

基于 Sinkhorn 的微小目标检测损失函数

李文生¹
LI Wensheng

摘要

微小目标检测在目标检测任务中是一个非常具有挑战性的任务。因为小目标物体在图片中只拥有极少的像素信息。观察到基于标准 IoU 的指标在检测微小目标时,对预测边界框与真实框之间的偏移非常敏感,同时无法对预测边界框与真实框之间无重叠情况的优化。通过计算预测框与真实框之间的分布距离提出了一个新的微小目标检测的评估指标。具体来说,将边界框建模为二维高斯分布,然后提出一种称为归一化 Sinkhorn 距离的新度量,计算预测框与真实框二维高斯分布之间的距离。所提出的归一化 Sinkhorn 距离可以取代 IoU 度量。实验表明,使用归一化 Sinkhorn 距离指标时,所提方法产生的性能比标准基线高 6.7 个 AP 点,比最先进的竞争对手高 6.0 个 AP 点。

关键词

目标检测;微小目标检测;损失函数

doi: 10.3969/j.issn.1672-9528.2026.05.001

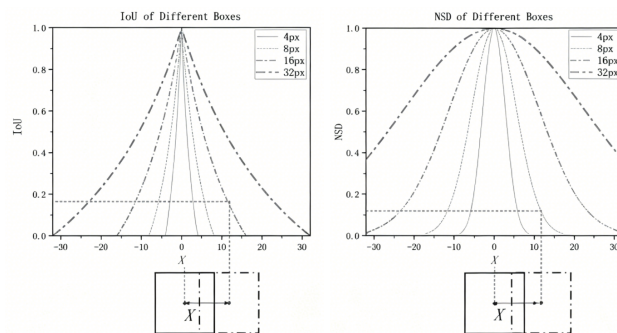
0 引言

目标检测在计算机视觉研究领域是人们主要的研究方向,随着卷积神经网络的发展,在自然图像数据中(例如 MS COCO^[1], Rascal VOC^[2])许多模型(Faster R-CNN^[3]、YOLO^[4]、SSD^[5])都取得了巨大的成功。然而现有的目标检测网络无法在微小目标检测中取得令人满意的检测结果。

主流算法普遍将交并比 IoU 作为评估指标,本文研究发现,航拍场景以微小目标为主,该度量指标难以适配此类检测任务。如图 1(a)所示,显示了不同边界框尺寸的 IoU 曲线,每条曲线都是通过保持边界框沿水平方向移动来绘制的。可以观察到边界框的尺寸越小,其曲线下降的速度越快,这显示了 IoU 度量对具有离散位置偏差的对象不具有尺度不变性。同时也可以发现每个尺寸的 IoU 曲线在横坐标的绝对值大于尺寸时,其 IoU 值便始终为 0。即如果真实框与预测框,两个框不相交,那么就无法衡量两个框的距离,同时值为 0。常用 IoU 作为优化函数与度量函数,当作为优化函数时,可以反向传播。也可以使用优化器进行优化得到最优值。如果 IoU 的值为 0,计算梯度的时候就无法使用优化器进行传播。

为了解决 IoU 在小目标检测中存在的缺陷,在本文中提出了一个新的评价标准归一化 Sinkhorn 距离(normalize sinkhorn distance, NSD),评价边界框两个中心点之间的归一化欧几里得距离,来缓解 IoU 在边界框上的问题。图 1(b)展示了不同边框尺寸的 NSD 曲线。可以看到当两个边框的距

离在逐渐变大时,NSD 值的下降速度比 IoU 的下降速度慢,尤其是当物体尺寸较小时。



(a) 不同边界框尺寸的 IoU 曲线 (b) 不同边框尺寸的 NSD 曲线

图 1 基于 IoU 和基于 NSD 的检测器的比较

这项工作的主要贡献可以概括如下:

- (1) 本文分析了微小物体上不同度量的属性,并提出了一种称为归一化 Sinkhorn 距离(normalize sinkhorn distance, NSD)的新度量,以克服微小物体上传统度量的弱点。
- (2) 将 NSD 应用于 RPN 和 NMS 模块中的正负分配模块。实验结果表明,提出的 NSD 比基线取得了显著的改进,并在 AI-TOD 数据集上获得了最先进的结果。

