MRIM-LIG at ImageCLEF 2016 Scalable Concept Image Annotation Task

Maxime Portaz[1,2,*], Mateusz Budnik[1,2,*], Philippe Mulhem[1,2,*], Johann Poignant[1,2,*]

[1] Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2] CNRS, LIG, F-38000 Grenoble, France
[*] Firstname.Lastname@imag.fr

Abstract. This paper describes the participation of the the MRIM research Group of the LIG laboratory in the ImageCLEF scalable concept image annotation subtask 1. We made use of a classical framework to annotate the 500K images of this task: we tuned an existing Convolutional Neural Network model to learn the 251 concepts and to locate bounding boxes of such concepts, and we applied a specific process to handle faces and face parts. Because of time constraints, we fully processed 35% of the full corpus (i.e. 180K images), and partially the remaining images of the corpus. For our first participation to this task, the results obtained show that we have to manage the localization in a more effective way.

Keywords: Convolutional Neural Networks, Landmark face detection, ImageNet, TRECVID

1 Introduction

The first participation of the MRIM group from the LIG laboratory at the ImageCLEF 2016 [7] scalable concept image annotation subtask 1 [3] is presented. Our approach was to use a classical framework based on face detection [8] followed by facial landmarks detection [6] for faces and face parts (eyes, nose and mouth), and to rely on convolutional neural networks [4] for each of the 251 concepts.

The ImageCLEF 2016 scalable concept image annotation subtask 1 consists of finding the location of 251 classes of objects in a corpus of 500K images. This task is challenging because of the difficulty of finding accurate location of objects in large sets of images. The objective is to assign a maximum of 100 bounding boxes per image, each bounding box being associated to one or more of the 251 concepts proposed. It is also possible to provide a confidence value for each of the tagging defined. The visual concepts defined for this subtasks do not match fully with concepts coming from the well known ImageNet database [1], so specific work has to be done to be able to tackle these concepts.

Because of time needed to process the whole corpus, we fully processed around 35% of the full image corpus (i.e. 180K images), and partially the remaining of the corpus. The results obtained are then negatively impacted by this partial processing.

The rest of this paper is organized as follows. In section 2, we define our approach: we mainly rely on convolutional neural networks for "classical" concepts, with a specific process dedicated to faces. Then, in section 3, we detail the results obtained, as well as some additional elements dedicated to analyzing our results in more detail. We conclude in section 5.

2 Proposed Approach

2.1 Overview

The overall process applied for detection and localization of concepts in images is described in figure 1. We generate possible bounding boxes, then apply Convolutional Neural Networks for each of the 251 concepts. For face and face part detection, we use face and facial landmarks detection. Such approaches have been successfully used by several participants during the 2015 campaign of ImageCLEF concept annotation task. We finally rank all the labeled bounding boxes by score or by size, depending on the run. This ranking is used as filtering to reduce the number of boxes per image, as we take only up to 100 boxes for each image (a limit chosen by the organizers).

2.2 Convolutional Neural Networks

We used a Deep Residual Convolutional Neural Network (ResNet) with 152 layers presented by Microsoft in the ImageNet'16 challenge [4]. The network was finetuned to match the 251 labels from ImageClef. Only the final layer was retrained.

Data Processed
A first step in the learning process was to map, when possible, the 251 CLEF concepts into concepts from existing image collections, namely the ImageNet concepts. From the full set C of 251 concepts, 224 are mapped directly to ImageNet concepts, and for each of the 27 remaining concepts we acquired 4519 images from Bing API using the concept name as query. We do not filter manually the resulting set of images.

As described in figure 2, we also define a second set of images to increase the quality of the concept detection. This second set also includes both Bing API and the validation set (2000 images, 10000 tagged bounding boxes) provided by the organizers of the task.

