ORIGINAL PAPER

Designing effective power law-based loss function for faster and better bounding box regression

Diksha Aswal[1] · Priya Shukla[1] · G. C. Nandi[1]

Received: 23 September 2020 / Revised: 22 March 2021 / Accepted: 23 April 2021 / Published online: 27 May 2021

Abstract

Effective bounding box regression is essential for running any real-time object detection algorithm with acceptable accuracy. The currently available loss functions have issues like high computations, and sometimes they suffer from a subtle problem of plateau for non-overlapping bounding boxes, as the resultant bounding boxes are found to be far from the ground truth. In the present investigation, we have proposed a loss function with a new power-law term introduced in it for the normalized distance, which converges as fast as the Complete Intersection over Union (CIoU), but turns out to be computationally much faster than the Intersection over Union (IoU) and Generalised IoU (GIoU). The proposed function is simpler than CIoU. The incorporated power term has been optimized based on the corresponding computational time and on the sum of errors simulated for about multi-million cases, the details of which have been elaborated in the paper. The proposed Absolute IoU (AIoU) loss function has been successfully implemented and tested using the state-of-the-art object detection algorithms, such as You Only Look Once (YOLO) and Single Shot Multibox Detector (SSD) and is found to achieve significant performance improvement, using well-known metric Average Precision (AP), indicating the effectiveness of our approach.

Keywords Object detection · Bounding box regression · Absolute IoU · Distance IoU · Complete IoU

Abbreviations

G	Ground truth box
P	Predicted box
$\rho(p, g)$	Distance between the centre (p, g) of the boxes
c	Diagonal distance between the two opposite end points of the smallest box enclosing the G and P.
$\Re_{DIoU}$	Penalty term
α	Prefactor
N	Power law index
AP	Average precision
BB	Bounding box
x,y	Centre coordinates of the bounding box.
w, h	Width and height of the bounding box, respectively.
IoU	Intersection over Union
DIoU	Distance IoU
CIoU	Complete IoU
AIoU	Absolute IoU

1 Introduction

Machine vision is an essential tool for automating various tasks in the industry as well as in the household chores. Machine vision is very useful in various applications like identification [26], classification [9] and localization [1] which usually involves image capturing and image processing to identify the action based on extracted features. Nowadays, almost all fields of engineering require machine vision for autonomous control and inspection of machines. For example, an electronics industry uses automated inspection to examine the quality of the components in the production line. Such kind of automation, when applied in a Just In Time (JIT) manufacturing environment, improves the production rate and reduces rejection losses due to the detection of early defects. Moreover, machine vision is acting as a vector (medium) in enhancing machine intelligence as well

✉ Priya Shukla
priyashuklalko@gmail.com

Diksha Aswal
diksha.aswal94@gmail.com

G. C. Nandi
gcnandi@iiita.ac.in
https://sites.google.com/site/gcnandi/

[1] Center of Intelligent Robotics, Indian Institute of Information Technology, Allahabad, Prayagraj, U.P. 211015, India

as artificial intelligence level. In this context, it is pertinent to say that the machine intelligence (MI), being a subset of artificial intelligence (AI), is widely being used for intelligent control, and decision making in the manufacturing as well as numerous service industries, whereas artificial intelligence is showing tremendous application potentials in cognitive science, deep learning, language understanding, autonomous car driving, experience-based learning which may revolutionize the society as we live in. In many applications in both MI and AI, vision-based object detection is a kernel and plays a very important role in identifying and locating objects in an image, video or scene. It is increasingly being used to count objects in a scene, determine their precise locations and track them with accurate labelling. Its application is increasing in surveillance and security [16], traffic monitoring [13], video recognition [30], robot vision [14], face detection and recognition [3] to name a few. Even it is proving to be an important tool for autonomous driver less cars for its real-time collision-free outdoor navigation. Many object detection algorithms are invented which uses Bounding Box (BB), which is a rectangular structure superimposed over an image including all important features of a particular object staying within it. It is one of the simplest and less time taking techniques of image annotation. Having said that, it also suffers from imprecise localization and detection for which a popular technique known as BB regression is used where the BB regressors are trained to regress from either region proposals or fixed anchor boxes to nearby bounding boxes of a pre-defined target object classes. Initially, l_n norm is adopted as loss for the BB regression. But as suggested by [20,27], there is no strong correlation between-norms, and improving their IoU values. IoU loss is proposed to improve IoU metric.

$$\mathcal{L}_{IoU} = 1 - IoU \tag{1}$$

IoU is a popular metric also called as Jaccard index used for calculating the similarity between the two shapes where, G is the ground truth box and P is the predicted box. However, IoU loss works only for overlapping boxes. For non-overlapping boxes, it does not provide any information about whether the boxes are nearer or farther. The moving gradient for such cases is not available. Rezatofighi et al. [20] propose new method to overcome the problem of IoU loss called Generalised IoU (GIoU), by adding one more term into the $\mathcal{L}_{IoU}$.

$$\mathcal{L}_{IoU} = 1 - \frac{G \cup P}{G \cap P} \tag{2}$$

$$\mathcal{L}_{GIoU} = 1 - IoU + \frac{C - U}{C} \tag{3}$$

Here, C is the smallest box enclosing both the ground truth and the predicted box and U = $G \cup P$. The added term will make the box to move towards ground truth box in the non-overlapping cases. Handling slow convergence and poor regression for some cases, Zheng et. al [29] propose a new method, Distance IoU (DIoU). Instead of focusing on area terms to add penalty into the IoU loss, they focus on the distance between the ground truth box and the predicted box. This penalty term directly helps to minimize the distance between the centre of the two BB. The DIoU loss function is shown in (4) where p and g are the central points of the boxes and the penalty term tries to minimize the distance between the two boxes.

$$\mathcal{L}_{DIoU} = 1 - IoU + \frac{\rho(p,g)^2}{c^2} \tag{4}$$

For considering the aspect ratio as an important geometric factor, a new loss function has been proposed and named as Complete IoU (CIoU) as shown in (5). CIoU loss takes three geometric properties into account such as overlap area, central point distance and aspect ratio which leads to faster convergence and better performance.

