CEA LIST's participation to the Scalable Concept Image Annotation task of ImageCLEF 2015

Etienne Gadeski, Hervé Le Borgne, and Adrian Popescu

CEA, LIST, Laboratory of Vision and Content Engineering, France
etienne.gadeski@cea.fr, herve.le-borgne@cea.fr, adrian.popescu@cea.fr

Abstract. This paper describes our participation to the ImageCLEF 2015 scalable concept image annotation task. Our system is only based on visual features extracted with a deep Convolutional Neural Network (CNN). The network is trained with noisy web data corresponding to the concepts to detect in this task. We introduce a simple concept localization pipeline that provides the localization of the detected concepts, among the 251 concepts, within the images.

Keywords: image annotation, classification, concept localization, convolutionnal neural networks.

1 Introduction

The scalable concept image Annotation subtask of ImageCLEF 2015 [1] is described in detail in [2]. The system we propose only relies on visual features extracted with a deep Convolutional Neural Network (CNN). In addition to the specific training of this CNN for this task we propose a concept localization pipeline which uses the spatial information that CNNs offer. We are able to propose a single concept annotation and localization framework for the 251 concepts. Section 2 outlines our method in both training and testing steps. In Sect. 3 we present our experiments, report our results and discuss them.

2 Method

In this section we detail the training and testing frameworks 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, action recognition, etc [3, 4].

2.1 Training

Data. We have crawled a set of $\approx$ 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 we believe it is not a big issue for training a deep CNN [5, 6]. We only used this additional data to train a 16-layer CNN [7]. We used 90% of the dataset for training and 10% for validation.

Network Settings. The network is 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, shearing, rotation, contrast adjustment, etc. It is known that data augmentation leads to better models [8] and reduces overfitting. Finally, the network takes a 224×224 RGB image as input and produces 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 [9].

2.2 Testing

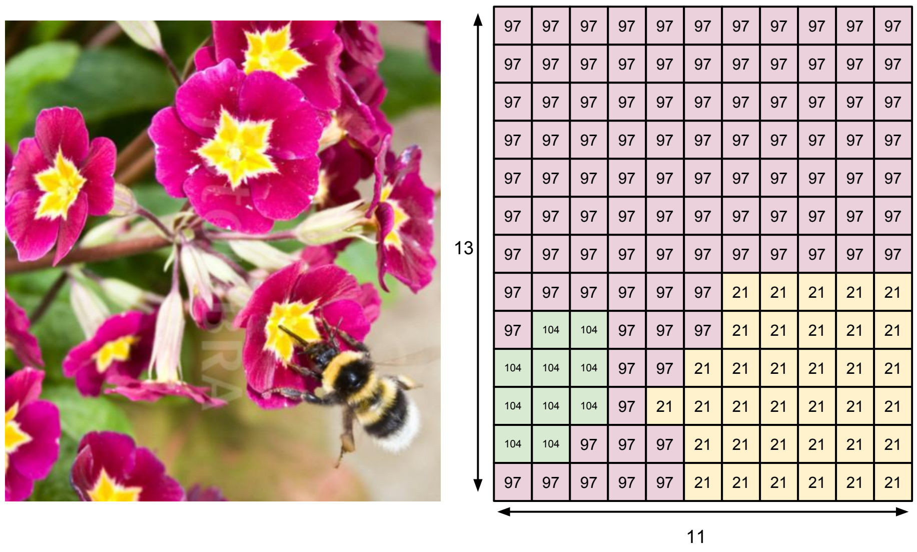
Feature Extraction. We convert the fine-tuned CNN model to a fully convolutional network. To do so, we explicitly convert the first fully-connected layer to a 7×7 convolution layer. The last two fully-connected layers are then converted to 1×1 convolution layers. We are able to do such a conversion because fully connected layers are nothing more than 1×1 convolutional kernels and a full connection table. This conversion allows us to feed images of any size ($H \times W$ pixels) to the network. In practice, any image with a side smaller than 256 pixels is isotropically rescaled so that its smallest side equals 256. In consequence, the output of the last layer (*i.e.* the classifier) is no longer a single vector with 251 dimensions but rather a spatial map $\mathbf{S}$ of $M \times N \times 251$ dimensions. The network has five 2×2 pooling layers, therefore it has a global stride of 32 pixels. We have:

$$M = \left\lceil \frac{H}{32} - 7 + 1 \right\rceil \text{ and } N = \left\lceil \frac{W}{32} - 7 + 1 \right\rceil . \tag{1}$$

For instance, with a 512×384 image the network will produce a $10 \times 6 \times 251$ feature map.