1 方法

本文基于 Sinkhorn 距离设计了一种在航拍微小目标检测中具有更好性能的度量。可以反映分布之间的距离,即使分布之间没有重叠的部分。实际上是对预测框与真实框进行数学建模,将其转化为对应高斯分布。计算高斯分布之间的距

离就可以反映真实框与预测框之间的关系。

1.1 边界框的高斯分布建模

使用 IoU 在评价边界框之间的相似性的过程中，默认了边界框内的像素是完全对应目标物体，即目标完全在目标框内且边界框内完全是目标物体。即边界框内的像素在参与计算时，每个像素在进行计算时的权重是符合均匀分布的，每个像素的权重是相同的。但是对于真实世界的目标来说，并非严格矩形，边界框周围往往存在一些背景像素。在这些边界框中，前景像素和背景像素分别集中在边界框的中心和边界上。

因此，在计算两个边界框相似度时，应该给予边界框中心点像素更高的权重，并且像素的重要性从中心到边界递减。本文选用维高斯分布的概率密度函数来描述边界框内不同像素的权重。

对于水平边界框 $R = (c_x, c_y, w, h)$ ，其中 (c_x, c_y) 、 w 和 h 分别表示中心坐标、宽度和高度。其内切椭圆方程可表示为：

$$\frac{(x - \mu_x)^2}{\sigma_x^2} - \frac{(y - \mu_y)^2}{\sigma_y^2} = 1 \quad (1)$$

式中： (x, y) 表示椭圆的中心坐标； (σ_x, σ_y) 表示沿 X 轴和 Y 轴的半轴长度。因此， $\mu_x = c_x$ ， $\mu_y = c_y$ ， $\sigma_x = w$ ， $\sigma_y = h$ 。

二维高斯分布的概率密度函数用公式表示为：

$$f(p|\mu, \Sigma) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right) \quad (2)$$

式中： p 表示像素坐标 (x, y) ； μ 表示变量的均值； Σ 表示协方差矩阵，描述各维变量之间的相关度。当：

$$(x - \mu)^T \Sigma^{-1}(x - \mu) = 1 \quad (3)$$

内切椭圆方程将可以转化为二维高斯分布的密度等值线。

因此，水平边界框 $R = (c_x, c_y, c_x, w, h)$ 可以建模为二维高斯分布 $N = (\mu, \Sigma)$ ，其中：

$$\mu = [c_x, c_y]^T, \Sigma = \left[\left[\left(\frac{w}{2} \right)^2, 0 \right]^T \left[0, \left(\frac{h}{2} \right)^2 \right]^T \right] \quad (4)$$

所以，对两个边界框相似度的评价可以转换为计算两个二维高斯分布之间的距离。

1.2 归一化高斯 Sinkhorn 距离

使用来自最优传输理论的 Sinkhorn 距离来计算两个分布之间距离。

对于两个二维高斯分布 $\mu_1 = \mathcal{N}(m_1, \Sigma_1)$ 和 $\mu_2 = \mathcal{N}(m_2, \Sigma_2)$ ， μ_1 和 μ_2 之间的二阶 Sinkhorn 距离定义为：

$$D_M^\lambda(N_p, N_g) = \inf \mathbb{E}(\|N_p - N_g\|_2^2)^{1/2} - \frac{H(N_p, N_g)}{\lambda} \quad (5)$$

$$H(N_1, N_2) = - \sum P_{N_1, N_1} \log P_{N_1, N_1} \quad (6)$$

式中： P_{N_1, N_1} 是真实框 $\mu_1 = \mathcal{N}(m_1, \Sigma_1)$ 与预测框 $\mu_2 = \mathcal{N}(m_2, \Sigma_2)$ 的熵。参数 λ 负责调整熵对结果的影响程度， λ 越大，熵的影响越小。

