Real-Time Brand Logo Recognition

Leonardo Bombonato(✉), Guillermo Camara-Chavez, and Pedro Silva

Federal University of Ouro Preto, Ouro Preto, Minas Gerais, Brazil
leonardobombonato@gmail.com

Abstract. The increasing popularity of Social Networks makes change the way people interact. These interactions produce a huge amount of data and it opens the door to new strategies and marketing analysis. According to Instagram (https://instagram.com/press/) and Tumblr (https://www.tumblr.com/press), an average of 80 and 59 million photos respectively are published every day, and those pictures contain several implicit or explicit brand logos. The analysis and detection of logos in natural images can provide information about how widespread is a brand. In this paper, we propose a real-time brand logo recognition system, that outperforms all other state-of-the-art methods for the challenging FlickrLogos-32 dataset. We experimented with 5 different approaches, all based on the Single Shot MultiBox Detector (SSD). Our best results were achieved with the SSD 512 pretrained, where we outperform by 2.5% of F-score and by 7.4% of recall the best results on this dataset. Besides the higher accuracy, this approach is also relatively fast and can process with a single Nvidia Titan X 19 images per second.

Keywords: Computer vision · Brand logo recognition
Deep learning · CNN

1 Introduction

Brand logos are graphic entities that represent organizations, goods, *etc.* Logos are mainly designed for decorative and identification purposes. A specific logo can have several different representations, and some logos can be very similar in some aspects. Logo classification in natural scenes is a challenging problem since they often appear in various angles and sizes, making harder to extract keypoints especially due to significant variations in texture, poor illumination, and high intra-class variations (Fig. 1). The automatic classification of those logos gives to the marketing industry a powerful tool to evaluate the impact of brands. Marketing campaigns and medias can benefit with this tool, detecting unauthorized distributions of copyright materials.

Several techniques and approaches were proposed in the last decades for object classification, such as Bag of Visual Words (BoVW), Deep Convolutional Neural Networks (DCNN), feature matching with RANSAC, *etc.* The most successful approaches in logo classification were based on the BoVW model

M. Mendoza and S. Velastín (Eds.): CIARP 2017, LNCS 10657, pp. 111–118, 2018.
https://doi.org/10.1007/978-3-319-75193-1_14

Fig. 1. This figure exemplifies the challenges of classifying logos in natural scenes, such as high intra-class variation, warping, occlusion, rotation, translation and scales.

and DCNN. Most recent approaches in logo recognition use a Region-Based Convolution Neural Network (RCNN) [15], this network introduces a selective search to find candidates. This approach has shown great results comparing to others proposed in previous researches. Even though the good results achieved by RCNN, it is really hard to train and test due to the characteristics of selective search that generates several potential bounding boxes categorized by a classifier. After classification, post-processing is used to refine the bounding boxes, eliminating duplicate detection, and re-scoring the boxes.

In this paper, we propose an approach for logo detection based on a deep learning model. Our results outperform state-of-the-art approach results, achieving a higher accuracy on the FlickrLogos-32 dataset. Our proposal uses transfer learning to improve the logo image representations, being not only more accurate but also faster in logo detection, processing 19 images per second using a NVidia Titan X card.

2 Related Works

Different approaches for logo recognition have been proposed through the last years. A few years ago, only shallow classifiers have been proposed to solve this challenge [3,18,19]. But with the increasing popularity of deep learning frameworks and because of its success in image recognition, some researches started to come up [2,8].

The first successful approach in logo recognition was based on contours and shapes in images with a uniform background. Francesconi *et al.* [6] proposed an adaptive model using a recursive neural network, the authors used the area and the perimeter of the logo as features.

After 2007, with the popularization of SIFT, countless applications started to use it, due to its robustness to rotation and scale transformations and partially invariant to occlusions. Many approaches based their proposals on SIFT descriptors [3,17–20].

RANSAC became a popular learning module for object recognition since its use in Lowe's research work [13]. Lowe used this technique to compare matched descriptors and find outlines, eliminating the false positives matches and thus

locating the object. In logo recognition, some researchers explored this method and achieved significant results, e.g. [3,20].

Deep Convolutional Neural Networks are on the trends in computer vision and especially in object recognition. Recent approaches in logo recognition applied this technique, achieving impressive results like [5,7,8,15].