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho(p,g)^2}{c^2} + \alpha\nu \tag{5}$$

In DIoU loss, the third term, i.e. $\frac{\rho^2(p,g)}{c^2}$ plays a vital role in the object detection as it enables the loss function to respond for the non-overlapping cases as shown in (4).

The value of IoU for the non-overlapping cases is zero and is invariant to any change in the BB position. The third penalty term takes care of such cases with the value of $\frac{\rho}{c}$ being tending towards zero as the ground truth box g and predicted box p moves close to each other. Further, it is important to note that as the two boxes moves farther and farther the sensitivity of the third term towards the change in the position of p decreases as represented in Fig. 1 between ρ and $\frac{\rho}{c}$. The derivative of the third term takes very small value when the g and p are far apart. Additionally, the small value of derivative will result in slow and poor convergence. Our analysis raises important questions towards the applicability of DIoU for the non-overlapping cases with the g and p farther apart as the value of penalty term is near to 1, and the derivative takes very small value. Further, the applicability of CIoU [29] has additional challenges as the fourth term mentioned in (5), which is supposed to converge aspect ratio of the BB to the ground truth box. The challenges are mentioned below:

1. The introduction of 4^{th} term definitely increases the computation time as calculating the trigonometric functions are computationally more expensive as compared to a mono or binomial expression.
2. For the non-overlapping bounding and ground truth box case, there is no point of changing the aspect ratio of the BB, if in the final state the BB is not converging to

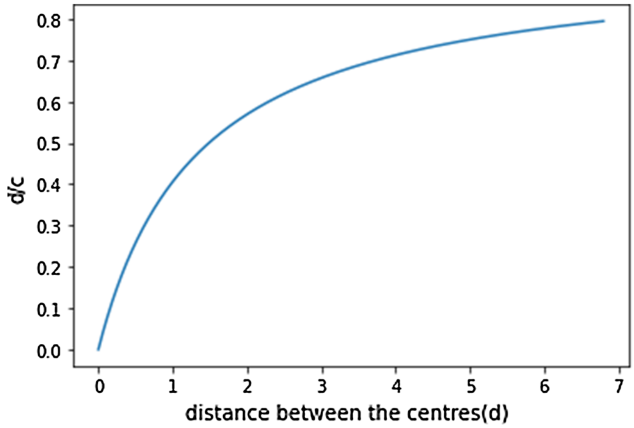


Fig. 1 ρ vs $\frac{\rho}{c}$

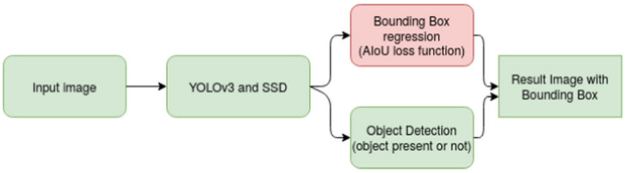

Input image
YOLOv3 and SSD
Bounding Box regression (AIoU loss function)
Object Detection (object present or not)
Result Image with Bounding Box

Fig. 2 Flow diagram

the ground truth box. Additionally, the optimization of two parameters (x,y), versus four parameters (x,y,w,h), is much easier as it may happen that the system might end up in local minimum due to changes in (w,h), values and the global minimum which has more dependence on (x,y) is never attained. In four parameters, (x,y) jointly represent the center coordinates of the BB, whereas w and h represent width and height of the BB, respectively.

3. For overlapping cases, the convergence of the aspect ratio of BB to that of the ground truth box is already being accounted by the 2nd term. The working of fourth term is redundant in the overlapping case and must be quantified in order to use it.

In this paper, we have tried to address these challenges and proposed a systematic framework to alleviate the problems with the current best known loss functions (DIoU and CIoU) and have designed a new loss function for BB regression which we are calling as Absolute IoU (AIoU). The design has been tested with the state-of-the-art object detection algorithms and found to achieve significant performance increase as compared to any known metrics, proving the effectiveness of our design. The outline of our proposed work is presented in Fig. 2, where AIoU is our proposed loss function for BB regression.

Major contributions of the paper

- A loss function with a new power-law term, introduced in it for the normalized distance, which converges as fast as the existing CIoU, but turns out to be computationally more efficient than the IoU and GIoU has been proposed.
- An optimized range of the power law index N for highest possible convergence with minimum computation time has been proposed.
- Extensive simulation experiments have been performed which show that the proposed loss function, which is much simpler than the existing ones, when used with the state-of-the-art algorithms like YOLO and SSD, provides improved results in some cases as stipulated in Tables 1 and 2 .

