MindLab at ImageCLEF 2014: Scalable Concept Image Annotation

Jorge A. Vanegas, John Arevalo, Sebastian Otálora, Fabián Páez, Santiago A. Pérez-Rubiano, and Fabio A. González

MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
{javanegasr, jearevaloo, jsotaloram, fmpaezri, saaperezru, fagonzalezo}@unal.edu.co

Abstract. This paper describes the participation of the MindLab research group of Universidad Nacional de Colombia at the ImageCLEF 2014 Scalable Concept Image Annotation challenge. Our strategy mainly relies in finding a good visual representation based on deep convolutional neural networks. Despite the simplicity of the proposed classifier which allows to deal with the large-scale nature of this task, we can achieve good performance (our proposed approach achieved the best MAP) thanks to the rich visual representation based on learned features.

Keywords: ImageCLEF, Visual Features, Convolutional Neural Networks, Multi-label Annotation.

1 Introduction

This paper describes the participation of MindLab research group of Universidad Nacional de Colombia in the 2014 version of the Scalable Concept Image Annotation challenge at ImageCLEF [6,1]. Our first motivation was to evaluate the use of learned features via deep convolutional neural networks (DCNN). The main strategy is based on transfer learning [7], by using in this domain a neural network trained in another similar domain. Current state-of-the-art results on ImageNet, the largest image classification challenge, are based on a DCNN trained in a supervised fashion. Moreover, in the last years multiple works using DCNN significantly improve upon the best performance in the literature for multiple image databases, showing the promising potential of systems based on DCNN [5]. The success of DCNN is attributed to their capability to learn a rich mid-level image representation. Some works have shown that it is possible to learn to extract this rich mid-level representation from one domain and use this knowledge to improve the performance in other related domain [4].

In this work we proposed a transfer learning approach by using a convolutional network that was trained over a million of images of the ImageNet dataset to enrich the visual representation of the images from this particular domain.

The rest of the paper is organized as follows: Section 2 describes the characteristics of the dataset; Section 3 describes our multi-label annotation approach; Section 4 presents the experimental results; and finally, Section 6 presents some concluding remarks.

2 The Dataset

The dataset is composed by a subset of images extracted from a database of millions of images downloaded from the Internet. For each image, the corresponding web page that contained the image is available offering a set of unstructured and noisy related text. This is a large training dataset composed by 500,000 images with meta-data but without labels. To evaluate the proposed systems, two sets with different list of images and corresponding concepts are provided:

Development Set. This set is annotated and composed by 1,000 images labeled with 107 different concepts.

Test Set. Is an unlabeled set composed by 4,122 unique images and 207 possible concepts.

To validate the scalability of the proposed systems, the list of concepts are different for the development and test sets, moreover, within each set the list of concepts will not be the same for all images.

3 Multi-label Annotation Model

3.1 Visual Representation

Although several sets of pre-processed visual features are provided by the challenge organizers, our strategy is based on building our own visual representation based on DCNN. And, for this strategy, we rely in the theory of transfer learning which is based in the ability of a system to recognize and apply knowledge learned in previous domains to novel domains, which share some commonality.

We use the Yangqing Jia et al. [2] (Caffe) pretrained network to represent images. Caffe is an open source implementation of the winning convolutional network architecture of the ImageNet challenge proposed by Krizhevsky et al. [3]. This network was trained over a million of images annotated with 1,000 ImageNet classes.

This convolutional network has 60 million parameters and has an architecture composed by eight layers: five convolutional layers and three fully-connected. The output of the last fully-connected layer is the input for a 1000-way soft-max layer which produces a distribution over the 1000 classes (normalized scores).

Each image is scaled so that the smallest dimension has 227 pixels preserving the original aspect ratio. This raw scaled image is given as input to the network. We used the last fully-connected layer activations, composed by 4096 neurons, as the visual representation for each image.