CNN Processing
One specificity of our proposal is to define a two-step learning process (basically two finetuning stages) as a way to increase the effectiveness of the concept detection. The CNN network comes pre-trained on the ImageNet dataset [1]. We used two validation sets: a) the first one is the set provided by the organizers of

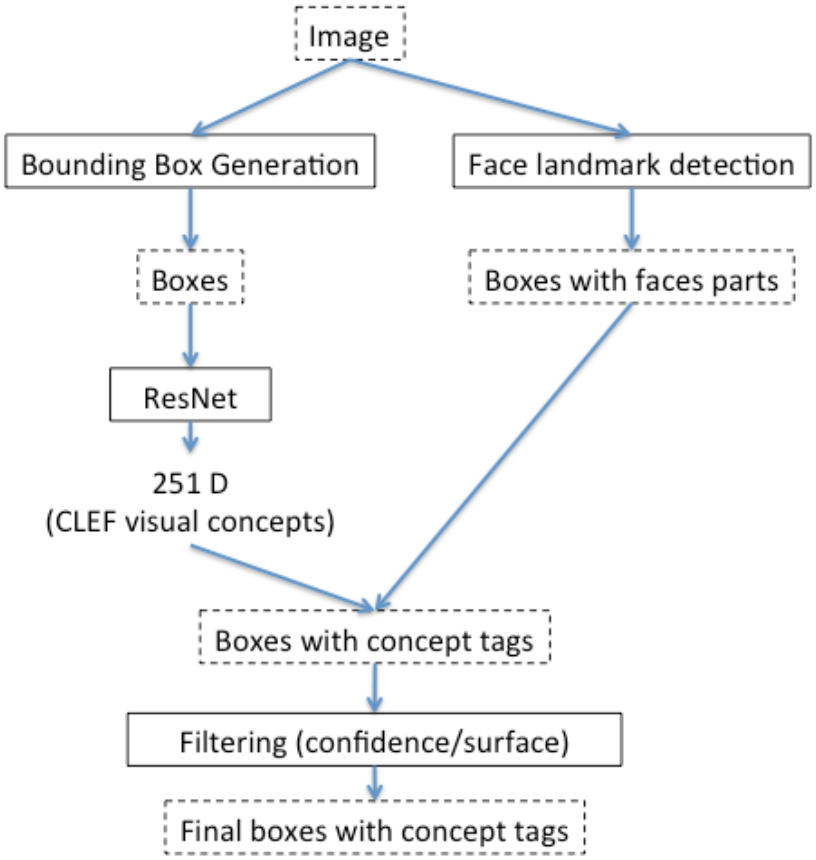

Fig. 1. MRIM-LIG Annotation System

the ImageCLEF task, and b) a second one that we defined to assess the quality of the training on "clean" images. The first finetuning step is evaluated on these two validation sets. While during the second learning step the first set (a) is used for training as well as some additional images (which were crawled from the Internet) for the concepts with the lowest recognition rate. After the second finetuning, the system is tested only on the (b) validation set. In other words:

- On our first set of training images, learn the last layer of CNN, then evaluate (success@1 success@5) on the two validation sets;
- During the second learning stage, for the low quality recognition concepts, we generate the second set of 200 additional training images per concept. As described above, we also add the validation set (a) provided by CLEF. We retrain the network on this combined and extended set.

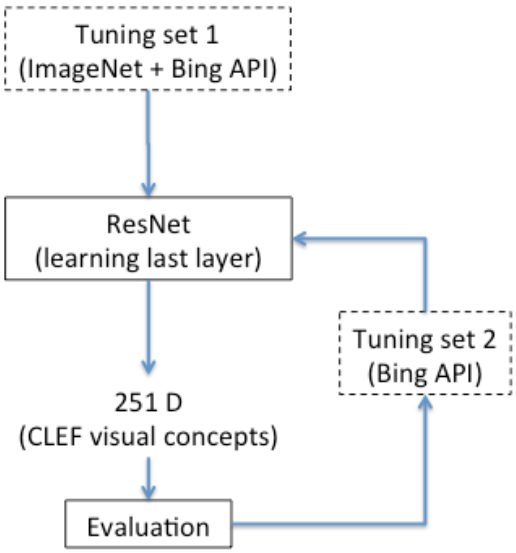

Fig. 2. MRIM-LIG two steps learning.

At the end of these two steps, we obtained the results presented in table 1. The two first rows of this table present the results after the first tuning step. The remaining two rows give the results after the second phase of finetuning. The second step seems to significantly increase the performance on the Bing validation set.

