CNRS TELECOM ParisTech at ImageCLEF 2016 Scalable Concept Image Annotation Task: Overcoming the Scarcity of Training Data

Hichem SAHBI

CNRS TELECOM ParisTech, Paris-Saclay University
46 rue Barrault, 75013 Paris, France
hichem.sahbi@telecom-paristech.fr

Abstract. We introduce our participation at the ImageCLEF 2016 scalable concept detection and localization task. As in ImageCLEF 2015, this edition focuses on generating not only annotations (concept detection) but also localizing concepts into a large image collection. In our runs, we focus mainly on concept detection; our solution is purely visual and based on deep features combined with standard linear support vector machines (SVMs) built on top of well enriched training sets. Starting from loosely labeled training sets, we propose an algorithm that learns the statistical dependencies between concepts and allows us to enrich the labels of these training sets, resulting into more effective SVMs for image annotation.

Keywords: label enrichment, SVMs, deep learning, image annotation

1 Introduction

Automatic image annotation is one of the major challenges in computer vision and machine learning. It consists in learning intricate relationships between keywords (a.k.a concepts/labels/categories) and training images, in order to assign list of keywords to newly observed visual contents (see for instance [1–5]). These concepts may either correspond to well defined physical entities (pedestrians, cars, etc.) or to high level, fine-grained notions resulting from the interaction of many entities into scenes (parties, fights, etc.). In both cases, image annotation is challenging due to the perplexity when assigning concepts to scenes especially when the number of possible concepts is taken from a large vocabulary, when training data are scarce and also when analyzing highly semantic and variable content.

Early image annotation techniques are content-based (e.g. [6–9]). They model straightforward "concept-image" relationships and learn how to assign concepts to new images; they first describe image observations using visual features[1], treat each concept as an independent class, and then train the corresponding

[1] either handcrafted such as color, texture, etc. or learned such as deep features [10]

concept-specific classifier to identify (separately) images belonging to that concept using a variety of machine learning and inference techniques, either generative or discriminative [11–16, 9, 17–24, 10, 25–36]. Extensions of these methods achieve structured output predictions [37, 38] by modeling not only "concept-image" relationships, but also "concept-concept" dependencies [39–42]. Indeed, concepts in image annotation are usually interdependent, i.e., the presence of one concept *may tell us* something about the presence of another one; for instance the presence of the concept "sea" usually implies the presence of other concepts such as "sky" or "sand". Hence modeling the statistical dependencies between concepts (both for training and inference) is crucial and this is usually achieved with graphical models and markov/conditional random fields [16]. Relationships between concepts can also be modeled by extracting (hand-crafted or learned) mid-level characteristics which are common to different concepts. This has recently received a particular attention in the context of deep networks and transfer learning [43, 10, 44]. However, the lack of labeled data may severely limit the usability of these methods and requires solutions in order to learn from few shots. Hence, learning from common data and characteristics is valuable in order to overcome the scarcity of training data especially when handling image annotation problems with a large number of concepts.

In this paper, we describe the participation of "CNRS-TELECOM ParisTech" at the ImageCLEF 2016 Scalable Concept Image Annotation Task [45, 46]. Our solution focuses mainly on concept detection; it combines effective deep features with SVM classifiers. As training data are scarce, we propose a solution that enriches the labels of these training data. This solution is based on measuring the statistical correlation between concepts in the training set and makes it possible to propagate labels to larger training sets. Note that our solution does not require the use of the meta-data associated to training and test data; indeed it is purely visual. In spite of this, the proposed runs are competitive.

2 Our Concept Prediction and Localization at a Glance

Our concept detection and localization results are obtained according to the two following steps:

i) Holistic concept detection: this step is achieved using global (holistic) visual features. For that purpose, we train "one versus all" SVMs for each concept, in order to detect whether that concept exists in a given test image (see extra details in Section 4.2).

ii) Blind concept localization: concept localization is achieved *blindly*, i.e., without observing the content of a given test image. In contrast to our last year participation [47], we did not investigate heuristics for concept localization and we use the whole image dimensions as bounding boxes; this turns out to be sufficient for many concepts as discussed in Section 4.2. So in this participation,

Fig. 1. (Top) Sample of pictures taken from the ImageCLEF2016 database (dev set). (Bottom) Sample of external pictures collected from the web; the leftmost picture belongs to the category "beach", while the middle and rightmost pictures belong to the categories "anchor" and "apple" respectively. It is clear that these pictures can also be assigned to categories ("sea", "sand", "cloud"), ("boat"), ("tree"), etc. as these concepts are highly correlated with the concepts "beach", "anchor" and "apple" respectively. So these pictures can be reused to train the classifiers of these concepts.

we focus on the first step only (i.e., Holistic concept detection) and mainly issues about enriching training datasets.

