Image annotation and two paths to text illustration

Hervé Le Borgne, Etienne Gadeski, Ines Chami, Thi Quynh Nhi Tran, Youssef Tamaazousti, Alexandru Lucian Ginsca, and Adrian Popescu

CEA, LIST, Laboratory of Vision and Content Engineering, France
herve.le-borgne@cea.fr, ines.chami@cea.fr, etienne.gadeski@cea.fr, thiquynhnhi.tran@cea.fr, youssef.tamaazousti@cea.fr, alexandru.ginsca@cea.fr, adrian.popescu@cea.fr

Abstract. This paper describes our participation to the ImageCLEF 2016 scalable concept image annotation main task and Text Illustration teaser. Regarding image annotation, we focused on better localizing the detected features. For this, we identified the saliency of the image to collect a list of potential interesting places into the image. We also added a specific human attribute detector that boosted the results of the best performing team in 2015. For the text illustration, we proposed two complementary approaches. The first one relies on semantic signatures that give a textual description of an image. This description is further matched to the textual query. The second approach learns a common latent space, in which visual and textual features are directly comparable. We propose a robust description, as well as the use of an auxiliary dataset to improve retrieval. While the first approach only uses external data, the second one was mainly learned from the provided training dataset.

1 Introduction

This paper describes our participation to the ImageCLEF 2016 [27] scalable concept image annotation main task (IAL: image annotation and localization) and Text Illustration teaser that are described in detail in [5].

Regarding image annotation, we improved our 2015 system [4] and focused on better localizing the detected features. In 2015, we proposed a concept localization pipeline which uses the spatial information that CNNs offer. To improve this, we identified the saliency of the image to collect a list of potential interesting places from the image, then detected the concepts found in these boxes. We also added a specific human attribute detector that boosted the results of the best performing team in 2015.

For text illustration, we proposed two complementary approaches. The first one relies on semantic signatures that give a textual description of an image. This description is further matched to the textual query. The second approach relies on the learning of a common latent space, in which visual and textual features are directly comparable, using a robust description and an auxiliary dataset to

improve retrieval. While the first approach only uses external data, the second one was mainly learned from the provided training dataset.

This manuscript is organized as follows. Section 2 deals with our participation to the image annotation and localization subtask, while Section 3 is dedicated to the text illustration teaser. Each time, we discuss some limits of the task itself that are important to better understand the results. Then we present the method(s) and finally comment the results of the campaign.

2 Image annotation task

2.1 Dataset limitation

As highlighted last year by the team that obtained the best results [15], the development set of the image annotation subtask (and it is probably the case for the test set as well) suffers from severe limitations due to the crowd-sourcing ground-truth annotation. They explain the annotation are inconsistent, incomplete, sometimes incorrect and there are even some cases that are "impossible" according to the assumptions of the task (*e.g.* the fact there are at most 100 concepts per image).

It seems these issues have not been addressed in the 2016 development set, thus the results are still subject to some limitations. However, on the other hand, the task is thus consistent with last year's results and we can directly compare the improvement from one year to another.

2.2 Method

In this section, we detail the training and testing frameworks that we used. Our method is based upon deep CNNs which have lately shown outstanding performances in diverse computer vision tasks such as object classification, localization and action recognition [20, 7].

Training

Data. We collected a set of roughly $251,000$ images ($1,000$ images per concept) from the Bing Images search engine. For each concept we used its name and its synonyms (if present) to query the search engine. This dataset is of course noisy but some works showed it is not a big issue to train a deep CNN [6, 29]. We used this additional data to train a 16-layer CNN [21] and the 50-layer ResNet [10]. We used 90% of the dataset for training and 10% for validation.

Network Settings. The networks were initialized with ImageNet weights. The initial learning rate is set to 0.001 and the batch size is set to 256. The last layer (the classifier) is trained from scratch, *i.e.* it is initialized with random weights sampled from a Gaussian distribution ($\sigma = 0.01$ and $\mu = 0$) and its learning rate is 10 times larger than for other layers. During training, the dataset is

enhanced with random transformations: RGB jittering, scale jittering, contrast adjustment, JPEG compression and flips. It is known that data augmentation leads to better models [30] and reduces overfitting. Finally, the networks take a 224×224 RGB image as input and produce 251 outputs, *i.e.* the number of concepts. The models were trained on a single Nvidia Titan Black with our modified version of the `Caffe` framework [14].

