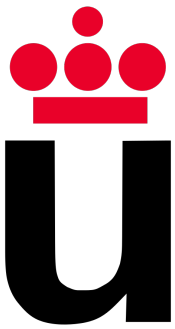

Master's thesis

Object detection with deep learning techniques

Marcos Pieras Sagardoy

Móstoles, 27 de Junio de 2016

Contents

1 Introduction **3**

2 Object detection **3**
2.1 Object detection pre-Deep learning 3
2.2 Object detection with deep learning 3

3 Datasets for object detection **3**
3.1 Pascal Visual Objects Classes 3
3.1.1 VOC07 3
3.1.2 VOC12 4
3.2 ImageNet 4
3.3 COCO 4
3.3.1 Dataset comparison 5

4 Evaluate **7**

1 Introduction

2 Object detection

2.1 Object detection pre-Deep learning

2.2 Object detection with deep learning

[4]

3 Datasets for object detection

This section describes the most common datasets used in object detection tasks.

Throughout the history of computer vision research datasets have played a critical role. They not only provide a means to train and evaluate algorithms, they drive research in new and more challenging directions. In order to accomplish this, they provide:

- a collection of challenging images and high quality annotation.
- an standard evaluation methodology, so the performance of the algorithms can be compared.

In the next subsections, we will explain the main datasets for object detection task.

3.1 Pascal Visual Objects Classes

The Pascal Visual Object Classes (VOC) challenge [2] is a benchmark in visual object category recognition and detection. Organised annually from 2005 to 2012, the challenge and its associated dataset has become accepted as one of the most benchmark for object detection. All the images are from the flickr consumer photographs website and annotated with the Mechanical turk tool.

The most popular editions of the challenge for object detection are those from years 2007 and 2012.

3.1.1 VOC07

The challenge of the year 2007, it contains 10 thousand images in the trainval and test sets, with almost 12 thousand objects. This was one the first datasets for object detection before the era deep learning. Also, it is very useful for researchers, due it has 2.5 mean object per image and it is very challenging. In the figure 1 we can observe the distribution of images and objects instances.

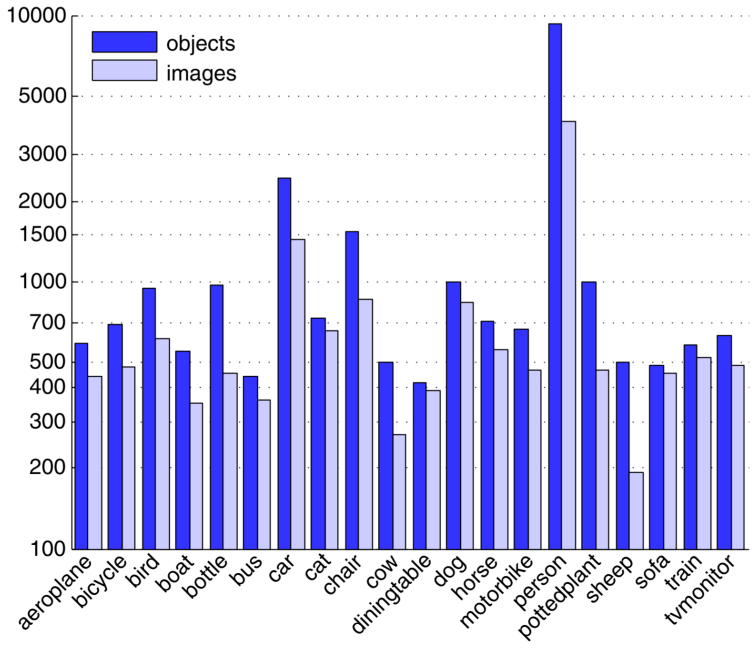

Figure 1: Distribution of VOC07 dataset.

3.1.2 VOC12

The 2012's edition is also, one of the most used dataset in object detection tasks. It increases the volume of images of the 2007 edition up to 10 thousand images on trainval and test set and similar quantity of instances per image.

3.2 ImageNet

ImageNet project [1] with the challenge ImageNet Large Scale Visual Recognition Challenge [ILSVRC] was the first large-scale database, temporally developed to supply the deep learning techinques, eager of feed with tons of images. ImageNet aims to populate the majority of the 80000 synsets of WordNet with an average of 500-1000 clean and full resolution images. The collection was based on the query of that words on several image search engines and human refined on the Amazon Mechanical Turk platform.

