual features to find image neighbors of queries (section 2), several models of tag (section 3 to 5), the way we learnt visual models (section 6) and finally the decision process (section 7).

1.1 baseline system

A baseline system based on the co-occurence of concepts and tags of the visual neighbors of each query is provided[7]. Each image I of the development set has to be annotated according to concepts $C_p, p = 1...95$. Such an image has K_v neighbors in the train test, according to a visual descriptor (csift BoV provided). Each of these neighbors ($k = 1 \dots K_v$) has T_k tags with a given score $(t_{k,1}, s_{k,1} \dots t_{k,i}, s_{k,i} \dots t_{k,T_k}, s_{k,T_k})$. Each of these tags is described with $N_{k,i}$ weighted concepts $(C_{k,i,1}, W_{k,i,1} \dots C_{k,i,j}, W_{k,i,j} \dots C_{k,i,N_{k,i}}, W_{k,i,N_{k,i}})$. Thus, the score of concept C_p for image I is:

$$Score_I(C_p) = \frac{1}{K_v} \sum_{k=1}^{K_v} \frac{\sum_{i=1}^{T_k} s_{k,i} W_{k,i,C_p}}{\sum_{i=1}^{T_k} s_{k,i}} \tag{1}$$

In practice, each tag is described by the same number $K_{concepts}$ of concepts (default: 6).

2 Searching visual neighbors

In the original system, visual neighbors are provided and said to be found with a C-SIFT based descriptor. We tested several alternative methods.

Descriptors are bag of visterms. SIFT local descriptors are densely extracted every 3 pixels. The bag are coded using local soft coding [3] and max pooling. Then two different spatial pyramid matching scheme [2] are used: $1 \times 1 + 3 + 2 \times 2$ for BoV_1 and $1 \times 1 + 2 \times 24 \times 4$ for BoV_2. Further details on bag-of-visterm design can be found in [6]. Several distances were tested on these vectors to find the neighbours. The histogram intersection (HI) distance implemented as:

$$Dist_{HI}(x - y) = 1 - \frac{1}{D} \sum_{i=1}^{D} \frac{min(x_i, y_i)}{max(x_i, y_i)} \tag{2}$$

and the classical $L1$ distance:

$$Dist_{L1}(x - y) = \frac{1}{D} \sum_{i=1}^{D} |x_i - y_i| \tag{3}$$

Results are shown in table 1, showing a non-significant improvement with the BoV_1 signature and the $L1$ distance.

System	mAP
K	32
Provided baseline	24.235
Random neighbors	17.878
BoV_1 HI	23.830
BoV_2 HI	23.468
BoV_1 L1	24.305
BoV_2 L1	23.229

Table 1. Result of the baseline system with several methods to search the $K = 32$ visual neighbors

3 Using a FlickR-based tag model

We used a FlickR-based tag model built from the selection of the 95 concepts ($Flickr_{95}$) and another one built from $30,000$ wikipedia concept ($Flickr_{30k}$). See [5, 8] for details about the way similarities are computed for these models.Note that both models were built from the FlickR tags. The $Flickr_{95}$ tag model was injected into the system provided, in conjunction with two different visual models. The mAP is reported in table 2. Note this performance measure should be independant from the parameter $K_{concepts}$ that was fixed to 6. The F-measure

	K_{visual}	8	16	32	64	128
Tag	Visual					
co-occurrence	csift	24.71(*)	24.77	24.24	23.63	23.10
co-occurrence	$BoV_1 + L1$	25.01(*)	25.08	24.31	23.60	22.80
$FlickR_{95}$	csift	25.08	25.92	26.61	27.33	27.51
$FlickR_{95}$	$BoV_1 + L1$	25.96	27.30	28.16	28.18	27.67
$Flickr_{30k}$	csift	30.25	29.46	29.23	28.80	28.44
$Flickr_{30k}$	$BoV_1 + L1$	31.05	30.25	29.50	29.07	28.48

Table 2. Result (mAP) of the system with different tag models and method to search the visual neighbors. (*) the mAP grows with $K_{concepts}$ here; result reported with $K_{concepts} = 6$

only uses the annotation decisions and is computed in two ways, one by analyzing each of the testing samples and the other by analyzing each of the concepts. Results are reported on table 3 and 4.