3 Training Datasets and Label Enrichment

Besides training data provided in ImageCLEF 2016, we collected automatically an *external* training set using the "googlebot-image" crawler. This external set consists in 42,272 images belonging to the 251 concepts of ImageCLEF. No post processing of these images was achieved (the whole content is used in order to train our SVM models and without localizing the concepts in images). Figs. 1, 2 show a sample of those images as well as the distribution of the number of images per concept. We also use the 2,000 images of the dev set provided by the ImageCLEF 2016 organizers as an *internal* training set in order to train and tune the parameters of our SVM models. All images are described using the coefficients of the FC7 layer of the pre-trained VGG network in [48].

Label enrichment As some concepts are rare, we use the 2,000 images of the dev set in order to enrich the labels of all the training set. The idea is to transfer the knowledge about the co-occurrence of some labels using a simple principle: *given two concepts c and c', if c, c' are highly correlated, then the presence of one of these two concepts in a given training image implies the presence of the other concept.*

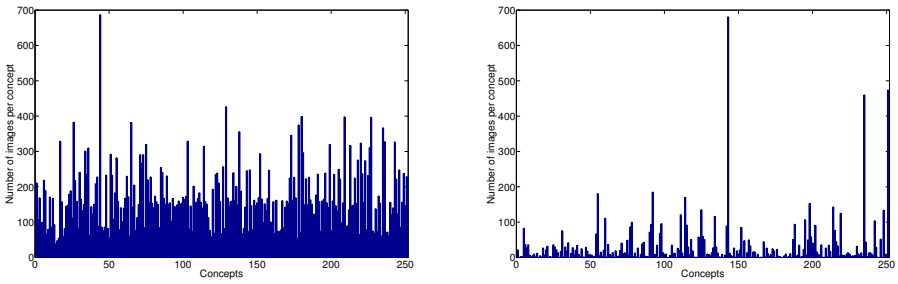

Fig. 2. Number of images per concept on the external training set (left) and on the ImageCLEF16 dev set (right).

In order to implement this principle, we define the asymmetric co-occurrence between two concepts as follows

$$\mathbf{C}(c'|c) = \frac{\sum_{i=1}^{N} \mathbf{Y}_{ic}\mathbf{Y}_{ic'}}{\sum_{i=1}^{N} \mathbf{Y}_{ic}}, \qquad c, c' = 1 \dots K, \tag{1}$$

here N is the size of the dev set and K is the number of concepts ($N = 2,000$, $K = 251$ in practice) and $\mathbf{Y} \in \mathbb{R}^{N \times K}$ is a matrix whose entry $\mathbf{Y}_{ic} = 1$ iff the concept c is present into image $\mathcal{I}_i$ and $\mathbf{Y}_{ic} = 0$ otherwise. As external images are collected using "individual keywords" as queries, they have a single label per image and cannot be used to learn these co-occurrences. In contrast, dev set images have multiple labels and are used instead. Hence, labels in the external set are enriched as follows $\forall c, c' \in \{1, \dots, K\}$, $\forall i \in \{N+1, \dots, N+N'\}$ ($N' = 42,272$), if $\mathbf{Y}_{ic} = 1$ and $\mathbf{C}(c'|c) \geq \sigma$ then $\mathbf{Y}_{ic'} \leftarrow 1$ (see Section 4 about the tuning of σ).
Fig (1, bottom) shows an example of this label enrichment process, where the presence of concept "apple" implies the presence of the concept "tree". Note that this enrichment process could also be achieved as a post processing step (i.e., after image annotation), however due to shortage of time, this issue has not been investigated.

4 ImageCLEF 2016 Evaluation

The targeted task is, again, concept detection and localization: given a picture, the goal is to predict which concepts (classes) are present into that picture and the coordinates of the bounding boxes surrounding these concepts.

4.1 ImageCLEF 2016 Collection

A very large amount of images was gathered by the organizers, and using associated web pages, tags and meta-data were also provided. This set includes more

than 500k images with only 2k images with known ground truth (i.e., labels and bounding boxes are given). These images belong to 251 concepts (see example in Fig. 1). In our runs, each image is again described with a visual feature vector corresponding to the FC7 layer of the VGG pretrained network. Note that the parameters of this network are not fine-tuned on training data and concepts.