In 2016, the project collects more than 10 million of annotated images with 1000 classes. Although its main purpose is image classification, it has an object detection challenge with 200 categories with over a 1 million images with objects annotated.

3.3 COCO

The Microsoft Common Objects in Context also known as COCO dataset [3], is a dataset that address the three core research problems in scene understanding:

- detecting non-iconic views of objects, for many dataset most of the objects have an iconic representation, they appear unobstructed , near the center of the photo and with their canonical shape. So in this dataset, they included images to struggle

the object recognition task, like objects in the background, partially occluded, amid clutter. Therefore, reflecting the composition of actual everyday scenes.

- contextual reasoning between objects, nowadays natural images contain multiple objects, and their identity can only be solved using context, due to small size or ambiguos appearance in the image, so in this dataset, images contain scenes rather isolated objects.

- the precise 2D localization of objects, also the detailed spatial understanding of object layout will be a core component of an image understaning system, so this dataset struggle to do so.

So, the three main tasks of this challenge are object classification, object detection and semantic scene labelling. This dataset contains 91 object categories, with 2.5 million labelled object instances in 328 thousand images, labeled with the Amazon Mechanical Turk tool.

3.3.1 Dataset comparison

The datasets from Pascal challenge are very useful to test object detection algorithms, their quantity is very handy (a few thousands of images) and contains a challenging quantity of objects per image, very interesting for the algorithms. But its little amount of images does not permit to train a network on this dataset, although it can be used to finetune the network.

The datasets for the ImageNet challenge are not used to much in object detection tasks, it contains a few instances per image, this not encourage researcher to use it. Although, it is very utilize to extract features and then finetune your net.

The COCO datasets, is the most recent one, is the one focus on object recognition, and the detection suppose a challenge due the objects are in common places and are very challenging to detect. And it is very interesting to due of the quantity of instances per image.

The COCO challenge contains 91 object categories with 82 of them having more than 5 thousand labeled instances. In total the dataset has 2.5 million labeled instances in 328 thousand images. In contrast to ImageNet dataset, COCO has fewer categories but more instances per category. Also, it has more instances per category than the VOC dataset. We can observe that difference in the chart 2. This fact aid in learning detailed object models capable to chance the variability and also its 2D location.

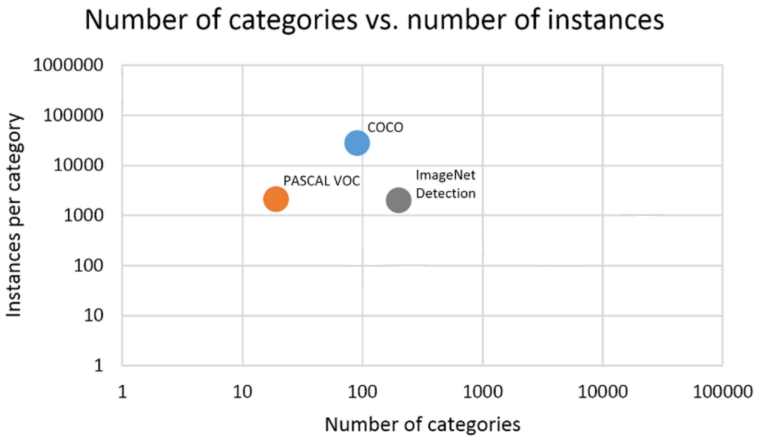

Figure 2: Distribution of pascal.

In addition, another prominent feature of the COCO over the other two, is the number of labelled instances per image which may aid in learning contextual information. This difference can be seen in 3 chart.

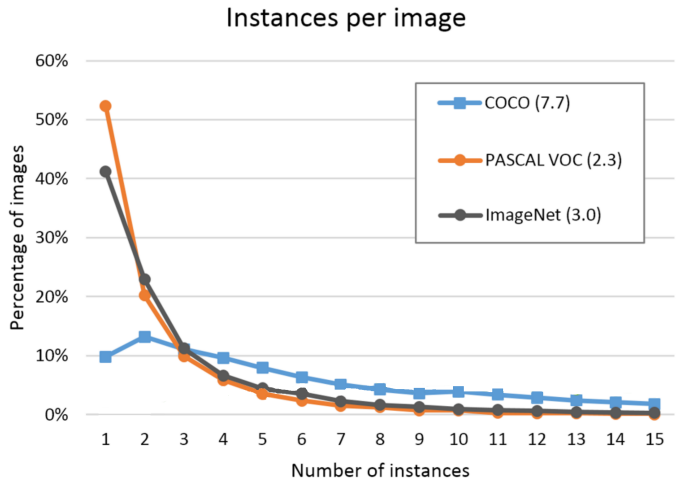

Figure 3: Distribution of pascal.

