

IMPROVED HARD EXAMPLE MINING APPROACH FOR SINGLE SHOT OBJECT DETECTORS

Aybora Köksal* Önder Tuzcuoğlu* Kutalmış Gökalp İnce* Yoldaş 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 γ 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

This study is funded by ASELSAN Inc. The codes are available at github.com/aybora/yolov5Loss

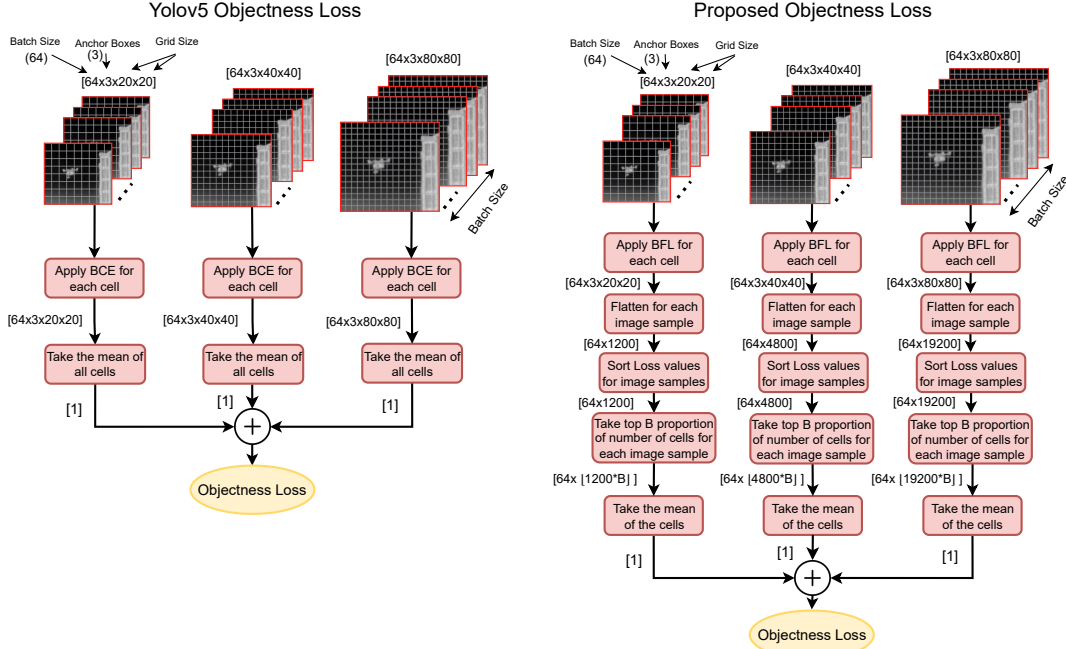


Fig. 1. YOLOv5 objectness loss (left) vs. proposed combined loss (right). For the proposed loss, firstly Balanced Focal Loss is applied for each cell instead of BCE, then for each feature map, detections are sorted with respect to their loss values. Finally, top B detection which has highest loss values are selected for loss calculation and backpropagation.

get the feature map. Then, for each detection, the loss value is calculated. After the non-maximum suppression (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. Quantita-

tive 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).

$$CE(p) = -\log(p) \quad (1)$$

$$FL(p) = -\alpha(1-p)^\gamma \log(p) \quad (2)$$

In (2), γ is the focus parameter, while α 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 γ 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, ξ . It should be noted that ξ 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 Dataset¹ 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/>

3818 frames are selected for the training set, whereas 2313 for the validation, and 1517 is selected for the test set.

As an ablation study, the methods are compared *head-to-head* according to their detection performance (True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN)) considering objectness score (confidence) and Intersection-over-Union (IoU) metric. More specifically, for any detection, if confidence > 0.5 and IoU > 0.5 , that detection is accepted as a TP. If confidence > 0.5 but IoU < 0.5 , that is a FP. If there is a ground truth detection whose confidence < 0.5 , then ground truth object becomes a FN. If confidence < 0.5 and there is no ground truth, that is a TN.