4.2 Submitted Runs

All our submitted runs (discussed below) are based on SVM training. For each concept, we trained "one-versus-all" SVM classifiers; we use many random folds (taken from training data) for multiple SVM training and we use these SVMs in order to predict the concepts on the test set. We repeat this training process, for each concept, through different random folds from the training set and we take the average scores of the underlying SVM classifiers. This makes classification results less sensitive to the sampling of the training set. Given a test image x, a concept c is declared as present into x iff $f_c(x) > \tau$, here $f_c(x) = \frac{1}{L} 1_{\{g_\ell(x)>0\}}$ and $g_\ell()$ is an SVM classifier trained on a random fold of positive and negative data (in practice $L = 10$; see also Tab. 1 for the setting of τ).

Our ten submitted runs correspond to the combination of five dataset enrichment strategies (see columns of Tab. 1 and section 3) and two datasets used for SVM training (external and ImageCLEF16/dev set). For all the submitted runs, performances are evaluated, by the organizers, using a variant of the Jaccard measure; the latter is defined as the intersection over union of bounding boxes provided in the submitted runs and those in the ground truth. Mean average precision (MAP) measures – based on different percentages of bounding box overlaps – are given for each concept and also averaged through different concepts (see our results in Tables 2, 3, 4, 5, 6). Details about these measures can be found in the ImageCLEF 2016 website[2]. In contrast to our last year participation, we do not address the issue of bounding box (BB) generation; our bounding boxes cover the whole areas of the test images. We expect further improvement of performances if we consider the BB generation heuristics used last year (as already shown in [47]).

From all these tables, we observe the following issues:

- For all the runs shown in table 2, we observe that combining external data with the ImageCLEF16 dev set provides a clear gain compared to the use of external data only; this may be explained by the fact that the ImageCLEF16 dev set has (possibly) a similar distribution compared to ImageCLEF16 test set, and this makes it possible to adapt training parameters (mainly the SVM weights) to the conditions of the test data. In contrast, the use of external data only does not allow to adapt these SVM parameters appropriately.

[2] http://www.imageclef.org/2016/annotation#Results

Table 1. Different runs as submitted to the ImageCLEF 2016 Challenge. RA stands for row adaptive, i.e., the threshold is set for each test image in order to guarantee that the maximum number of detections per image is ≤ 100.

Datasets \ Enrichment	No	Yes	Yes	Yes	Yes
		$\sigma = 0.01$	$\sigma = 0.01$	$\sigma = 0.1$	$\sigma = 0.75$
	$\tau = 0.00$	τ (RA)	$\tau = 8.00$	$\tau = 0.00$	$\tau = 0.00$
External	TAB.0.1.res	TAB.0.4.res	TAB.0.4.1.res	TAB.0.3.res	TAB.0.5.res
External +ImageCLEF16 (dev)	TAB.1.1.res	TAB.1.4.res	TAB.1.4.1.res	TAB.1.3.res	TAB.1.5.res

- Tables 3-6 show a subset of concepts whose performances improve after the enrichment process w.r.t the baseline (i.e., "TAB.1.1.res" vs. "TAB.1.3.res" in table 3 and "TAB.1.1.res" vs. "TAB.1.5.res" in table 4); again, and as already described in table 1, "TAB.1.1.res" is a baseline run that uses external and ImageCLEF16 data without enrichment while runs "TAB.1.3.res" and "TAB.1.5.res" rely on the enrichment process. From all these tables we observe a clear gain in performance especially for concepts with a high correlation factor[3]. This is predictable as concepts with high correlation factors (such as "arm", "shirt", "shoe" in tables 3, 4) co-occur with many other concepts and hence inherit larger training subsets. Some concepts even with small correlation factors (such as "apron", "cup") also benefit from the enrichment process, with a relatively smaller gain.

- The same behavior also occurs when considering the performance with 50% overlap (see tables 5, 6). We also notice that concepts which are usually centered in pictures (such as "motorcycle", "kitchen", "shirt") are relatively well localized using our simple blind localization. Other difficult concepts (such as "cat" in table 6) get substantial improvement. It is also clear that better concept detection implies better localization results (see again tables 3, 4 vs. tables 5, 6).

