International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 7 July 2023, pp: 53-57 www.ijaem.net ISSN: 2395-5252

Focal Loss for Dense Object Detection

[1.]Dr. W. T. Chembian, [2]Nibarajith.R, [3]Sukumar.K, [4]Praveen Kumar.P, [5]Prasannaa.S

[1] Associate Professor [2,3,4,5] Students,
Department of Computer Science and Engineering.
VEL Tech High Tech Dr.Rangarajan Dr.Sakunthala Engineering College

Date of Submission: 25-06-2023 Date of Acceptance: 07-07-2023

ABSTRACT: A two-stage approach popularized by R-CNN is the foundation for the most accurate object detectors to date. In this method, a classifier is applied to a small number of candidate object locations. On the other hand, although one-stage detectors that are used over a regular, dense sampling of potential object locations have the potential to be faster and simpler, their accuracy has so far lagged behind that of two-stage detectors. We investigate the reason for this in this paper. We discover that the primary reason lies in the extreme foreground-background class imbalance that occurs during dense detector training. By reshaping the standard cross entropy loss so that it down weights the loss assigned to well-classified examples, we propose to address this class imbalance. The novel Focal Loss keeps the large number of easy negatives from overwhelming the detector during training by concentrating training on a small number of hard examples. RetinaNet, a straightforward dense detector, was designed and trained in order to assess the efficacy of our loss. RetinaNet surpasses all current state-of-the-art two-stage detector accuracy when trained with the focal loss, as shown by our findings, while maintaining the speed of previous one-stage detectors.

Keywords: Computer vision, object detection, machine learning, convolutional neural networks

I. INTRODUCTION

A proposal-driven two-stage mechanism is the foundation of the current object detectors. The R-CNN framework made a method that uses a convolutional neural network to first classify each candidate location as one of the sparse foreground classes or a background location popular. Due to a number of improvements, this two-stage framework consistently achieves top accuracy on the challenging COCO benchmark.

Even though two-stage detectors work well, the obvious question is: Could the same level of precision be achieved by a straightforward one-stage detector? One stage detectors are used to perform a regular, dense sampling of object locations, scales, and aspect ratios. Positive signs can be seen in recent research on one-stage detectors like YOLO and SSD. When compared to the two-stage methods that are currently in use, these detectors are faster and have an accuracy of between 10% and 40%.

This paper does more than that: A one-stage object detector that matches the most recent COCO AP of more complex two-stage detectors, such as Faster R-CNN Mask R-CNN or Feature Pyramid Network (FPN) variants, is presented by us for the first time. In order to achieve this result, we propose a brand-new loss function that eliminates class imbalance during training, which is the primary obstacle that prevents one-stage detectors from achieving modern accuracy.

We present a new loss function as a more effective alternative to previous approaches to addressing class imbalance in this paper. The scaling factor of the loss function, which is a dynamically scaled cross entropy loss, decreases to zero as confidence in the correct class increases. During preparing, this scaling factor naturally can consequently down-weight the commitment of simple models and immediately center the model around hard models.

II. PROPOSED METHODOLOGY

Our proposed Focal Loss outperforms the alternatives of hard example mining or training with sampling heuristics, which were the prior state-of-the-art approaches for training one-stage detectors. Last but notleast, we show that our results are comparable to those of other instantiations that do not depend on the exact form of the focal loss.We create the RetinaNet, a straightforward one-stage object detector, in order to demonstrate the effectiveness of the proposed focal loss.RetinaNet is accurate and works well; The best single-model results from both one-stage

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 7 July 2023, pp: 53-57 www.ijaem.net ISSN: 2395-5252

and two-stage detectors that were previously published were outperformed by our best model, which is based on a ResNet-101-FPN backbone, runs at 5 frames.

III. RELATED WORK

Classic Object Detectors

Typical detectors for objects. A classifier is used on a dense image grid in the sliding-window paradigm, which has a long and varied history. LeCun et al.'s classic work is one of the first successes. who used handwritten digit recognition with convolutional neural networks. Face detection models based on boosted object detectors were widely adopted by Viola and Jones. Effective methods for detecting pedestrians were developed as a result of the introduction of HOG and integral channel features. DPMs achieved top results on PASCAL for many years and assisted in expanding dense detectors to more general object categories. With the resurgence of deep learning, two-stage detectors quickly came to dominate object detection, whereas the sliding-window approach was the leading detection paradigm in traditional computer vision.