Localizing concepts. We provide two approaches to detect the concepts and localize them.

The first method, named **FCN**, is the same as described in [4]. It is a simple and efficient framework where the concept detection and localization are done simultaneously with a unique forward pass of the image to process. More in-depth information about this framework can be found in [4].

The second method is based upon the generic object detector **EdgeBoxes** [31], which takes an image as input and produces R regions where visual objects are likely to appear (*objectness* detection). In our experiments, we extracted a maximum of 100 regions per image then fed each one to the CNN models. We finally kept the concept that had the highest probability among the 251 concepts. Therefore this framework outputs R predictions per image.

In addition to these methods, we also used a face detector [26] to categorize more precisely the faces, eyes, noses and mouths. We extracted those features on all images and aggregated the results to the boxes detected by the CNN frameworks. It was our belief that it would slightly boost our accuracy since these kind of "objects" are quite hard to capture even with a good generic object detector.

Combination of runs We also combined some runs by simply concatenating detected the boxes. When the number of boxes was above the allowed limit (100) we randomly removed some of them (this case was very rare).

2.3 Results

We submitted ten runs to the campaign, with settings that allow to measure the benefit of different choices. We studied the influence of three parameters:

- our last year's method used to localize the concepts [4] compared to the use of EdgeBoxes [31]
- the CNN architecture, by comparing VGG [21] and ResNet [10] that obtained good results at the ILSVRC campaign in 2014 and 2015.
- the use of a face part detector [26]

Results of individual runs are reported in Table 1.

On the ILSVRC challenge, VGG had 7.3% classification error and ResNet obtained 5.7%. On the 2016 ImageCLEF dataset, we obtain similar results with VGG and ResNet when we use FCN to localize the concepts and VGG is better

Table 1: Results of individual runs submitted in terms of mAP (×100) with 0.5 overlap. We report results including the facepart detection (*face*) or not (*raw*).

		FCN	EdgeBoxes
VGG	raw	-	32.16
	face	37.66	31.75
ResNet	raw	37.35	27.18
	face	37.84	24.14

Table 2: Results (mAP ×100) for two individual runs and four concepts concerned by the face part detector.

	VGG + EdgeBoxes		ResNet + FCN	
	raw	face	raw	face
mouth	100	54	2	55
eye	100	36	33	37
nose	25	28	7	42
face	67	75	50	76

with EdgeBoxes. Regarding the VGG-based scores, we noticed that our results are about 8 points better than last year, showing the benefit of the new learning process.

The use of a face part detector does not significantly improve our results and they are even lower when we use the EdgeBoxes-based localization. This is quite surprising since a similar process boosted the results of last year's best performing team from 30.39 to 65.95 [15]. However, regarding these results, we should notice this "boost" was observed on the test dataset only (on the development set, the performances were more or less the same with and without the body part detection). It is also hard to explain how the results can increase by 35 points while the body-part detector deals with less than 10 classes among 250. Following a discussion with the organizers of the campaign, it seems that there was a bug in the evaluation script (fixed since then, and probably reported on this year's overview [5]) and that detecting body parts is finally not very interesting, making our results in line with other participants' ascertainment. However, there are still some unexpected results with regards to the concepts that are directly concerned with face part detection. In Table 2, we report the results for four of these concepts and two different settings (results of ResNet+EdgeBoxes are similar to VGG+EdgeBoxes). With FCN, the behavior is in line with expectation. On the contrary with EdgeBoxes, the concepts *mouth* and *eye* are perfectly detected, that is quite unlikely. Although there are obvious issues with EdgeBoxes as explained below, this strange result may be due to a remaining bug in the evaluation script.

The most disappointing result is that the use of the EdgeBoxes-based localization gives globally lower results than the FCN one. A possible reason is that EdgeBoxes generates much more boxes than FCN and that a significant part leads to wrong concept estimation, hence penalizing the global score.