Table 1 Result of YOLOv3 [18]. The training is done on PASCAL-VOC [5] dataset

Loss / Evaluation	AP50	AP75	AP
CIoU [29]	0.6032	0.5428	0.5287
AIoU	0.5926	0.5432	0.49267

Table 2 Result of SSD [12]. The training is done on PASCAL-VOC [5] dataset

Loss/evaluation	AP50	AP75	AP
CIoU [29]	0.77	0.5435	0.5001
AIoU	0.769	0.5434	0.5003

2 Related work

2.1 Object detection

In the past, a number of classical object detection algorithms have been developed based on feature extractions and classifications [2,23,24].They are good for relatively small datasets. With the tremendous advancement of computation power and availability of huge dataset, Deep learning techniques are playing a vital role in handling object detection tasks with much better accuracy. In [7], authors proposed the Region-based Convolutional-Neural-Networks (R-CNN) which extracts regions from the image using selective-search algorithm [25] which are then fed to the Convolutional-Neural-Networks (CNN) for feature extraction. R-CNN is being further modified in [19] as Faster-RCNN and used in [28] to detect oil tanks with the proposed hierarchical approach to minimize false-alarm rates. In [24], authors have introduced the Polarimetric-SAR (PolSAR) segmentation for image which abolishes the need of parameter initialization and gives the better performance with significant noise resistance. With the above-mentioned research for images, researchers have also explored video processing [4,22]. Moreover, Song et al. [21] proposed a method with temporal feature aggregation for detecting multi-scale pedestrian objects which works well with the small scale pedestrian. A topological line is applied to pedestrian instances. However, none of the above-mentioned researches focus on developing a loss function which may help tuning learning parameters for efficient BB regression.

2.2 Loss function

There are many loss functions for object detection like focal loss [11](Lin et al. 2017), focuses training on a sparse set of hard examples and prevents the vast number of easy negatives. Pang et al. [15] proposed balanced loss for classification and bounding box regression. They apply balanced L1 loss, balanced IoU sampling which bring a significant improvement in the challenging dataset like PASCAL-VOC [5]. The regression of rectangular boxes is still the most popular manner in the state-of-the-art object detection algorithms [18], [6,8,12] A direct way of getting the regression is used initially. Redmon et al. in YOLOv1 [17] calculated the square root of the BB size for prediction. R-CNN [7] has used selective-search algorithms [25] to calculate the location and sizes using target boxes for formulating the parameters of BB. Ren et al. [19] come up with an idea of dense anchor boxes and convergences of the bounding boxes to the target boxes for better results. There are some methods [10,27] that try to incorporate IoU to improve the performance of the BB regression. But they have some shortcomings for non-overlapping bounding boxes. Rezatofighi et al. [20] propose a method which covers almost all the short comings in the previous IoU-based loss functions. However, here the convergence of losses are slow and hence, there is a scope of better box regression which is explained by Zheng et. al. [29]. They adopt better method which they call DIoU and CIoU losses which has faster convergence and better regression accuracy. In the present investigation, we further propose to make CIoU more simpler, keeping all its good properties intact, which can cover farther cases also. The developmental details of our proposed loss function have been discussed in subsequent section.

3 Methodology

3.1 Preliminaries

The target variable in an object detection is defined as: y = $[p_c, b_x, b_y, b_h, b_w, c_1, c_2, c_3, c_4]$ where p_c is the confidence of object, i.e. the classes present in the image called as confidence score. In object detection, target location is positioned by creating a BB around it. It is a rectangular box having x, y axes coordinates for the upper-left corner and x, y axes coordinates for the lower-right corner of the rectangle. The BB consists of parameters like $[b_x, b_y, b_w, b_h]$. Anchor Boxes help to detect objects of different sizes, scales and overlapping objects. The network model does not directly predict the BB. It predicts the probability for each anchor box. IoU is the ratio of overlapping area of target box or ground truth box ‘g’ to the predicted box ‘p’.

$$IoU = \frac{Areaof\,Overlap}{Areaof\,Union} \tag{6}$$

IoU and confidence are the metric that are used as a criteria for deciding whether the prediction is true positive or false positive.

3.2 Proposed approach

The BB detection metric and the corresponding loss functions have evolved to account for all possible cases, such as better regression of loss function for overlapping cases (Loss GIoU) and then incorporation of additional terms in loss function to define sensitivity for non-overlapping cases (Loss DIoU and CIoU). Even though Zheng et al. [29] claim that the CIoU BB loss function is the first complete method to regress the loss, it suffers from a subtle weakness of the reduced sensitivity of the loss function to the change in parameter values (x, y, w, h) when the bounding and ground truth boxes are farther apart. In this research, a novel approach has been proposed which serves as a solution to the existing problem defined earlier. The validation and robustness of the proposed approach over the existing methods are thoroughly quantified through simulation experiments similar to that proposed by Zheng et al. [29]. The approach here is to modify the third term $\frac{\rho(p,g)}{c}$ of the known DIoU loss function in order to alleviate the problem of slow and non-convergence of BB to the ground truth when the boxes are farther apart. The slow convergence happens as the value of $\frac{\rho(p,g)}{c}$ approaches to 1 when the boxes are farther apart. As a result of a value closer to 1 for $\frac{\rho(p,g)}{c}$, the derivative of the $\frac{\rho^2(p,g)}{c^2}$ (third term) takes a very small value when evaluated with respect to ‘(x, y, w, h)’ or ‘d/c’. The proposed approach consists of the loss function for BB regression similar to DIoU with the third term being changed to $\frac{\rho^2(p,g)}{c^2} * \alpha$ and named as “Absolute IoU (AIoU),” where the value of α would affect the derivative sensitivity of this penalty term, convergence time and total error after convergence. The new term can also be written as the modification to the third term of DIoU with an alpha pre-factor as : $(\frac{\rho(p,g)}{c})^{(N-2)} * (\frac{\rho(p,g)}{c})^2$. In order to determine the value of N, first the applicability of the framework must be tested. For the purpose of proof of concept, value of 2 to 8 can be chosen to see the affect on the sensitivity term $(\frac{\rho(p,g)}{c})^N$ with the change in ρ/c value. The increase in value of N, the derivative of $(\frac{\rho(p,g)}{c})^N$ with respect to any of the box parameters (x, y, w, h) would take larger value for higher values of

