基于注意力特征金字塔的轻量级目标检测算法

赵义飞，王　勇
（北京工业大学 信息学部，北京 100124）

摘　要： 基于深度学习的目标检测算法因其模型复杂度和对计算能力的要求，难以部署在移动设备等低算力平台上。为了降低模型的规模，提出一种轻量级目标检测算法。该算法在自顶向下的特征融合的基础之上，通过添加注意力机制构建特征金字塔网络，以达到更细粒度的特征表达能力。该模型以分辨率为 320×320 的图像作为输入，浮点运算量只有 0.72 B，并在 VOC 数据集上取得了 74.2%的 mAP，达到了与传统单阶段目标检测算法相似的精度。实验数据表明，该算法在保持了检测精度的同时显著降低了模型运算量，更适合低算力条件下的目标检测。

关键词： 目标检测；特征金字塔；注意力机制；轻量级算法
中图分类号： TN98；TP391　　**文献标识码：** A　　DOI：10.16157/j.issn.0258-7998.211320

中文引用格式： 赵义飞，王勇. 基于注意力特征金字塔的轻量级目标检测算法[J].电子技术应用，2021，47(10)：33-37.
英文引用格式： Zhao Yifei，Wang Yong. Lightweight object detection algorithm based on attention feature pyramid network[J]. Application of Electronic Technique，2021，47(10)：33-37.

Lightweight object detection algorithm based on attention feature pyramid network

Zhao Yifei，Wang Yong
（Faculty of Information Technology，Beijing University of Technology，Beijing 100124，China）

Abstract： Object detection algorithms based on deep learning are difficult to deploy on low computing power platforms such as mobile devices due to their complexity and computational demands. In order to reduce the scale of the model，this paper proposed a lightweight object detection algorithm. Based on the top-down feature fusion，the algorithm built a feature pyramid network by adding an attention mechanism to achieve more fine-grained feature expression capabilities. The proposed model took an image with a resolution of 320×320 as input and had only 0.72 B FLOPs，achieved 74.2% mAP on the VOC dataset and the accuracy is similar to traditional one-stage object detection algorithms. Experimental data shows that the algorithm significantly reduces the computational complexity of the model，maintains the accuracy，and is more suitable for object detection with low computing power.

Key words： object detection；feature pyramid；attention mechanism；lightweight algorithm

0 引言

目标检测是计算机视觉的关键组成部分之一，旨在探索统一框架下人类视觉认知过程的模拟和行人检测、人脸识别、文本检测等特定应用场景下视觉任务的完成。2012 年，Krizhevsky 等[1]提出的 AlexNet 将卷积神经网络应用在了图像分类算法之中并取得了惊人的效果，从此基于深度学习的卷积神经网络算法开始取代传统的基于人工特征的算法，成为了计算机视觉领域的主流研究方向。

目前基于深度学习的目标检测算法可分为单阶段检测算法和两阶段检测算法两类。单阶段目标检测算法以 SSD[2]和 Yolo[3-5]系列算法为代表，是一种通过在卷积神经网络提取的特征图上设置锚点，并对每个锚点上预设的不同大小和长宽比例的边界框进行检测的方法。两阶段目标检测算法以 RCNN[6-8]系列算法为代表，先在特征图上采用额外步骤生成候选区域，再对候选区域进行检测。与单阶段算法相比，两阶段算法一般拥有更高的检测精度，但由于增加了额外的运算量，检测速度也相对较低。

基于深度学习的目标检测算法拥有很好的性能，同时也有着更高的模型复杂度和计算量。实际应用中，由于功耗、成本等限制，这些模型是难以直接部署的。为了使目标检测任务能在移动设备等难以提供高额算力的硬件上更好地完成，本文提出了一种基于轻量级网络的检测算法，并通过引入注意力机制使得原始特征图具有更细粒度的特征表达能力，从而获得更好的检测效果。