Last, we evaluated the combination of runs, as reported in Table 3. Once again, the results are quite disappointing since the more we combine, the lower the results are. Since the combination of runs is a simple concatenation of the boxes found by each run, it is not clear to us how the results can decrease (the mAP should be at least better than the weaker run). It probably results from the way the results are evaluated but unfortunately the exact method used is not available.

Table 3: Results for three combinations of runs.

Methods combined	mAP (0.5 overlap) $\times 100$
All VGG-based	32.76
All ResNet-based	27.11
All (both above)	21.7

3 Text illustration

3.1 Task realism

The task consists of matching textual queries to images *without* using the textual features derived from the web page the images belong to, although these last ones are available since they are part of the $500k$ noisy dataset. In practice, the queries were mainly obtained by removing the HTML tags from these web pages, and retaining all the remaining text. It thus raises an issue with regards to the realism of the task.

Indeed, when one wants to illustrate a text in practice, she would submit the *interesting part* of the text as query to the system. It does not make sense to add some noisy data in the query such as that coming from the generic task bar as in the query `--/--diUdSrlGyv7zF4` that starts with:

> Taakbalk Navigation Subnavigation Content home Who is who organisational chart contact intranet nederlands zoekterm: Navigatie About K.U.Leuven Education Research Admissions Living in Leuven Alumni Libraries Faculties, Departments & Schools International cooperation Virology - home Current Labmembers Former Labmembers Research Projects Publications Contact Us Where To Find Us Courses (...)

Of course, it is hard for the organizers to extract this "relevant text" at a large scale, since there are $180,000$ queries. However, this could be part of the task: if the query was the actual HTML page, the system could include an automatic search of the relevant text by using the DOM structure of the query.

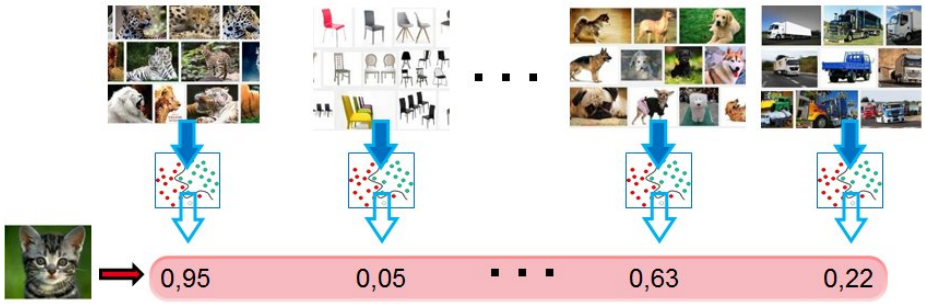


Fig. 1: The *semantic signature* principle: an image is described in terms of likelihood to be similar to some concepts.

3.2 Semantic signature approach

Specific object detectors have been developed for a long time, to be able to recognize *e.g.* faces [28], pedestrians [3] or buildings [18]. More recently, it has been proposed to use a set of object or concept detectors as image descriptors [24, 17], introducing the "semantic features". With this approach, images are described into a fixed size vector space as it is the case with Bag-of-visual words, Ficher Kernels [13] or even when one uses the last fully connected layer of a CNN as a feature [20]. However, contrary to these approaches, each dimension of a semantic feature is associated to a precise concepts that makes sense for a human (Fig. 1). The "semantic signature approach" to text illustration thus consists in:

- (i) extracting relevant concept from images of reference
- (ii) expressing the corresponding concepts with words and index them
- (iii) matching the query to the index textually

During the campaign, we tested several alternatives for each of these steps.

Regarding step (i), our system is based on recently published work [23, 22], that is itself an extension of the Semfeat descriptor [6]. Relying on powerful mid-level-features such as [21], this semantic feature is a large set of linear classifiers that are built automatically. The authors showed that keeping a small part of the K most active concepts and setting the others to zero (*sparsification*) led to a more efficient descriptor for an image retrieval task. However, there are two limitation to this approach: first, the value of K has to be fixed in advance; secondly, sparsification is not efficient in a classification context, in the sense that the performance obtained is below that of the mid-level feature it is built on. For these reasons, [23] proposed to compute K for each image independently, with regard to the actual content of the image. The principle is to keep the "dominant concepts" only, when we are confident on their detection.