- From all these results it is clear that the run "TAB.1.5.res" is better than the run "TAB.1.3.res" as the former is more conservative (i.e., threshold σ is relatively high) while the latter is less conservative and benefits from larger training sets. Finally, figures 3, 4 show the concepts for which we obtained the best results among different participants in ImageCLEF16.

5 Conclusion

We discussed in this paper, our participation at the ImageCLEF 2016 Scalable Concept Image Annotation Task. In our runs, concept detection is based on deep

[3] The correlation factor of a concept is defined as $F_\sigma(c) = \sum_{c'=1}^{K} 1_{\{\mathbf{C}(c'|c) \geq \sigma\}}$.

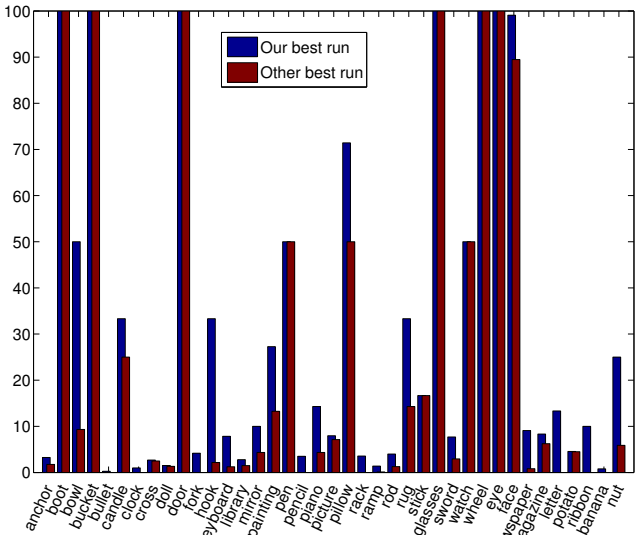

Fig. 3. This figure shows the concepts for which we outperform other participants' runs (blue bars: our best performances, red bars: other participants' best performances on these concepts). These performances correspond to 0 % overlap.

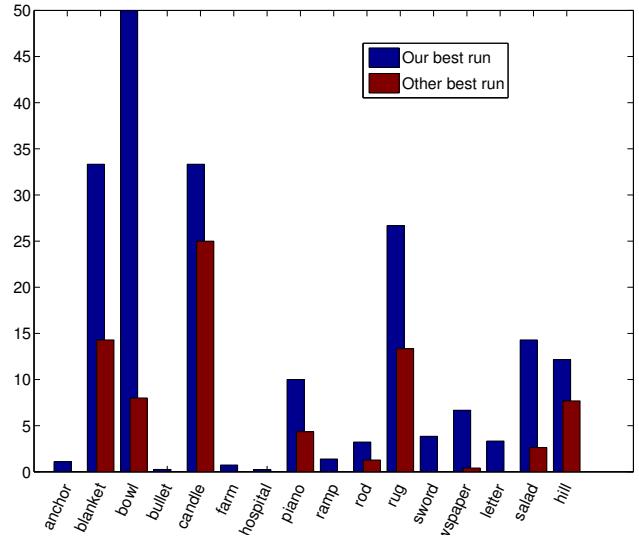

Fig. 4. This figure shows the concepts for which we outperform other participants' runs (blue bars: our best performances, red bars: other participants' best performances on these concepts). These performances correspond to 50 % overlap.

Table 2. Performances (in %) of our different concept detection and localization results (taken from ImageCLEF 2016 results).

Runs \ Overlap	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100 %
External Only											
CNRS/TAB.0.1.res	**19.62**	15.67	12.01	9.78	8.13	6.73	5.77	4.83	3.86	2.81	1.65
CNRS/TAB.0.5.res	19.39	**15.89**	**12.38**	**10.08**	**8.39**	**6.88**	**5.90**	**4.95**	3.90	2.75	1.64
CNRS/TAB.0.3.res	17.31	14.27	11.53	9.53	8.00	6.77	5.89	4.93	**3.97**	**2.84**	**1.81**
CNRS/TAB.0.4.res	10.59	7.43	5.71	4.64	3.73	3.05	2.67	2.25	1.87	1.37	0.91
CNRS/TAB.0.4.1.res	10.25	7.12	5.41	4.33	3.48	2.85	2.49	2.10	1.72	1.25	0.82
External + CLEF16 dev											
CNRS/TAB.1.1.res	**24.75**	**21.89**	**18.32**	**15.14**	**12.83**	**11.11**	**9.62**	**7.71**	**6.13**	**4.13**	**2.42**
CNRS/TAB.1.5.res	21.53	19.44	16.48	13.66	11.56	9.96	8.66	6.85	5.41	3.38	2.06
CNRS/TAB.1.3.res	16.85	13.72	10.98	9.06	7.58	6.36	5.63	4.74	3.79	2.67	1.64
CNRS/TAB.1.4.res	10.29	7.03	5.06	3.99	3.30	2.70	2.35	2.01	1.67	1.29	0.83
CNRS/TAB.1.4.1.res	9.79	6.55	4.80	3.74	3.09	2.53	2.17	1.91	1.57	1.19	0.74

