Open Set Logo Detection and Retrieval

Paper ID 666??

Keywords: Logo Detection, Logo Retrieval, Logo Dataset, Trademark Retrieval, Open Set Retrieval, Deep Learning.

Abstract: Current logo retrieval research focuses on closed set scenarios. We argue that the logo domain is too large for this strategy and requires an open set approach. To foster research in this direction, a large-scale logo dataset, called Logos in the Wild, is collected and released to the public. A typical open set logo retrieval applications is, for example, judging the effectiveness of advertisement in sports event broadcasts. Given a query sample in shape of a logo image, the task is to find all further occurrences of this logo in a set of images or videos. Currently common logo retrieval approaches are unsuitable for this task because of their closed world assumption. Thus, an open set logo retrieval method is proposed in this work which allows searching for previously unfamiliar logos by a single query sample. A two stage concept with separate logo detection and comparison is proposed where both modules are based on task specific Convolutional Neural Networks (CNNs). If trained with the Logos in the Wild data, significant performance improvements are observed, especially compared with state-of-the-art closed set approaches.

1 INTRODUCTION

Automated search for logos is a desirable task in visual image analysis. A key application is the effectiveness measurement of advertisements. Being able to find all logos that match a query, for example, a logo of a specific company, in images allows to assess the visual frequency and prominence of logos in TV broadcasts. Typically, these broadcasts are sports events where sponsorship and advertisement is very common. This requires a flexible system where the query can easily be defined and switched according to the current task. Especially, also previously unseen logos should be found if one query sample is available. This requirement excludes basically all current logo retrieval approaches because they make a closed world assumption where all searched logos are known beforehand. Instead, this paper focuses on open set logo retrieval where only one sample image of a logo is available.

Consequently, a novel processing strategy for logo retrieval based on a logo detector and a feature extractor is proposed as illustrated in figure 1. Similar strategies are known from other open set retrieval tasks, such as face or person retrieval (Bäuml et al., 2010; Herrmann and Beyerer, 2015). Both, the detector and the extractor are task specific CNNs. For detection, the Faster R-CNN framework (Ren et al., 2015) is employed and the extractor is derived from classification networks for the ImageNet challenge (Deng et al., 2009).

The necessity for open set logo retrieval becomes

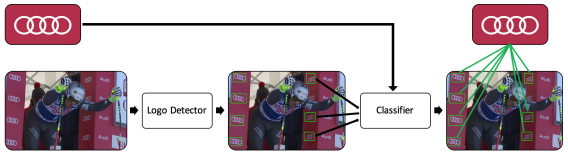

Figure 1: Proposed logo retrieval strategy.

obvious when having a look at the diversity and amount of existing logos and brands[1]. The METU trademark dataset (Tursun et al., 2017) contains, for example, over half a million different brands. Given this number, a closed set approach where all different brands are pre-trained within the retrieval system is clearly inappropriate. This is why our proposed feature extractor generates a discriminative logo descriptor, which generalizes to unseen logos, instead of a mere classification between previously known brands. The well-known high discriminative capabilities of CNNs allow to construct such a feature extractor.

One challenge for training a general purpose logo detector lies in appropriate training data. Many logo or trademark dataset (Eakins et al., 1998; Hoi et al., 2015; Tursun et al., 2017) only contain the original logo graphic but no in-the-wild occurrences of these logos which are required for the target application. The need for annotated logo bounding boxes in the images limits the number of suitable datasets. Exist-

[1]The term brand is used in this work as synonym for a single logo class. Thus, a brand might also refer to a product or company name if an according logo exists.

ing logo datasets (Joly and Buisson, 2009; Kalantidis et al., 2011; Romberg et al., 2011; Letessier et al., 2012; Bianco et al., 2015; Su et al., 2016; Bianco et al., 2017) with available bounding boxes are often restricted to a very small number of brands and mostly high quality images. Especially, occlusions, blur and variations within a logo type are only partially covered. To address these shortcomings, the novel Logos in the Wild dataset is collected and made publicly available [2].

The contributions of this work are threefold:

- A novel open set logo detector which can detect previously unseen logos.
- An open set logo retrieval system which needs only a single logo image as query.
- The introduction of a novel large-scale in-the-wild logo dataset.