For step (ii), we considered two sets of concepts. The first one, based on WordNet, contains 17,467 concepts, each being described by its main synset (one word). The second set is that collected automatically in [6]. It contains

around $30,000$ concepts derived from Flickr groups, each described by three words.

For the textual matching step (iii), we considered classical inverse indexing, computing the query-document similarity from the "weight/score" associated to each indexed document.

3.3 Text-image common space approach

The design of common latent spaces has been proposed for a while [16, 19], in particular in the case of textual and visual modality [12, 32, 11, 2, 8]. Given two modalities, let say a visual and a textual modality described by their respective features, the general idea is to learn a latent common sub-space of both feature spaces, such that visual points are directly comparable to textual ones. One of the recent popular approach is to use Canonical Correlated Analysis (CCA), in particular in its kernelised version (KCCA) [9].

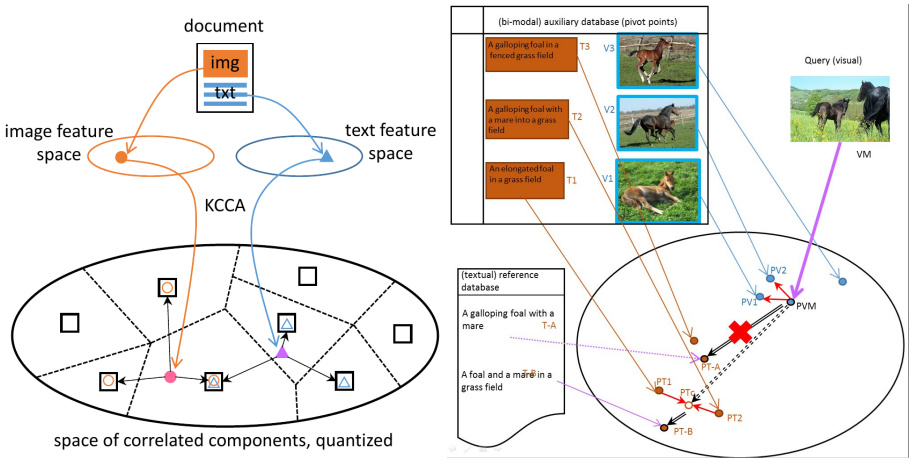
Let us consider N data samples $\{(x_i^T, x_i^I)\}_{i=1}^N \subset \mathbb{R}^{d_T} \times \mathbb{R}^{d_I}$, simultaneously represented in two different vector spaces. The purpose of CCA is to find maximally correlated linear subspaces of these two vector spaces. More precisely, if one notes $X^T \in \mathbb{R}^{d_T}$ and $X^I \in \mathbb{R}^{d_I}$ two random variables, CCA simultaneously seeks directions $w_T \in \mathbb{R}^{d_T}$ and $w_I \in \mathbb{R}^{d_I}$ that maximize the correlation between the projections of x^T onto w_T and of x^I onto w_I,

$$w_T^*, w_I^* = \underset{w_T, w_I}{\arg\max} \frac{{w_T}' \, C_{TI} \, w_I}{\sqrt{{w_T}' \, C_{TT} \, w_T \, {w_I}' \, C_{II} \, w_I}} \tag{1}$$

where C_{TT}, C_{II} denote the autocovariance matrices of X^T and X^I respectively, while C_{TI} is the cross-covariance matrix. The solutions w_T^* and w_I^* are found solution of an eigenvalue problem. The d eigenvectors associated to the d largest eigenvalues define maximally correlated d-dimensional subspaces in $\mathbb{R}^{d_T}$ and respectively $\mathbb{R}^{d_I}$. Even though these are linear subspaces of two different spaces, they are often referred to as "common" representation space.

KCCA aims to remove the linearity constraint by using the "kernel trick" to first map the data from each initial space to the reproducing kernel Hilbert space associated to a selected kernel and then looking for correlated subspaces in these RKHS.