Two Stage Detectors

In object detection, the current standard is a two-stage approach. Selective Search was the first to use this method, and it starts with a small set of candidate proposals that should include all objects and get rid of most negative locations. The proposals are divided into foreground classes and background at the second stage. R-CNN upgraded the second-stage classifier to a convolutional network, resulting in significant accuracy gains and heralding the modern era of object detection. R-CNN's speed has increased over time thanks to the use of learned object proposals. The Quicker R-CNN structure was made when proposition age and the second-stage classifier were consolidated into a solitary convolutional network by District Proposition Organization (RPN).

RetinaNet Detector

RetinaNet is a single, unified network with a backbone network and two task-specific subnetworks. The backbone, a ready-to-use convolutional network, generates a convolutional feature map for the entire input image. The first subnet performs convolutional object classification on the backbone's output; In the second subnet, convolutional bounding box regression is carried out. The two subnetworks have a clear design that we suggest explicitly for thick, one-stage discovery. Even though there are numerous options for these components' specifics, the majority of design parameters, as demonstrated by the experiments, are not particularly sensitive to precise values.

Training Dense Detection

Along with a variety of advancement systems, we conduct a variety of tests to investigate the behavior of the misfortune capability for thick identification. For each and every experiment, we build a Feature Pyramid Network on top of 50 or 101 ResNets. For all ablation studies, we use a 600-pixel image scale for training and testing.

Our first attempt to train RetinaNet without altering the initialization or learning strategy makes use of standard cross entropy loss. This fails quickly because the network diverges during training. However, in order to simulate initializing the final layer of our model so that the prior probability of detecting an object is p 14%, the following steps can be taken: 01 makes it possible to learn. From this initialization, training RetinaNet with ResNet- 50 yields a respectable AP of 30.2 on COCO. We use p 14 because the precise value of p has no effect on the results: 01 for each test.

The use of the a-balanced CE loss was our next attempt to improve learning. Creating a 14: 75 results in a gain of 0.9 AP points. To better comprehend the focal loss, we investigate the empirical loss distribution of a coupled model. For this, we use our default ResNet-101 600-pixel model with 36.0 AP that was trained with g 14 2. Using this model, a large number of random images are sampled to determine the predicted probability for 107 negative windows and 105 positive windows. From that point onward, we standardize the misfortune so it approaches one and register FL for these examples independently for up-sides and negatives. The cumulative distribution function (CDF) can be plotted and the normalized loss can be used to sort the loss from lowest to highest for both positive and negative samples.

In Relation to Accuracy: With larger backbone networks, accuracy rises, but inference speeds slow. The shorter image side is used to set the input image scale, which is the same. We plot the speed-to-accuracy trade-off curve for Retina-Net and compare it to more recent methods using public numbers on COCO test-dev. The plot demonstrates that RetinaNet, made possible by our focal loss, is superior to all other methods when the low-accuracy regime is removed. In comparison to the recently released ResNet-101-FPN Faster R-CNN, RetinaNet with ResNet-101-FPN and a 600-pixel image scale (for simplicity, we refer to it as

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 7 July 2023, pp: 53-57 www.ijaem.net ISSN: 2395-5252

RetinaNet-101-600) runs in 172 milliseconds per image. Because it uses larger scales, RetinaNet is faster and more accurate than any two-stage method. In terms of faster runtimes, ResNet-50-FPN is superior to ResNet-101-FPN only at one operating point (500 pixel input). The high frame rate regime cannot be addressed in this work because it will likely require special network design. We note that after publication, faster and more precise results can now be obtained using a modified Faster R-CNN.