3 Deep Convolutional Neural Networks

Artificial Neural Network is a classification machine learning model where the layers are composed of interconnected neurons with learned weights. These weights are learned by a training process. Convolutional Neural Network (CNN) is a type of feedforward artificial neural network and a variation of a multilayer perceptron. A neural network with three or more hidden layers is called deep network.

Transfer Learning. In a CNN each layer learns to "understand" specific features from the image. The first layers usually learn generic features like edges and gradients, the more we keep forwarding in the layers, the more specific the features the layer detects. In order to "understand" these features, it is necessary to train the network, adjusting the net weights according to a predefined loss function. If the network weights initiate with random values, it requires much more images and training iterations compared to using pretrained weights. The use of net weights trained with other dataset is called "fine-tuning" and it demonstrates to be extremely advantageous compared to training a network from scratch [4]. This technique is useful when the number of training images per class is scarce (e.g. 40 images for this problem), which makes it hard for the CNN to learn. Furthermore, transfer learning also speeds up the training convergence [16].

Data Augmentation. Training a DCNN requires lots of data, especially very large/deep networks. When the dataset does not provide enough training images, we can add more images using data augmentation process. This process consists of creating new synthetic images, that simulate different view angles, distortions, occlusions, lighting changes, etc. This technique usually increases the robustness of the network resulting in better results.

3.1 Single Shot MultiBox Detector

Single Shot MultiBox Detector (SSD [11,12]) makes predictions based on feature maps taken at different stages, then it divides each one into a pre-established set of bounding boxes with different aspect ratios and scales. The bounding boxes adjust itself to better match the target object. The network generates scores using a regression technique for estimate the presence of each object category in each bounding box. The SSD increases its robustness to scale variations by concatenating feature maps from different resolutions into a final feature map. This network generates scores for each object category in each bounding box and produces adjustments to the bounding box that better match the object shape. At the end, a non-maximum suppression is applied to reduce redundant detections (Fig. 2).

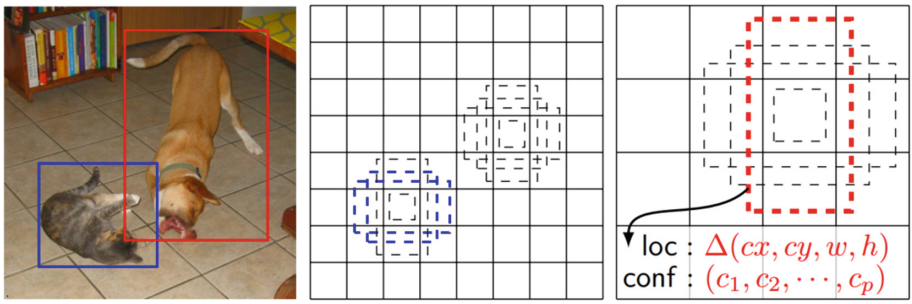

Fig. 2. (a) The final detection produced by the SSD. (b) A feature map with 8×8 grid. (c) A feature map with 4×4 grid and the output of each box, the location and scores for each class. Image extracted from [11].

SSD Variants. The SSD approach uses a base network to extract features from images and use them in detection layers. The extra layers in the SSD are responsible for detecting the object. There are some differences between SSD 300/500 and SSD 512. The SSD 512 is an upgrade of SSD 500, the improvements are presented as follows:

1. The pooling layer (*pool6*) between fully connected layers (*fc6* and *fc7*) was removed;
2. The authors added convolutional layers as extra layers;
3. A new color distortion data augmentation, used for improving the quality of the image, is also added;
4. The network populates the dataset by getting smaller training examples from expanded images;
5. Better proposed bounding boxes by extrapolating the image's boundary.

4 Our Approaches and Contributions

Logos detection can be considered a sub problem of object detection since they usually are objects with a planar surface. Our approaches are based on the SSD framework since it performs very well in object detection. We explore the performance of SDD model on logo images domain. We analyze the impact of using pretrained weights, rather than training from scratch, with the technique called transfer learning. We compare different implementations of the SSD and we also explore the impact of warping image transformations to meet the shape requirements of the SSD input layer.

4.1 Transfer Learning Methodology

To use the transfer learning technique was necessary to redesign the DCNN. This re-design remaps the last layer, adapting the class labels between two different

datasets. Therefore, all convolution and pooling layers are kept the same, and the last fully-connected layers (responsible for classification) are reorganized for the new dataset. For logos detection, the fine-tuning was made over a pretrained network, trained for 160.000 iterations on PASCAL VOC2007 + VOC2012 + COCO datasets [12].

