Karlsruhe Institute of Technology

Fraunhofer IOSB

Logo retrieval in mass data using deep learning

Master thesis

KIT - Karlsruhe Institute of Technology
Fraunhofer IOSB - Fraunhofer Institute of Optronics,
System Technologies and Image Exploitation

Andras Tüzkö

June 14, 2017

Main Advisor:	Dipl.-Inform. Christian Herrmann
Co-Advisor:	Dipl.-Inform. Daniel Manger

Interaktive Echtzeitsysteme

Statement of authorship

I hereby declare that I have produced this work by myself except the utilities known to the supervisor, that I have labeled all used utilities completely and detailed and that I have labeled all material that has been taken with or without modification from the work of others.

Karlsruhe, June 14, 2017 Andras Tüzkö Andras Tüzkö

Contents

1 Introduction

Advertising with static logos is one of the most important marketing methods. A very effective way to reach a lot of people with these static logos is, to sponsor sport teams or to buy advertising spaces in sport events broadcasted on the TV. However, the prices of these surfaces mean huge expenses for the advertiser. This is the reason why the need for logo appearance statistics of sport videos arises. In particular there is a desire for quantitative measurement of the proportional size of the logo to the screen, and of the time one particular logo is visible on the screen. This data is then used to judge the cost efficiency for the specific logo placement, i.e. to be able to decide on which sport event to advertise, with which size of logo, and where to place it.

In this work a system for logo retrieval with proposal based object detection and classification will be presented. The system consists of a logo detector, and a classifier used for feature extraction. The logo detector is a faster region based convolutional neural network [Ren15] trained to recognize logos on images. The features of the proposed regions are extracted with a ResNet neural network [He15]. To recognize logos in videos, the videos will be cut into frames, and then the system will be run on every image.

The challenge of this task is manifold. The first problem is that the logos in these videos are far from being perfectly clear. They can be partially occluded, blurred - if the camera is moving fast, perspectively transformed, rotated and can have various coloring, suiting well to the design of the shirt or the arena. In addition, there is a problem with the ambient illumination variation just as for other computer vision tasks. Second challenge is the large variety of different brand's logos. This makes the detection of logos very challenging. Furthermore, there are only a few smaller publicly available datasets, with bounding box annotated logos, and the majority of the images are adjusted to ensure a good visibility of the logos on them, not like the frames of the sport videos.

In the decade before, hand-crafted feature extraction was prevalent in computer vision tasks. It needed an expert to create such a system, and it yielded often only mediocre results. Deep learning methods for computer vision problems are dominant since the success of convolutional neural networks in 2012 [Kri12]. The great improvement of deep learning methods is, compared to earlier systems (e.g. SIFT [Low04], HOG [Dal05] features), to learn how to extract features automatically. The development of deep nets is mainly powered by the annually organized ImageNet classification challenge [Rus15]. Since the aim of this contest to classify an object, which is filling out the majority of an image, the location of the particular object is irrelevant. To be able to classify and recognize objects which have a much smaller size relative to the size of the whole image, region based classification can be utilized.

The rest of this thesis is organized as follows. Section 2 reviews the related work within image retrieval, object detection and logo retrieval. In Section 3 the proposal based object detection with convolutional neural networks will be introduced. Section 4 describes the logo retrieval system.

Afterwards, Section 5 includes evaluation and comparison of the system with another logo retrieval method. Finally, the last section concludes the work and gives prospects on future work.

2

Related Work

2.1 Image Retrieval

Many technics outside the scope of deep learning exists for images retrieval from videos. SIFT features [Low04] with bag-of-visual-words were used to efficiently get translation invariant descriptors around keypoints by Zisserman [Siv03]. HOG [Dal05] SIFT, HOG blockwise orientation histograms

2.2 Object Detection

Viola Jones

2.3 Logo Retrieval

3

Proposal Based Object Detection and Classification

In this section the theoretical overview of the logo retrieval system will be presented. First of all the fully convolutional networks will be introduced in section 3.1. Section 3.2 explains region proposal systems for generating candidate object locations on an image. Afterwards, the section 3.3 describes region based convolutional neural networks for object detection and classification. The improvement of this method, the fast region based convolutional neural networks will be detailed in the section 3.3.2. Folowing this, the further development, the faster region based convolutional neural networks will be reviewed in the section 3.3.3.

3.1 Fully Convolutional Neural Networks

A neural network is fully convolutional, if it does not contain any fully connected layers. Firstly Matan et.al. used FCNs for recognizing strings of digits. Long et. al. proposed [Lon14] how to transform a deep neural network with fully connected classifier layers at the end, to a fully convolutional network. For this purpose the fully connected layers at the end of the network are to be converted to convolutional layers.