α	AP	AP_{50}	AP_{75}
.10	0.0	0.0	0.0
.25	10.8	16.0	11.7
.50	30.2	46.7	32.8
.75	31.1	49.4	33.0
.90	30.8	49.7	32.3
.99	28.7	47.4	29.9
.999	25.1	41.7	26.1

(a) **Varying α for CE loss ($\gamma = 0$)**

γ	α	AP	AP_{50}	AP_{75}
0	.75	31.1	49.4	33.0
0.1	.75	31.4	49.9	33.1
0.2	.75	31.9	50.7	33.4
0.5	.50	32.9	51.7	35.2
1.0	.25	33.7	52.0	36.2
2.0	.25	**34.0**	**52.5**	**36.5**
5.0	.25	32.2	49.6	34.8

(b) Varying γ for FL (w. optimal α)

#sc	#ar	AP	AP_{50}	AP_{75}
1	1	30.3	49.0	31.8
2	1	31.9	50.0	34.0
3	1	31.8	49.4	33.7
1	3	32.4	52.3	33.9
2	3	**34.2**	**53.1**	**36.5**
3	3	34.0	52.5	**36.5**
4	3	33.8	52.1	36.2

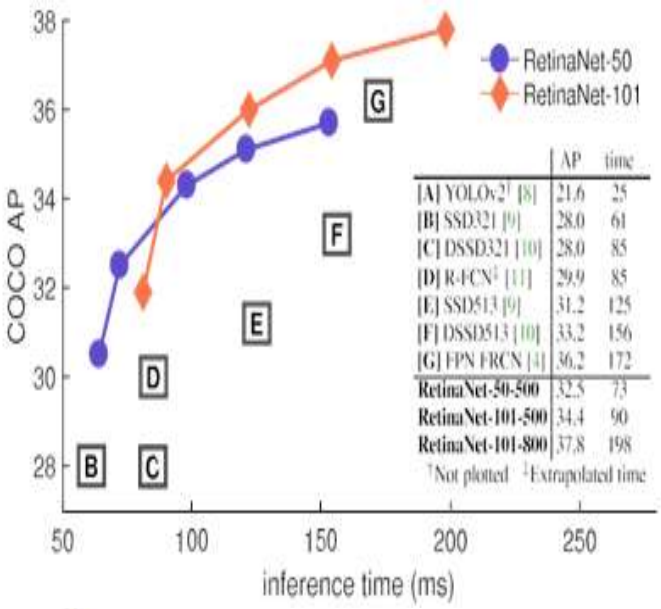
(c) Varying anchor scales and aspects

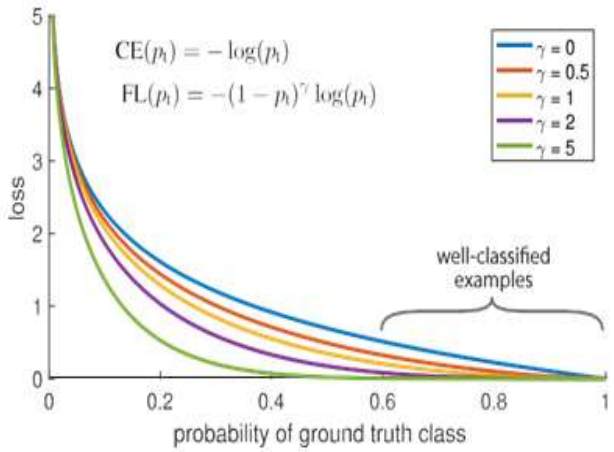
depth	scale	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	time
50	400	30.5	47.8	32.7	11.2	33.8	46.1	64
50	500	32.5	50.9	34.8	13.9	35.8	46.7	72
50	600	34.3	53.2	36.9	16.2	37.4	47.4	98
50	700	35.1	54.2	37.7	18.0	39.3	46.4	121
50	800	35.7	55.0	38.5	18.9	38.9	46.3	153
101	400	31.9	49.5	34.1	11.6	35.8	48.5	81
101	500	34.4	53.1	36.8	14.7	38.5	49.1	90
101	600	36.0	55.2	38.7	17.4	39.6	49.7	122
101	700	37.1	56.6	39.8	19.1	40.6	49.4	154
101	800	37.8	57.5	40.8	20.2	41.1	49.2	198

International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 7 July 2023, pp: 53-57 www.ijaem.net ISSN: 2395-5252

method	batch size	nms thr	AP	AP_{50}	AP_{75}
OHEM	128	.7	31.1	47.2	33.2
OHEM	256	.7	31.8	48.8	33.9
OHEM	512	.7	30.6	47.0	32.6
OHEM	128	.5	32.8	50.3	35.1
OHEM	256	.5	31.0	47.4	33.0
OHEM	512	.5	27.6	42.0	29.2
OHEM 1:3	128	.5	31.1	47.2	33.2
OHEM 1:3	256	.5	28.3	42.4	30.3
OHEM 1:3	512	.5	24.0	35.5	25.8
FL	n/a	n/a	**36.0**	**54.9**	**38.7**