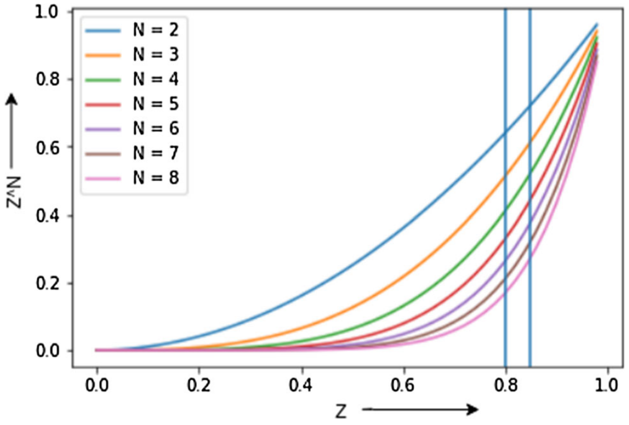


Fig. 3 z vs z^N

$\frac{\rho(p,g)}{c}$. For smaller values of $\frac{\rho(p,g)}{c}$ such as 0.1, the concept of higher N value would not be useful. Now, at this point, it is important to remember that the functioning of the penalty term is to response when there is a non-overlapping case. For cases where the ρ/c value is small, most of the time the cases will come under overlapping category and hence the convergence of the boxes would be taken care by the second IoU term. As the derivative of $(\frac{\rho(p,g)}{c})^N$ is dependent upon the $(\frac{\rho(p,g)}{c})^{N-1}$, as shown in (7), the higher value of N might result in much faster convergence for the non-overlapping cases where, $d/c = \frac{D^2}{C^2}$ and $G = d/c$.

$$f'(x) = (N/2) * (d/c)^{((N/2)-1)} * (G(x).dx) \tag{7}$$

So, the power of $\frac{\rho(p,g)}{c}$ can change the number of iteration in which loss converges and the distance upto which $\frac{\rho(p,g)}{c}$ will remain sensitive.

In the above graph, $z = \frac{\rho(p,g)}{c}$. As z is approaching towards 1, it is more steeper for higher power of z. The change in value of z is more, when power is higher as shown in Fig. 3. Two vertical lines corresponding to $z_1 = 0.8$ and $z_2 = 0.85$ shows that the change from z_1 to z_2 is more when N is higher. As higher the value of N, more changes will be noticed. Higher changes will lead to faster convergence of parameters. But the value N cannot be randomly taken as higher as possible. So, here we have investigated what is the rational behind selecting N? Basically there are two criteria for selecting N, i.e. **convergence time** and **error**. Considering the third term $(\frac{\rho(p,g)}{c})$, i.e. penalty term for now, where $D = \rho(p, g)$, distance between the centre of the boxes and C is the diagonal distance of the smallest box covering predicted box p and ground truth box g, there are four parameters (x, y, w, h) which help in converging p to g. Now, taking derivative of the function (8) w.r.t x, we get an expression which has been shown in (7).

$$F = (\frac{D}{C})^N \tag{8}$$

Now, the whole concept should be that, the d/c value should change as we change the value of N so that we get the highest possible convergence. The higher value of N will be leading to more computation time as shown in (7). Convergence time is the product of convergence steps and computation time of each step. As the value of N is increased, the number of convergence step will be less but after a certain value of N computation time will start dominating the metric and convergence time starts increasing. As discussed above, with increase in the value N, the chances of error keep decreasing, but we have to make sure for optimum point of convergence time error must be lower. So, to get the optimum point where the convergence time and the error is low, simulation experiments are performed. Another important aspect proposed is that, for the non-overlapping BB and ground truth BB case, there is no point of changing the aspect ratio of the BB, if in the final state the BB is not converging to the ground truth box.

3.3 Simulation experiment

Adapting the simulation experiments for DIoU loss function from DIoU and CIoU, here all the cases of different distances, scales and aspect ratios are taken. During experiments, 7 ground truth boxes with same centre and area $(area = 1)$ are considered, and hence, all the ground truth boxes are the unit boxes with different aspect ratio, i.e. 1:4, 1:3, 1:2, 1:1, 2:1, 3:1 and 4:1. **Distance:** 500 points are chosen uniformly around the centre with radius 5 to place the anchor boxes. The anchor boxes are distributed uniformly at 500 points with 7 different aspect ratio and 7 different areas. Forty-nine different boxes are kept fixed at each point. **Scale:** For each point, 7 different areas are set as 0:5, 0:67, 0:75, 1, 1:33, 1:5 and 2. **Aspect Ratio:** For the given point, 7 aspect ratios are chosen that are same as the target boxes (i.e. 1:4, 1:3, 1:2, 1:1, 2:1, 3:1 and 4:1). All the 7 × 7 × 7 × 500 boxes are to be converged to the target boxes. Total 7 × 7 × 7 × 500 = 171,500 regression cases will take place.