4.2 Our Proposal Approaches

We explored 5 different approaches, Table 1 shows all different setups. All networks were trained for 100.000 iterations using the Nesterov Optimizer [14] with a fixed learning rate of 0.001. The SSD 300 and SSD 500 were only explored using pretrained weights because they were easily surpassed by the SSD 512. The approach SSD 500 AR was an attempt to reduce the warp transformation of the input image since in the training and testing phase, the SSD needs to fit the input image into a square resolution.

Table 1. Our five proposal approaches

Acronym	Method	Training details	Extra
SSD 300	SSD 300	Pretrained	
SSD 500	SSD 500	Pretrained	
SSD 500 AR	SSD 500	Pretrained	Preserving aspect ratio
SSD 512 FS	SSD 512	From scratch	
SSD 512 PT	SSD 512	Pretrained	

5 Experiments

We evaluate and analyze our approaches on FlickrLogos-32 dataset [19]. Our experiments ran on the Caffe deep learning framework [9] and using 2× Nvidia Tesla K80. We first describe the dataset, then we compare the performance of our approaches and finally we compare our results to state-of-the-art methods in logo recognition.

5.1 DataSet

FlickrLogos-32 (FL32) is a challenging dataset and the most promising approaches in logo recognition experimented their proposals on it. This dataset was proposed by Romberg [19], many approaches evaluated their performances on this dataset [3,5,8,18]. Romberg also defined an experimental protocol, splitting the dataset into training, validation and testing sets. In all approaches, we strictly follow this protocol. Table 2 shows the distribution between, train, validation and test sets. We have used P1 + P2 (except no-logos) for training and P3 for testing.

Table 2. Evaluation protocol table

Subset	Description	Images	Sum
P1	Hand-picked images, single logo, clean background	10 per class	320
P2	Images showing at least a single logo under various views	30 per class	3960
	Non-logo images	3000	
P3	Images showing at least a single logo under various views	30 per class	3960
	Non-logo images	3000	
Total			8240

5.2 Comparison of Our Approaches

All the 5 different approaches are represented in the left chart of Fig. 3, while the right figure shows the metrics for our best approach, the SSD pretrained. Analyzing the figure we can see that the approaches SSD 300, SSD 500 and SSD 500 AR achieved poor results if compared to the SSD 512. We see that in all cases using pretrained weights resulted in better performance. Analyzing only the best result, SSD 512 PT, we see that we achieve our best F-score with a threshold of 90.

5.3 Comparison Against Other Researches

The comparison among other researches and our best result (the SSD 512 with pretrained weights) can be seen in the Table 3. Analyzing the results we can see that our method outperforms by 2.5% the F-score and by 7.4% the recall

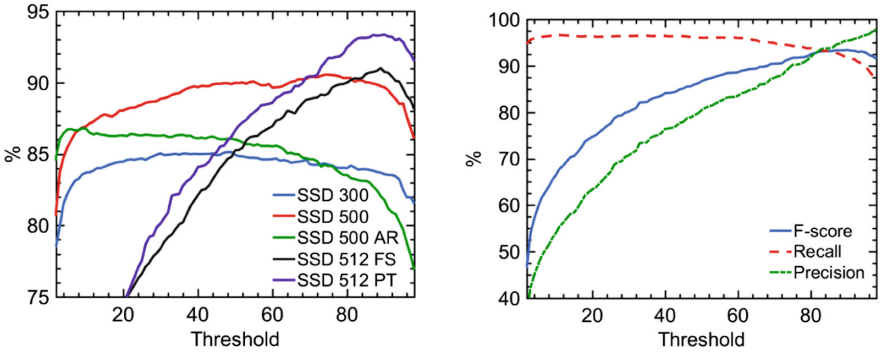

Fig. 3. The left image shows the F-score of out 5 different approaches. The right image shows the F-score, Precision and Recall of the best approach (AR - Preserving aspect ratio, FS - From scratch and PT - Fine-tuning of a pretrained model).

of the state-of-the-art. The high recall achieved is due to the fact that the SSD uses some of its extra layers to estimate the object location and also it can well generalize the object. The approach proposed by Li et al. [10] achieved such high precision due to the process of feature matching that eliminates false positive matches.