In this space, textual and visual documents are directly comparable, thus it is possible to perform cross-modal retrieval [12, 11, 2, 8]. However, it has been recently found that the learned common space may not represent adequately all data [25]. It has thus be proposed a more robust representation data within the common space consisting in coding the original visual and textual point with respect to a codebook (Figure 2a). This method, named Multimedia Aggregated Correlated components (MACC) is detailed in [25]. Another contribution that "'compensates" the defaults of the representation space is to project a bi-modal auxiliary dataset into the common space and use the known text-image connections of this dataset as a "pivot" to link *e.g.* a textual query to an appropriate image of the reference database (Figure 2b).



(a) robust description (b) pivot principle

Fig. 2: Illustration of (a) the approach of [25] to describe robustly the bi-modal documents into a common description space (b) the pivot principle using an auxiliary database

MACC considers the use of two datasets. A first training dataset $\mathcal{T}$ is used to learn the common space with a KCCA. Due to computational issues, the number of documents that can be used to learn this space is limited to few ten thousands. Hence, at a quite large scale such as that of the text illustration subtask, it is important to use a second auxiliary dataset $\mathcal{A}$ to "compensate" the possible limitation of the initial learning.

Once the settings of the common latent space are chosen, the principle of the text illustration is quite straightforward, as illustrated in Figure 3. Regarding the images of the reference database, we extract the same feature as that used during learning, namely the FC7 fully connected layer of VGG [21]. This vector is projected on the latent space and the MACC signature is computed then stored into the reference database.

Regarding the textual query, we process the raw text in order to fix the defaults identified in Section 3.1. We first remove the stopwords using the Stanford NLTK package [1] that contains lists of stopwords in several languages. As months, days and numbers are not included in the stopwords list from NLTK, we also filtered them out as they are hard to illustrate and might add noise to our model. Additionally, we also removed words containing special characters such as """' that are often found in noisy words. Furthermore, we combined the stopwords list with a part of speech tagger developed in the NLTK library. NLTK Pos Tag categorizes our set of words and labels each word according to its grammatical properties. In order to take the more descriptive words, we choose to keep only nouns (proper and common) and adjectives.

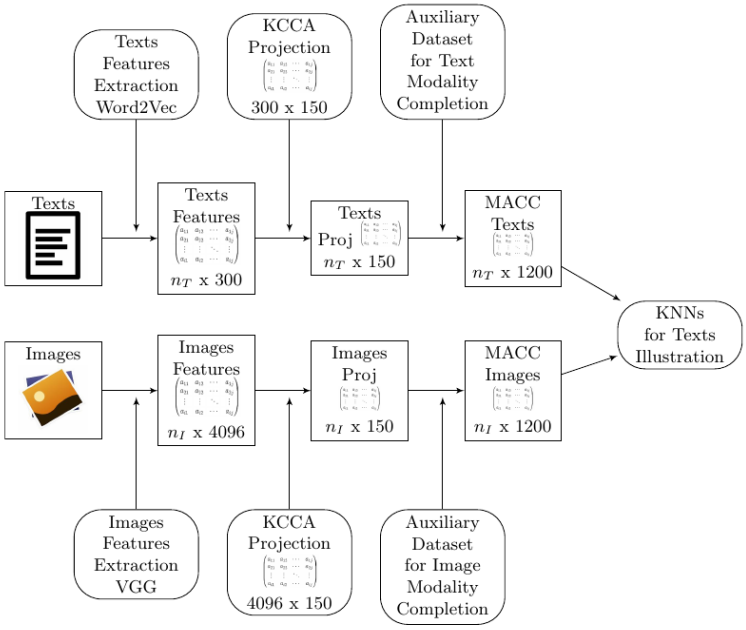


Fig. 3: General principle to illustrate a textual query from a purely visual database, using a common space and MACC representation.

For each word i of the resulting vocabulary, we extract a word2vec vectorial representation t_i. Then, we compute a weight w_i equal to its tfidf value. For a document d, we select the k terms in the textual description that have the largest weights. We then compute a unique vector v_d, representing d, from the k selected words describing a document, weighting each t_i with its corresponding weight w_i, resulting into the Weighted Arithmetic Mean (WAM) $v_d = \frac{\sum_{i=1}^{n} w_i t_i}{\sum_{i=1}^{n} w_i}$.