1 特征金字塔网络

基于深度学习的目标检测算法中，SSD[2]算法通过多级检测的方式在不同尺度的原始特征图上各自进行预测，而 FPN[9]则进一步通过一种自顶向下的简单的特征融合方式从原始特征图中重构出特征金字塔，将网络浅层的强位置信息与深层的强语义信息相结合。PANet[10]

在 FPN 的基础上添加了自底向上路径聚合的结构，利用网络浅层的位置信息增强了整个特征层次，缩短了浅层与深层特征之间的信息路径。EfficientDet[11]提出了一种新型的 BiFPN 结构，在对 PANet 进行了简化之后，增加了 shortcut 结构并引入了加权策略，通过对原有特征图整体赋予不同权重，更灵活地实现了多尺度特征融合。DetectoRS[12]则通过将原始 FPN 融合后的输出作为输入重新返回到模型中再次进行计算的方式实现了一种 Recursion-FPN 的结构，并在目标检测、实例分割等多个领域达到了最高精度。

2 本文方法

选择低计算量的神经网络作为特征提取骨架，可以显著提升算法的检测速度，降低算法部署的成本需求。本文以 Cream[13]分类模型为基础，采用单阶段的方式进行检测。Cream 模型的主体结构由 NAS(Neural Architecture Search)获得，相比人工设计的网络，具有更好的特征提取性能和更少的浮点运算次数。

2.1 注意力特征金字塔网络

文献[9]通过采用最邻近插值法将深层尺寸较小的特征图放大至与相邻的前一层相等的方式将深层特征图融合进浅层达到重建特征图的目的，这种方法简化了计算，但不具备学习能力。文献[11]通过对不同层的特征图整体赋予权重的方式，实现了一种更为灵活的加权特征融合 BiFPN。文献[14]对卷积神经网络中特征通道之间的相互依赖关系进行显式建模，采用了一种全新的特征重标定的策略，将注意力机制引入到了计算机视觉分类算法当中。这种注意力机制使得模型能够通过学习的方式自动获取不同特征通道的重要程度，并相应地对来自不同通道的特征进行抑制或者提升，从而将更有效的特征向后传递，达到提升算法精度的目的。受文献[11]和[14]启发，本文提出了注意力特征金字塔网络(Attention-Feature Pyramid Network，AFPN)，通过在通道维度进行重新标定的方式对原始特征图对特征金字塔的贡献进行更为细粒度的建模。以 Cream604 为例，本文将 Cream604 中下采样倍数为 8 倍、16 倍、32 倍的特征图分别记为 P8、P16 和 P32，并额外添加一个 64 倍的下采样层，特征图记为 P64。模型的整体结构如图 1 所示。

由于不同层的特征图在分辨率和通道数上都不一致，在进行特征融合时，要对参与构建特征金字塔的特征图进行调整。文献[11]指出不同的输入特征图对特征金字塔的贡献是不均等的，并通过添加权重系数来表示原始特征图对其所参与构建的特征金字塔层的贡献，通过分别将与原始特征图相邻的下一层和上一层特征图按照先后次序融入的方式构建新的特征图。AFPN 同样考虑了输入特征图对特征金字塔的贡献不同，采取了一种更为细化的特征融合的方式。AFPN 以低层的特征图为主体，采用自顶向下的方式将深层的特征图向浅层融合，将深层富含的语义信息与浅层的位置信息相结合，以求得到更好的检测效果。为了简化计算，AFPN 采取最邻近插值的方式对深层特征图进行上采样。由于通过插值进行上采样的方式不需要进行计算，因而没有可供进行学习的参数。AFPN 首先在上采样特征图后面添加了额外的卷积模块，增加了上采样特征图的特征表达能力。随后，AFPN 采用注意力机制对低层的特征图和经过上采样的高层特征图进行建模，让网络通过学习的方式自动获取原始特征图在通道维度对新特征的重要程度。如图 2 所示，以大小为 96×20×20 的特征图 P16 为例，经 AFPN 中的特征融合得到 P16 out 的过程可以用如下公式描述：

