计算机科学
COMPUTER SCIENCE

基于改进Yolov4-tiny的轻量型目标检测算法

窦智, 胡晨光, 梁竞一, 郑李明, 刘国奇

引用本文

窦智, 胡晨光, 梁竞一, 郑李明, 刘国奇基于改进Yolov4-tiny的轻量型目标检测算法[J]. 计算机科学, 2023, 50(6A): 220700006-7.

DOU Zhi, HU Chenguang, LIANG Jingyi, ZHENG Liming, LIU Guoqi. Lightweight Target Detection Algorithm Based on Improved Yolov4-tiny [J]. Computer Science, 2023, 50(6A): 220700006-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700006

基于改进 Yolov4-tiny 的轻量型目标检测算法

窦 智[1] 胡晨光[1] 梁竞一[1] 郑李明[2] 刘国奇[1]
1 河南师范大学计算机与信息工程学院 河南 新乡 453007
2 金陵科技学院机电工程学院 南京 211169

摘 要 面向视频的深度学习算法运算复杂度较高,难以满足实时性要求,严重影响了其在边缘计算和实时系统中的应用。轻量化网络成为了研究热点之一,针对大型网络的轻量化网络显著降低了原网络的参数规模,提升了检测速度,但检测精度难以满足工业需求。针对上述问题,文中提出了一种改进的目标检测轻量化网络,在保持小参数规模的前提下,有效提高了检测性能。文中在 YOLOv4-tiny 骨干网络中添加 VIT(Vision Transformer)结构,利用多头自注意力机制使网络可以提取更深层次的物体特征;使用简化后的 Bi-FPN,将两检测通道改为三检测通道,增加注意力融合机制,提高模型对图片特征的利用率,提高网络对不同尺寸大小目标的检测精度;使用 Ghost 卷积替换传统卷积操作,降低网络计算复杂度,减少网络参数。在 COCO 数据集上进行实验,实验结果表明,在保持网络规模不变的情况下,改进后的算法相比 YOLOv4-tiny 原网络检测精度取得了明显提升,可同时满足边缘计算及实时系统对深度网络轻量化和准确度的要求。

关键词:目标检测;轻量化网络;多头自注意力机制;加权特征融合
中图法分类号 TP391

Lightweight Target Detection Algorithm Based on Improved Yolov4-tiny

DOU Zhi[1],HU Chenguang[1],LIANG Jingyi[1],ZHENG Liming[2] and LIU Guoqi[1]
1 School of Computer and Information Engineering,Henan Normal University,Xinxiang,Henan 453007,China
2 School of Mechanical and Electrical Engineering,Jinling University of Science and Technology,Nanjing 211169,China

Abstract Video-oriented deep learning algorithms have high computational complexity and are difficult to meet real-time requirements,which seriously affects their applications in edge computing and real-time systems. Lightweight networks have become one of the research hotspots. Lightweight networks for large networks significantly reduce the scale of the original network parameters and improve the detection speed,but the detection accuracy is had to meet industrial needs. In view of the above problems,this paper proposes an improved lightweight target detection network,which can effectively improve the detection performance while maintaining a small parameter scale. In this paper,the vision transformer(VIT) structure is added to the YOLOv4-tiny backbone network,and the multi-head self-attention mechanism enables the network to extract deeper object features. Using the simplified Bi-FPN,the two detection channels are changed to three detection channels,and the attention mechanism is introduced in the feature map fusion node to improve the model's utilization of image features and the network's detection accuracy for objects of different sizes. Using Ghost convolution to replace traditional convolution operations,so as to reduce network computational complexity and network parameters. Experimental results on the COCO dataset show that the improved algorithm has significantly improved the detection accuracy of the original YOLOv4-tiny network while keeping the network scale unchanged,it can simultaneously meet the requirements of edge computing and real-time systems for the lightweight and accuracy of deep networks.