The ImageCLEF validation set was included in the training set at the second stage of tuning. That is why a surprisingly strong result (denoted with "*"), compared to the first tuning, is obtained: it does not generalize and was included just for illustrative purposes.

Table 1. Post-tuning evaluation results

Tuning set	Validation set	success@1	success@5
1	ImageCLEF	0.1523	0.3418
1	Bing	0.6521	0.8435
2	ImageCLEF*	0.6290	0.6290
2	Bing	0.7337	0.9333

Concept Localization

We used the work of Uijlings, van de Sande, Gevers and Smeulders [5] to perform selective search to define bounding box detection. The idea is mainly to define *a*

priori a set of bounding boxes that are expected to contain one visual concept. The selective search use Felzenswalb algorithm [2] for image segmentation. In our runs, we use a width for Gaussian kernel of 0.8, and a scale factor of 500. The minimum size for a box is set to 200 pixels. These constant give a average of 517 boxes per image. Each of these boxes will be used as an input image on which the CNN will be applied to detect objects.

Actual Processing Achieved
Due to time constraints, we applied the full process to 180k images: selective search and clustering of bounding boxes, and CNN detection on each of the selected boxes. On average, the number of boxes generated per image is 517. For each of the remaining images (320K images), we applied detection on: a) the full image, and b) a small subset of the initial boxes selected randomly. On average, the number of boxes generated per image for each remaining image is 8. Overall, we processed 95 millions of boxes for our submissions.

2.3 Face Detection

The detection and localization of parts of faces is achieved through a two step process:

- Frontal faces are detected using the "classical" Viola and Jones approach [8] based on cascade of simple Haar-like features;
- Then 8 facial landmarks [6] are detected on these faces. They correspond to the 2 mouth corners, 4 eye canthus, the tip of the nose and the center of the face. We used then simple heuristics to define faces, eyes, noses and mouths bounding boxes based on these landmarks.

All images of the ImageCLEF corpus are processed using the above steps. With such process, at least one faces is detected on 64642 of the 510K images (12.7% of the whole corpus). A total of 91102 faces "boxes" are detected on these images.

3 Evaluation Results

The runs submitted by the MRIM-LIG team are the following:

- **RUN1_LIG_DLo**: Annotation using the Convolutional Neural Network described in part 2.2, with a ranking of the bounding boxes according to the confidence value;
- **RUN2_LIG_DLo**: Annotation using the CNN described in part 2.2, with a ranking of the bounding boxes according to the surfaces of the boxes;
- **RUN3_LIG_Fo**: Annotation of the face parts only, using the Viola/Jones approach described in part 2.3;
- **RUN4_LIG_DLF**: Annotation using both the CNN and face parts detection, with a ranking of the bounding boxes according to the confidence value;
- **RUN5_LIG_DLF**: Annotation using both the CNN and face parts detection, with a ranking of the bounding boxes according to the surfaces of the boxes;

3.1 Official Results

The official MAP at 0% overlap and MAP at 50% overlap results of our runs are presented in table 2. We find that the run RUN5 (that fuses the face parts and deep learning results, ranking based on surfaces) achieves our best result (rank 11 for overlap 0, and rank 9 for overlap 0.5). At overlap 0.5, our second best result is RUN4 (that fuses the face parts and deep learning results, ranking based on confidence values). The difference between RUN5 and RUN4 are negligible. We suppose that comes from the fact that only 180K images where fully processed, and for the remaining ones we did not have more than 100 boxes, and the ranking only plays a role when we obtain more than 100 boxes. The same holds also for our runs RUN1 and RUN2 (based only on deep learning features).

Compared to the runs of other participants, we find that our general runs that integrate deep learning do not obtain very high results. This can be explained by the fact that, as mentioned before, the whole proposed process was applied only on 180K images of the 510K images of the corpus.

As expected, our run RUN3, that detects only face parts has a very low overall result, ranked 23 for both overlap 0 and overlap 0.5.