$$P16\ out = MBConvSE(P16) + MBConvSE(Upsample(P32)) \quad (1)$$

其中，Upsample 表示最邻近插值上采样操作，这部分没有参数，不具备学习能力；P32 的大小为 320×10×10，经过上采样后变为 320×20×20。MBConvSE 表示融合通道注意力机制的转置瓶颈[15]结构，具体如图 3 所示。其中 Conv 代表卷积运算(Convolution)，BN 代表批量归一化操作(Batch Normalization)，SE 代表基于通道的注意力模块(Squeeze-and-Excitation block)，而 SE 模块中的 Global Pooling 则仍然采用全局平均池化(Global Average Pooling)的方式进行运算。文献[15]指出，通过将特征图扩展至高维再降回低维的方式可以减少 ReLU 激活函数对特征图中所包含信息的损耗。AFPN 中的转置瓶颈模块选用

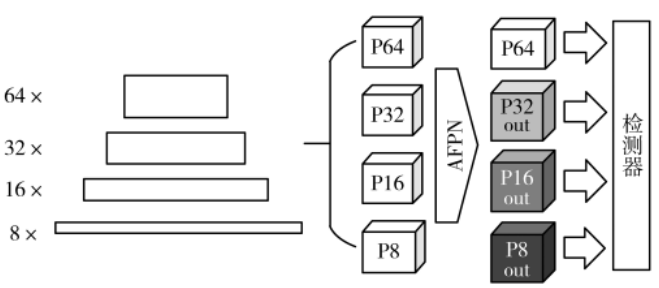

图 1 目标检测模型整体结构

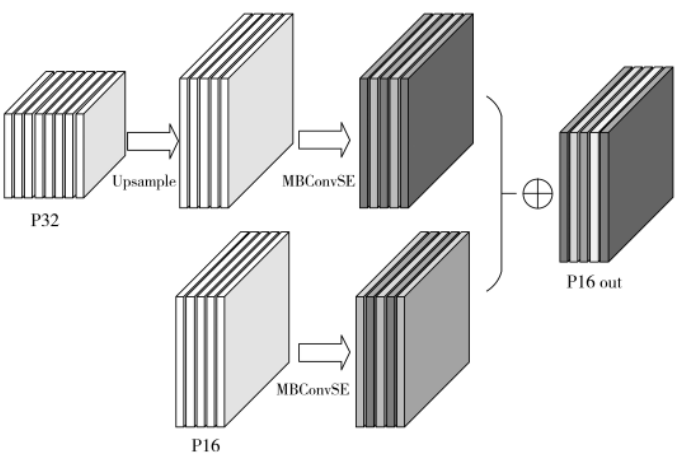

图 2 注意力特征金字塔网络结构

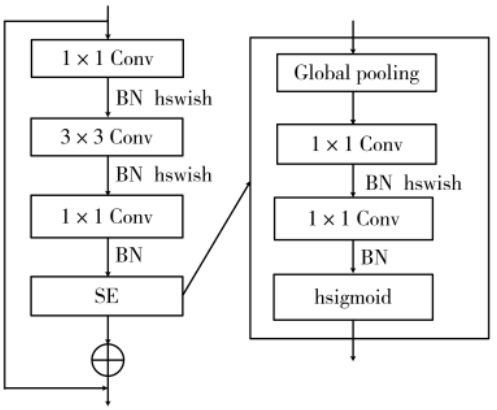

图 3 MBConvSE 模块

了更适合低算力设备的 hswish 激活函数，对于 P32，首先通过 1×1 卷积将特征图扩张至 640×20×20，再通过 3×3 卷积提取特征，最后通过 1×1 卷积降维至与 P16 相同的 96×20×20 以便进行特征融合；对于 P16，则先扩张为 320×20×20 再调整回 96×20×20。对于调整后的两组特征图，采用 element-wise add 而不是 concatenate 的方式以避免计算量的进一步增加。AFPN 中注意力机制的实现也针对计算量进行了优化，具体为将文献[14]中的 SE module 中的两个全连接层替换为 1×1 卷积并保持通道数不变，并将最后的 sigmoid 函数替换为 hsigmoid 激活函数。此外注意力机制的实现被直接融合进了转置瓶颈结构之中，与文献[14]中的标准 SE module 相比，这种方式可以在保持相似性能的同时进一步降低运算量。