Algorithm 1: Algorithm of calculating loss

```
Input: n = 171,500 are the anchor boxes at all the 500 scatter point around the centre (10,10) and radius = 5. $G_{i=1}^{i=7}$ are all the target boxes with 7 aspect ratios and area = 1. All the target boxes are fixed at centre (10,10).;
for $i \leftarrow 1$ to $7$ do
    for $j \leftarrow 1$ to $n$ do
        for $t \leftarrow 1$ to $T$ do
            $\eta = \begin{cases} 0.1, & t <= 0.8T \\ 0.01, & 0.8T < t <= 0.9T; \\ 0.001 & t > 0.9T \end{cases}$
            $\Delta P_j^{t-1}$ is gradient of $\mathcal{L}_{AIoU}(P_j^{t-1}, G_i)$ w.r.t $P_j^{t-1}$;
            $P_i^t = P_i^{t-1} + \eta(2 - IoU_i^{t-1})\Delta P_i^{t-1}$;
            $E(t,j) = E(t,j) + |P_j^t - G_i|$;
        end
    end
end
Return $E$
```

With the help of the loss function $\mathcal{L}$, the simulation of the BB regression for each case has been done. For the Predicted Box P_j, the current prediction is obtained by (9) where P_j^t is the Predicted Box at iteration t, ΔP_j^{t-1} is the gradient loss $\mathcal{L}$ w.r.t P_j^{t-1}. η is the learning step and gradient is multiplied by the term $(2 - IoU_i^{t-1})$ to accelerate the convergence. For each case, simulation is terminated when $T = 200$ and error curve is shown in Fig. 4.

$$P_j^t = P_j^{t-1} + \eta(2 - IoU_j^{t-1})\Delta P_j^{t-1} \tag{9}$$

For finding the optimum value of N, the error plot which is shown in Fig. 4 is combined with the error of the (x, y) only, while parameter (w, h) is not taken into considerations as of now, since there is no change in the values of w and h before the overlapping happens as described in Algorithm 1. The error E is calculated using (10) where c is the total cases and n is the total number of anchor boxes taken for each cases.

$$Error = \sum_{i=1}^{c}\sum_{j=1}^{n} E(i, j), \tag{10}$$

Fig. 4 shows that we get the minimum error and time at $N = 3$. The error is increasing with increase in N because with the larger value of N closer boxes will not converge faster and the changes will be very small which will lead to increase the error. However, when the boxes start overlapping the same algorithm of DIoU can be used. The increase in the convergence time after $N = 3$ shows that after a particular value of N computation time of convergence dominates the whole metric, for the loss function:

$$\mathcal{L}_{AIoU} = 1 - IoU + \Re_{DIoU} * \alpha \tag{11}$$

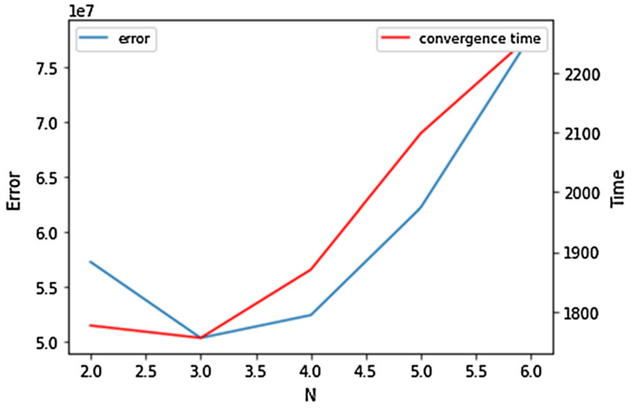


Fig. 4 175,100 regression cases are considered to get the optimum value of N. Convergence Time and Error for different values of N

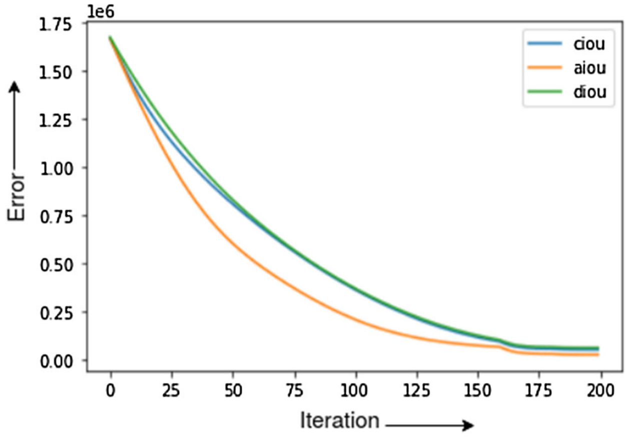


Fig. 5 Regression error summation curves for CIoU and AIoU (centres)

where:
$\Re_{DIoU} = \frac{\rho^2(p,g)}{c^2}$, $\alpha = (\frac{\rho(p,g)}{c})^{N-2}$, $N = 3$

For the cases where $IoU < threshold$, i.e. for the non-overlapping cases, the $\mathcal{L}_{AIoU}$ as shown in (11) is applied and only x, y are optimized. And for the cases where $IoU >= threshold$, $N = 2$ and w, h are also converged parallelly.

To prove our given assumption, the overall errors for both CIoU and AIoU method have been calculated and the regression sum error characteristics are shown in Fig. 5 and 6. The error sum curve Fig. 5 takes the summation of error of parameters (x, y), and Fig. 6 is the loss curve of parameters (w, h). In AIoU case, initially only two parameters (x, y) are optimized, when the boxes are close enough to each other (such that they overlap), subsequently, convergence of other two parameters (w, h) is also taken into consideration which results in faster convergence, as optimization of two parameters is taken place at a time rather than four parameters which is much easier to compute. CIoU has a fourth term for considering the consistency of aspect ratios for BB. But it also increases the computation time. Fig. 7 demonstrates the above-mentioned facts and therefore, converges all the four parameters faster than any other loss function.

As there is no point of optimizing width and height of the predicted box when boxes are not overlapped, it should be done only when the boxes are overlapped which obviously will decreases the computation time as well as reduce the fluctuations in the parameters.