As we do not have a predefined hard example set, we define hard examples as the distinguishing failures of alternative methods. The detections and misses which fall at the same cell are classified as TP-TP, TP-FN, TN-FP, FP-FP, FN-FN, as it can be observed in Figure 2. Since all of these methods are already the state-of-the-art, the frames for which both of these algorithms fail are counted as hard examples. Therefore, the aim of these experiments is to check TP-FN and TN-FP pairs. The reason is, if method A has correct outputs (TPs and TNs) for some FNs and FPs of method B, and method A has less incorrect outputs for TPs and TNs of method B, then it is reasonable to assume that method A is better for hard examples. This is an unsupervised performance evaluation approach, since the number of hard examples is unknown.

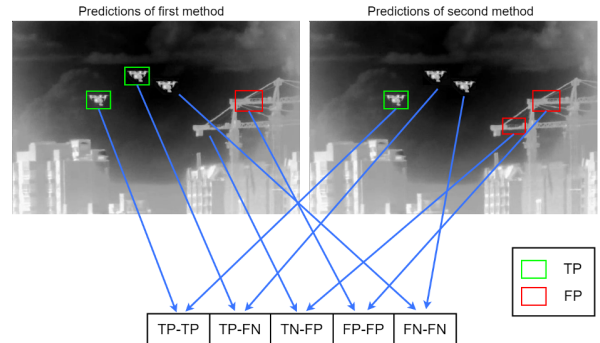


Fig. 2. A prediction confusion matrix pairing method on one frame for the comparison of two different training models. Image taken from 2021 Anti-UAV dataset¹.

Default loss vs Focal Loss: Original YOLOv5 loss function and Original Focal Loss are compared with $\gamma: 1.5$ and $\alpha: 0.25$. According to the results which are tabulated in Table 1, the baseline focal loss function degrades the algorithm by mistakenly converting 8.06% of the test set which was evaluated as TP into FN. Although it also converts some FNs into TPs and FP into TNs, the overall performance of hard examples is decreased by 6% on test set by implementing original focal loss into YOLOv5. Therefore, the baseline focal loss is not used for the rest of the experiments.

Default loss vs. Balanced Focal Loss: Original YOLOv5 loss function and the proposed Balanced Focal Loss are com-

Pairwise comparison of method pairs in terms of Number (#) and Percentage (%) of Frames in the Test Set

Table 1. M1: Default, M2: Focal Loss **Table 2.** M1: Default, M2: Bal. Focal Loss ($\xi : 30$)

M1	M2	# Fr.	Fr. %	M1	M2	# Fr.	Fr. %
FN	TP	9	0.59	FN	TP	63	4.14
FP	TN	17	1.12	FP	TN	14	0.92
TP	FN	122	8.06	TP	FN	12	0.79
TN	FP	7	0.46	TN	FP	16	1.05
FP	FP	4	0.26	FP	FP	7	0.46
FN	FN	215	14.20	FN	FN	161	10.57
TP	TP	1140	75.30	TP	TP	1250	82.07

Table 3. M1: Default, M2: LRM ($B : 0.35$) **Table 4.** M1: Default, M2: Combined ($\xi : 30, B : 0.35$)

M1	M2	# Fr.	Fr. %	M1	M2	# Fr.	Fr. %
FN	TP	75	4.92	FN	TP	83	5.47
FP	TN	14	0.92	FP	TN	16	1.06
TP	FN	5	0.33	TP	FN	9	0.59
TN	FP	16	1.05	TN	FP	9	0.59
FP	FP	7	0.46	FP	FP	5	0.33
FN	FN	149	9.78	FN	FN	141	9.30
TP	TP	1257	82.53	TP	TP	1253	82.65

Table 5. M1: LRM ($B : 0.35$, M2: Combined ($\xi : 30, B : 0.35$) **Table 6.** M1: Bal. Focal Loss ($\xi : 30$, M2: Combined ($\xi : 30, B : 0.35$)