Concept Localization. For each image, our network outputs a spatial feature map, $\mathbf{S}$, of size $M \times N \times 251$. In every of the $M \times N$ cells of this feature map, we obtain the probability for each concept to appear at this location. In our experiments, we chose to apply a max-pooling to each cell (over all 251 concepts) to only get the most probable concept within the location. Thus, we obtain a unique map giving the most probable concept at each cell location:

$$\mathbf{R} = \left\{ \forall i \in [1; M], \forall j \in [1; N], r_{i,j} : r_{i,j} = \arg\max_{c} \left(\mathbf{s}_{i,j}\left(c\right) \right) \right\}, \tag{2}$$



97	97	97	97	97	97	97	97	97	97	97
97	97	97	97	97	97	97	97	97	97	97
97	97	97	97	97	97	97	97	97	97	97
97	97	97	97	97	97	97	97	97	97	97
97	97	97	97	97	97	97	97	97	97	97
97	97	97	97	97	97	97	97	97	97	97
97	97	97	97	97	97	97	97	97	97	97
97	97	97	97	97	97	21	21	21	21	21
97	104	104	97	97	97	21	21	21	21	21
104	104	104	97	97	21	21	21	21	21	21
104	104	104	97	21	21	21	21	21	21	21
104	104	97	97	97	21	21	21	21	21	21
97	97	97	97	97	21	21	21	21	21	21

Fig. 1. Example of one image of 600×544 pixels (on the left) with its corresponding 13×11 map (on the right). The concept 97 corresponds to "flower", 21 to "bee" and 104 to "garden".

where $\mathbf{s}_{i,j}$ is the 251-dimensional vector at position (i, j) in $\mathbf{S}$. As illustrated in Fig. 1 this gives us an approximate but yet interesting localization of the different concepts within the image. To limit the number of regions where a concept is likely to appear, we further merge the neighboring cells which have that concept, by retaining the largest rectangular region containing all the contiguous cells with a given most probable concept. We map these regions coordinates to the original image coordinates. The final result can be seen in Fig. 2.

3 Experiments and Results

We report the results obtained for the two runs submitted to the campaign and propose ways of improvement to our framework.

3.1 Runs

Two runs were submitted for this task. We used two models, one per run. They only differ in the number of iterations accomplished in the training step. Results are reported in Table 1. Figure 3 shows the mean average precision (mAP) of our runs using varying percentage of overlap between our bounding boxes and the ground truth.

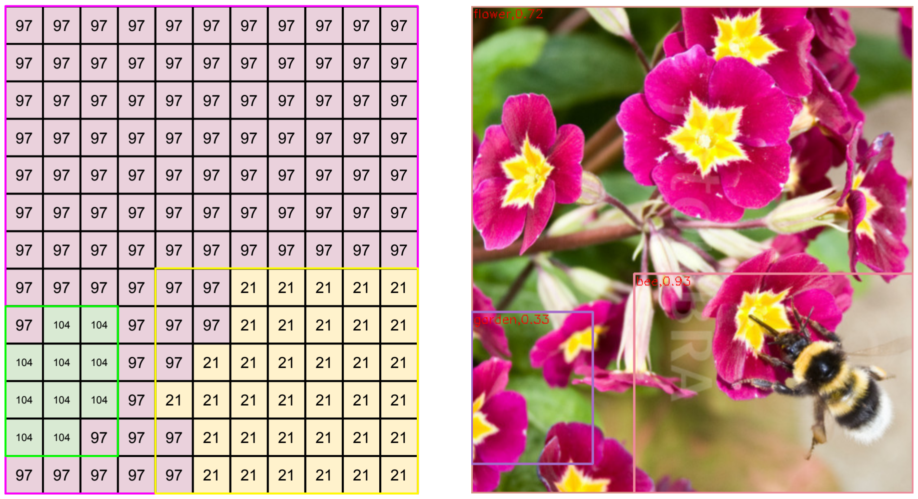


Fig. 2. On the left, we show the concepts regions proposed by our pipeline. The purple region corresponding to “flower” covers the whole area. The yellow region corresponding to “bee” is in the bottom right corner and there is also a green region corresponding to “garden” on the bottom left corner. We also display these mapped regions on the original image.