3.2 Text preprocessing

As text representation for images we processed the provided word-score features based on term frequency, DOM attributes and word distance to the image. This text is preprocessed using stop-words removal and stemming, generating a final list of words for each image that is used as textual annotation. Notice that this training set is noisy, that means that a lot of incorrect words could be associated to an image.

3.3 Label Assignment for training set

The process to assign labels to images is as follows: the list of query concepts are stemmed and compared to the list of words of the textual annotation obtained in 3.2, if the query concept is presented in the list of words assigned to an image, this concept is assigned as label to the corresponding image. After the label assignment process, if some image does not have any concept, then this image is removed from the original training set. This leads to a total of 383,815 filtered training images to train the annotation model for the development set composed by 107 concepts; and a total of 427,444 training images for the test set composed by 207 concepts.

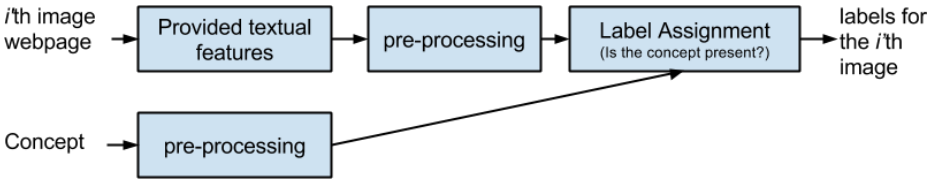


Fig. 1: Label Assignment process to images in the training set

3.4 Multi-label Annotation

Once we extracted a training set composed by the filtered images which are represented by the visual features generated through the convolutional neural network and annotated with its corresponding concepts, we trained a logistic regression model with multiple outputs, which produces a distribution over the 207 different concepts of the test set. Later, visual features are extracted for the test images and the annotations are predicted by using the trained logistic regression.

3.5 Decision

The logistic regression layer produces a distribution with denotes the probability of belonging to each concept, in order to give a final decision in the annotation,

it is necessary to define a threshold value. To define an appropriate value for this threshold we perform an exploration by using the development dataset. Figure 2 shows the results of the exploration, using three different strategies: 1) the output of the logistic regression is normalized by samples assigning 1 to the maximum value and 0 the minimum value 2a; 2) the output for each concept is normalized by setting 1 to the maximum value achieved among all samples (2b), and 3) the logistic regression output is used directly, without normalization (2c).