4 Window-restricted FlickR-based tag models

The tag model of each training document is built with a restriction of the word-image distance. In the original web page a training image has been found, the method consists in taking into account words that are less than a given distance

Tag	Visual	K_{visual} / $K_{concepts}$	8	16	32	64	128
co-occurrence	csift	2	11.92	12.49	11.54	10.81	10.32
		4	15.53	16.48	16.60	16.37	15.77
		6	17.90	18.96	18.60	17.88	17.54
		8	19.27	19.71	19.61	19.03	18.65
		10	20.09	20.14	19.86	19.59	19.23
co-occurrence	$BoV_1 + L1$	2	12.46	13.00	12.20	11.64	10.47
		4	16.21	16.78	16.33	15.48	14.96
		6	17.76	18.43	18.20	17.81	17.38
		8	19.07	19.60	19.29	19.06	18.85
		10	19.83	20.19	19.96	19.63	18.97
$FlickR_{95}$	csift	2	13.96	14.82	15.31	16.39	15.76
		4	16.56	17.47	18.38	18.98	18.93
		6	17.67	18.49	19.67	19.98	19.98
		8	18.19	19.17	19.82	20.48	20.72
		10	18.65	19.36	19.96	20.61	21.29
$FlickR_{95}$	$BoV_1 + L1$	2	14.83	15.72	16.66	16.25	15.80
		4	17.48	18.68	19.36	19.44	19.07
		6	18.64	20.04	20.73	21.01	20.90
		8	19.28	20.33	21.19	21.37	21.18
		10	19.38	20.51	21.20	21.68	21.68

Table 3. Result (mFsamp) of the system with different tag models and method to search the visual neighbors.

from the considered image. Moreover, we considered a lemmatized and non-lemmatized version of the models.

Results are comparable to those obtained with $Flickr_{30k}$ (around 0.31) but no improvement is actually observed.

5 Wikipedia-based tag models

Similarly to the FlickR-based tag models, tags are projected onto 1187980 wikipedia concepts. The concepts being ranked to the numbers of their incoming links, the representation can be pruned to a lower dimension.

6 Learning visual models

For a given tag-model, training images are sorted according to their score for each concept. Then we select positive and negative examples according to different strategies to learn linear SVM models from the BoV_1 signatures. Text model used was $Flickr^w_{30k}0$ i.e the FlickR tags projected on the $30k$ wikipedia concepts, with restriction a window of size 0.

Tag	Visual	K_{visual} / $K_{concepts}$	8	16	32	64	128
co-occurrence	csift	2	7.25	5.49	4.33	3.60	2.91
		4	11.33	9.13	8.47	7.81	6.11
		6	13.98	12.48	10.67	9.41	8.00
		8	15.34	13.17	11.87	10.46	9.39
		10	16.18	14.15	12.67	11.72	10.43
co-occurrence	$BoV_1 + L1$	2	8.28	7.95	6.01	5.24	3.00
		4	12.26	11.49	9.15	7.87	7.12
		6	14.40	13.14	11.62	10.38	8.95
		8	15.42	14.45	13.72	11.93	10.34
		10	16.21	15.70	14.40	13.03	11.18
$FlickR_{95}$	csift	2	15.11	15.74	15.35	15.39	13.79
		4	16.10	17.20	17.95	18.13	17.26
		6	16.30	17.54	18.91	19.34	18.78
		8	16.26	17.75	18.41	19.43	19.89
		10	16.35	17.61	17.96	19.16	20.53
$FlickR_{95}$	$BoV_1 + L1$	2	17.00	17.17	17.15	15.61	14.42
		4	18.64	19.08	19.16	17.98	17.05
		6	18.62	19.28	20.47	19.56	18.85
		8	18.25	19.33	20.56	20.00	18.85
		10	17.88	18.99	20.34	19.97	20.06

Table 4. Result (mFcnpt) of the system with different tag models and method to search the visual neighbors.

