Cluster-based Anchor Box Optimisation Method for Different Object Detection Architectures

Ana Filipa Sampaio
ana.sampaio@fraunhofer.pt
João Gonçalves
joao.goncalves@fraunhofer.pt
Luís Rosado
luis.rosado@fraunhofer.pt
Maria João M. Vasconcelos
maria.vasconcelos@fraunhofer.pt

Fraunhofer Portugal AICOS
Porto, Portugal

Abstract

In object detection frameworks based on deep learning, the pre-established anchor boxes are critical to ensure an adequate localisation of the objects that should be detected. As some datasets comprise objects of distinctive shapes and specific sizes, this work describes a methodology to adjust the anchor attributes to the dataset used for the task at hand.

For that, an analysis of the dataset's bounding box properties is performed, and *k*-means clustering is applied to identify the rectangular box scales and ARs that yield the best representation of the object dimensions and shapes existing in the dataset. The particularities of four popular object detection meta-architectures were taken into account to ensure that the output of the proposed method is fully compatible with the anchor box settings of different networks. The application of this methodology is illustrated using a private cervical cancer dataset.

1 Introduction

Recently, the popularity of deep learning models for object detection tasks has arisen, owing to their robustness and promising performances. These algorithms aim at localising the objects in each image in terms of rectangular bounding boxes that mark the region of the object, while also distinguishing their class [1]. Most model architectures devised for this purpose achieve the detection step through the proposal of object regions and the regression of their bounding box coordinates, with many resorting to anchor boxes, or box priors, to generate the object proposals [2, 3, 4].

Anchor boxes are bounding box templates extracted at pre-defined locations of the feature map of the convolutional neural networks (CNNs) that define the object candidates assessed by the network. Their dimensions may be directly set [5] or specified in terms of the scales and aspect ratios (ARs) combined to define the candidates extracted at each location [2, 3]. Due to their role in object proposal, anchor box settings are critical to ensure the reliable localisation of the objects in the image; hence, the anchor box scales and ARs ought to be carefully defined, bearing in mind the specificities of the annotated objects in the dataset.

Although the anchors considered in the most common object detection architectures are designed to encompass myriad object scales and shapes, in some scenarios the object proposals generated using generic anchors might not be able to match the objects that should be detected. Thus, this work presents a methodology to adjust the anchor properties to the type of objects existing in a specific dataset, enabling a more targeted object proposal procedure. Clustering is used to identify the most representative bounding box dimensions and shapes present in the dataset, which are mapped to the parameters of specific object detection CNNs, taking into account their design differences. Finally, the application of this methodology is demonstrated using a private cervical cancer dataset.

2 Methodology

The idea of exploiting dimension clusters to adjust the box priors used for object detection was already proposed in [5]. In that work, *k*-means clustering is applied to the bounding box width and height values of the training data to find several cluster centres, each associated with distinct anchor dimensions. The optimal number of anchors is established by finding the number of centres that allows a high average intersection over union (IoU) between the anchors and the ground truth boxes and does not increase substantially the computational complexity of the algorithm. The dimensions (height and width) that characterise the cluster centres are used directly to define the anchor boxes considered by YOLO.

However, in other object detection models (such as Faster R-CNN, SSD and RetinaNet), the size and shape of the anchor boxes are parameterised separately through the specification of several box scales and ARs, combined to determine the dimensions of the anchors extracted from the feature maps. Ergo, the proposed methodology applies the *k*-means algorithm in 3 distinct domains: the bi-dimensional height and width space (described above); and the domains of bounding box scales and ARs as separate variables, since this enables an easier adaptation to the way the anchor boxes are defined in the other meta-architectures. In this case, the within-cluster sum-of-squares distance metric is minimised to find the optimal clustering centres for each *k* value.

2.1 Aspect ratio and scale clustering

To find the optimal anchor shapes, the ARs of the dataset's bounding boxes are computed as the ratio between the width and the height of each bounding box. The anchor scales are computed as the ratio between the area of each bounding box and the area of the whole image. The *k*-means clustering algorithm is applied independently to the scale and AR values, finding the optimal cluster centres for each of these variables. For both properties, several *k* (number of cluster centres) values are tested and evaluated based on the sum of squared distances between each bounding box instance and its nearest cluster centre.