2 RELATED WORK

Current logo retrieval strategies are generally solving a closed set detection and classification problem. Eggert et.al. (Eggert et al., 2015) utilized CNNs to extract features from logos and determined their brand by classification with a set of Support Vector Machines (SVMs). Fast R-CNN (Girshick, 2015) was used for the first time to retrieve logos from images by Iandola et al. (Iandola et al., 2015) and achieved superior results on the FlickrLogos-32 dataset (Romberg et al., 2011). Furthermore, R-CNN, Fast R-CNN and Faster R-CNN were used in (Bao et al., 2016; Oliveira et al., 2016; Qi et al., 2017). As closed set methods, all of them use the same brands for training as for validation.

2.1 Open Set Retrieval

Retrieval scenarios in other domains are basically always considered open set, i.e., samples from the currently searched class have never been seen before. This is the case for general purpose image retrieval (Sivic and Zisserman, 2003), tattoo retrieval (Manger, 2012) or for person retrieval in image or video data where face or appearance-based methods are common (Bäuml et al., 2010; Weber et al., 2011; Herrmann and Beyerer, 2015). The reason is that these in-the-wild scenarios offer usually a too large and impossible to capture variety of object classes. In case of persons, a class would be a person identity resulting in a cardinality of billions. Consequently, methods are designed and trained on a limited set of classes and have to generalize to previously unseen classes. We argue that this approach is also required for logo retrieval because of the vast amount of existing brands and according logos which cannot be captured in advance. Typically, approaches targeting open set scenarios consist of an object detector and a feature extractor (Zheng et al., 2016). The detector localizes the objects of interest and the feature extractor creates a discriminative descriptor regarding the target classes which can than be compared to query samples.

2.2 Object Detector Frameworks

Early detectors applied hand-crafted features, such as Haar-like features, combined with a classifier to detect objects in images (Viola and Jones, 2004). Nowadays, deep learning methods surpass the traditional methods by a significant margin. In addition, they allow a certain level of object classification within the detector which is mostly used to simultaneously detect different object categories, such as persons and cars (Sermanet et al., 2013). The YOLO detector (Redmon et al., 2015) introduces an end-to-end network for object detection and classification based on bounding box regressors for object localization. This concept is similarly applied by the Single Shot MultiBox Detector (SSD) (Liu et al., 2016). Faster Region-Based Convolutional Neural Network (R-CNN) (Ren et al., 2015) introduces a Region Proposal Network (RPN) to detect object candidates in the feature maps and classifies the candidate regions by a fully connected network. Improvements of the Faster R-CNN are the Region-based Fully Convolutional Network (R-FCN) (Dai et al., 2016), which reduces inference time by an end-to-end fully convolutional network, and the Mask R-CNN (He et al., 2017), adding a classification mask for instance segmentation.

2.3 CNN-based Classification

AlexNet (Krizhevsky et al., 2012) was the first neural network after the conquest of SVMs, achieving impressive performance on image content classification and winning the ImageNet challenge (Deng et al., 2009). It consists of five convolutional layers, each followed by a max-pooling, which counted as a very deep network at the time. VGG (Simonyan and Zisserman, 2015) follows the general architecture of AlexNet with an increased number of convolutional layers achieving better performance. The inception architecture (Szegedy et al., 2015) proposed a multi-