features combined with linear SVMs trained on well enriched datasets. The enrichment process is based on measuring the co-occurrence of concepts and this makes it possible to reuse training images across correlated concepts. Observed results show that i) the enrichment process has a positive impact on performances especially for concepts with high correlations with others, and ii) the use of both external and provided ImageCLEF16 dev set enhances performances compared to the use of external data only; indeed, in spite of being relatively small, the provided dev set makes it possible to adapt the parameters of our SVM models to the distribution of dev and test data.

A future possible extension, of this work, is to make the enrichment process label dependent, i.e., how to mix and select different enrichment strategies for different concepts. Another possible extension is to achieve *late* label enrichment, as a post processing step, by augmenting annotation results on the test set using the same label enrichment strategy.

Acknowledgments. This work was supported in part by a grant from the Research Agency ANR (Agence Nationale de la Recherche) under the MLVIS project ANR-11-BS02-0017.

Table 3. This table shows for each concept, its correlation factor, the size of the initial training set, the size of the enriched set, the performance before enrichment (i.e., run "TAB.1.1.res") and after enrichment (i.e., run "TAB.1.3.res"). Again, for run "TAB.1.3.res", $\sigma = 0.1$ (see table 1). All these performances correspond to 0% overlap.

concepts	$F_\sigma(.)$	# initial set	# enriched set	perfs % (before enrichment)	perfs % (after enrichment)
butterfly	50	51	7398	0	2.17
barn	78	180	13007	0	0.18
bullet	19	112	2180	0	0.26
cup	6	80	306	0	0.33
fork	1	150	150	0	0.36
hospital	57	178	10894	0.43	0.57
keyboard	25	177	3515	3.70	3.82
mat	1	79	79	0	1.64
mirror	70	304	20072	2.04	6.67
pencil	4	194	519	0	1.79
shirt	**36**	**188**	**5939**	**48.98**	**58.97**
shoe	**76**	**356**	**20377**	**20.00**	**23.40**
sock	3	357	357	13.33	14.00
vase	22	171	3624	0	8.57
vest	1	75	75	3.70	6.06
tongue	43	181	6265	0	1.06
mouth	1	227	227	50.00	73.91
neck	1	72	72	64.29	72.16
foot	11	302	1541	12.50	18.69
arm	**44**	**159**	**7199**	**61.11**	**81.94**
magazine	19	184	2598	1.06	8.33
apple	16	246	2138	0	1.06
orange	95	379	17030	0	0.38
salad	64	199	11531	10.00	10.53
canal	50	125	7492	0.40	1.43
nut	30	85	4318	3.57	6.67

Table 4. This table shows for each concept, its correlation factor, the size of the initial training set, the size of the enriched set, the performance before enrichment (i.e., run "TAB.1.1.res") and after enrichment (i.e., run "TAB.1.5.res"). Again, for run "TAB.1.5.res", $\sigma = 0.75$ (see table 1). All these performances correspond to 0% overlap.