Since the number of weights of a neuron in a fully connected layer is defined by the shape of the data of the layer, the trained network can process only a fix-sized input. As a fully convolutional network does not have fully connected layer anymore, it has the advantage of being able to train and test with images of arbitrary sizes.

The outputs of such a network are two dimensional feature maps, which can be used as heatmaps per class. These convolutional maps can also be used directly for semantic segmentation, where each pixel of an image should be classified. Nowadays fully convolutional networks are essential part of state-of-the-art object detectors, yielding better performance, image size agnosticism, as well as shorter training and inference times.

3.2 Region proposal systems

To recognize different objects on an image, like logos, small regions should be considered. The easiest way to search for these locations is the exhaustive sliding window search, applied on multiple scales. Although, as section 2.2 presents, this induces a lot of computational costs. In order to reduce this computational burden, region proposal systems can be utilized. Region proposals are possible object locations on an image.

Earlier computer vision solutions used external proposal systems. This means that the proposals of every image should be pre-calculated before training or inference. One of the most popular region proposal methods is selective search [Uij13]. It merges neighbor regions according to a similarity score in a bottom-up fashion. It processes an image under 2s on the CPU, which precludes the possibility of real time applications. Edge Boxes [Zit14] are efficiently calculating the number of contours in a box, and ranking them according to that almost real time. Today, as section 3.3.3 introduces, the proposal system is already part of the neural network.

3.3 Region Based Convolutional Neural Networks

This section gives a brief overview about how faster region based convolutional neural networks evolved.

3.3.1 Regions with Convolutional Neural Network Features

Although this network is already historical, it is worth to mention it, because it helps to understand the improvements of the later systems. Region based convolutional neural networks [Gir13] consist of three separate systems. Firstly, region proposals are generated external with selective search. There will be altogether 2000 object positions considered. Secondly, each region of the possible object locations is warped to a size of 227x227, and then the feature vector of every single region is extracted with a CNN. The network is pretrained on the ImageNet dataset [Kri12], and then fine-tuned on the final classes. The network is run on every region proposal bounding boxes, to extract vectors with a fixed-size. These vectors will to be written to the disk. Thirdly, a set of class-specific linear SVM is used to classify the specific region.

3.3.2 Fast Region Based Convolutional Neural Network

Fast Region based CNNs [Gir16] are aimed to improve the classification accuracy and feature vector extraction speed of the interesting regions, generated also with selective search. An intermediate convolutional feature map is extracted from the whole input image with a fully convolutional neural network, also called as base network in [Ser13]. The output is a downscaled feature map, which is fed to the so called RoI (region of interest) pooling layer. This layer crops regions from the map according to the appropriate downscaled region proposals, and executes a modified version of max pooling on each regions, which results in a convolutional map with a fixed-shape, regardless the size of the region.

After the pooling, fully connected layers are used to calculate the final class probabilities and bounding box regressions for each region. The output of the bounding box regression are class specific small position and size adjustments, needed to refine the rough object locations.

The improvements of this method compared to the previous region based CNN introduced in section 3.3 are as follows:

- **Joint inference:** much shorter training and inference time is achieved by the lower computational redundancy of running convolutional layers on the whole image only once, rather then for every proposed regions.
- **One network:** the feature extraction and the classification happens in the same network. This has more advantages:
  - This results again in faster test and training times, due to the unnecessity of writing the extracted feature vectors to disk, which incidentally could require hundreds of gigabytes of storage [Gir16] for the VOC07 trainval set [Eve].
  - As the backpropagation is implemented through the RoI pooling layer, the whole network, together with the convolutional layers, can be trained jointly, against earlier implementations, like R-CNN [Gir13] or spatial pyramid pooling networks (SPPnet) [He14].

- **Minibatch from a few images:** Faster training speed is achieved by collecting a minibatch only from two images, rather than every region from different images. This method is proved to converging within similar times, despite the high correlated regions.

