IMPROVED HARD EXAMPLE MINING APPROACH FOR SINGLE SHOT OBJECT DETECTORS

Aybora Köksal⋆ Önder Tuzcuoğlu⋆ Kutalmış Gökalp İnce⋆ Yoldas Ataseven† A. Aydın Alatan⋆

* Center for Image Analysis (OGAM), Department of Electrical and Electronics Engineering
Middle East Technical University, Ankara, Turkey
† ASELSAN Inc., Ankara, Turkey

ABSTRACT

Hard example mining methods generally improve the performance of the object detectors, which suffer from imbalanced training sets. In this work, two existing hard example mining approaches (LRM and focal loss, FL) are adapted and combined in a state-of-the-art real-time object detector, YOLOv5. The effectiveness of the proposed approach for improving the performance on hard examples is extensively evaluated. The proposed method increases mAP by 3% compared to using the original loss function and around 1-2% compared to using the hard-mining methods (LRM or FL) individually on 2021 Anti-UAV Challenge Dataset.

Index Terms— hard example mining, loss rank mining, real time object detection

1. INTRODUCTION

Object detection performance has rapidly increased during the last decade by the utilization of Convolutional Neural Networks (CNN) for feature extraction. Even though most of the object detectors [1, 2, 3, 4] work well on common datasets, such as MS COCO [5], they usually suffer from two main problems: the imbalance between the number of background-foreground data and infrequent observation of trained foreground object representations in the test set, i.e. the tail problem.

In order to cope with the imbalance problem, some example mining methods are proposed for the two-stage object detectors [1, 2, 6, 7]. All of these methods are specific to the two-stage detectors, since they are based on the outputs of the RoI Pooling stage. Therefore, these methods are not applicable to the single shot object detectors. Unfortunately, the pioneering examples of single shot object detectors [3, 8, 9, 10], do not have a solution for this problem. Besides the imbalance problem, some appearances of a class might be rare in the dataset of interest. Such rare occurrences which lie at the tails of the appearance distribution are dominated by the rest of the dataset, and therefore, they are hard to learn.

Focusing on the hard examples of an imbalanced dataset is tried to be handled by bootstrapping by Sung [11]. The main idea of this approach is incrementally increasing the weight of the examples that trigger false alarms; this study is one of the prominent solutions for iterative learning. Later, bootstrapping ideas are also used in SVM, with the introduction of Latent SVM paradigm [12]. Finally, bootstrapping methods also became popular in object detection the object detection research by the utilization of SVM [1, 13, 14].

In order to mine a hard example via RoIs properly, Online Hard Example Mining (OHEM) method is introduced [15]. The idea suggests considering only the most beneficial RoIs for the backpropagation. RoIs which give the highest loss values are assumed to be the hardest examples, and therefore, the most beneficial ones. Hence, the aforementioned method selects B/N worst loss cases for training and discards the remaining during training. Although this novel approach is one of the most promising approaches in hard example mining, it is only applicable to two-stage networks, since it requires RoIs to work on.

Lin et. al. [16] introduced an inherent hard example mining method for a single shot object detector without sacrificing its real time performance. They introduced focal loss to use hard examples more effectively. The loss function is designed to make the detections with higher loss values more important in back propagation than the others by performing gamma correction with a $\gamma$ factor larger than 1. After its efficiency was observed, focal loss was also used in other one stage object detectors such as EfficientDet [17]. The idea was also applied to YOLOv3 on MS COCO dataset [5], but it did not increase the baseline performance [10]. Since the other state-of-the-art object detectors are working well with focal loss, it might be worth trying to modify focal loss in YOLO to make it work properly.

Based on the idea of OHEM, Yu et. al. introduce Loss Rank Mining (LRM) [18]. The method is applicable for single shot detectors, and it makes the object detector to focus on hard examples by filtering-out some easy examples on the feature map just before the detection stage. During training, as the first step, the input goes through the model backbone to
get the feature map. Then, for each detection, the loss value is calculated. After the non-maximum suppressor (NMS) stages, the loss values of these detections are sorted in descending order and the first $K$ detection results are selected and filtered. The rest of the detection values are not used during the training process. This idea might be applicable and beneficial to the current object detectors, if it can be implemented into their structure.

In the recent studies, several methods for hard example mining are also proposed. Jin et. al. [19] introduced an unsupervised hard example mining method for video sequences. Their approach suggests a template matching solution between consecutive frames. If the matched templates are not temporarily consistent, they are flagged as hard examples and the training continues iteratively. Wang et. al. [20] propose an adversarial network structure in order to create artificially occluded hard examples for the imbalance problem. Although both of these approaches are worth to mention, unfortunately, they are not automatic processes and they frequently need human intervention.

The proposed study combines two different hard example mining approaches and applies the resulting method on YOLOv5, which is one of the best-performing single shot object detectors. For that purpose, the focal loss is adapted to YOLOv5 and LRM (which is originally designed to work with a single feature map) is modified to work with multiple feature maps. Next, these two methods are combined to obtain a single loss function, as shown in Figure 1. Quantitative experiments are conducted to verify that the methods are mining the hard examples without manipulating the number of hard examples.

The proposed method differs from the previous work in the following points: a) Our modified focal loss approach increases YOLOv5 detection performance, which is not the case for the original focal loss implementation of YOLOv5; b) Our LRM structure filters the detections in each feature map separately, which is not achieved in the original LRM; c) The proposed approach combines the focal loss and LRM approaches into one novel loss function; d) Our performance evaluation method allows us to check whether the suggested approach increases the performance on hard examples, without defining them explicitly.