[2] `url://to.come`

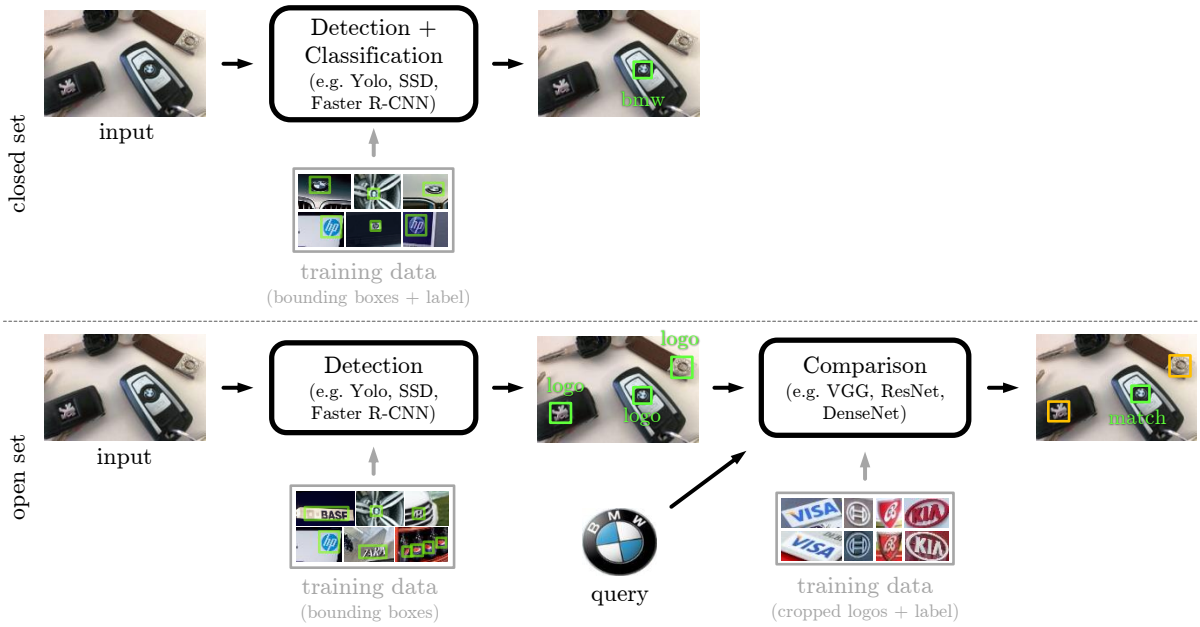

Figure 2: Comparison of closed and open set logo retrieval strategy.

path network module for better multi-scale addressing, but was shortly after superseded by the Residual Networks (ResNet) (He et al., 2015; He et al., 2016). They increase network depth heavily up to 1000 layers in the most extreme configurations by additional skip connections which bypass two convolutional layers. The recent DenseNet (Huang et al., 2016a) builds on a ResNet-like architecture and introduces "dense units". The output of these units is connected with every subsequent dense unit's input by concatenation. This results in a much denser network than a conventional feed-forward network.

3 LOGO DETECTION

The current state-of-the-art approaches for scene retrieval create a global feature of the input image. This is achieved by either inferring from the complete image or by searching for key regions and then extracting features from the located regions, which are finally fused into a global feature (Torii et al., 2015; Arandjelovic et al., 2016; Kalantidis et al., 2016). For logo retrieval, extraction of a global feature is counterproductive because it lacks discriminative power to retrieve small objects. Additionally, global features usually include no information about the size and location of the objects which is also an important factor for logo retrieval applications.

Therefore, we choose a two-stage approach consisting of logo detection and logo classification as figure 2 illustrates for the open set case. First, the logos have to be detected in the input image. Since currently almost only Faster R-CNNs (Ren et al., 2015) are used in the context of logo retrieval, we follow this choice for better comparability and because it offers a straight forward baseline method. Other state-of-the-art detector options, such as SSD (Liu et al., 2016) or YOLO (Redmon and Farhadi, 2016), potentially offer a faster detection at the cost of detection performance (Huang et al., 2016b).

Detection networks trained for the currently common closed set assumption are unsuitable to detect logos in an open set manner. By considering the output brand probability distribution, no derivation about occurrences of other brands are possible. Therefore, the task raises the need for a generic logo detector, which is able to detect all logo brands in general.

Baseline

Faster R-CNN consists of two stages, the first being an RPN to detect object candidates in the feature maps and the second being classifiers for the candidate regions. While the second stage sharply classifies the trained brands, the RPN will generate candidates that vaguely resemble any of the brands which is the case for many other logos. Thus, it provides an indicator whether a region of the image is a logo or not. The trained RPN and the underlying feature extractor network are isolated and employed as a baseline open set logo detector.

Table 1: Publicly available in-the-wild logo datasets in comparison with the novel Logos in the Wild dataset.

	dataset	brands	logo images	RoIs
public	BelgaLogos (Joly and Buisson, 2009; Letessier et al., 2012)	37	1,321	2,697
	FlickrBelgaLogos (Letessier et al., 2012)	37	2,697	2,697
	Flickr Logos 27 (Kalantidis et al., 2011)	27	810	1,261
	FlickrLogos-32 (Romberg et al., 2011)	32	2,240	3,404
	Logos-32plus (Bianco et al., 2015; Bianco et al., 2017)	32	7,830	12,300
	TopLogo10 (Su et al., 2016)	10	700	863
	combined	80 (union)	15,598	23,222
new	Logos in the Wild	872	11,054	32,850

Brand Agnostic

The RPN strategy is by no means optimal because it obviously has a bias towards the pre-trained brands and also generates a certain amount of false positives. Therefore, another option to detect logos is suggested which we call the brand agnostic Faster R-CNN. It is trained with only two classes: background and logo. We argue that this solution which merges all brands into a single class yields better performance than the RPN detector because of two reasons. First, in the second stage, fully connected layers preceding the output layer serve as strong classifiers which are able to eliminate false positives. Second, these layers also serve as stronger bounding box regressors improving the localization precision of the logos.