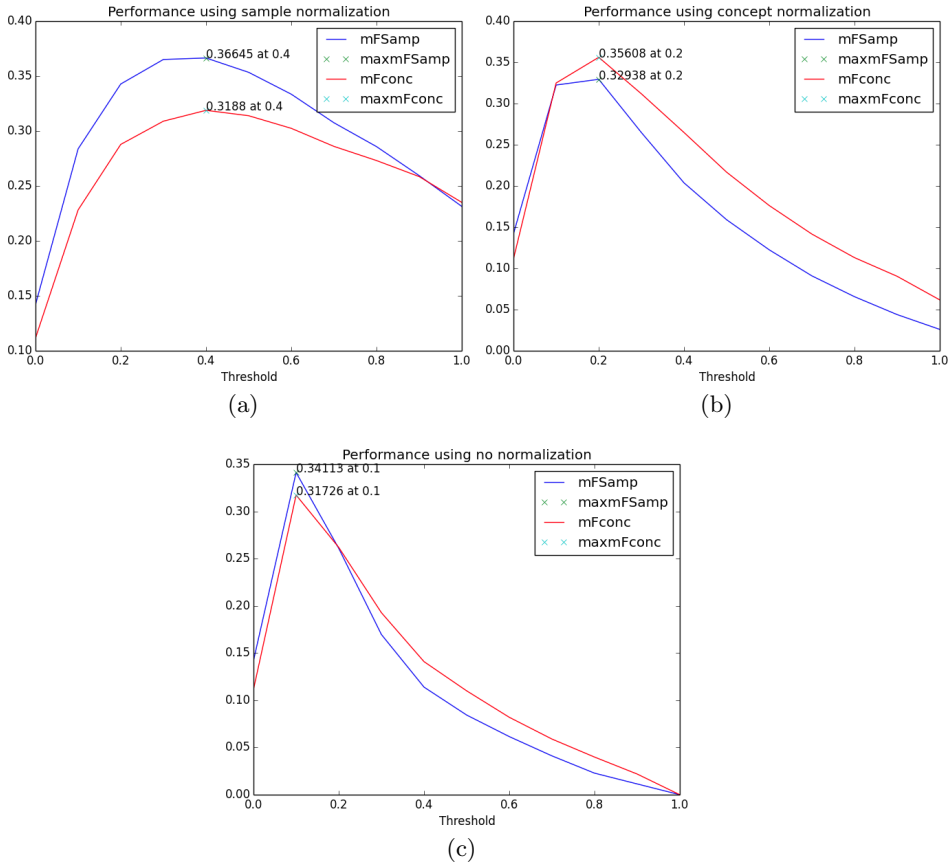


Fig. 2: Threshold exploration for three strategies: threshold based on sample normalization (2a), threshold based on concept normalization (2b) and threshold using no normalization (2c). Performance reported in mean F-measure for samples (mFSamp) and mean F-measure for concepts (mfConc).

The different performance curves in figure2 gives clues to select the score normalization method. Normalizing either by concept or sample give better re-

sults than no normalization at all. And among those two types of normalization, sample normalization yields a curve which is less sensitive to slight threshold variations, allowing more tolerance in the choice of the threshold.

4 Experimental results

We submitted 2 runs, where the only difference is the strategy used for threshold assigning:

Run 1 (MindLab_01): In this run we assigned the best threshold found in the exploration performed in the development set using per sample normalization. This threshold was used for all concepts.

Run 2 (MindLab_02): In this run the output for each concept is also normalized per sample as in the first run, but two types of thresholds are used. The best threshold found for the development concepts with sample normalization was used for the new concepts in the test set. For the concepts present in development and test, an specific threshold was used for each concept. This specific threshold was fine tuned for each development concept to achieve the best performance in the development images.

The official results of both submitted runs are reported in Table 1, also, the best result obtained among all the submissions is reported for comparison. As can be seen from the table, our strategy achieved a better result in MAP value than the best submission obtained among all the participants, but a more adequate strategy is required for selecting the final annotations. Figures 3 and 4 show the obtained results of precision and recall for both submissions grouping by concept or sample. When comparing performance of both submissions grouping by concept (figures 3a and 4a), an improvement in recall is evident for the second submission. This can be attributed to the specific threshold used for the concepts on the second submission. This improvement in recall is also present when comparing the performance of both submissions grouping by sample (figures 3b and 4b). But this comparison also reveals a drawback of the strategy used for the second submission, as precision drops significantly. This results bring forward the need to evaluate other strategies for threshold selection, which do not suffer this kind of disadvantages.

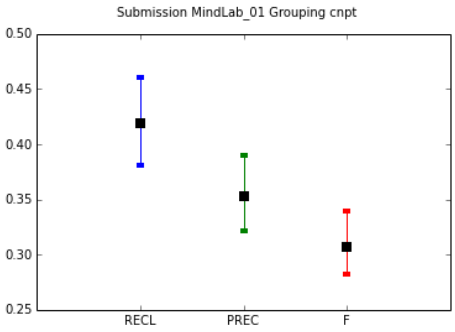
5 Conclusions

In this work, we proposed a method for multi-label annotation. Despite the simplicity of the proposed classifier which allows to deal with the large-scale nature of this task, we can achieve a good performance (our proposed approach achieved the best MAP) thanks to the richness of the visual representation based on learned features via deep convolutional neural networks.