2.2 AFPN 损失函数计算

AFPN 采用传统的单阶段目标检测算法的检测模式，即对特征图上产生的每个预测框进行检测。由于目标检测模型的训练一般是在包含多个类的数据集上进行的，因此损失函数的计算要在所有类上进行。记 x_i^p 用于表示第 i 个先验框与数据集中物体所处的真正位置在类别 p 上的匹配程度，当两者的交并比大于预设值时，认为该预测框为正样本，即 $i \in \text{Positive}$，此时 $x_i^p=1$；否则计为负样本，$x_i^p=0$，即 $i \in \text{Negative}$。文献[16]指出，正负样本比例失衡是导致单阶段目标检测算法精度较低的重要原因之一。采用难例挖掘的方式对整幅样本比例进行控制，通过将负样本的数量控制在正样本 3 倍左右，可以简单而有效地提高训练的稳定性。最终得到的 AFPN 的损失函数 Loss 由定位损失 L_l 和类别置信度损失 L_c 共同组成：

$$\text{Loss}=\frac{1}{N}(L_l(x_i^p,l,g)+L_c(x_i^p,c)) \tag{2}$$

其中，N 为数据集的总类别数；定位损失 $L_l(x_i^p,l,g)$ 由 smooth L_1 损失函数计算得到，具体计算方式如下：

$$L_l(x_i^p,l,g)=\sum_{i\in\text{Positive}}^{N}\sum_{m\in\{x,y,h,w\}}x_i^p\ \text{smooth}\ L_1(l_i^m-g^m) \tag{3}$$

其中，$x_i^p\in\{0,1\}$，l_i^m 和 g^m 分别代表对于物体 m 预测框给定的位置和数据集中所标注的目标真实位置，x、y、h、w 分别代表位置的中心坐标和坐标框的高宽，smooth L_1 的计算方式为：

$$\text{smooth}\ L_1=\begin{cases}0.5x^2, & |x|<1\\ |x|-0.5, & x<-1\ \text{or}\ x>1\end{cases} \tag{4}$$

类别置信度 $L_c(x_i^p,c)$ 由 Softmax 损失函数计算得到：

$$L_c(x_i^p,c)=-\sum_{i\in\text{Positive}}^{N}x_i^p\log(\hat{c}_i^p)-\sum_{i\in\text{Negative}}\log(\hat{c}_i^0) \tag{5}$$

其中，$x_i^p\in\{0,1\}$，$\hat{c}_i^0$ 代表被模型预测为不属于任何类，即认为是背景的概率；$\hat{c}_i^p$ 由 softmax 函数对计算得到：

$$\hat{c}_i^p=\frac{e^{c_i^p}}{\sum_p e^{c_i^p}} \tag{6}$$

其中，c_i^p 为模型对物体在每一类上所预测的概率。$\hat{c}_i^0$ 和 c_i^p 均由模型通过向前计算得到。

3 实验

3.1 数据集及评价指标

Pascal VOC 数据集是计算机视觉领域的经典数据集，共包含了人、猫、公共汽车等在内的 20 类物体。其中 VOC07 数据集共有 9 963 张图片，共包含了 24 640 个带标注的物体；VOC12 数据集中训练集和验证集共有 11 540 张图片，共包含了 27 450 个带标注的物体。将 VOC07 和 VOC12 数据集中的训练集和验证集作为总训练集，VOC07 数据集中的测试集作为总测试集，已成为目标检测算法经典的训练和评估方法。

由于目标检测问题本质上仍然是二分类问题，因此可将样例按所数据集中所标注的真实类别与模型所预测类别的组合划分为真正例(True Positive，TP)、假正例(False Positive，FP)、真反例(True Negative，TN)和假反例(False Negative，FN)4 种情况，所有情况组成的混淆矩阵如表 1 所示。