Table 2. Official overlap evaluation results

Run	MAP_0	rank (on 30)	MAP_0.5	rank (on 30)
RUN5	0.2084	11	0.1353	9
RUN2	0.2051	12	0.1317	11
RUN4	0.2030	13	0.1351	10
RUN1	0.1998	14	0.1309	12
RUN3	0.0123	23	0.0104	23

When considering the additional official measures related to the minimum number of boxes per image, we see a plateau above a minimum of 20 boxes. This shows that when a image has less than 20 boxes in the ground truth set our proposal has difficulty to find relevant concepts or boxes. This can be also attributed to the fact that we did not fully process the whole corpus, as explained earlier.

3.2 Detailed analysis of face parts results

Here we try to give additional insight into the results obtained when considering only the face elements from deep learning and predefined face extraction approaches [8, 6]. In table 3, we present the average precision results obtained for our overlap ranking approaches runs RUN2 (deep learning only), RUN3 (face parts only), and RUN5 (fusion), for the concepts *mouth*, *eye*, *nose* and $face$.

One interesting point that we get from table 3 is that, for the MAP at 0 and for the $face$ concept, the deep learning approach (RUN2) outperforms both the predefined detection (RUN3) and fusion (RUN5). We recall that face

Table 3. Face parts results

Run	MAP_0				$MAP\_0.5$			
	mouth	eye	nose	face	mouth	eye	nose	face
RUN2	0.1502	0.4053	0.1964	0.8947	0.08336	0.2955	0.1607	0,5722
RUN3	0.6787	0.7078	0.7172	0.8416	0.5578	0.4177	0.6941	0.8172
RUN5	0.2082	0.6699	0.7076	0.8663	0.1366	0.4198	0.6804	0.7216

is already a concept available in ImageNet. However, for the other concepts this is not the case. When the localization is evaluated, then the predefined detection outperforms the deep learning approach. When considering the fusion run (RUN5), we see that most of the time such fusion does not work properly as it does not seem to boost the results. The only case when the fusion outperforms the other runs is for MAP 0.5 for the eye, and the increment is marginal.

4 Current limitations of the scalable concept annotation task

After checking the official global results and the per concept results, we feel that:

- The size of the ground truth seems small: many concept results aP values are equal to 1 (or exactly 0.5, 0.25, etc.), leading to think that there are only very few ground truth regions defined for most concepts. A collaborative annotation interface open to participants may be a good idea to get more ground truth, leading to results that are more statistically valid. In this case, it should be possible to force a minimum number of examples for each concept in the ground truth;
- The ground truth is not released by the organizers after the official results. Even if we understand the reason why the organizers do that, such ground truth may be of a great help for the participant to study why and when their approach fail. Alternatively, a bigger and more representative validation set should be very helpful to participants;.
- Without obtaining the ground truth, we think that the number of boxes per concept in the ground truth should be released, so that participants may have cues about their results per concept;
- Even if the name of the task is "scalable concept annotation", we wonder if it should be possible to get, in addition to the existing measures, other measures that are able to focus on the runs submitted: limiting the evaluation on the concepts detected is already possible by averaging a posteriori the aP of a subset of concepts, but it is impossible for the participants that were not able for any reason to process all the images to evaluate the quality of such runs only on the subset of image processed.

5 Conclusion

For our first participation in the Image CLEF scalable concept detection, we used classical approaches based on convolutional networks as well as specific elements related to the detection of parts of faces. Selective search was applied on the images in a way to detect concepts from CNNs. Because only a subset (35%) of the whole corpus was fully processed, the official results we obtain are not as high as they could have been. We found that the fusion of predefined face part extraction and deep learning detection did not give positive results: such fusion has to be studied in more detail in the future. The elements related to the definition of localization has also to be studied in the future to allow fast detection of such boxes.

References

1. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
2. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.
3. 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 Challenge. In *CLEF2016 Working Notes*, CEUR Workshop Proceedings, Évora, Portugal, September 5-8 2016. CEUR-WS.org.
4. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
5. J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 2013.
6. M. Uřičář, V. Franc, and V. Hlaváč. Detector of facial landmarks learned by the structured output SVM. In *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, pages 547–556, 2012.
7. 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. M. J. Puigcerver, A. H. Toselli, J.-A. Snchez, and E. Vidal. General Overview of ImageCLEF at the CLEF 2016 Labs. Lecture Notes in Computer Science. Springer International Publishing, 2016.
8. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.