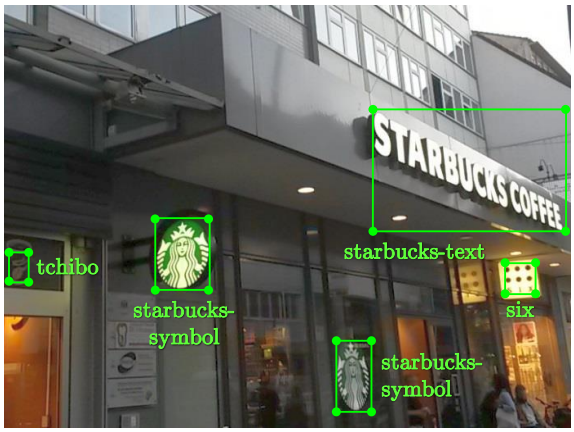

Figure 4: Annotations differentiate between textual and graphical logos.

4 LOGO COMPARISON

After logos are detected, the correspondences to the query sample have to be searched. For logo retrieval, features are extracted from the detected logos for comparison with the query sample. Thus, the logo feature vectors for the query image and the ones for the database are collected and normalized. Pair-wise comparison is then performed by cosine similarity.

In order to retrieve as many logos from the images as possible, the detector has to operate at a high recall. However, for difficult tasks, such as open set logo detection, high recall values induce a certain amount of false positive detections. The feature extraction step thus has to be robust and tolerant to these false positives.

Donahue et al. suggested that CNNs can produce excellent descriptors of an input image even in the absence of fine-tuning to the specific domain of the image (Donahue et al., 2015). This motivates to apply a network pre-trained on a very large dataset as feature extractor. Namely, several state-of-the-art CNNs trained on the ImageNet dataset (Deng et al., 2009) are explored for this task. To adjust the network to the logo domain and the false positive removal, the networks are fine-tuned on logo detections. The final network layer is extracted as logo feature in all cases.

Altogether, the proposed logo retrieval system consists of a class agnostic logo detector and a feature extractor network. This setup is advantageous for the quality of the extracted logo features because the extractor network has only to focus on a specific region. This is an improvement compared to including both logo detection and comparison in the regular Faster R-CNN framework which lacks generalization to unseen classes. We argue that the specialization in the regular Faster R-CNN to the limited number of specific brands in the training set does not cover the complexity and breadth of the logo domain. This is why a separate and more elaborate feature extractor is proposed even though this might appear like a less elegant solution.

5 LOGO DATASET

To train the proposed logo detector and feature extractor, a novel logo dataset is collected to supple-

Figure 3: Examples from the collected Logos in the Wild dataset.

ment publicly available logo datasets. A comparison to other public in-the-wild datasets with annotated bounding boxes is given in table 1. The goal is an in-the-wild logo dataset with images including logos instead of the raw original logo graphics. In addition, images where the logo represents only a minor part of the image are preferred. See figure 3 for a few examples of the collected data. Following the general suggestions from (Bansal et al., 2017), we target for a dataset containing significantly more brands instead of collecting additional image samples for the already covered brands. This is the exact opposite strategy than performed by the Logos-32plus dataset. Starting with a list of well-known brands and companies, an image web search is performed. Because most other web collected logo datasets mainly rely on Flickr, we opt for Google image search to broaden the domain. Brand or company names are searched directly or in combination with a predefined set of search terms, e.g., 'advertisement', 'building', 'poster' or 'store'.