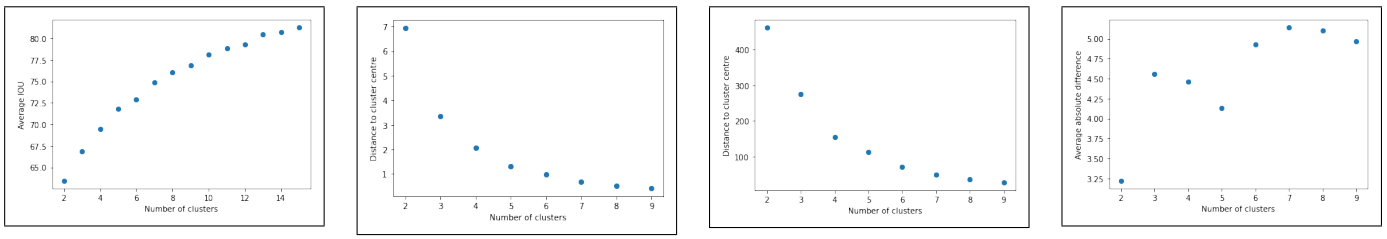
2.2 Selection of the optimal anchor scales

More cluster centres are expected to result in anchors more representative of the dimensions of the objects in the dataset, as verified in [5]; yet, the consideration of more bounding box scales and ARs implies the generation of many more object candidates during the training and execution of the algorithm, subsequently increasing its computational burden. Thus, the selection of the number of scales and ARs used in the detection algorithm is accomplished considering the **trade-off between the sum-of-squares distance** - representative of the intra-cluster variability, which should be minimised - **and the inherent computational complexity**.

In addition, the **design differences** of the current state of the art detection architectures should also be taken into account for the specification of the anchor box settings, since they might affect the anchors extracted in the object proposal step. To address these variations, four architectures - YOLO [5], Faster R-CNN [2], SSD [3] and RetinaNet [4] were examined.

One of the key disparities among these frameworks is associated with the convolutional layers from which the object proposals are retrieved: YOLO and Faster R-CNN apply the anchors to a single feature map, whereas SSD and RetinaNet propose boxes of different scales by extracting candidates from network layers of varying depths. Moreover, as aforementioned, the YOLO model contrasts with the remaining architectures by defining the anchor box dimensions directly, instead of setting the box priors through scale and AR combinations.

Even though SSD and RetinaNet both resort to feature maps of multiple depth levels to propose objects at different scales, in the SSD framework each extraction layer is associated with a single scale, whereas RetinaNet allows the specification of more than 1 sub-scale for each level. Therefore, in the SSD model, the number of feature maps used for anchor generation is equal to the number of object scales considered, and there is a direct mapping between the selected scales and the anchor extraction layers. Alternatively, in the RetinaNet framework, a feature pyramid



(a) (b) (c) (d)

Figure 1: Graphical representation of some metrics according to the k value used in the experiment: (a) average IoU between the cluster centres and the dataset's objects, for the width/height clustering; within-cluster sum-of-squares distance for the (b) scale and (c) aspect ratio values; (d) average absolute difference among the aspect ratio values in each set.

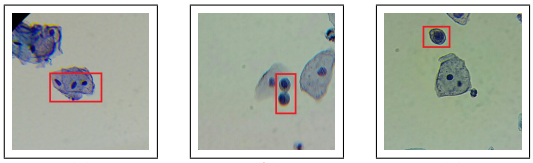
network is used to provide the convolutional layers that are the basis for anchor generation, being characterised by feature maps with a fixed consecutive resolution difference (a factor of 2), designated as octave levels. Accordingly, to take advantage of the scale values found through the proposed methodology, a careful correspondence between the selected scales and the architecture-specific parameters must be conducted.

2.3 Identification of the most discriminating aspect ratios

Given that the ARs influence the object shapes that will be more easily detected by the algorithm, the established values should be sufficiently discrepant to allow the examination of a diverse set of object shapes. To ensure this diversity, in the proposed approach, the choice of the number of ARs is based not only in the sum-of-squares distance, but also in the average absolute difference between the several AR values in each of the possible sets (inter-cluster variability).