As said above, the classical KCCA algorithm does only support some dozen thousand document to learn the latent common space. To get around this limitation, we first proceed to a selection of the training data, in order to build a corpus with a diversified vocabulary. To do so, we divide our train set of $300k$ documents into 10 groups of $30k$ documents each. We then clustered every group of textual features with K-Means and compute 100 clusters per group. To build a $20k$ documents corpus for instance, we select for each group 20 random documents per cluster (20K = $n_{groups} n_{clusters} n_{doc}$ = 10x100x20). Similarly, we build diversified sets for the pivotal basis by selecting a certain number of random documents per cluster.

3.4 Submission and Results

We submitted four individual runs and three runs that merge them differently. Some synthetic results are presented in Table 4. Globally, the recall at 100 is

low, that is explained by the difficulty of the task as well as the noise in the queries.

Table 4: Results of the seven run submitted.

Method	Recall @ 100
CBS + wordnet	1.44
CBS + flickrgroup	1.74
Wam5	2.47
Wam7	4.33
MergeA	4.47
MergeB	4.51
MergeC	4.50

We run the method described in Section 3.2 with a semantic signature computed with CBS and both the WordNet and FlickRgroups vocabularies. We obtained better results with the smaller vocabulary of WordNet. The basic classifiers of the Wornet-based semantic features are learned with "cleaner" annotated images then those based on the FlickRGroups. However, since the original Semfeat paper [6] showed better or similar results with both types of classifiers, we suggest that in the current task the (small) difference of performance may be due to a better coverage of the vocabulary with respect to the queries.

The approach based on the common space and the MACC representation lead to significantly better results. A first run `Wam5` used $22k$ images for $\mathcal{T}$, $64k$ for $\mathcal{A}$ and 5 best words were retained to build the training and testing textual features. In the second run `Wam7` we used the same training dataset to learn the common space but $\mathcal{A}$ was extended to $164k$ documents while we retained up to $30k$ words to build the textual training features. For the textual testing features, we kept only 10 words to build the WAM, because of the noisy aspect of the query data.

The experiments we run on a development dataset (not reported) extracted from the $300k$ development images suggest that a large part of the improvement between `Wam5` and `Wam7` is due to the growth of the auxiliary dataset $\mathcal{A}$.

By merging several runs, the results are marginally improved. The run `mergeA` concatenates 10 best results of `CBS+WordNet` and `Wam5` with 80 best of `Wam7`. We then also consider run `wam8` similar to `Wam7` but its testing textual queries that were built with five best words (instead of 10 for `Wam7`). The run `mergeB` merges the 10 best results of `CBS+WordNet`, `Wam5` and $Wam8$ with the 70 best of `Wam7`. Finally, `mergeC` concatenates the 5 best results of `CBS+FlickRGroups`, the 10 best results of `CBS+WordNet` and `Wam5`, the 15 best of `Wam8` and 60 best of `Wam7`. While the settings are quite different between the three merging methods, the results are similar, showing that the results is mainly due to the first answers of `Wam7`.

4 Conclusion

We presented the results to the Image Annotation and Localization subtask and the Text Illustration teaser. The results to IAL are good in comparison to other participants, but the contribution proposed in 2016 did not lead to significantly better results than our 2015 system. It is partially due to the fairer evaluation of the task, but since the exact method of evaluation is not released, it is hard to fully interpret the results, in particular why the combination of runs decreases the mAP. Regarding the Text Illustration teaser, we proposed two methods based on recently published work. The results are globally low, due to the difficulty of the task in general[1] and the very noisy queries in particular.