For each search result, the first N images are downloaded, where N is determined by a quick manual inspection to avoid collecting too much garbage. After removing duplicates, this results in 4 to 608 images per searched brand. These images are then one-by-one manually annotated with logo bounding boxes or sorted out if unsuitable. Images are considered unsuitable if they contain no logos or fail the in-the-wild requirement, which is the case for the original raw logo graphics. Taken pictures of such logos and advertisement posters on the other hand are desired to be in the dataset. Annotations distinguish between textual and graphical logos as well as different logos

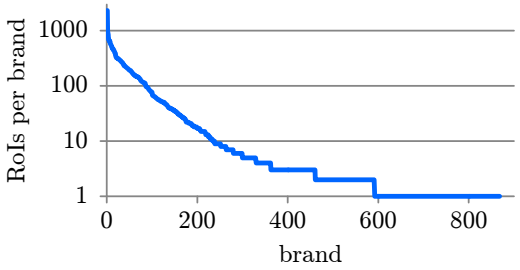

Figure 5: Distribution of number of RoIs per brand.

from one company as exemplary indicated in figure 4. Altogether, the current version of the dataset, contains 872 brands with 32,850 annotated bounding boxes. 238 brands occur at least 10 times. An image may contain several logos with the maximum being 118 logos in one image. The full distributions are shown in figures 5 and 6.

The collected Logos in the Wild dataset exceeds the size of all related logo datasets as shown in table 1. Even the union of all related logo datasets contains significantly less brands and RoIs which makes Logos in the Wild a valuable large-scale dataset. As the annotation is still an ongoing process, different dataset revisions will be tagged by version numbers for future reference. Note that the numbers in table 1 are the current state (v2.0) whereas detector and feature extractor training used a slightly earlier version with numbers given in table 2 (v1.0) because of the required time for training and evaluation.

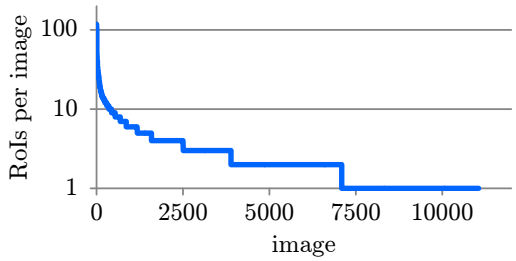

Figure 6: Distribution of number of RoIs per image.

Table 2: Train and test set statistics.

phase	data	brands	RoIs
train	public	47	3,113
	public+LitW v1.0	632	18,960
test	FlickrLogos-32 test	32	1,602

6 EXPERIMENTS

The proposed method is evaluated on the test set benchmark of the public FlickrLogos-32 dataset including the distractors. Additional application specific experiments are performed on an internal dataset of sports event TV broadcasts. The training set consists of two parts. The union of all public logo datasets as listed in table 1 and the novel Logos in the Wild (LitW) dataset. For a proper separation of train and test data, all brands which are present in the FlickrLogos-32 test set are removed from the public and LitW data. 10 percent of the remaining images are set aside for network validation in each case. This results in the final training and test set sizes listed in table 2.

In the first step, the detector stage alone is assessed. Then, the combination of detection and comparison for logo retrieval is evaluated. Detection and matching performance is measured by the Free-Response Receiver Operating Characteristic (FROC) curve (Miller, 1969) which denotes the detection or detection and identification rate versus the number of false detections.

In all cases, the CNNs are trained until convergence. Due to the diversity of applied networks and differing dataset sizes, training settings are numerous and optimized in each case with the validation data. Convergence occurs after 200 to 8,000 training iterations with a varying batch-size of 1 for the Faster R-CNN detector, 7 for the DenseNet161, 18 for the ResNet101 and 32 for the VGG16 training due to GPU memory limitation.

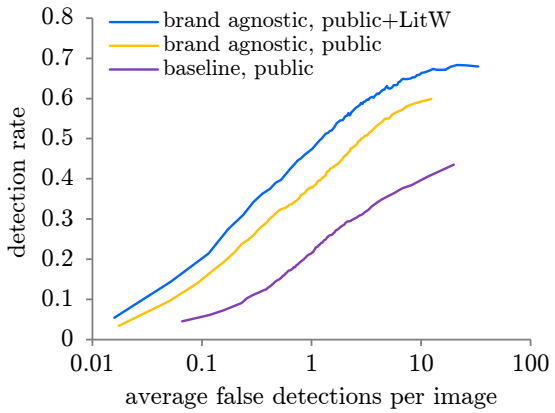

Figure 7: Detection FROC curves for the FlickrLogos-32 test set.