Table 1. Results for the Image Concept detection and localization task.

Run	**# iterations**	**mAP 0% overlap**	**mAP 50% overlap**
all_img.bb.run1	26,000	0.420532	0.271942
all_img.bb.run2	34,000	0.445121	0.294559

3.2 Discussion

We note that, although there is only a small gap of 8,000 iterations of stochastic gradient descent between the two models, the second one performs significantly better (+2.5%) than the first one. Unfortunately, due to the lack of time, we could not have trained the model further and we believe it would provide a non negligible improvement.

Regarding the localization, our method is quite efficient but it limits its ability to detect several similar objects forming a compact group into the image, such as a *cow herd*.

However, we think there are several ways of improving our method. The first one would be to directly train the CNN in a “fully convolutional” way to improve the localization of concepts that can appear at different scales, therefore we could also improve the robustness of our method by extracting the spatial feature map at different scales.

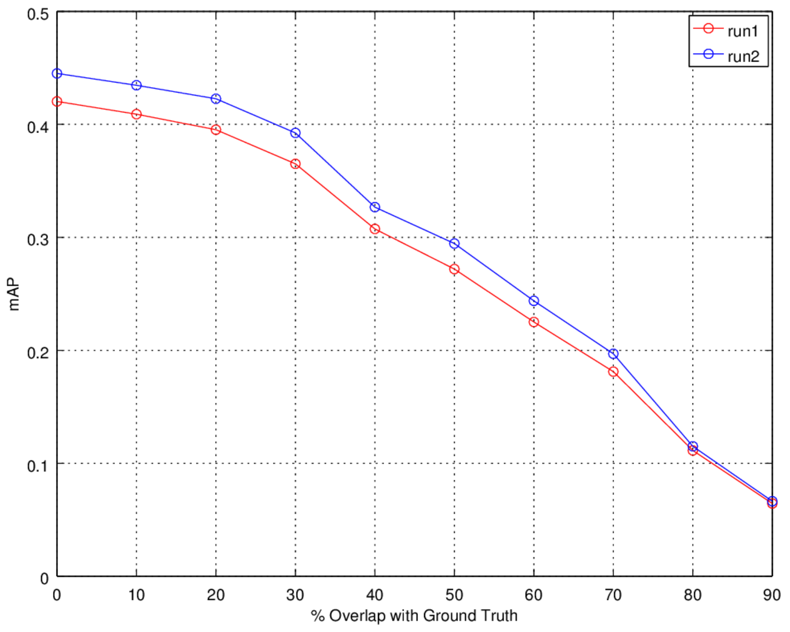


Fig. 3. The mean average precision (mAP) of our two runs are reported with several percentage of overlap between our predicted bounding boxes and the ground truth.

4 Conclusion

We proposed a simple and effective framework to annotate and localize concepts within images. The results obtained in this campaign are an encouraging step for us toward building a better framework for concept annotation and localization.

References

1. Villegas, M., Müller, H., Gilbert, A., Piras, L., Wang, J., Mikolajczyk, K., de Herrera, A.G.S., Bromuri, S., Amin, M.A., Mohammed, M.K., Acar, B., Uskudarli, S., Marvasti, N.B., Aldana, J.F., del Mar Roldán García, M.: General Overview of ImageCLEF at the CLEF 2015 Labs. Lecture Notes in Computer Science. Springer International Publishing (2015)
2. Gilbert, A., Piras, L., Wang, J., Yan, F., Dellandrea, E., Gaizauskas, R., Villegas, M., Mikolajczyk, K.: Overview of the ImageCLEF 2015 Scalable Image Annotation, Localization and Sentence Generation task. In: CLEF2015 Working Notes. CEUR Workshop Proceedings, Toulouse, France, CEUR-WS.org (September 8-11 2015)
3. 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)

4. 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)
5. 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)
6. Vo, P.D., Ginsca, A.L., Le Borgne, H., Popescu, A.: Effective training of convolutional networks using noisy web images. In: 13th International Workshop on Content Based Multimedia Indexing (CBMI). (2015) Prague, Czech Republic.
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
8. Wu, R., Ya, S., Shan, Y., Dang, Q., Sun, G.: Deep image: Scaling up image recognition. CoRR **abs/1501.02876** (2015)
9. 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)