Moreover, the COCO dataset uses images from non-canonical point view, allowing to the algorithm to be robust to everyday views. This feature can be observed in the plot 4, in this plot we can observe differences views of the same category. And clearly the coco's images is the most uniconic representation.

(a) Pascal VOC. (b) ImageNet. (c) COCO.

Figure 4: Distribution of pascal.

Finally, the table 1 summarizes the main statistics of the dataset stated previously.

	VOC07	**VOC12**	**ImageNet [2014]**	**Coco [2015]**
trainval set	5011	11540	476688	165482
test set	4952	10991	40152	81434
Number of classes	20	20	200	80
Mean obj per image	2.5	2.4	1.1	7.2
Number person instances	4690	8566	-	300000

Table 1: Datasets tables

4 Evaluate

In order to compare the performance of the different datasets, each challenges establish a clear measure. In this thesis, we used the interpolated average precision (AP), used in the Pascal VOC challenge (based on [5]).

For each class, the precision/recall curve is computed from a method's ranked output.

- Recall, is defined as the proportion of all positives examples ranked above a given rank.
- Precision is the proportion of all examples above the rank which are from the positive class.

The AP summarises the shape of the precision/recall curve, and is defined as the mean precision at a set of eleven equally spaced recall levels [0,0.1,...,1]:

$$AP = \frac{1}{11} \sum_{r \epsilon (0,...,1)} p_{interp}(r)$$

The precision at each recall level r is *interpolated* by taking the maximum precision measured for a method for which the corresponding recall exceeds r:

$$p_{interp}(r) = max_{\hat{r}:\hat{r}>r} p(\hat{r})$$

The authors justified this measurment as a way to reduce the impact of the 'wiggles' in the precision/recall curve, caused by small variations in the ranking of examples. In the figure 5, we can observe this effect on the curve.

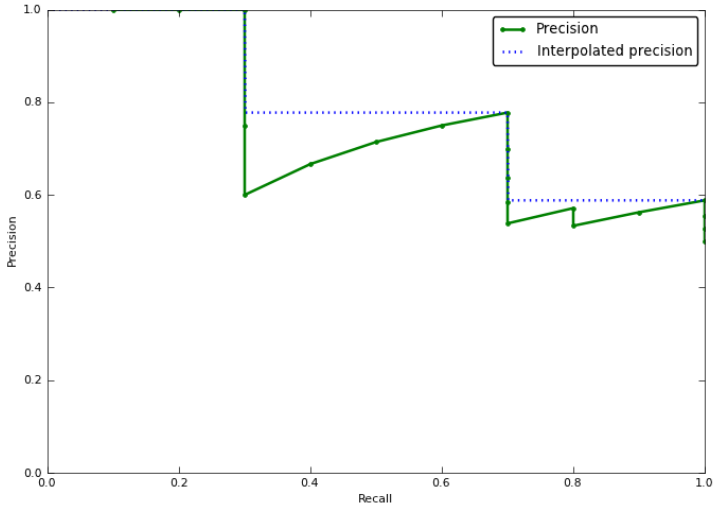

Figure 5: Comparision interpolated and normal curve.

In addition, detections were assigned to ground truth objects and judged to be true/false positives by measuring bounding box overlap. To be considered a correct detection, the area of overlap a_0 between the predicted bounding box B_p and ground truth bounding box B_{gt} must exceed 0.5 by the formula:

$$a_0 = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}$$

where $B_p \cap B_{gt}$ denotes the intersection of the predicted and ground truth bounding boxes and $B_p \cup B_{gt}$ their union. The treshold of 50 % was set deliberately low to account for inaccuracies in bounding boxes in the ground truth data. Multiple detections of the same object in an image were considered false detections.

Finally, we want to point out, even we don't take into account in our implementation, setting the threshold IoU to a value of 0.5 could cause misdetections of small objects [1], they propose an adaptive setting of that threshold based on the size of the ground truth and so detect correctly small objects. In practice this change only affects 5.5% of objects in the detection validation set.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[2] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

[3] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[5] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.