In summary, our proposed AIoU loss function is found to do BB regression much faster and has proved to gain significant performance improvements as compared to the DIoU and CIoU loss functions. The main reason behind the performance improvement of the AIoU over DIoU and CIoU is the reconfiguration and modification of the third penalty term. The higher power of $\frac{\rho(p,g)}{c}$ ($N = 3$ in AIoU versus $N = 2$ for DIoU) plus first shifting predicted box towards ground truth and then working over its aspect ratio, ensured that BB approaches the ground truth box at a much faster rate without any plateau in the derivative value when the boxes are farther apart. The achieved fast convergence in the case of AIoU is due to two following reasons.

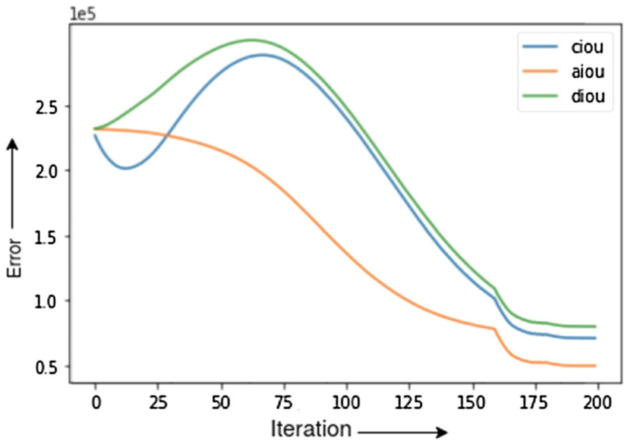
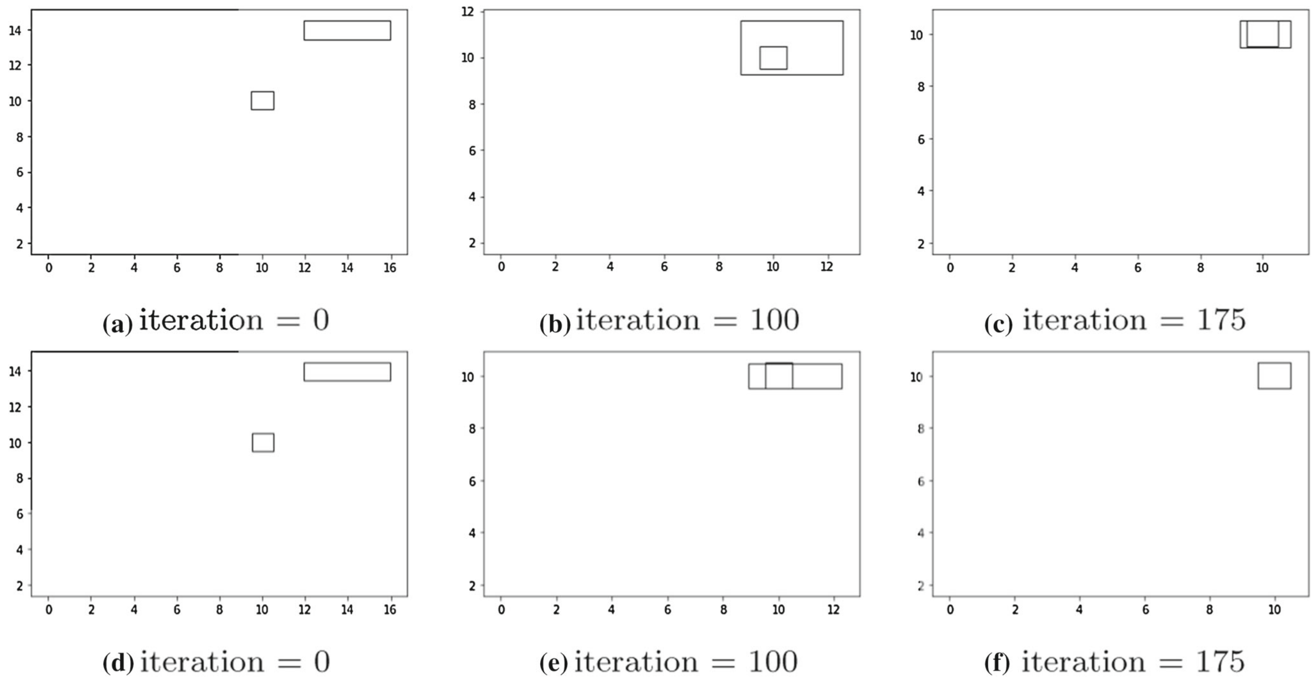


Fig. 6 Regression error summation curves for CIoU and AIoU (width, height)

- First, making higher value of power, for the third term, enhances the response (derivative) of the loss function to the change in regression parameters (x, y, w and h).
- Second, the fourth term (aspect ratio) in CIoU is not required and has been dropped in the AIoU as the proposed AIoU loss brings the boxes closer without changing the aspect ratio and once the boxes start to overlap, the aspect ratio converges to the true value due to the IoU term.

Further, the experiment results prove the authenticity of the proposed methods with the existing methods and paves the way for implementing the newly proposed AIoU loss function into the state-of-the-art object detection algorithms, such as YOLOv3 and SSD.



(a) iteration $= 0$ **(b)** iteration $= 100$ **(c)** iteration $= 175$

(d) iteration $= 0$ **(e)** iteration $= 100$ **(f)** iteration $= 175$

Fig. 7 Simulation Experiments run with both the AIoU and CIoU loss functions,(a), (b) and (c) show how the CIoU loss function is working. (d), (e) and (f) show that AIoU converges the box by iteration 175^{th}

4 Experimental results

The proposed AIoU loss function has been implemented on some state-of-the-art object detection algorithms such as one stage detection algorithms (YOLOv3 [18] and SSD [12]). Two popular datasets used in the experiments are VOC2007 & VOC2012 [5]

4.1 YOLOv3 on PASCAL VOC

YOLOv3 is the most effective object detection algorithm till now. It gives very fast and accurate results. PASCAL VOC is one of the important dataset for designing object detection and recognition based models. YOLOv3 is trained on PASCAL VOC dataset for both CIoU and AIoU loss functions. VOC2007 & VOC2012 both are used in combination for training. So total 16,551 images from 20 classes are trained and the test data are VOC 2007, which consists of 4,952 images. For the comparison purposes, AP50, AP75 and AP have been calculated and used as performance metrices as shown in Table 1.