The network is trained with a multi-task loss function for classification and bounding box regression, defined as:

$$L(p,u,t^u,v) = L_{cls}(p,u) + \lambda[u \geq 1]L_{loc}(t^u,v) \tag{3.1}$$

where p is the computed class probabilities, u is the groundtruth class, t^u is the predicted bounding box offsets for every classes, and v is the groundtruth bounding box position and size. Since the probabilities are calculated with softmax as usual, the log loss is used for the classification error: $L_{cls}(p,u) = -log p_u$. For bounding box regression loss, a smooth version of L1 loss is used, which is defined as follows:

$$L_{loc}(t^u,v) = \sum_{i\in(x,y,w,h)} smooth_{L_1}(t_i^u, v_i) \tag{3.2}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{3.3}$$

is As the background class has the

The network is trained for K+1 classes, where K is the number of object classes, and the background is also modelled as a separate class. During training, the positive examples are chosen, regarding the intersection over union (IoU) value to the groundtruth. This value is widely used for measuring the overlapping between regions, regardless the actual size of the regions. The calculation between two regions, R_1 and R_2 is as follows:

$$\frac{area(R_1 \cap R_2)}{area(R_1 \cup R_2)} \tag{3.4}$$

For positive training examples there are thoose regions applied, which have an IoU with the groundtruth at least 0.5. For the background class are the examples with IoU [0.1,0.5) used.

3.3.3 Faster Region Based Convolutional Neural Network

A great disadvantage of the Fast R-CNN method is, that the region proposals are generated externally. Girshick et.at. introduces Faster R-CNN [Ren15], which generates the interesting regions within the neural network nearly cost-free (10ms pro image). This system consists of a region proposal system and a Fast R-CNN object detector.

Region Proposal Network

The convolutional feature map, extracted by the base network, is processed by the RoI pooling layer. Additionally a thin fully convolutional network, the region proposal network (RPN), is also utilized on the convolutional maps, to generate the proposals. Reference boxes, called "anchors" are generated at every position of the conv feature map, in different scales and different aspect ratios. This ensures the scale invariance of the objects. A convolutional layer with 3x3 kernel extracts a fixed-size vector from every window of the conv map, where in each window 9 anchors are considered. As the fully convolutional network iterates through the conv map in a sliding window fashion, translation invariance is granted. An objectness score and bounding box offset is then calculated for every anchor, by different classifiers and regressors, specialized in a specific scale and aspect ratio.

Training

As the RPN is a class agnostic object detector, it should detect all the types of objects, which the network is trained on. For this purpose, the class informations, during training of RPN, can be discarded. Positive examples are collected from the proposals with an IoU higher than 0.7 with any groundtruth. Negative label is assigned for regions, which have an IoU, lower than 0.3.

Since the objective of the RPN is the same as for Fast R-CNN, namely to classify regions and regress bounding box coordinates, the loss functions for training the RPN can the same multi-task loss, which are used for training a Fast R-CNN, detailed in section 3.3.2.

4 Logo Retrieval System

4.1 Logo Datasets

The hunger of deep learning method for training data is well-known. As the publicly available logo datasets are quite small, a better training result can be achieved if the datasets are merged. The different datasets with the number of brands, images and bounding box rois can be seen in table 4.1. The total number of brands means the number of different brands altogether.

There are also trademark datasets available having a much greater cardinality [Tur17]. The images of this dataset contain however only the logo of a company, without any context of the logo. This dataset turned out to have no use for region based deep learning methods, since this approach needs to learn to distinguish between objects to be learned and the background. The network was trained with the fusion of FlickrLogos-32 and the trademark dataset, and tested with the evaluation method of FlickrLogos-32.

Table 4.1: Publicly available logo datasets with with bounding box annotations

	Number of brands	**Number of logo images**	**Number of ROIs**
BelgaLogos	37	1321	2697
Flickr Logos 27	27	810	1261
FlickrLogos-32	32	$70 \cdot 32 = 2240$	3404
Logos-32plus	32	7830	12300
TopLogo10	10	$10 \cdot 70 = 700$	863
Total (union)	**80**	12 901	20 525

4.2 Logo Detection

supervised pre-training on a large auxiliary dataset (ILSVRC), followed by domain- specific fine-tuning on a small dataset (PASCAL), is an effective paradigm for learning high-capacity CNNs when data is scarce. RCNN paper [Gir13]

4.3 Logo Comparison

Krizhevsky?s CNN can be used (without fine- tuning) as a blackbox feature extractor, yielding excellent performance on several recognition tasks work by Donahue et al. [Don13]

5

Evaluation

5.1 Logo Detection

5.2 Logo Retrieval

6

Conclusion

6.1 Summary

6.2 Future Work

Bibliography

[Dal05] DALAL, Navneet und TRIGGS, Bill: Histograms of Oriented Gradients for Human Detection, in: *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, IEEE Computer Society, Washington, DC, USA, S. 886–893, URL `http://dx.doi.org/10.1109/CVPR.2005.177`