6.1 Detection

As baseline, the state-of-the-art closed set logo retrieval method from (Su et al., 2016) based on Faster R-CNN is employed and trained on the public portion of the training data. It is naively adapted to open set detection by using the RPN scores as detections. This skips the closed set classification part of the network which is trained on different logos than should be detected on the test set. The proposed brand agnostic logo detector is first trained on the same public data for comparison. All Faster R-CNN detectors are based on the VGG16 network. The results in figure 7 indicate that the proposed brand agnostic strategy is superior by a significant margin.

Further improvement is achieved by combining the public training data with the novel logo data. Adding LitW as additional training data improves the detection results with its large variety of additional training brands. This confirms findings from other domains, such as face analysis, where wider training datasets are preferred over deeper ones (Bansal et al., 2017). This means it is better to train on additional different brands than on additional samples per brand. As direction for future dataset collection, this suggests to focus on additional brands.

6.2 Retrieval

For the retrieval experiments, the Faster R-CNN based state-of-the-art closed set logo retrieval method from the previous section serves again as baseline. Now the full network is applied and the logo class probabilities of the second stage are interpreted as feature vector which is then used to match previously unseen logos. For the proposed open set strategy, the best logo detection network from the pre-

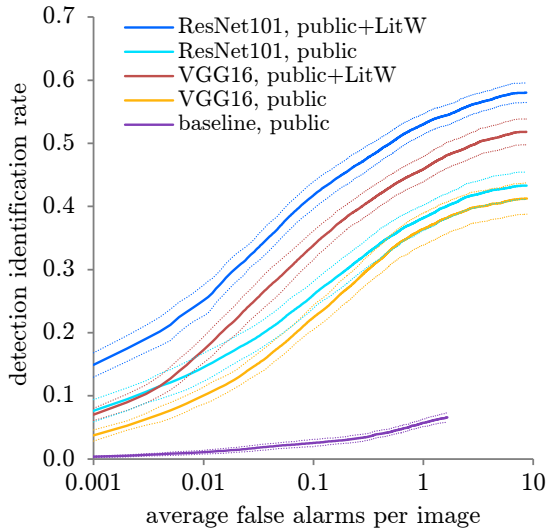

Figure 8: Detection+Classification FROC curves for the FlickrLogos-32 test set. Including dashed indicators for one standard deviation. DenseNet results are omitted for clarity, refer to table 3 for full results.

vious section is used in all cases. Detected logos are described by the feature extraction network outputs where three different state-of-the-art classification architectures, namely VGG16 (Simonyan and Zisserman, 2015), ResNet101 (He et al., 2015) and DenseNet161 (Huang et al., 2016a), serve as base networks. All networks are pretrained on ImageNet and afterwards fine-tuned either on the public logo train set or the combination of the public and the LitW train data.

FlickrLogos-32

In 10 iterations, each of the 10 FlickrLogos-32 train samples for each brand serves as query sample. This allows to assess the statistical significance of results similar to a 10-fold-cross-validation strategy. Figure 8 shows the FROC results for the trained networks including indicators for the standard deviation of the measurements. The detection identification rate denotes the amount of ground truth logos which are correctly detected and are assigned the correct brand. While the baseline method is only able to find a minor amount of the logos, our best performing approach is able to correctly retrieve 25 percent of the logos if tolerating only one false alarm every 100 images. As expected, the more recent network architectures provide better results. Also, including the LitW data in the training yields a significant boost in performance. Specifically, the larger training dataset has a larger impact on the performance than a better network architecture. Table 3 compares our open set results

Table 3: FlickrLogos-32 test set retrieval results.

setting	method	*map*
open set	baseline, public (Su et al., 2016)	0.036
	VGG16, public	0.286
	ResNet101, public	0.327
	DenseNet161, public	0.368
	VGG16, public+LitW	0.382
	ResNet101, public+LitW	**0.464**
	DenseNet161, public+LitW	0.448
closed set	BD-FRCN-M (Oliveira et al., 2016)	0.735
	DeepLogo (Iandola et al., 2015)	0.744
	Faster-RCNN (Su et al., 2016)	0.811
	Fast-M (Bao et al., 2016)	**0.842**

with closed set results from the literature in terms of the mean average precision (*map*). We achieve more than half of the closed set performance in terms of *map* with only one sample for a brand at test time instead of dozens or hundreds of brand samples at training time. In addition, our approach is not limited to the 32 FlickrLogos brands but generalizes with a similar performance to further brands. In contrast, the closed set approaches hardly generalize as is shown by the baseline open set method. This is the second best closed set approach only retrained on out-of-test brands.