(d) **FL** *vs.* **OHEM** baselines (with ResNet-101-FPN)





IV. CONCLUSION

In this study, one-stage object detectors can't beat the best two-stage methods because of class imbalance, according to our findings. We propose the focal loss, which adds a modulating term to the cross entropy loss, to focus learning on hard negative examples. Our system is direct and especially suitable. We report extensive experimental analysis that demonstrates its speed and accuracy at the cutting edge and demonstrate its effectiveness by creating a fully convolutional one-stage detector.

REFERENCES

[1]. R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich featurehierarchies for accurate object detection and semanticsegmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,2014, pp. 580–587.

[2]. R.Girshick,"FastR-CNN,"inProc.Int.Conf.Comput.Vis.,2015.

[3]. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towardsreal-time object detection with region proposal networks," inProc. Neural Inf. Process. Syst., 2015.

[4]. T.Y.Lin,P.Dollár,R.Girshick,K.He,B.Hariharan,andS.Belongie,"Featurepyramidnetworksforobjectdetection,"in Proc.IEEEConf.Comput.Vis.PatternRecognit.,2017.

[5]. K.He,G.Gkioxari,P.Dollár,andR.Girshick," MaskR-CNN,"in Proc.IEEEInt.Conf.Comput.Vis.,2017.

[6]. T.-Y.Lin,M.Maire,S.Belongie,J.Hays,P.Perona,D.Ramanan, P. Dollár,and C. L. Zitnick, "Microsoft COCO: Common objectsin context," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[7]. J.Redmon,S.Divvala,R.Girshick,andA.Farhadi,"Youonlylookonce:Unified,realtimeobjectdetection,"inProc.IEEEConf.Com-put. Vis. Pattern Recognit., 2016.

[8]. J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger,"in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017.

[9]. W.Liu,D.Anguelov,D.Erhan,C.Szegedy,S. Reed,C.-Y.Fu,and A.C. Berg, "SSD: Single shot multibox detector," in Proc. Eur.Conf. Comput. Vis., 2016.

[10]. C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD:Deconvolutional single shot detector," arXiv: 1701.06659, 2016.

[11]. J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection viaregion-based fully convolutional networks," in Proc. Neural Inf.Process. Syst., 2016.

[12]. J.R.Uijlings,K.E.vandeSande,T.Gevers,andA.W.Smeulders,"Selectivesearchforobjectrecognition,"Int.J.Comput.Vis.,vol. 104, pp. 154–171, 2013.

[13]. C.L.ZitnickandP.Dollár,"Edgeboxes:Locatingobjectproposalsfrom edges," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 391–405.

[14]. P. O. Pinheiro, R. Collobert, and P. Dollar, "Learning to segmentobject candidates," in Proc. Neural Inf. Process. Syst., 2015.

[15]. P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár,"Learning torefine object segments," in Proc. Eur. Conf. Comput. Vis., 2016.

[16]. A.Shrivastava,A.Gupta,andR.Girshick,"Trainingregion-basedobject detectors with online hard example mining," in Proc. IEEEConf. Comput. Vis. Pattern Recognit., 2016.

[17]. K.-K. Sung and T. Poggio, "Learning and example selection forobject and pattern detection," in MIT A.I. Memo No. 1521, 1994.

[18]. H. Rowley, S. Baluja, and T. Kanade, "Human face detection invisual scenes," Carnegie Mellon Univ., Pittsburgh, PA, USA,Tech. Rep. CMU-CS-95–158R, 1995.

[19]. P.ViolaandM.Jones,"Rapidobjectdetection usingaboostedcas-cade of simple features," in Proc. IEEE Conf. Comput. Vis. PatternRecognit., 2001, pp. I-511–I-518.

[20]. P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascadeobjectdetectionwithdeformablepartmodels,"inProc.IEEEConf.Comput. Vis. Pattern Recognit., 2010, pp. 2241–2248.