concepts	$F_\sigma(.)$	# initial set	# enriched set	perfs % (before enrichment)	perfs % (after enrichment)
wolf	8	103	1325	2.08	4.55
deer	3	149	149	47.27	50.00
airplane	8	413	644	47.22	48.89
apron	**9**	**173**	**318**	**8.00**	**12.50**
basket	1	310	310	2.63	4.17
bathtub	13	172	2794	8.33	14.29
cabinet	9	463	7042	25.93	26.47
cap	9	251	9625	8.14	8.33
computer	1	142	142	14.63	17.65
cup	**6**	**80**	**80**	**0**	**13.04**
drum	7	369	1405	8.51	16.28
farm	15	244	6557	0.57	0.74
flag	7	155	391	8.70	9.86
fork	1	150	150	0	4.17
helmet	**32**	**176**	**1156**	**11.11**	**18.80**
keyboard	25	177	770	3.70	6.67
kitchen	**8**	**174**	**433**	**10.71**	**17.65**
mask	6	214	939	1.75	1.82
mat	1	79	79	0	0.80
microphone	**8**	**187**	**769**	**10.00**	**21.43**
mirror	12	304	8928	2.04	3.39
motorcycle	**8**	**170**	**340**	**43.33**	**75.00**
painting	**10**	121	121	10.53	16.67
pencil	4	194	194	0	2.00
piano	**7**	**70**	**469**	**7.41**	**14.29**
ramp	17	308	4143	0.82	1.18
shirt	**10**	**188**	**1098**	**48.98**	**58.77**
stadium	1	68	68	28.00	35.29
sword	14	106	705	3.23	3.70
toilet	1	152	152	25.00	27.27
towel	19	148	148	9.09	13.64
tractor	13	280	2219	12.90	15.38
tray	**18**	**91**	**650**	**0**	**10.00**
vase	22	171	1134	0	3.03
vest	1	75	75	3.70	7.09
wall	**2**	**79**	**248**	**66.67**	**95.00**
mouth	1	227	227	50.00	75.36
foot	**11**	**302**	**550**	**12.50**	**32.76**
arm	**18**	**159**	**2361**	**61.11**	**77.94**
newspaper	1	163	163	6.67	9.09
book	8	126	126	10.00	12.50
salad	16	199	2586	10.00	10.53
beer	19	141	195	9.09	16.67
beach	17	70	239	4.00	4.17
ribbon	6	331	331	3.23	10.00
valley	2	313	313	17.74	21.43
male_child	**26**	**75**	**1152**	**11.78**	**20.00**

Table 5. This table shows for each concept, its correlation factor, the size of the initial training set, the size of the enriched set, the performance before enrichment (i.e., run "TAB.1.1.res") and after enrichment (i.e., run "TAB.1.3.res"). Again, for run "TAB.1.3.res", $\sigma = 0.1$ (see table 1). All these performances correspond to 50% overlap.

concepts	$F_\sigma(.)$	# initial set	# enriched set	perfs % (before enrichment)	perfs % (after enrichment)
ball	56	241	9539	0	0.13
barn	78	180	13007	0	0.18
basket	1	310	310	0	0.40
bottle	76	115	12848	0	0.45
box	30	90	3371	0	0.68
bullet	19	112	2180	0	0.26
cap	82	251	22114	0	0.08
flag	27	155	4162	0	0.10
helmet	32	176	4923	0	0.28
ladder	4	330	786	0	0.19
mat	1	79	79	0	0.14
microphone	36	187	6117	0	0.24
necktie	94	485	24349	0	0.13
pillow	53	297	9049	0	2.38
scarf	16	151	2087	0.61	0.87
shirt	**36**	**188**	**5939**	**12.24**	**16.24**
shoe	76	356	20377	0	1.42
stick	122	232	21878	0	0.28
toilet	1	152	152	0	0.70
towel	39	148	6440	0	0.29
vest	1	75	75	0	0.76
wheel	45	169	8093	0	0.33
eye	33	192	4424	0	0.12
face	32	238	4769	2.78	3.95
radio	39	202	6649	0	0.20
book	8	126	742	0	0.12
letter	71	271	12167	0	3.23
wine	29	256	4209	0	0.35
canal	50	125	7492	0.40	1.43
femalechild	12	79	1232	2.30	2.48

Table 6. This table shows for each concept, its correlation factor, the size of the initial training set, the size of the enriched set, the performance before enrichment (i.e., run "TAB.1.1.res") and after enrichment (i.e., run "TAB.1.5.res"). Again, for run "TAB.1.5.res", $\sigma = 0.75$ (see table 1). All these performances correspond to 50% overlap.