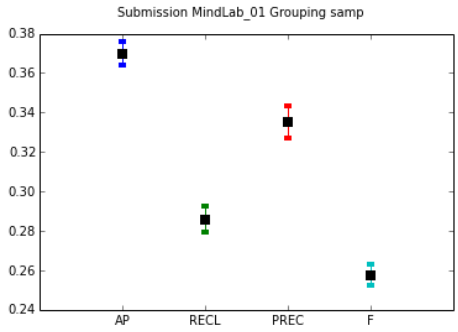
The experimental results showed that a good performance can be achieved by applying knowledge from other similar domain (transfer learning).

Table 1: Performance measures of the submitted runs for the Scalable Concept Image Annotation task

Run	Position	MF-samples (%)	MF-concepts (%)	MAP-samples (%)
MindLab_01	8	25.8 [25.2–26.3]	30.7 [28.2–34.0]	**37.0 [36.4–37.6]**
MindLab_02	10	24.8 [24.2–25.3]	31.7 [29.2–34.8]	**37.0 [36.4–37.6]**
	1	**37.7 [37.0–38.5]**	**54.7 [50.9–58.3]**	36.8 [36.1–37.5]

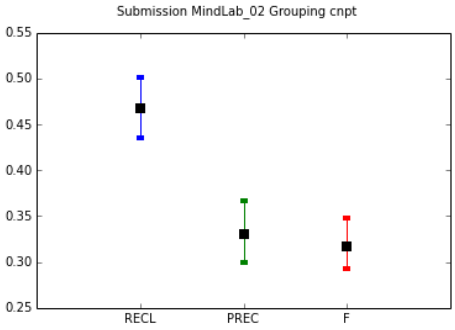


(a) Results grouping by concept

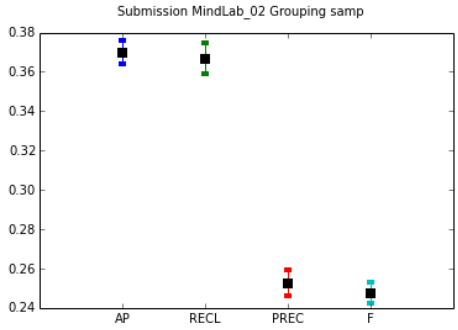


(b) Results grouping by sample

Fig. 3: Results for submission MindLab_01. Average Precision (AP), Recall (RECL), Precision (PREC) and F-measure values are reported for samples (3b) and concepts (3a)



(a) Results grouping by concept



(b) Results grouping by sample

Fig. 4: Results for submission MindLab_02. Average Precision (AP), Recall (RECL), Precision (PREC) and F-measure values are reported for samples (4b) and concepts (4a)

Acknowledgments

This work was partially funded by project Multimodal Image Retrieval to Support Medical Case-Based Scientific Literature Search, ID R1212LAC006 by Microsoft Research LACCIR and Jorge Vanegas and John Arevalo also thanks for doctoral grant supports Colciencias 617/2013. Sebastian Otálora also thanks Colciencias for its support through the grant "Jóvenes Investigadores 2012" in call 566.

References

1. Barbara Caputo, Henning Müller, Jesus Martinez-Gomez, Mauricio Villegas, Burak Acar, Novi Patricia, Neda Marvasti, Suzan Üsküdarlı, Roberto Paredes, Miguel Cazorla, Ismael Garcia-Varea, and Vicente Morell. ImageCLEF 2014: Overview and analysis of the results. In *CLEF proceedings*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2014.
2. Yangqing Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.
3. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.
4. M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.
5. Jurgen Schmidhuber. Multi-column deep neural networks for image classification. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 3642–3649, Washington, DC, USA, 2012. IEEE Computer Society.
6. Mauricio Villegas and Roberto Paredes. Overview of the ImageCLEF 2014 Scalable Concept Image Annotation Task. In *CLEF 2014 Evaluation Labs and Workshop, Online Working Notes*, 2014.
7. Fuzhen Zhuang, Ping Luo, Hui Xiong, Yuhong Xiong, Qing He, and Zhongzhi Shi. Cross-domain learning from multiple sources: A consensus regularization perspective. *IEEE Transactions on Knowledge and Data Engineering*, 22(12):1664–1678, 2010.