4.2 SSD on PASCAL VOC

The PyTorch implementation has been used in line with the Liu et. al. [12] paper. Resnet50 model has been used as a backbone for the VOC dataset. The model has been trained with around 232 epochs and each epoch having around 514 iterations, so around 120K iterations in total have been used for training the model . Initially, smooth l_n norm is used as BB regression which are not functional with IoU metric. So, better loss has been required. And there should be appropriate loss function for BB regression. For faster convergence, a default momentum term (0.9) has been used. It has been observed that for dense anchor boxes, taking higher value of weights gives more power to the regression, thus improving the performance. The default weight for box regression has been fixed to 5 as done for DIoU. For the comparison purposes, AP50, AP75 and AP have been calculated and used as performance metrices as shown in Table 2.

After measuring the performance of our proposed method over the metric, Average Precision (APs), as shown already in Table 1 and 2 , we randomly selected some images and applied our loss function, AIoU, for computing BB and found out that it was correctly able to attach the bounding boxes unambiguously, good enough for the object detection purposes as shown in Fig. 8.

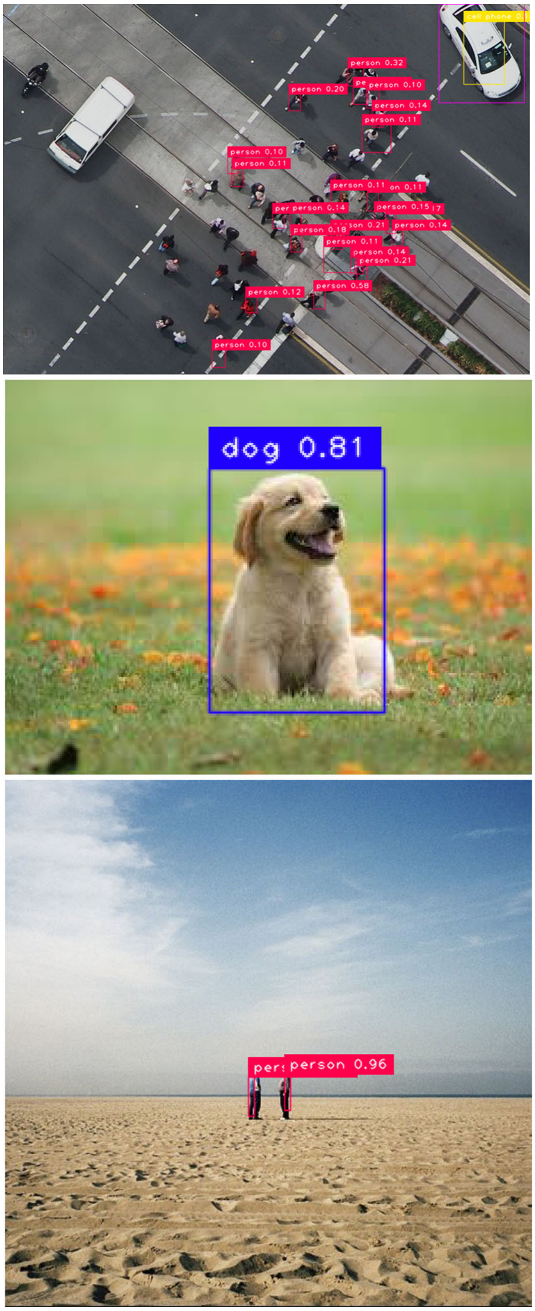


Fig. 8 Detection Results using the proposed AIoU loss function

5 Conclusion and future work

As it is well known, object detection algorithms highly rely on the regression of the bounding boxes. Here, we proposed and established a novel idea towards getting a faster and better BB regression. In addition to the introduction of α in the third term, the aspect ratio of CIoU loss function has been dropped in the proposed approach. By minimizing the distance between the boxes by multiplying it with a factor makes it more simpler. Instead of calculating the aspect ratio by taking another term in the loss function, we have proposed an idea of optimizing the width and height of the boxes only when they start overlapping. Firstly, the higher value of the power for the third term enhances the response (derivative) of the loss function to the changes in the regression parameters (x, y, w, h). Secondly, the fourth term (aspect ratio) in CIoU is not required at all and has been dropped in the AIoU as the proposed AIoU loss brings the boxes closer without changing the aspect ratio and once the boxes start to overlap as the aspect ratio converges to the true value due to the presence of IoU term. Our proposed AIoU loss function has given a significant improvement over CIoU in some cases. For YOLOv3 as well as SSD, the AP's values are better for our method (AIOU) such as 0.5926, 0.5432 (marginally nearer with 0.5428), 0.49267 and 0.769, 0.5434, 0.5003, respectively, as shown in Tables 1 and 2 . The proposed loss function, being good and simple, would take much less computation time in converging the bounding boxes and would work much better for many farther (distant) cases.