The simplest strategy consisted in setting the 100 most similar images as positive and the 500 least similar as negative. It leaded to a mAP of 0.219 on the devel queries.

A second strategy consisted to select as positive all images above a given score (0.8) and as negative all those below a smaller threshold (0.1). Given that classes were then strongly ubalanced, we limited negative image to nine times the positive on for each SVM model, leading to a mAP of 0.207. When the number of negative samples are forced to be equal to the positive ones, the mAP is 0.212.

A last strategy was tested, for wich the choice of the images was based on the visual coherency [4]. The 1000 most similar images to each concept are resorted according to their VCscore and the 100 best are thus selective as positive examples. Negative examples are chosen as the 1000 least similar images to the concept. Although promising, this approach leaded to a mAP of 0.209 only.

We thus finally decided to keep the first and simplest strategy.

7 Decision

The decision value (0/1) is taken independently on each query, according to its similarity to the 95 or 116 concepts. We compute the average (μ) and the

standard deviation (σ) of the scores and set the decision to 1 for all concepts having a score above $\mu + \sigma$.

8 Participation to the campaign

8.1 Submitted runs

We submitted five runs to the campaign, based on the obsvation made during preliminary experiments:

Run 1: we computed the FlickR-based tag model and merged it with the visual similarities. The weights were respectively 0.8 and 0.2.

Run 2: we added to Run 1 a Wikipedia-based tag model with a representation pruned to $5,000$ dimensions.

Run 3: we added to Run 2 a FlickR-based tag model with a representation pruned to $50,000$ dimensions.

Run 4: similar to Run 3 with a visual model selected on scores

Run 5: similar to Run 4 with a FlickR-based tag model with a representation pruned to $200,000$ dimensions.

8.2 Results

	devel set			test set		
	mAP	MF-sample	MF-concepts	mAP	MF-sample	MF-concepts
Run 1	34.6	28.7	23.6	29.4	23.0	19.0
Run 2	39.6	30.2	24.6	33.6	24.2	20.1
Run 3	40.4	31.8	25.3	34.1	25.2	20.2
Run 4	40.3	32.2	26.1	34.2	26.0	21.2
Run 5	39.2	31.6	25.4	33.6	25.7	21.0

Table 5. Result of our five runs to the campaign.

Results are quite close from each other and around 10 points (in term of mAP) below the best run of the campaign.

References

1. B. Caputo, H. Muller, B. Thomee, M. Villegas, R. Paredes, D. Zellhofer, H. Goeau, A. Joly, P. Bonnet, J. Martinez Gomez, I. Garcia Varea, and M. Cazorla. Imageclef 2013: the vision, the data and the open challenges. In *Proc CLEF 2013, LNCS*, 2013.
2. Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

3. Lingqiao Liu, Lei Wang, and Xinwang Liu. In Defense of Soft-assignment Coding. In *IEEE International Conference on Computer Vision*, 2011.
4. Débora Myoupo, Adrian Popescu, Hervé Le Borgne, and Pierre-Alain Moëllic. Multimodal image retrieval over a large database. In *Proceedings of the 10th international conference on Cross-language evaluation forum: multimedia experiments*, CLEF'09, pages 177–184, Berlin, Heidelberg, 2010. Springer-Verlag.
5. Adrian Popescu and Gregory Grefenstette. Social media driven image retrieval. In *ACM International Conference on Multimedia Retrieval*, pages 33:1–33:8, 2011.
6. Aymen Shabou and Hervé Le Borgne. Locality-constrained and spatially regularized coding for scene categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3625, 2012.
7. Mauricio Villegas, Roberto Paredes, and Bart Thomee. Overview of the imageclef 2013 scalable concept image annotation subtask. In *CLEF 2013 working notes*, 2013.
8. Amel Znaidia, Aymen Shabou, Adrian Popescu, Hervé Le Borgne, and Céline Hudelot. Multimodal feature generation framework for semantic image classification. In *ICMR*, International Conference on Multimedia Retrieval, ICMR '12, Hong Kong, China, June 5-8, 2012, page 38, 2012.