SportsLogos

In addition to public data, target domain specific experiments are performed on TV broadcasts of sports events. In total, this SportsLogos test set includes 298 annotated frames with 2,348 logos of 40 brands. In comparison to public logo datasets, the logos are usually significantly smaller and cover only a tiny fraction of the image area as illustrated in figure 9. Besides perimeter advertising, logos on clothing or equipment of the athletes and TV station or program overlays are the most occurring logo types. Overall, the results in this application scenario are slightly worse than in the FlickrLogos-32 benchmark with a drop in *map* from 0.464 to 0.354 for the best performing method, as indicated in figure 10. The baseline approach takes the largest performance hit showing that closed set approaches not only generalize badly to unseen logos but also to novel domains. In contrast, the proposed open set strategy shows a relatively stable cross-domain performance. Training with LitW data again improves the results significantly.

7 CONCLUSIONS

The limits of closed set logo retrieval approaches motivate the proposed open set approach. By this, gen-

Figure 9: Example football scene with small logos in the perimeter advertising.

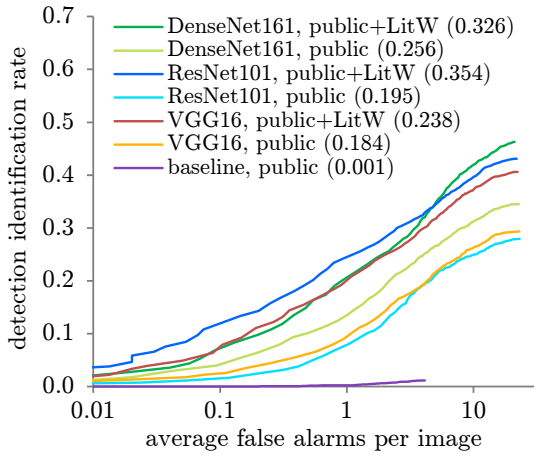

Figure 10: Detection+Classification FROC curves for the SportsLogos test set, *map* is given in brackets.

eralization to unseen logos and novel domains is improved significantly in comparison to a naive extension of closed set approaches to open set configurations. Due to the large logo variety, open set logo retrieval is still a challenging task where trained methods benefit significantly from larger datasets. The lack of sufficient data is addressed by introduction of the large-scale Logos in the Wild dataset. Despite being bigger than all other in-the-wild logo datasets combined, dataset sizes should probably be scaled even further in the future. Adding the Logos in the Wild data in the training improves the mean average precision from 0.368 to 0.464 for open set logo retrieval on FlickrLogos-32.

REFERENCES

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307.

Bansal, A., Castillo, C., Ranjan, R., and Chellappa, R. (2017). The Do's and Don'ts for CNN-based Face Verification. *arXiv preprint arXiv:1705.07426*.

Bao, Y., Li, H., Fan, X., Liu, R., and Jia, Q. (2016). Region-based CNN for Logo Detection. In *International Conference on Internet Multimedia Computing and Service*, ICIMCS'16, pages 319–322, New York, NY, USA. ACM.

Bäuml, M., Bernardin, K., Fischer, M., Ekenel, H., and Stiefelhagen, R. (2010). Multi-pose face recognition for person retrieval in camera networks. In *International Conference on Advanced Video and Signal-Based Surveillance*. IEEE.

Bianco, S., Buzzelli, M., Mazzini, D., and Schettini, R. (2015). Logo recognition using cnn features. In *International Conference on Image Analysis and Processing*, pages 438–448. Springer.

Bianco, S., Buzzelli, M., Mazzini, D., and Schettini, R. (2017). Deep learning for logo recognition. *Neurocomputing*, 245:23–30.

Dai, J., Li, Y., He, K., and Sun, J. (2016). R-FCN: Object Detection via Region-based Fully Convolutional Networks. *arXiv preprint arXiv:1605.06409*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Con-*

ference on Computer Vision and Pattern Recognition, pages 2625–2634. IEEE.

Eakins, J. P., Boardman, J. M., and Graham, M. E. (1998). Similarity retrieval of trademark images. *IEEE multimedia*, 5(2):53–63.

Eggert, C., Winschel, A., and Lienhart, R. (2015). On the Benefit of Synthetic Data for Company Logo Detection. In *ACM Multimedia Conference*, MM '15, pages 1283–1286, New York, NY, USA. ACM.