3 Application to a private cervical cancer dataset

The approach described was applied to a private dataset comprised of 1489 microscopic images in total, acquired from liquid-based cervical cytology samples of 21 patients with a μSmartScope device [6]. This dataset includes 2436 bounding box annotations of abnormal regions (indicative of cervical lesions, illustrated in fig. 2), provided by a clinical expert from Hospital Fernando Fonseca. As the dataset had been previous split in training and test subsets according to a 80%/20% ratio, only images from the training set were analysed to establish the anchor settings.



(a) (b) (c)

Figure 2: Examples of images from the cervical cancer dataset with the bounding boxes of abnormal cells outlined (in red).

The range of k values tested in the clustering experiments was limited to a maximum of 15 for the dimensional clustering case and of 9 for the scale and AR studies to limit the number of object candidates generated. The results obtained for the several ks are depicted in fig. 1. For each k value, the final anchor dimensions, scales and ARs were obtained from the coordinates of the corresponding cluster centres.

As expected, more cluster centres lead to anchors more representative of the dimensions of the objects in the dataset, associated with a lower sum-of-squares distance error and a higher IoU metric. However, it is important to select a number of cluster centres associated with a reasonable number of anchors. Hence, adequate values for the **anchor dimensions used in YOLO** (directly defined in the image domain) would be the ones obtained for $k = 9$, for instance, resulting in the anchors of normalised dimensions $(0.36, 0.38)$, $(0.30, 0.20)$, $(0.15, 0.39)$, $(0.19, 0.19)$, $(0.3, 0.58)$, $(0.30, 0.29)$, $(0.23, 0.28)$, $(0.67, 0.67)$, $(0.57, 0.28)$. An appropriate choice for the **scale values** could be the scales of the cluster centres for $k = 6$ (0.06,0.13,0.25,0.44,0.66,0.91), since these would produce a restricted number of object proposals while keeping the same number of feature maps used in the original implementation, which is an advantage when pre-trained models are used.

The **selection of the ARs** should be grounded not only in the intra-cluster variability, but also considering how well-separated a cluster is from other clusters. Even though the ARs reported for $k = 7 - 9$ yielded larger differences, their consideration would increase the computational burden of the model. As the ARs clustered for $k = 6$ (0.68, 1.18, 1.90, 3.63, 7.47, 14.55) exhibit an average difference metric similar to the ones produced by more ARs, these would be suitable for the dataset analysed.

4 Discussion and conclusions

This work presents a method to optimise the localisation step in object detection networks, achieved through the adjustment of the anchor boxes' settings to the properties of the dataset used. The performed analysis addressed the factors that may influence the establishment of the anchors, in particular the similarity between the extracted anchors and the dataset's objects, the computational complexity of the model, the variety of anchor shapes and the ability to implement the anchors of choice in the existing detection models. Additionally, in studies that rely on pre-trained networks for fine-tuning, for architectures whose number of layers is directly associated with the anchor scales extracted, the number of object proposal layers should be the same as in the original model, to fully take advantage of the pre-trained weights.

Nonetheless, the experiments reported still correspond to exploratory work and further tests ought to be conducted. Future work should include the examination of the impact of the anchors' setup in the final detection performance through the comparison of the adjusted anchor settings with the default ones, as well as a characterisation of the computational burden yielded by some of the possible anchor configurations. Different clustering approaches, as well as more informative distance metrics for cluster validation, should also be explored.

Acknowledgements

This work was done under the scope of "CLARE: Computer-Aided Cervical Cancer Screening",project with reference POCI-01-0145-FEDER-028857 and financially supported by FEDER through Operational Competitiveness Program – COMPETE 2020 and by National Funds through Fundation for Science and Tecnhology FCT/MCTES.

References

[1] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review,"

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks,"

[3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," vol. 9905, pp. 21–37.

[4] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," pp. 1–1.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger,"

[6] L. Rosado, P. T. Silva, J. Faria, J. Oliveira, M. J. M. Vasconcelos, D. Elias, J. M. C. da Costa, and J. S. Cardoso, "μSmartScope: Towards a fully automated 3d-printed smartphone microscope with motorized stage," in *Biomedical Engineering Systems and Technologies*, Communications in Computer and Information Science, pp. 19–44, Springer International Publishing.