concepts	$F_\sigma(.)$	# initial set	# enriched set	perfs % (before enrichment)	perfs % (after enrichment)
cat	**8**	**173**	**679**	**13.64**	**22.22**
deer	3	149	149	21.82	22.92
fish	**10**	**148**	**1562**	**6.25**	**18.18**
airplane	8	413	644	40.28	46.67
apron	**9**	**173**	**318**	**4.00**	**12.50**
basket	1	310	310	0	0.51
bathtub	13	172	2794	8.33	14.29
bottle	28	115	5489	0	0.45
box	30	90	338	0	3.12
computer	1	142	142	3.66	3.92
cup	6	80	80	0	0.39
drum	7	369	1405	4.26	6.98
farm	15	244	6557	0.57	0.74
flag	7	155	391	0	1.41
helmet	32	176	1156	0	0.85
kitchen	**8**	**174**	**433**	**10.71**	**17.65**
motorcycle	**8**	**170**	**340**	**43.33**	**75.00**
necktie	9	485	15719	0	0.31
picture	6	236	319	2.17	2.70
pillow	8	297	2023	0	0.90
ramp	17	308	4143	0.82	1.18
scarf	16	151	1018	0.61	1.32
shirt	**10**	**188**	**1098**	**12.24**	**20.18**
shoe	8	356	6901	0	0.40
stadium	1	68	68	26.00	35.29
stick	9	232	6042	0	0.27
sword	14	106	705	3.23	3.70
toilet	1	152	152	0	9.09
tractor	13	280	2219	3.23	3.85
train	3	102	102	9.09	9.30
tray	**18**	**91**	**650**	**0**	**10.00**
vest	1	75	75	0	1.57
wheel	9	169	753	0	0.56
ear	23	58	593	0	0.38
head	10	378	378	4.94	8.06
book	8	126	126	0	0.51
letter	20	271	4560	0	0.28
beach	17	70	239	4.00	4.17
valley	2	313	313	9.68	14.29
femalechild	12	79	79	2.30	3.85
male_child	**26**	**75**	**1152**	**2.17**	**5.00**

References

1. A. Torralba, K.P. Murphy, and W.T. Freeman, "Sharing visual features for multiclass and multiview object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) vol 25, issue 5*, 2007.
2. F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 1719–1726.
3. J. Li and J. Wang, "Real-time computerized annotation of pictures," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 6, pp. 985–1002, 2008.
4. V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," *In: Proc. of NIPS*, 2004.
5. S. Moran and V. Lavrenko, "A sparse kernel relevance model for automatic image annotation," *International Journal of Multimedia Information Retrieval*, vol. 3, no. 4, pp. 209–229, 2014.
6. C. Carson, M. Thomas, S. Belongie, J.M. Hellerstein, and J. Malik, "Blobworld: A system for region-based image indexing and retrieval," *In Third International Conference on Visual Information Systems*, pp. 509–516, 1999.
7. J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *In Proceedings of the Beyond Patches workshop, in conjunction with CVPR2006*, 2006.
8. J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," *In: Proc. of ACM SIGIR, pp. 119-126*, 2003.
9. G. Carneiro, A.B. Chan, P.J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 3, pp. 394–410, 2007.
10. A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1106–1114.
11. K. Barnard, P. Duygululu, D. Forsyth, D. Blei, and M. Jordan, "Matching words and pictures," *The Journal of Machine Learning Research*, 2003.
12. H. Sahbi, "Network-dependent image annotation based on explicit context-dependent kernel maps," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 625–630.
13. D.M. Blei and M.I. Jordan, "Modeling annotated data," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, New York, NY, USA, 2003, SIGIR '03, pp. 127–134, ACM.
14. O. Yakhnenko and V. Honavar, "Annotating images and image objects using a hierarchical dirichlet process model," in *Proceedings of the 9th International Workshop on Multimedia Data Mining: held in conjunction with the ACM SIGKDD 2008*. ACM, 2008, pp. 1–7.
15. J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. on PAMI*, vol. 25, no. 9, pp. 1075–1088, 2003.
16. X. He, R.S. Zimel, and M.A. Carreira, "Multiscale conditional random fields for image labeling," *In CVPR*, 2004.
17. F. Monay and D. GaticaPerez, "Plsa-based image autoannotation: Constraining the latent space," *in Proc. of ACM International Conference on Multimedia*, 2004.

