改进 Cascade R-CNN 的舰船检测模型

欧奕昱[1,2] 左育莘[2] 杨锐[2] 李秀[1]

摘要 本文以 Cascade R-CNN 为基本架构，先采用 HTC 网络在 COCO 数据集上预训练，再将参数迁移学习至本模型中，然后使用多尺度训练(SNIP)的方法，并逐步融合 soft-NMS、Res2Net、可变卷积网络(DCN)、Adam 优化算法和学习率热身等技巧提高了检测器的检测效果。

关键词 Cascade R-CNN, HTC, SNIP

The ship detection model of Improved Cascade R-CNN

Yimin Ou[1,2] Yushen Zuo[2] Rui Yang[2] Xiu Li[1]

Abstract This paper uses Cascade R-CNN as the basic architecture. Firstly, the HTC network was pre-trained on the COCO data set, and then the pre-trained parameters were transferred into our model. By the multi-scale training SNIP, we gradually integrated some techniques, such as soft-NMS, Res2Net, Variable Convolutional Network (DCN), Adam optimization algorithm and learning rate warm-up, into our model so that the detection effect could be improved distinctively.

Key words Cascade R-CNN, HTC, SNIP

1 引言

近年来人工智能发展迅速，并应用到了各个领域，极大地满足了人民的美好生活需要。将人工智能与海洋技术相互融合，实现海洋感知，加速海洋技术智能化，是坚持陆海统筹，发展海洋强国的必然要求。随着科技的进步，计算机硬件设备的性能提升，深度学习崭露锋芒。在目标检测领域，深度学习技术更是大放异彩。基于深度学习的目标检测方法主要分为两大类。第一类是基于候选区域的多阶段目标检测算法，如R-CNN[1]、FsterRCNN[2] 、SPP-Net[3]、Cascade RCNN[7]等。两阶段主要为候选区域生成和使用神经网络对目标进行分类和定位，这种方式精度较高，但是速度偏慢。第二类是SSD[4]、YOLO[5]等单阶段目标检测算法，此类算法主要通过回归模型直接对目标进行类别预测和位置回归，虽然速度有了极大地提升，但是检测的精度下降严重。

1. 中国科学院自动化研究所高技术创新中心 北京 100080 2. 中国科学院自动化研究所模式识别国家重点实验室 北京 100080 3. 中国科学院自动化研究所《国际自动化与计算杂志》编辑部 北京 100080 4. 中国科学院自动化研究所《自动化学报》编辑部 北京 100080

1. Hi-Tech Innovation Centre, Institute of Automation, Chinese Academy of Sciences, Beijing 100080 2. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080 3. Editorial Office of *International Journal of Automation and Computing*, Institute of Automation, Chinese Academy of Sciences, Beijing 100080 4. Editorial Office of *Acta Automatica Sinica*, Institute of Automation, Chinese Academy of Sciences, Beijing 100080

1 相关工作

1) 用混合任务级联网络（HTC）在COCO数据集上进行预训练，并将学好的模型迁移到本次比赛所使用的模型中。
2) 使用Albu数据增强算法[6]对给定的海洋舰船数据集进行扩充。
3) 分别使用ResNet101和Res2Net101两种主干网络（backbone）对给训练集采用交叉折验证的方法进行训练和验证。

2 模型和方法描述

本次比赛我们主要使用多阶段级联RCNN（Cascade RCNN），该网络是一种基于Faster RCNN的双阶段检测模型。我们首先使用HTC算法模型进行预训练，HTC模型可以看成是Cascade RCNN模型再添加一个分割的分支，因此我们用该模型在COCO数据集上进行检测加分割的多任务训练，使得网络能够提取丰富的语义特征。其次，我们将HTC模型中的主干网络架构从ResNet换成Res2Net，并使用Res2Net-101网络，来提升模型的性能。然后，我们将上一步得到的参数迁移到Cascade R-CNN架构上，使用舰船数据集进行训练。此外，我们还使用了Soft-NMS，Albu数据增强算法，多尺度训练与测试SNIP，DCN（可变形卷积网络）等方法来提升检测性能。同时，我们还使用了N-fold策略来保证模型的鲁棒性。其详细的描述如下：