[Don13] DONAHUE, Jeff; JIA, Yangqing; VINYALS, Oriol; HOFFMAN, Judy; ZHANG, Ning; TZENG, Eric und DARRELL, Trevor: DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. *CoRR* (2013), Bd. abs/1310.1531, URL `http://arxiv.org/abs/1310.1531`

[Eve] EVERINGHAM, M.; VAN GOOL, L.; WILLIAMS, C. K. I.; WINN, J. und ZISSERMAN, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[Gir13] GIRSHICK, Ross B.; DONAHUE, Jeff; DARRELL, Trevor und MALIK, Jitendra: Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR* (2013), Bd. abs/1311.2524, URL `http://arxiv.org/abs/1311.2524`

[Gir16] GIRSHICK, Ross; DONAHUE, Jeff; DARRELL, Trevor und MALIK, Jitendra: Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* (2016), Bd. 38(1): S. 142–158, URL `http://dx.doi.org/10.1109/TPAMI.2015.2437384`

[He14] HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing und SUN, Jian: Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *CoRR* (2014), Bd. abs/1406.4729, URL `http://arxiv.org/abs/1406.4729`

[He15] HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing und SUN, Jian: Deep Residual Learning for Image Recognition. *CoRR* (2015), Bd. abs/1512.03385, URL `http://arxiv.org/abs/1512.03385`

[Kri12] KRIZHEVSKY, Alex; SUTSKEVER, Ilya und HINTON, Geoffrey E: ImageNet Classification with Deep Convolutional Neural Networks, in: F. Pereira; C. J. C. Burges; L. Bottou und K. Q. Weinberger (Herausgeber) *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc. (2012), S. 1097–1105, URL `http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf`

[Lon14] LONG, Jonathan; SHELHAMER, Evan und DARRELL, Trevor: Fully Convolutional Networks for Semantic Segmentation. *CoRR* (2014), Bd. abs/1411.4038, URL `http://arxiv.org/abs/1411.4038`

[Low04] Lowe, David G.: Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision* (2004), Bd. 60(2): S. 91–110, URL `http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94`

[Ren15] Ren, Shaoqing; He, Kaiming; Girshick, Ross und Sun, Jian: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in: C. Cortes; N. D. Lawrence; D. D. Lee; M. Sugiyama und R. Garnett (Herausgeber) *Advances in Neural Information Processing Systems 28*, Curran Associates, Inc. (2015), S. 91–99, URL `http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks`
`pdf`

[Rus15] Russakovsky, Olga; Deng, Jia; Su, Hao; Krause, Jonathan; Satheesh, Sanjeev; Ma, Sean; Huang, Zhiheng; Karpathy, Andrej; Khosla, Aditya; Bernstein, Michael; Berg, Alexander C. und Fei-Fei, Li: ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* (2015), Bd. 115(3): S. 211–252

[Ser13] Sermanet, Pierre; Eigen, David; Zhang, Xiang; Mathieu, MichaÃ«l; Fergus, Rob und LeCun, Yann: OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *CoRR* (2013), Bd. abs/1312.6229, URL `http://dblp.uni-trier.de/db/journals/corr/corr1312.html#SermanetEZMFL13`

[Siv03] Sivic, Josef und Zisserman, Andrew: Video Google: A Text Retrieval Approach to Object Matching in Videos, in: *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, IEEE Computer Society, Washington, DC, USA, S. 1470–, URL `http://dl.acm.org/citation.cfm?id=946247.946751`

[Tur17] Tursun, Osman; Aker, Cemal und Kalkan, Sinan: A Large-scale Dataset and Benchmark for Similar Trademark Retrieval. *CoRR* (2017), Bd. abs/1701.05766, URL `http://arxiv.org/abs/1701.05766`

[Uij13] Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T. und Smeulders, A.W.M.: Selective Search for Object Recognition. *International Journal of Computer Vision* (2013), URL `http://www.huppelen.nl/publications/selectiveSearchDraft.pdf`

[Zit14] Zitnick, Larry und Dollar, Piotr: Edge Boxes: Locating Object Proposals from Edges, in: *ECCV*, European Conference on Computer Vision, URL `https://www.microsoft.com/en-us/research/publication/edge-boxes-locating-object-proposals-from-edges/`

List of Figures

List of Tables

Acknowledgment

»Physics is to mathematics as sex is to masturbation«

R.P. Feynman

»In der Informatik geht es genauso wenig um Computer wie in der Astronomie um Teleskope.«

Dijkstra