但是， $D_M^\lambda(N_p, N_g)$ 表示的两个分布之间的距离，不能直接用来作为两个分布之间的相似度（即 0 到 1 之间的值）。因此使用其指数形式归一化并获得称为归一化 Sinkhorn 距离（NSD）的新度量：

$$\text{NSD}(N_p, N_g) = \exp\left(-\frac{\sqrt{D_M^\lambda(N_p, N_g)}}{C}\right) \quad (7)$$

$$C = \sqrt{\frac{\sum_{i=1}^M \sum_{j=1}^{N_i} w_{ij} \times h_{ij}}{\sum_{i=1}^M N_i}} \quad (8)$$

式中： M 表示某个数据集中的图像数量； N_i 表示第 i 张图像中标记的边界框数量； w_{ij} 、 h_{ij} 表示第 i 个图像中第 j 个边界框的宽度和高度。

与常规的评价指标相比，NSD 在检测微小目标的任务中具有以下优点：

（1）尺度不变性：NSD 的实质是计算两个分布之间的距离，而相同形状但边长不同的目标的分布是相同的。所以目标的大小不会对 NSD 的计算产生影响。

（2）对位置偏移的平滑性：可以测量任意位置边界框（包括没有重叠的边界框与互相包含的边界框）。

2 实验

2.1 实验设置

本文提出的方法在对微小目标的检测进行实验时，将使用 AI-TOD^[6] 这一航空图像数据集。AI-TOD 数据集拥有 28 036 张航空图片，包含 700 624 个对象实例，图片的大小为 800 px×800 px。AI-TOD 有 8 个实例对 airplane (AI)，bridge (BR)，storage-tank (ST)，ship (SH)，swimming-pool (SP)，vehicle (VE)，person (PE)，wind-mill (WM)。AI-TOD 中最大的物体小于 64 px，AI-TOD 中 86% 的物体小于 16 px，平均像素为 12.8 px，远小于其他数据集。如 PASCAL VOC (156.6 px)、MS COCO (99.5 px)。为了获得与最先进方法相比的最终性能，并分别在训练集和测试集上训练和评估最终模型。

使用 ImageNet^[7] 预训练的 ResNet-50^[8] 和 FPN 作为主干。所有实验均基于 MMDetection^[9] 代码库，并在具有 NVIDIA RTX3090 GPU 的计算机上进行训练。所有模型均使用随机梯度下降 (SGD) 优化器进行 12 个周期的训练，动量为 0.9，权重衰减为 0.000 1，批量大小为 4。将初始学习率设置为

0.005，并在 epoch8 和 11 处对其进行衰减。在推理阶段，使用预设分数 0.05 来过滤背景边界框，并应用 IoU 阈值 0.5 的 NMS 来生成前 100 个置信边界盒子。除非另有说明，所有实验均使用上述训练和推理参数。

此外，使用 AP(平均精度)指标来评估所提出方法的性能。首先，精确率和召回率分别为：

$$\text{Precision} = \frac{TP}{TP + FP}$$
(9)

$$\text{Recall} = \frac{TP}{TP + FN}$$
(10)

式中：TP、FP 和 FN 分别表示真阳性（true positive）、假阳性（false positive）和假阴性（false negative）的数量。将以 Recall 为图像的横坐标，以对应的 Precision 为纵坐标，得到 P-R 图。

2.2 与其他损失函数的对比

在这项工作中，在同一基本网络（即 Faster R-CNN）上重新实现了四个指标，即 IoU、GIoU^[10]、CIoU^[11]和 DIoU，并且实现了本文所提出的 NSD。以比较其在微小目标检测任务上的性能。

GIoU、DIoU 和 CIoU 在原始工作中用作损失函数，但这些指标在本工作中同时用于正和负锚分配模块以及 NMS 模块。具体来说，正负锚点分配模块是为训练 RPN 而设计的，RPN 将为每个锚点分配一个二进制类标签。当锚点在真实边界框中具有最高度量值或在任何真实边界框中具有高于 0.7 的度量值时，会为该锚点分配正标签。同时，如果所有真实边界框的度量值都低于 0.3，会为负锚点分配一个负标签。非正负锚点对训练没有贡献。NMS 模块旨在通过度量值过滤掉冗余边界框。但需保持阈值参数与 Faster R-CNN 相同。