表 1 预测结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

查准率(Precision)和查全率(Recall)可由表中的混淆矩阵计算得出，二者分别定义为：

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}=\frac{\text{TP}}{\text{all detections}} \tag{7}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}=\frac{\text{TP}}{\text{all ground truths}} \tag{8}$$

其中，all detections 表示所有预测框的数量，all ground truths 表示数据集中所有标注物体的数量。对于 VOC 数据集，首先计算所有不同的 Recall 值，然后将每个 Recall 值对应的大于等于该 Recall 的最大 Precision 值进行求和平均，即得到了 AP(Average Precision)值。对数据集中包含的所有类的 AP 值取平均数即得到了 mAP(mean Average Precision)。mAP 作为最常用的检测指标之一，被广泛应用于图像分类、目标检测等计算机视觉领域。

3.2 数据预处理

文献[2]指出，通过数据增强的方式对数据集中的图像进行预处理，可以有效提升模型的检测性能。本文采取了随机裁剪、镜像翻转、图像扩张、随机添加噪声、随机调整亮度等方式对 VOC 数据集中的图像进行了增强处理，以提高模型的鲁棒性。将经过数据增强处理的图像缩放到 320×320 像素，并将图像对应的标注信息进行相应调整，即得到了模型输入值。

3.3 模型超参数设置

本文采用 mini-batch 梯度下降法，设置 batch size 为 32，并使用 momentum 优化器对梯度下降过程进行优化，momentum 系数设置为 0.9，L2 正则化系数设置为 5×10^{-4}。模型共在 VOC 数据集上进行 120 000 次迭代，初始学习率设置为 1×10^{-3}，并采用动态调整学习率的方式，具体为在迭代进行到 80 000 次和 100 000 次时将学习率调整为 1×10^{-4} 和 1×10^{-5}，从而更助于找到模型最优解。

3.4 实验结果

表 2 比较了本文所设计的模型与目前部分主流算法在 VOC 数据集上的表现。

表 2 不同算法在 VOC 数据集上的表现

算法	图像大小	mAP/%	Flops/B
SSD300[2]	300×300	74.3	15
YOLO V3[5]	448×448	76.9	65.8
FasterRCNN+Resnet[8]	600×1 024	79.8	112
AFPN	320×320	74.2	0.72
Tiny YOLOv3[5]	416×416	58.4	5.52
YOLO Nano[17]	416×416	69.1	4.57

可以看出，对于复杂检测模型，以 SSD300 为例，本文设计的算法在保持了与其相似输入图像尺寸和精度的同时，将浮点运算量降低到了 1/20 以下，部分图像的检测结果对比如图 4 所示。对于其他小型检测模型，以 YOLO Nano 为例，本文的算法依然占有精度和运算量的优势。

4 结论

本文针对传统目标检测算法模型复杂度高、计算量大的问题，通过对特征融合方式进行研究，提出了一种基于注意力特征融合的轻量级检测算法。实验结果证明了本文的算法的有效性，对于移动端等低算力平台，本文的算法具有更强的适用性。

图 4 AFPN(左)和 SSD300(右)检测结果对比

参考文献

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems, 2012: 1097-1105.

[2] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//European Conference on Computer Vision. Springer, 2016: 21-37.

[3] 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: 779-788.

[4] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7263-7271.

[5] REDMON J, FARHADI A. Yolov3: an incremental improvement[J]. arXiv preprint arXiv: 1804.02767, 2018.

[6] 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.

[7] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[8] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems, 2015:

91-99.
[9] LIN T,DOLLAR P,GIRSHICK R,et al.Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017.
[10] LIU S,QI L,QIN H,et al.Path aggregation network for instance segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018.
[11] TAN M,PANG R,LE Q.EfficientDet:scalable and efficient object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2020.
[12] QIAO S,CHEN L,YUILLE A.DetectoRS::detecting objects with recursive feature pyramid and switchable atrous convolution[J].arXiv preprint arXiv:2006.02334,2020.
[13] PENG H,DU H,YU H,et al.Cream of the crop:distilling prioritized paths for one-shot neural architecture search[J].arXiv preprint arXiv:2010.15821,2010.
[14] HU J,SHEN L,SUN G,et al.Squeeze-and-excitation networks[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,42(8):2011-2023.
[15] SANDLER M,HOWARD A,ZHU M,et al.Mobilenetv2:inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018:4510-4520.
[16] LIN T,GOYAL P,GIRSHICK R,et al.Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision,2017:2980-2988.
[17] WONG A,FAMUORI M,SHAFIEE M,et al.YOLO nano:a highly compact you only look once convolutional neural network for object detection[J].arXiv preprint arXiv:1910.01271,2019.

(收稿日期:2021-01-24)

作者简介:
赵义飞(1995-),通信作者,男,硕士研究生,主要研究方向:深度学习、目标检测,E-mail:723935706@qq.com。
王勇(1974-),男,博士,副教授,主要研究方向:并行与分布式计算。

扫码下载电子文档

(上接第28页)

与实现[D].北京:北京工业大学,2018.
[7] AI and compute[R].OpenAI,2018.05.16.
[8] 盎司财经.区块链搭建"生产关系"新模式,较人工智能更具"革命性"[EB/OL].(2018-03-1 3)[2020-12-15].http://www.sohu.com/a/22543 2740_99987131.
[9] 潘吉飞,黄德才.区块链技术对人工智能的影响[J].计算机科学,2018,45(11A):53-37.
[10] Bytom WhitePaper V1.0[R].BYTOM,2017,06.
[11] 赵泓维.对于医疗人工智能企业算力问题,英伟达打出关键一招[EB/OL].(2019-01-29)[2021-01-19].http://www.chidaolian.com/article-23081-1.
[12] 中国区块链技术和应用发展白皮书(2016)[R].中国区块链技术和产业发展论坛,2016.
[13] 链门户.什么是智能合约?智能合约真的智能吗?[EB/OL].(2018-08-14)[2021-01-20].http://www.lianmenhu.com/blockchain-5574-6.
[14] 李庆华.智慧合约——智能合约安全问题的AI解决方案[EB/OL].(2018-05-04)[2021-01-14].https://cloud.tencent.com/developer/news/202428.
[15] 区块链头条.JarvisPlus创始人兼CEO吴骞:让每个人可以用自然语言来使用区块链和智能合约[EB/OL].(2018-07-30)[2021-01-16].https://www.sohu.com/a/24424 7557_100112552.

(收稿日期:2021-02-22)

作者简介:
张伟娜(1985-),通信作者,女,硕士,工程师,主要研究方向:人工智能技术应用、产业发展应用,E-mail:zhwn0704@163.com。
黄蕾(1975-),女,博士,高级工程师,主要研究方向:人工智能产业发展应用。
张箴(1985-),男,本科,初级工程师,主要研究方向:人工智能产业发展应用。

扫码下载电子文档

(上接第32页)

现初步研究[J].图书馆杂志,2018,37(11):90-98.
[12] 吉久明,施陈炜,李楠,等.基于GloVe词向量的"技术——应用"发现研究[J].现代情报,2019,39(4):13-22.
[13] FARMER W J,RIX A J.Evaluating power system network inertia using spectral clustering to define local area stability[J].International Journal of Electrical Power and Energy Systems,2022,134(3):107404.
[14] 耿丽君.韵律形态学研究综述[J].成都理工大学学报(社会科学版),2020,28(1):98-104.
[15] 杨飘,董文永.基于BERT嵌入的中文命名实体识别方法[J].计算机工程,2020,46(4):40-45.

(收稿日期:2021-03-01)

作者简介:
杨政(1987-),男,硕士研究生,高级工程师,主要研究方向:电力文本分析与应用、电网数字化以及网络安全。
尹春林(1991-),男,助理工程师,主要研究方向:自然语言处理、迁移学习。
李慧斌(1984-),男,博士,副教授,主要研究方向:计算机视觉、图像处理与模式识别、深度学习、自然语言处理。

扫码下载电子文档

版权声明