M1	M2	# Fr.	Fr. %	M1	M2	# Fr.	Fr. %
FN	TP	23	1.52	FN	TP	34	2.24
FP	TN	17	1.12	FP	TN	15	0.99
TP	FN	19	1.25	TP	FN	11	0.73
TN	FP	8	0.53	TN	FP	6	0.40
FP	FP	6	0.40	FP	FP	8	0.53
FN	FN	131	8.64	FN	FN	139	9.17
TP	TP	1313	86.55	TP	TP	1302	85.94

pared by using $\gamma : 1.5, \alpha : 0.25$ and $\xi : 30$. According to the results which are shown in Table 2, Balanced Focal Loss successfully converts 4.14% of the test set to TP which was FN in default loss function, while lost its 0.79% of TP to FN. Therefore, the performance of hard examples is improved by more than 3% with the usage of Balanced Focal Loss.

Default loss vs. LRM: Original YOLOv5 loss function and LRM are compared with $B : 0.35$. According to the results in Table 3, 4.92% of the test set evaluated as FN are converted to TP, while 0.33% of them are lost. In the overall evaluation, the performance of the hard examples is increased by around 4.50% with LRM.

Default loss vs. Combined: Original YOLOv5 loss is compared with our Combined loss approach with $\gamma : 1.5, \alpha : 0.25$ and $\xi : 30$ and $B : 0.35$. According to the results in Table 4, 5.47% of the test set evaluated as FN are converted to TP,

while 0.59% of them are lost. Therefore, the performance of the hard examples are increased by around 5% with LRM and Balanced Focal Loss combined.

LRM vs Combined: Proposed LRM is compared with our Combined loss approach with $\gamma : 1.5, \alpha : 0.25$ and $\xi : 30$ and $B : 0.35$. According to the results which are presented in Table 5, 1.52% of the test set evaluated FN are converted to TP, while 1.25% of them are lost. Moreover, 1.12% of the set which was FP are converted to TN, while 0.53% are lost. Therefore, using the balanced focal loss in addition to LRM increases the overall hard example performance by 1%.

Balanced Focal Loss vs Combined: Proposed Balanced Focal Loss is compared with our Combined loss approach with $\gamma : 1.5, \alpha : 0.25$ and $\xi : 30$ and $B : 0.35$. According to the results in Table 6, 2.24% of the test set which was FN are converted to TP, while 0.73% of them are lost. Moreover, 0.99% of the set evaluated as FP are converted to TN, while 0.40% are lost. Overall, using LRM in addition to Balanced Focal Loss increases the hard example performance by 2%.

After the head-to-head comparison of algorithms for hard example performance, it is time to check their overall performance by using standard object detection metrics. For that purpose, precision, recall and mAP@.5 metrics are used.

The results are tabulated in Table 7. For these experiments, the parameters are kept constant with following values: $\alpha = 0.25, \gamma = 1.5, \xi = 30, B = 0.35$. Table 7 indicates that the prior experiments are coherent with the experiments on object detection metrics, our LRM and Balanced Focal Loss combined approach outperforms all the other loss selections in terms of Precision, Recall and mAP@.5 metrics.

Table 7. Performance evaluation of the baseline and proposed methods in terms of Precision (%), Recall (%), mAP_{0.5} (%) and mAP_{0.5:0.95} (%).

Method	Prec.	Rec.	mAP _{0.5}	mAP _{0.5:0.95}
Default	98.0	85.4	90.4	53.3
Focal Loss	93.6	84.3	90.4	53.3
Bal. Focal Loss	98.2	89.6	92.6	55.9
LRM	98.0	90.0	93.2	57.3
Combined	98.3	91.0	93.5	56.1

4. CONCLUSION

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%.

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, lorenzomamma, 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 on-line 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, “Efficientdet: 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-rcnn: 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 Koksall, 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.