2.1 区域提议网络

2.1.1 RPN 与多阶段级联

经典的Faster RCNN框架使用区域提议网络（RPN）生成检测框来实现目标定位，极大的提升了检测速度[2]。区域提议网络(RPN)的基本原理为：使用多尺度方法，为ResNe101卷积网络输出特征图的每个像素点生成9个候选框（anchor），将这些候选框作为初始的检测框，然后使用分类器提取出正候选框（positive anchor），再使用回归网络对正候选框进行微调，接着根据分类器的分数选择一定数量的候选框，并用图像边界限定候选框范围、剔除尺寸小的候选框，最后对他们使用非极大值抑制（NMS），输出剩余的候选区域（proposal）。训练阶段，根据人为设定的交并比（IoU）赋予这些候选区域正负标签。为了减小计算量，从最后的选出的候选区域中以一定的比例采样出固定数量的候选区域作为感兴趣区域（RoI）。测试阶段，RPN网络根据分类器的分数选择与训练阶段相同的候选区域作为RoI。最终将RPN网络得到的RoI送入RoI池化层变换到统一大小，并使用主干网络进行分类与回归得到结果。

因为单个RPN网络训练阶段和测试阶段存在

候选区域的质量与分布不匹配问题。通常训练阶段因为使用了IoU区分正负候选框，使提议区域的质量较测试阶段提议区域的质量高，所以模型会出现过拟合现象。为了解决这种不匹配问题，我们使用了流行的多阶段级联法(Cascade R-CNN)[7]，其原理如图1所示。在此级联结构中设置IoU值逐步递增，逐步选择更佳的区域。本次阶段目标的回归依赖于上一阶段回归的结果，几个级联的阶段之间逐步微调，候选区域的IoU值被逐步提高，使测试阶段的数据集与IoU值高的检测器之间不会出现严重的不匹配问题。

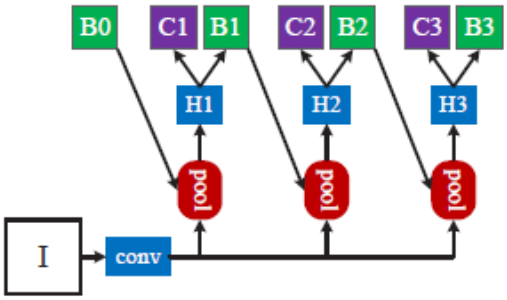

图1. Cascade-RCNN示意图[7]

2.1.2 soft-NMS

我们在筛选重叠候选框时采用了改进方法，未采用非极大值抑制算法（NMS），而是使用了soft-NMS算法。采用非极大值抑制这种方法删除有重叠的候选框时，可能会删除必要的候选框。假阳性（FP）最有可能出现在与得分最高的框重叠区域较多的检测框中，所以为了降低假阳性出现的几率，不盲目的删除所有IoU大于阈值的框，我们采用了soft-NMS算法。此算法通过降低重叠区域候选框的置信度得分来筛选区域[8]。如果一个检测框与得分最高的检测框有大部分重叠，它会获得很低的置信度，如果小部分重叠，那么它的置信度不会受到很大的影响。这种做法不会影响到AP的值。NMS与 soft-NMS算法的效果对比如图2所示。可以看到NMS算法遗漏了一只长颈鹿。采用该方法后，将可以有效提升检测结果。

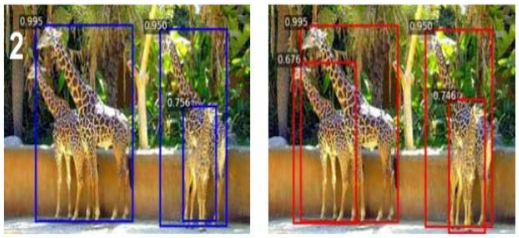

图2. NMS效果图(左) 与 soft-NMS效果图(右)的对比图[8]