如表 1 所示，发现基于 IoU 的指标（即 DIoU、CIoU、GIoU 和 IoU）在 AP 上比 NSD 更差，认为准确度差距是由于在训练阶段难以将高质量的正锚点分配给真实样本造成的。

表 1 NSD 与基于 IoU 的损失函数在 AI-TOD 验证集上的比较

Function	AP	AP _{0.5}	AP _{0.75}	AP _{vt}	AP _t	AP _s	AP _m
DIoU	11.4	25.6	6.9	0.0	6.9	24.4	32.2
CIoU	10.8	24.5	6.9	0.0	6.7	24.5	31.7
GIoU	10.4	23.4	5.4	0.0	5.5	20.7	29.4
IoU	11.1	26.3	7.6	0.0	7.5	25.3	33.6
NSD(Our)	17.6	43.2	11.0	2.9	17.6	26.5	34.5

注：最好的指标加粗。

一方面，对于低信噪比的微小物体，其边界可能与背景混淆，导致正负样本分配模块中心距比宽度和高度更重要。另一方面，基于 IoU 的指标对微小物体很敏感。因此，使用仅考虑边界框之间所在分布之间距离的 NSD 度量可以获得更好的结果。

2.3 主要试验

将在 AI-TOD 数据集上，验证 NSD 的有效性。在这个实验中，使用 AI-TOD 的训练集进行训练，使用测试集进行验证。为了验证 NSD 可以应用于任何的检测器并提高性能，选择单阶段的检测器（即 RetinaNet^[12]、ATSS^[13]）和多阶段的检测器（即 Faster R-CNN、Cascade R-CNN^[14]、DetectoRS^[15]）。同时为了确保提出的方法在最先进的网络中也有效果，选用目前最先进的小目标检测网络 DCFL^[16]，SAH^[17] 测试，测试结果如

表 2 AI-TOD 测试集上基线和 NSD（带*）的定量比较

Method	Backbone	AP	AP ₅₀	AP ₇₅	AP _{vt}	AP _t	AP _s	AP _m
RetinaNet ^[2]	ResNet-50	4.7	13.6	2.1	2.0	5.3	6.4	7.6
ATSS ^[3]	ResNet-50	12.9	30.5	8.4	2.0	11.0	19.5	29.0
FasterR-CNN ^[4]	ResNet-50	11.1	26.3	7.6	0.0	7.2	23.3	33.6
CascadeR-CNN ^[5]	ResNet-50	13.8	30.8	10.4	0.0	10.5	25.5	35.8
DetectoRS ^[6]	ResNet-50	14.8	32.8	14.3	0.0	10.5	28.1	38.1
DCFL ^[10]	ResNet-50	25.4	58.9	17.9	10.5	24.3	35.4	40.3
SAHI ^[11]	ResNet-50	23.6	55.6	16.3	9.5	22.5	33.7	39.6
Ours:								
RetinaNet*	ResNet-50	9.2	24.9	5.0	3.3	9.8	13.4	16.8
ATSS*	ResNet-50	13.7	33.5	8.9	2.3	11.6	21.2	31.5
FasterR-CNN*	ResNet-50	17.6	33.2	11.0	2.9	9.6	26.5	34.5
CascadeR-CNN*	ResNet-50	18.6	44.3	13.0	3.5	17.6	26.9	36.5
DetectoRS*	ResNet-50	20.6	49.0	14.5	6.5	19.5	29.6	38.3
DCFL*	ResNet-50	27.9	60.5	25.6	11.2	25.4	38.5	46.9
SAHI*	ResNet-50	28.6	55.9	26.4	10.6	26.5	41.5	47.5