As a future work, this loss function is required to be experimented with more Data sets to test for its generalization and robustness. For now, we have implemented this only on the PASCAL VOC dataset. More state-of-the-art object detection algorithms can be implemented using AIoU loss function to further explore the performance of the proposed approach

Acknowledgements We are very thankful to the Central Computing Facility (CCF) of Indian Institute of Information Technology, Allahabad for providing us the necessary GPU access.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Abdallah, A.A., Saab, S.S., Kassas, Z.M.: A machine learning approach for localization in cellular environments. In: 2018 IEEE/ION position, location and navigation symposium (PLANS), pp. 1223–1227 (2018). https://doi.org/10.1109/PLANS.2018.8373508
2. Akbarizadeh, G.: A new statistical-based kurtosis wavelet energy feature for texture recognition of SAR images (2012). https://doi.org/10.1109/TGRS.2012.2194787
3. Chen, W., Huang, H., Peng, S., Zhou, C., Zhang, C.: Yolo-face: a real-time face detector. Vis. Comput. (2020). https://doi.org/10.1007/s00371-020-01831-7
4. Davari, N., Akbarizadeh, G., Mashhour, E.: Intelligent diagnosis of incipient fault in power distribution lines based on corona detection in uv-visible videos. In: IEEE Transactions on Power Delivery pp. 1–1 (2020). https://doi.org/10.1109/TPWRD.2020.3046161
5. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**, 303–338 (2009)
6. Fu, C., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: DSSD : Deconvolutional single shot detector. CoRR **abs/1701.06659** (2017). arXiv:1701.06659
7. Girshick, R.B., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR **abs/1311.2524** (2013). arXiv:1311.2524
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322
9. Ikonomakis, E., Kotsiantis, S., Tampakas, V.: Text classification using machine learning techniques. WSEAS Trans. Comput. **4**, 966–974 (2005)
10. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.) Computer Vision—ECCV 2018, pp. 816–832. Springer, Cham (2018)
11. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. CoRR **abs/1708.02002** (2017). arXiv:1708.02002
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.E., Fu, C., Berg, A.C.: SSD: single shot multibox detector. CoRR **abs/1512.02325** (2015). arXiv:1512.02325
13. Lorenčí-k, D., Zolotová ,I.: Object recognition in traffic monitoring systems. In: World Symposium on Digital Intelligence for Systems and Machines (DISA), pp. 277–282 (2018). https://doi.org/10.1109/DISA.2018.8490634
14. Mahajan, M., Bhattacharjee, T., Krishnan, A., Shukla, P., Nandi, G.C.: Robotic grasp detection by learning representation in a vector quantized manifold. In: International Conference on Signal Processing and Communications (SPCOM), pp. 1–5 (2020). https://doi.org/10.1109/SPCOM50965.2020.9179578
15. Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., Lin, D.: Libra R-CNN: towards balanced learning for object detection. CoRR **abs/1904.02701** (2019). arXiv:1904.02701
16. Raghunandan, A., Mohana, Raghav, P., Aradhya, H.V.R.: Object detection algorithms for video surveillance applications. In: International Conference on Communication and Signal Processing (ICCSP), pp. 0563–0568 (2018). https://doi.org/10.1109/ICCSP.2018.8524461
17. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: You only look once: Unified, real-time object detection (2015). arXiv:1506.02640
18. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. CoRR **abs/1804.02767** (2018). arXiv:1804.02767
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1 (2015). https://doi.org/10.1109/TPAMI.2016.2577031
20. Rezatofighi, S.H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I.D., Savarese, S.: Generalized intersection over union: A metric and A loss for bounding box regression. (2019). arXiv:1902.09630
21. Song, T., Sun, L., Xie, D., Sun, H., Pu, S.: Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation. (2018). arXiv:1807.01438
22. Tang, P., Wang, C., Wang, X., Liu, W., Zeng, W., Wang, J.: Object detection in videos by high quality object linking. IEEE Trans. Pattern Anal. Mach. Intell. **42**(5), 1272–1278 (2020). https://doi.org/10.1109/TPAMI.2019.2910529
23. Tirandaz, Z., Akbarizadeh, G.: A two-phase algorithm based on kurtosis curvelet energy and unsupervised spectral regression for segmentation of sar images (2016). https://doi.org/10.1109/JSTARS.2015.2492552
24. Tirandaz, Z., Akbarizadeh, G., Kaabi, H.: Polsar image segmentation based on feature extraction and data compression using weighted neighborhood filter bank and hidden markov random field-expectation maximization. Measurement **153**, 10732 (2020). https://doi.org/10.1016/j.measurement.2019.107432
25. van de Sande, K.E.A., Uijlings, J.R.R., Gevers, T., Smeulders, A.W.M.: Segmentation as selective search for object recognition.

In: International Conference on Computer Vision, pp. 1879–1886 (2011). https://doi.org/10.1109/ICCV.2011.6126456

26. Wäldchen, J., Mäder, P.: Machine learning for image based species identification. Methods Ecol. Evol. **9**, 1 (2018). https://doi.org/10.1111/2041-210x.13075
27. Yu, J., Jiang, Y., Wang, Z., Cao, Z., Huang, T.S.: Unitbox: An advanced object detection network. CoRR **abs/1608.01471** (2016). arXiv:1608.01471
28. Zalpour, M., Akbarizadeh, G., Alaei-Sheini, N.: A new approach for oil tank detection using deep learning features with control false alarm rate in high-resolution satellite imagery (2020). https://doi.org/10.1080/01431161.2019.1685720
29. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D.: Distance-iou loss: Faster and better learning for bounding box regression (2019). arxiv:1911.08287
30. Zhou, H., Meng, D., Zhang, Y., Peng, X., Du, J., Wang, K., Qiao, Y.: Exploring emotion features and fusion strategies for audio-video emotion recognition, pp. 562–566 (2019). https://doi.org/10.1145/3340555.3355713

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.