2.2 卷积神经网络

2.2.1 Res2Net101

我们模型的主干网络采用Res2Net101，它是对ResNet101的改进。相比于普通网络ResNet101在每三层之间增加了短路机制，形成了残差学习[9]，其中BN和ReLU都提前的单层方式是ResNet残差单元常采用的方式[10]。卷积网络可以通过残差学习无损的传播梯度，这种方式解决了深度网络梯度消失或梯度爆炸问题，使得网络基本不在出现退化现象。这也是ResNet优于VGG16等主干网络的主要原因。如图3所示，Res2Net通过在单个残差块内构造分层的残差类连接实现了以更细粒度（granular level）表示多尺度特征的功能，并增加每个网络层的感受野（receptive fields）范围[11]。这种方式可以增加卷积网络学习的信息量，能明显的提高模型的分类效果。

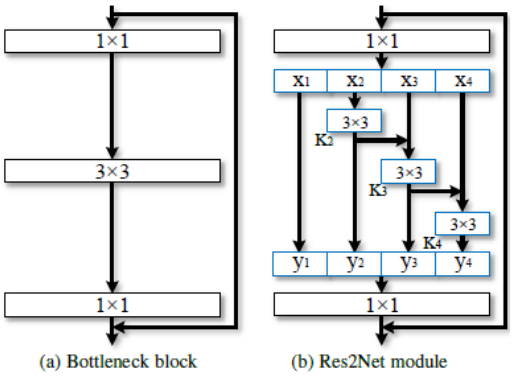

Fig. 2: Comparison between the bottleneck block and the proposed Res2Net module (the scale dimension $s = 4$).

图3. Res2Net分层残差连接示意图[11]

2.2.2 可变卷积网络 DCN

由于视角的变化，同样的物体在图像中可能发生刚体甚至非刚体形变，从而呈现出不同的大小与姿态。当大量这样的数据出现在测试数据集中时，模型的泛化性能会降低，所以我们使用了具有学习空间几何形变能力的可变卷积网络（DCN）[12]，它的做法是为每个采样点增加一个偏置，使采样点发生偏移。可变形卷积核的大小和位置可以根据当前需要识别的图像内容进行动态调整，从而改变了感受野。如图4所示，其直观效果就是不同位置卷积核的采样点位置会根据图像内容发生自适应变化，从而契合不同物体的形状、大小等几何形变。采用可变卷积网络之后，可以有效的提高模型精度。

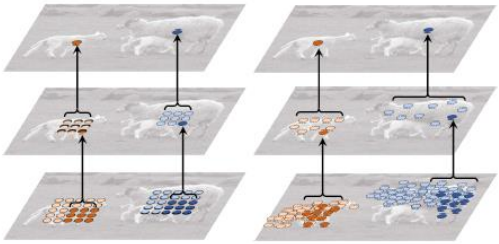

图 4.　可变卷积示意图[12]

2. 2. 3 Adam 优化算法

深度学习领域优化算法的选择会很大程度上影响模型的效果，采用不同的优化算法，可能会导致截然不同的效果。Adam优化算法结合了AdaGrad和RMSProp两种优化算法的优点，综合考虑了梯度的一阶矩和二阶矩，通过自适应学习率计算出更新步长[13]。其实现简单，计算高效，可以应用于大规模数据和参数场景，对陌生数据集有较好的适应能力，所以在我们采用了这种优化算法。

2.3　训练方法

2. 3. 1 HTC 预训练

为了让卷积网络提取的特征包含更好的语义信息来表征目标，我们使用了HTC算法模型先进行预训练，其示意图如图5所示。HTC模型采用了一种多任务多阶段的混合级联结构，并将语义分割模块引入到整体框架中[14]。并且还在每一个相邻阶段的分割分支上增加了一条连接，这个连接实现了对各阶段逐渐调整和增强。另外，因为分割是在像素级别上对全图进行精细的分类，所以分割后的特征具有很强的空间位置信息，同时对前景和背景有较强的分辨能力，它的引入能进一步提升检测效果。