注：最好的指标加粗。

表 2 所示。

从表 2 可以看到，Faster R-CNN、Cascade R-CNN、DetectoRS 在 AP_{vt} 指标上为 0。但是将 IoU 替换为 NSD 时，AP_{vt} 的指标都有了显著的增加，分别达到了 2.9、3.55、6.5。这表明 NSD 可以提升部分网络对微小目标的检测性能。DCFL 与 SAHI 都是目前最先进的微小目标检测网络，在加入 NSD 后，评价指标依然得到了提升，AP 提升了 2.5、5。在评价指标 AP_{vt} 上分别提升了 0.7 与 1.1。在实验中，DCFL 在加入 NSD 后在 AP₅₀ 指标上得到了最优值 60.5。DCFL 在添加 GDW 后在 AP₅₀，AP_{vt} 取得了最好的效果，分别为 60.5、11.2。SAHI 在加入 NSD 都在除 AP₅₀、AP_{vt} 上，其余指标上都获得最好的效果，AP 为 28.6，AP₇₅ 为 26.4，

AP_T 为 26.5, AP_s 为 41.5, AP_m 为 47.5。在 Faster R-CNN, Cascade R-CNN, DetectoRS 添加 NSD 后, AP_v 与 AP_t 的提升幅度都是比较大的, 但是 AP_s 与 AP_m 的提升幅度较小。这说明 NSD 专注于微小目标检测, 对于小目标检测具有良好的效果。

3 结语

在本文中, 观察到 IoU 指标在对于检测微小目标的检测器中存在问题, IoU 指标对于微小目标的位置偏移极其敏感, 这严重影响了检测器对于微小目标的检测, 使得检测器的性能的提高极其困难。为了解决这个问题, 提出了一种称为归一化 Sinkhorn 距离 (NSD) 的新度量来测量微小物体的边界框之间的相似性。将 NSD 嵌入到基于锚的检测器的标签分配、非极大值抑制和损失函数中以取代原始的 IoU 度量。实验结果表明, 所提方法可以大幅提高微小物体检测性能, 并在 AI-TOD 数据集上达到最先进的水平。

参考文献:

- [1] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//Computer Vision-ECCV 2014. Berlin:Springer, 2014:740-755.
- [2] EVERINGHAM M, ALI ESLAMI S M, GOOL V L, et al. The pascal visual object classes challenge: a retrospective[J]. International journal of computer vision, 2015, 111: 98-136.
- [3] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6): 1137-1149.
- [4] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 779-788.
- [5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Computer Vision – ECCV 2016. Berlin: Springer, 2016: 21-37.
- [6] WANG J W, YANG W, GUO H W, et al. Tiny object detection in aerial images[C]//2020 25th International conference on pattern recognition (ICPR). Piscataway: IEEE, 2021: 3791-3798.
- [7] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115: 211-252.
- [8] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.
- [9] CHEN K, WANG J Q, PANG J M, et al. MMDetection: open MMLab detection toolbox and benchmark[PP/OL]. (2019-06-17)[2026-04-15]. <https://doi.org/10.48550/arXiv.1906.07155>.
- [10] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 658-666.
- [11] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: faster and better learning for bounding box regression[C]//Proceedings of the AAAI conference on artificial intelligence. Alto: AAAI, 2020: 12993-13000.
- [12] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. Piscataway: IEEE, 2017: 2980-2988.
- [13] ZHANG S F, CHI C, YAO Y Q, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[PP/OL]. (2020-06-20)[2026-04-15]. <https://doi.org/10.48550/arXiv.1912.02424>.
- [14] CAI Z W, VASCONCELOS N. Cascade R-CNN: delving into high quality object detection[PP/OL]. (2017-12-03)[2026-04-15]. <https://doi.org/10.48550/arXiv.1712.00726>.
- [15] QIAO S Y, CHEN L C, YUILLE A. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. Piscataway: IEEE, 2021: 10213-10224.
- [16] XU C, DING J, WANG J W, et al. Dynamic coarse-to-fine learning for oriented tiny object detection[PP/OL]. (2023-04-18)[2026-04-15]. <https://doi.org/10.48550/arXiv.2304.08876>.
- [17] AKYON F C, ALTINUC S O, TEMIZEL A. Slicing aided hyper inference and fine-tuning for small object detection[PP/OL]. (2022-10-24)[2026-04-15]. <https://doi.org/10.48550/arXiv.2202.06934>.

【作者简介】

李文生 (1992—), 男, 河南平顶山人, 硕士, 研究方向: 计算机应用。

(收稿日期: 2025-09-15 修回日期: 2026-04-28)