Girshick, R. (2015). Fast R-CNN. In *International Conference on Computer Vision*.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask R-CNN. *arXiv preprint arXiv:1703.06870*.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027*.

Herrmann, C. and Beyerer, J. (2015). Face Retrieval on Large-Scale Video Data. In *Canadian Conference on Computer and Robot Vision*, pages 192–199. IEEE.

Hoi, S. C. H., Wu, X., Liu, H., Wu, Y., Wang, H., Xue, H., and Wu, Q. (2015). LOGO-Net: Large-scale Deep Logo Detection and Brand Recognition with Deep Region-based Convolutional Networks. *CoRR*, abs/1511.02462.

Huang, G., Liu, Z., and Weinberger, K. Q. (2016a). Densely Connected Convolutional Networks. *CoRR*, abs/1608.06993.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al. (2016b). Speed/accuracy trade-offs for modern convolutional object detectors. *arXiv preprint arXiv:1611.10012*.

Iandola, F. N., Shen, A., Gao, P., and Keutzer, K. (2015). DeepLogo: Hitting Logo Recognition with the Deep Neural Network Hammer. *CoRR*, abs/1510.02131.

Joly, A. and Buisson, O. (2009). Logo retrieval with a contrario visual query expansion. In *ACM Multimedia Conference*, pages 581–584.

Kalantidis, Y., Mellina, C., and Osindero, S. (2016). Cross-dimensional weighting for aggregated deep convolutional features. In *European Conference on Computer Vision*, pages 685–701. Springer.

Kalantidis, Y., Pueyo, L., Trevisiol, M., van Zwol, R., and Avrithis, Y. (2011). Scalable Triangulation-based Logo Recognition. In *ACM International Conference on Multimedia Retrieval*, Trento, Italy.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, pages 1097–1105. Curran Associates, Inc.

Letessier, P., Buisson, O., and Joly, A. (2012). Scalable mining of small visual objects. In *ACM Multimedia Conference*, pages 599–608. ACM.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer.

Manger, D. (2012). Large-scale tattoo image retrieval. In *Canadian Conference on Computer and Robot Vision*, pages 454–459. IEEE.

Miller, H. (1969). The FROC Curve: a Representation of the Observer's Performance for the Method of Free Response. *The Journal of the Acoustical Society of America*, 46(6(2)):1473–1476.

Oliveira, G., Frazão, X., Pimentel, A., and Ribeiro, B. (2016). Automatic Graphic Logo Detection via Fast Region-based Convolutional Networks. *CoRR*, abs/1604.06083.

Qi, C., Shi, C., Wang, C., and Xiao, B. (2017). Logo Retrieval Using Logo Proposals and Adaptive Weighted Pooling. *IEEE Signal Processing Letters*, 24(4):442–445.

Redmon, J., Divvala, S. K., Girshick, R. B., and Farhadi, A. (2015). You Only Look Once: Unified, Real-Time Object Detection. *CoRR*, abs/1506.02640.

Redmon, J. and Farhadi, A. (2016). YOLO9000: better, faster, stronger. *arXiv preprint arXiv:1612.08242*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

Romberg, S., Pueyo, L. G., Lienhart, R., and van Zwol, R. (2011). Scalable Logo Recognition in Real-world Images. In *ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 25:1–25:8, New York, NY, USA. ACM.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *CoRR*, abs/1312.6229.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, pages 1470–1477. IEEE.

Su, H., Zhu, X., and Gong, S. (2016). Deep Learning Logo Detection with Data Expansion by Synthesising Context. *CoRR*, abs/1612.09322.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Conference on Computer Vision and Pattern Recognition*, pages 1–9. IEEE.

Torii, A., Arandjelovic, R., Sivic, J., Okutomi, M., and Pajdla, T. (2015). 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817.

Tursun, O., Aker, C., and Kalkan, S. (2017). A Large-scale Dataset and Benchmark for Similar Trademark Retrieval. *arXiv preprint arXiv:1701.05766*.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Weber, M., Bäuml, M., and Stiefelhagen, R. (2011). Part-based clothing segmentation for person retrieval. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 361–366. IEEE.

Zheng, L., Zhang, H., Sun, S., Chandraker, M., and Tian, Q. (2016). Person Re-identification in the Wild. *CoRR*, abs/1604.02531.