References

1. Bird, S., Klein, E., Loper, E.: Natural language processing with Python. " O'Reilly Media, Inc." (2009)
2. Costa Pereira, J., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. TPAMI 36(3), 521–535 (2014)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: In CVPR. pp. 886–893 (2005)
4. Gadeski, E., Le Borgne, H., Popescu, A.: Cea list's participation to the scalable concept image annotation task of imageclef 2015. In: CLEF2015 Working Notes. CEUR Workshop Proceedings (2015)
5. Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2016 Scalable Concept Image Annotation Task. In: CLEF2016 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Évora, Portugal (September 5-8 2016)
6. Ginsca, A.L., Popescu, A., Le Borgne, H., Ballas, N., Vo, P., Kanellos, I.: Large-scale image mining with flickr groups. In: 21th International Conference on Multimedia Modelling (MMM 15) (2015)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
8. Gong, Y., Ke, Q., Isard, M., Lazebnik, S.: A multi-view embedding space for modeling internet images, tags, and their semantics. IJCV 106(2), 210–233 (Jan 2014)
9. Hardoon, D.R., Szedmak, S.R., Shawe-Taylor, J.R.: Canonical correlation analysis: An overview with application to learning methods. Neural Comput. 16(12), 2639–2664 (2004)
10. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV (2015)
11. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research pp. 853–899 (2013)
12. Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. IJCV 100(2), 134–153 (Nov 2012)

[1] the other participant to the task obtained outstanding results around 80%. If they actually used the same data as us, we'll of course revised the interest of our methods!

13. Jegou, H., Perronnin, F., Douze, M., Sanchez, J., Perez, P., Schmid, C.: Aggregating local image descriptors into compact codes. Transactions on Pattern Analysis and Machine Intelligence 34(9), 1704–1716 (2012)
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
15. Kakar, P., Wang, X., Chia, A.Y.S.: Automatic image annotation using weakly labelled web data. In: CLEF2015 Working Notes (2015)
16. Li, D.: Multimedia content processing through cross-modal association. In: Proc. ACM international conference on Multimedia. pp. 604–611. ACM Press (2003)
17. Li, L.j., Su, H., Fei-fei, L., Xing, E.P.: Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: NIPS (2010)
18. Malobabić, J., Le Borgne, H., Murphy, N., O'Connor, N.E.: Detecting the presence of large buildings in natural images. In: Content-Based Multimedia Indexing (CBMI), 2005 International Workshop on (2005)
19. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML. pp. 689–696 (2011)
20. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: An astounding baseline for recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2014)
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
22. Tamaazousti, Y., Le Borgne, H., Hudelot, C.: Diverse concept-level features for multi-object classification. In: International Conference on Multimedia retrieval (ICMR'16). New York, USA (June 2016)
23. Tamaazousti, Y., Le Borgne, H., Popescu, A.: Constrained local enhancement of semantic features by content-based sparsity. In: International Conference on Multimedia retrieval (ICMR'16). New York, USA (June 2016)
24. Torresani, L., Szummer, M., Fitzgibbon, A.: Efficient object category recognition using classemes. In: European Conference on Computer Vision. ECCV (2010)
25. Tran, T.Q.N., Le Borgne, H., Crucianu, M.: Aggregating image and text quantized correlated components. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, USA (June 2016)
26. Uřičář, M., Franc, V., Thomas, D., Sugimoto, A., Hlaváč, V.: Real-time Multi-view Facial Landmark Detector Learned by the Structured Output SVM. In: BWILD '15: Biometrics in the Wild 2015 (IEEE FG 2015 Workshop) (2015)
27. Villegas, M., Müller, H., García Seco de Herrera, A., Schaer, R., Bromuri, S., Gilbert, A., Piras, L., Wang, J., Yan, F., Ramisa, A., Dellandrea, E., Gaizauskas, R., Mikolajczyk, K., Puigcerver, J., Toselli, A.H., Snchez, J.A., Vidal, E.: General Overview of ImageCLEF at the CLEF 2016 Labs. Lecture Notes in Computer Science, Springer International Publishing (2016)
28. Viola, P., Jones, M.: Robust real-time object detection. In: IJCV (2001)
29. Vo, P.D., Ginsca, A.L., Le Borgne, H., Popescu, A.: Effective training of convolutional networks using noisy web images. In: CBMI (2015), prague, Czech Republic
30. Wu, R., Ya, S., Shan, Y., Dang, Q., Sun, G.: Deep image: Scaling up image recognition. CoRR abs/1501.02876 (2015)
31. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: ECCV. European Conference on Computer Vision (September 2014)
32. Znaidia, A., Shabou, A., Le Borgne, H., Hudelot, C., Paragios, N.: Bag-of-multimedia-words for image classification. In: ICPR. pp. 1509–1512. IEEE (2012)