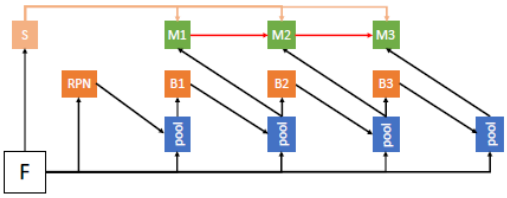

图5.　HTC框架示意图[14]

2. 3. 2 基于图像金字塔的尺度归一化方法

在anchor机制中，anchor的大小和长宽比是固定的，因此一些尺度很大或很小的物体没办法被分类到前景中。这种现象会影响检测器的效果，所以我们使用了基于图像金字塔的尺度归一化方法SNIP[15]进行多尺度训练和多尺度测试，以便提高检测器的精度。SNIP的原理如图6所示，训练和测试时，将图片缩放至不同的分辨率，需要检测的目标便也同时被缩放，它们将会以不同的比例出现在检测器中。虽然单个检测器在高分辨率时不容易检测到大目标，在低分辨率时不容易检测到小目标，但是将上述缩放的图片送入并行的检测器后，这些大目标和小目标总可以被检测到。

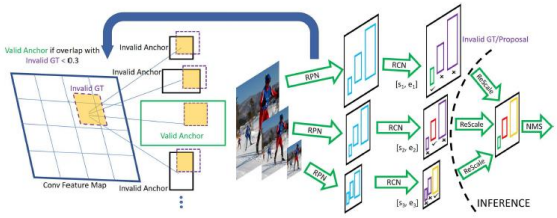

图6.　SNIP原理图[15]

2. 3. 3 学习率热身（warm up）

根据文献知道，在刚开始训练的时候，所有的参数都是随机值，离最终的结果偏离比较大，如果直接使用较大的学习率可能会造成数值不稳定[16]。因此可以先用一个较小的学习率，为模型热身，当训练过程稳定后再调回学习率，以便让模型的训练过程更稳定。

3　实验结果

首先我们采用soft-NMS改进的Cascade RCNN网络，然后通过上文所述方法逐步改进模型。采用多尺度训练并逐步改善模型的过程中所得的结果如表1所示。

表1. 逐步改善的模型训练结果表

模型	mAP(%)	训练时间(epoch)
Cascade RCNN	64.68	12
Cascade RCNN+Albu	64.32	12
Cascade RCNN+Albu+DCN	66.36	12
Cascade RCNN+Albu+DCN+Res2Net	67.09	20

mAP及训练时间变化过程如图8所示。

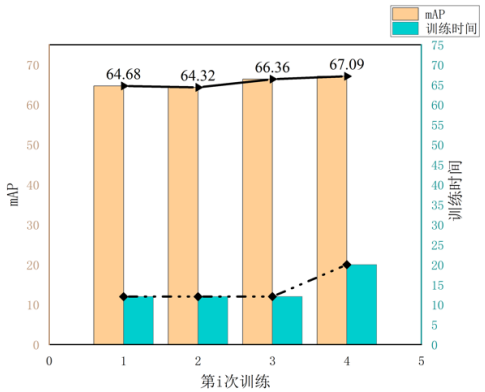

图7.　逐步改善的模型训练结果变化图

图7中1、2、3和4次训练所采用的模型为表1中1、2、3和4行标注的模型。从图中可以看到采用了Albu数据增强后，mAP几乎没有下降。这说明了数据集中的图像的对比度等图像性质变化较大。另