## 2. PROPOSED METHOD

The proposed approach has a combined hard example mining structure which uses Balanced Focal Loss and Loss Rank Mining. Both of these methods are proposed in a way these can be used individually or combined.

### 2.1. Balanced Focal Loss

The popular cross entropy loss function is shown in (1), whereas original Focal loss function, proposed by Lin et. al. [16], is given in (2).
\[ CE(p) = -\log(p) \]  
\[ FL(p) = -\alpha(1 - p)^\gamma \log(p) \]

In (2), \( \gamma \) is the focus parameter, while \( \alpha \) denotes the correction parameter. The original YOLOv5 implementation already has a flag for activation of focal loss. However, it generally decreases the performance of YOLOv5, since \( \gamma \) factor makes the value of the objectness loss negligibly small so that it becomes insignificant with respect to the box regression loss. Therefore, it should be scaled appropriately in order to make these two loss values comparable. In our work, focal loss is weighted by an additional balancing parameter, \( \xi \). It should be noted that \( \xi \) is weight of the objectness loss in the overall loss function which is aimed to be higher than 1.

2.2. Loss Rank Mining

In its original paper [18], this method is used with YOLOv2 [9], which has one feature map for object detection. Since YOLOv5 uses three feature maps for small, medium and large objects, the original method is also modified in such a way that it works with all three feature maps.

In the original LRM structure, first \( K \) detection results with the highest amount of loss are selected. In our method, first \( B \) (rank factor) detections are selected for each feature map. The comparison between the proposed combined objectness loss structure and the original YOLOv5 loss is illustrated in Figure 1 and the method can be summarized as follows for each feature map:

1. Through each mini-batch, Balanced Focal Loss is applied to cells for obtaining loss values of the detections.
2. By flattening the three-dimensional cell structure, loss values of each image sample are concatenated into different vectors separately.
3. Loss values of each image are sorted by a value.
4. From the sorted loss vectors, the top \( B \) proportion of the number of cells for each image sample is selected.
5. The mean of each selected loss is taken individually.
6. These averages are summed to form the objectness loss.

3. EXPERIMENTS

Throughout the experiments, YOLOv5s [4] is used as the baseline detector. 2021 Anti-UAV Challenge Dataset\(^1\) is used during the experiments. In order to speed-up the training phase, the training set is generated with 1/20 of the original frame rate. Removing adjacent video frames is known not to cause a significant drop in the detection performance [21].

\(^1\)https://anti-uav.github.io/dataset/
Two hard example mining methods are modified and adapted on a state-of-the-art object detector, YOLOv5. The experiments clearly indicated that although the original focal loss degrades the precision of YOLOv5, the proposed Balanced Focal Loss corrects such inaccuracies and improves the overall performance. Similarly, LRM structure is modified to integrate with YOLOv5 architecture and the experiments demonstrate a meaningful increase in mAP scores. Finally, Balanced Focal Loss and LRM methods are combined and the final object detection performance is calculated as 93.5% mAP, improving the baseline performance by 3.1%.

4. CONCLUSION
5. REFERENCES

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[2] Ross Girshick, “Fast r-cnn,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, “You only look once: Unified, real-time object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[4] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, Ayush Chaurasia, TaoXie, Liu Changyu, Abhiram V, Laughing, tkianai, yxNONG, Adam Hogan, lorenzomammmana, AlexWang1900, Jan Hajek, Laurentiu Diaconu, Marc, Yonghye Kwon, oleg, wanghaoyang0106, Yann Defretin, Aditya Lohia, ml5ah, Ben Milanko, Benjamin Fineran, Daniel Khromov, Ding Yiwei, Doug, Durgesh, and Francisco Ingham, “ultralytics/yolov5,” Apr. 2021.

[5] 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, vol. abs/1405.0312, 2014.

[6] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” CoRR, vol. abs/1506.01497, 2015.

[7] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun, “R-fcn: Object detection via region-based fully convolutional networks,” in Advances in neural information processing systems, 2016, pp. 379–387.

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

[9] Joseph Redmon and Ali Farhadi, “Yolo9000: Better, faster, stronger,” 2016.

[10] Joseph Redmon and Ali Farhadi, “Yolov3: An incremental improvement,” CoRR, vol. abs/1804.02767, 2018.

[11] Kah-Kay Sung, “Learning and example selection for object and pattern detection,” 1996.

[12] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, “Object detection with discriminatively trained part-based models,” IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pp. 1627–1645, 2009.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 37, no. 9, pp. 1904–1916, 2015.

[14] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders, “Selective search for object recognition,” International journal of computer vision, vol. 104, no. 2, pp. 154–171, 2013.

[15] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick, “Training region-based object detectors with online hard example mining,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 761–769.

[16] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” CoRR, vol.abs/1708.02002, 2017.

[17] Mingxing Tan, Ruoming Pang, and Quoc V Le, “Eficientdet: Scalable and efficient object detection,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.

[18] Hao Yu, Zhaoning Zhang, Zheng Qin, Hao Wu, Dongsheng Li, Jun Zhao, and Xicheng Lu, “Loss rank mining: A general hard example mining method for real-time detectors,” in 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018, pp. 1–8.

[19] SouYoung Jin, Aruni RoyChowdhury, Huaizu Jiang, Ashish Singh, Aditya Prasad, Deep Chakraborty, and Erik Learned-Miller, “Unsupervised hard example mining from videos for improved object detection,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 307–324.

[20] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta, “A-fast-rnn: Hard positive generation via adversary for object detection,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2606–2615.

[21] Aybora Koksal, Kutalmis Gokalp Ince, and Aydin Alatan, “Effect of annotation errors on drone detection with yolov3,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2020.