18. Y. Wang and S. Gong, "Translating topics to words for image annotation," *In: Proc. of ACM CIKM*, 2007.
19. E. Chang, K. Goh, G. Sychay, and G. Wu, "Cbsa: content-based soft annotation for multimodal image retrieval using bayes point machines," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 1, pp. 26–38, 2003.
20. H. Zhang, A.C. Berg, M. Maire, and J. Malik, "Svm-knn: Discriminative nearest neighbor classification for visual category recognition," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. IEEE, 2006, vol. 2, pp. 2126–2136.
21. M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 309–316.
22. A. Makadia, V. Pavlovic, and S. Kumar, "A new baseline for image annotation," in *Computer Vision–ECCV 2008*, pp. 316–329. Springer, 2008.
23. T. Mei, Y. Wang, X.S. Hua, S. Gong, and S. Li, "Coherent image annotation by learning semantic distance," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.
24. D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 8, pp. 1371–1384, 2008.
25. C. Cusano, G. Ciocca, and R. Schettini, "Image annotation using svm," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2003, pp. 330–338.
26. Y. Gao, J. Fan, X. Xue, and R. Jain, "Automatic image annotation by incorporating feature hierarchy and boosting to scale up svm classifiers," *in Proc. of ACM MULTIMEDIA*, 2006.
27. H. Sahbi, J.Y. Audibert, and R. Keriven, "Context-dependent kernels for object classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 4, pp. 699–708, 2011.
28. H. Sahbi, "Cnrs-telecom paristech at imageclef 2013 scalable concept image annotation task: Winning annotations with context dependent svms.," in *CLEF (Working Notes)*, 2013.
29. P. Vo and H. Sahbi, "Transductive kernel map learning and its application to image annotation," in *BMVC*, 2012, pp. 1–12.
30. H. Sahbi and X. Li, "Context based support vector machines for interconnected image annotation (the saburo tsuji best regular paper award)," in *In the Asian Conference on Computer Vision (ACCV)*, 2010.
31. M. Jiu and H. Sahbi, "Deep kernel map networks for image annotation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1571–1575.
32. M. Jiu and H. Sahbi, "Laplacian deep kernel learning for image annotation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 1551–1555.
33. H. Sahbi, "Imageclef annotation with explicit context-aware kernel maps," *International Journal of Multimedia Information Retrieval*, vol. 4, no. 2, pp. 113–128, 2015.
34. P. Vo and H. Sahbi, "Transductive inference & kernel design for object class segmentation," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 2173–2176.

35. M. Jiu and H. Sahbi, "Semi supervised deep kernel design for image annotation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 1156–1160.
36. H. Sahbi, J.Y. Audibert, J. Rabarisoa, and R. Keriven, "Context-dependent kernel design for object matching and recognition," in *Computer Vision and Pattern Recognition, CVPR*. IEEE, 2008, pp. 1–8.
37. B. Taskar, V. Chatalbashev, D. Koller, and C. Guestrin, "Learning structured prediction models: A large margin approach," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 896–903.
38. I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in *Journal of Machine Learning Research*, 2005, pp. 1453–1484.
39. S. Nowozin and C.H. Lampert, "Structured learning and prediction in computer vision," *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3–4, pp. 185–365, 2011.
40. L. Bertelli, T. Yu, D. Vu, and B. Gokturk, "Kernelized structural svm learning for supervised object segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2153–2160.
41. P. Vo and H. Sahbi, "Modeling label dependencies in kernel learning for image annotation," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 5886–5890.
42. P. Vo and H. Sahbi, "Contextual kernel map learning for scene transduction," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 3797–3801.
43. P. Duygulu, K. Barnard, J.F.G. deFreitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," *In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97-112. Springer, Heidelberg*, 2002.
44. S.J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.
45. A. Gilbert, L. Piras, J. Wang, F. Yan, A. Ramisa, E. Dellandrea, R. Gaizauskas, M. Villegas, and K. Mikolajczyk, "Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task," in *CLEF2016 Working Notes*, Évora, Portugal, September 5-8 2016, CEUR Workshop Proceedings, CEUR-WS.org ¡http://ceur-ws.org¿.
46. M. Villegas, H. Müller, A. García Seco de Herrera, R. Schaer, S. Bromuri, A. Gilbert, L. Piras, J. Wang, F. Yan, A. Ramisa, E. Dellandrea, R. Gaizauskas, K. Mikolajczyk, J. Puigcerver, A.H. Toselli, J.A. Snchez, and E. Vidal, "General Overview of ImageCLEF at the CLEF 2016 Labs," in *' '*, Lecture Notes in Computer Science. Springer International Publishing, 2016.
47. H. Sahbi, "Cnrs telecom paristech at imageclef 2015 scalable concept image annotation task: Concept detection with blind localization proposals," in *CLEF 2015 Evaluation Labs and Workshop, Online Working Notes*, 2015.
48. K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference*, 2014.