外，在采用了可变卷积之后，训练过程的mAP提升了2个点，说明数据集中的数据由于拍摄角度等问题存在形变。

为了保证在未知数据集上有较好的表现，我们采用交叉折验证方法分别检测主干网络不同时模型在数据集上的鲁棒性，其结果如图8所示。

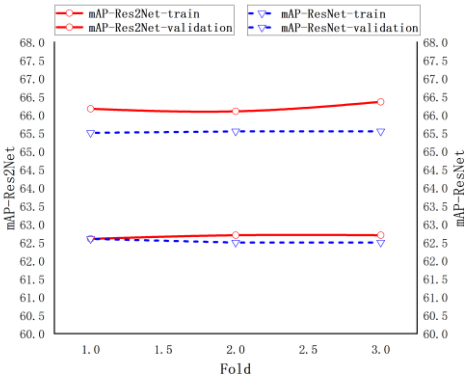

图8. mAP值随Fold的变化图

当主干网络为Res2Net101时，检测器的mAP值变化如图中红色线条所示，主干网络是ResNet101时，检测器的mAP值变化如图中蓝色线条所示，可以看出Res2Net101的检测表现较好，故最终模型的主干网络采用Res2Net101。但是只有训练集和测试集的数据分布相似时模型才能较好的工作，所以我们分别用测试集在主干网络为ResNet101和Res2Net101的模型中进行测试。ResNet101的测试mAP值为66.36%，Res2Net101测试的mAP值为67.09%。其差值和两模型在训练时mAP的差值接近，说明训练集和测试集数据分布相似。

4 结论

本模型参考(SNIP)多尺度训练和多尺度测试并逐步增加技巧，提升算法的精度。得出如下结论:

1) 在本数据集上使用Albu方法进行数据增强，对模型的效果提升不多，但是可以增加鲁棒性。
2) 在本模型上使用可变卷积DCN对平均精度的均值mAP提升较多，为2.04%，而且训练时间并没有增加，均为12epoch。
3) 本模型使用Res2Net101主干网络后，精度也有一定的提升，同时我们设置更长的训练时长（20epoch）来得到更好的模型。
4) 通过对数据集做了分析和切分，做了可靠的对比实验。
5) 因为有可靠的数据集切分，所以我们可以在本地做大量测试，不必理会上传次数和上传时间。

References

1. Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580-587.
2. Girshick R. Fast R-CNN[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1440-1448.
3. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2014, 37(9): 1904-1916.
4. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C] //Proceedings of the European Conference on Computer Vision, 2016: 21-37.
5. Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 779-788Wang F Y, Fundamental Issues in Research of Computing with Words and Linguistic Dynamic Systems. *Acta Automatica Sinica* (Periodical style), 2005, **31**(6): 844—852
6. Buslaev A, Iglovikov V I, Khvedchenya E, et al. Albumentations: fast and flexible image augmentations[J]. Information, 2020, 11(2): 125.
7. Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6154-6162.
8. Bodla N, Singh B, Chellappa R, et al. Improving object detection with one line of code. CoRR (2017)[J]. arXiv preprint arXiv:1704.04503.
9. He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.
10. He K , Zhang X , Ren S , et al. Identity Mappings in Deep Residual Networks[J]. 2016.
11. Gao S , Cheng M M , Zhao K , et al. Res2Net: A New Multi-scale Backbone Architecture[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, PP(99):1-1.
12. Dai J , Qi H , Xiong Y , et al. Deformable Convolutional Networks[J]. 2017.
13. Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
14. Chen K, Pang J, Wang J, et al. Hybrid task cascade for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2019: 4974-4983.

15. Singh B, Davis L S. An analysis of scale invariance in object detection snip[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3578-3587.
16. He T , Zhang Z , Zhang H , et al. Bag of Tricks for Image Classification with Convolutional Neural Networks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019.

作者一 欧奕旻，硕士研究生，主要研究方向为深度学习和目标检测。E-mail: oyml9@mails.tsinghua.edu.cn

(**FIRST Author-Yimin Ou**, enrolled postgraduate, the main research area is deep learning and target detection.)