Table 3. Comparison of out best approach with the methods in the state-of-the-art (HCF - Hand Crafted Feature, DL - Deep Learning)

Method	Method	Year	Precision	Recall	F1
Romberg et al. [19]	HCF	2011	0.981	0.610	0.752
Revaud et al. [17]	HCF	2012	0.980	0.726	0.841
Romberg et al. [18]	HCF	2013	0.999	0.832	0.908
Li et al. [10]	HCF	2014	**1.000**	0.800	0.890
Bianco et al. [1]	DL	2015	0.909	0.845	0.876
Eggert et al. [5]	DL	2015	0.996	0.786	0.879
Oliveira et al. [15]	DL	2016	-	-	0.890
Bianco et al. [2]	DL	2017	0.976	0.676	0.799
Our	DL	2017	0.954	**0.919**	**0.933**

6 Conclusion

In this work, we investigated the use of DCNN, transfer learning and data augmentation on logo recognition system. The combination among them has shown that DCNN is very suitable for this task, even with relatively small train set it provides greater recall and f-score. A relevant contribution of this paper is the use of data augmentation combined with transfer learning to surpass the lower data issue and allow to use deeper networks. These techniques improve the performance of DCNN in this scenario. The results of our approach reinforce the robustness of DCNN approach, which surpasses the F1-score literature results.

Acknowledgements. The authors thank UFOP and funding Brazilian agency CNPq.

References

1. Bianco, S., Buzzelli, M., Mazzini, D., Schettini, R.: Logo recognition using CNN features. In: Murino, V., Puppo, E. (eds.) ICIAP 2015. LNCS, vol. 9280, pp. 438–448. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23234-8_41
2. Bianco, S., Buzzelli, M., Mazzini, D., Schettini, R.: Deep learning for logo recognition. arXiv preprint arXiv:1701.02620 (2017)
3. Boia, R., Florea, C.: Homographic class template for logo localization and recognition. In: Paredes, R., Cardoso, J.S., Pardo, X.M. (eds.) IbPRIA 2015. LNCS, vol. 9117, pp. 487–495. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19390-8_55

4. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. arXiv preprint arXiv:1405.3531 (2014)
5. Eggert, C., Winschel, A., Lienhart, R.: On the benefit of synthetic data for company logo detection. In: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, pp. 1283–1286. ACM (2015)
6. Francesconi, E., Frasconi, P., Gori, M., Marinai, S., Sheng, J.Q., Soda, G., Sperduti, A.: Logo recognition by recursive neural networks. In: Tombre, K., Chhabra, A.K. (eds.) GREC 1997. LNCS, vol. 1389, pp. 104–117. Springer, Heidelberg (1998). https://doi.org/10.1007/3-540-64381-8_43
7. Hoi, S.C., Wu, X., Liu, H., Wu, Y., Wang, H., Xue, H., Wu, Q.: Logo-net: large-scale deep logo detection and brand recognition with deep region-based convolutional networks. arXiv preprint arXiv:1511.02462 (2015)
8. Iandola, F.N., Shen, A., Gao, P., Keutzer, K.: Deeplogo: hitting logo recognition with the deep neural network hammer. arXiv preprint arXiv:1510.02131 (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)
10. Li, K.W., Chen, S.Y., Su, S., Duh, D.J., Zhang, H., Li, S.: Logo detection with extendibility and discrimination. Multimedia Tools Appl. **72**(2), 1285–1310 (2014)
11. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.: SSD: single shot multibox detector. arXiv preprint arXiv:1512.02325 (2015)
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
14. Nesterov, Y.: A method of solving a convex programming problem with convergence rate O(1/k2) (1983)
15. Oliveira, G., Frazão, X., Pimentel, A., Ribeiro, B.: Automatic graphic logo detection via fast region-based convolutional networks. arXiv preprint arXiv:1604.06083 (2016)
16. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. knowl. Data Eng. **22**(10), 1345–1359 (2010)
17. Revaud, J., Douze, M., Schmid, C.: Correlation-based burstiness for logo retrieval. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 965–968. ACM (2012)
18. Romberg, S., Lienhart, R.: Bundle min-hashing for logo recognition. In: Proceedings of the 3rd ACM International Conference on Multimedia Retrieval, pp. 113–120. ACM (2013)
19. Romberg, S., Pueyo, L.G., Lienhart, R., Van Zwol, R.: Scalable logo recognition in real-world images. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, p. 25. ACM (2011)
20. Yang, F., Bansal, M.: Feature fusion by similarity regression for logo retrieval. In: 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), p. 959. IEEE (2015)