Keywords Object detection,Lightweight network,Multi-head self-attention mechanism,Weighted feature fusion

1 引言

基于深度学习的目标检测方法主要分为两类。一类是基于回归的一阶段目标检测,以 yolo 系列[1]和 SSD[5]为代表。此类算法通过密集采样,直接在图像上生成若干候选框,对这些候选框的类别概率、框位置信息进行回归预测并生成预测框。另一类是基于候选区域的二阶段目标检测,以 Faster R-CNN[6]和 Mask R-CNN[7]为代表。

此类算法先找到一定数量的候选框,然后对各个候选框进行定位与分类。一阶段目标检测相比二阶段目标检测有较低的时间复杂度,但精确度较低,对于简单场景下的目标检测,通常选用一阶段目标检测算法。

YOLOv4(You Only Look Once version 4)算法是在 YOLOv3 的基础上改进的一阶段目标检测算法,该方法的检测性能与同时期的目标检测算法相比有较为明显的优势,YOLOv4 的主干提取特征网络使用 CSPDarknet53 与 PAN-

基金项目:国家自然科学基金(U1904123,61901160)
This work was supported by the National Natural Science Foundation of China(U1904123,61901160).
通信作者:窦智(2015160@htu.edu.cn)

et[8]结构，检测精度高，但对硬件配置要求较高，检测速度较慢，在移动端和工业设备上部署的成本较高。YOLOv4-tiny[9]作为 YOLOv4 的简化版，以精简的网络结构换取了较高的检测速度和可移植性，是深度学习应用于智能边缘计算领域的有效解决方案，但 YOLOv4-tiny 提取图像中目标特征的能力有所欠缺，检测精度远低于 YOLOv4，亟需在保证轻量化的前提下进一步提高性能。

为提高轻量化网络的检测性能，文献[9]在 YOLOv4-tiny 骨干网络中添加 Max Module 结构，以获取更多的局部特征，并构建自上而下的多尺度融合网络，以提高网络的检测精度。文献[11]使用 Mish 激活函数替换原网络的 Leaky Relu 激活函数，并添加空间金字塔池化层，提高了网络对目标尺寸大小的敏感度。文献[12]提出了一种 ECSP Block 增强模块，增强网络的学习能力，用卷积层替换最大池化层，降低下采样操作引起的细节特征丢失，并构建双通道特征融合金字塔网络，进一步提升了网络性能。文献[13]在 YOLOv4-tiny 网络的基础上添加空间金字塔池化(Spatial Pyramid Pooling，SPP)模块，融合了图像的局部和全局特征，增强了网络的准确定位能力，并在 YOLOv4-tiny 原网络的 3 个最大池化层和新增 SPP 模块后各添加一个 1|1 的卷积模块，减少了网络的参数，提高了网络的运算速度。文献[14]先将 ResNet-D 模块和调整后的 Res-CBAM 融入 YOLOv4-tiny 模块的骨干网络中，并将 YOLOv4-tiny 骨干网络中的 CSPOSANet 模块替换为 ResNet-D 模块，以减少模型所需的计算量(Floating Point Operations，FLOPs)；然后，将调整后的 Res-CBAM 特征融合方式替换为通道内堆叠模式，实现辅助分类器功能；最后，使用 5 种不同感受尺度的特征进行预测，并通过合并预测框来优化结果的显示。与原始的 YOLOv4-tiny 模型相比，该方法的检测速度得到了一定的提升。此外，文献[15-18]也基于 YOLOv4-tiny 网络进行了一定的改进，改进方法在某些特定的应用领域优于原网络。上述方法虽然对网络性能做了不同的改进，但在不同的应用场景下，其优化的效果不稳定。本文综合考虑 YOLOv4-tiny 在目标检测领域的优化方式，提出了一种改进的 YOLOv4-tiny 目标检测算法，并采用公开数据集 COCO2017 作为测试用例进行了详细的对比实验，实验结果验证了本文方法的优越性。为解决轻量化网络在具体应用场景中检测精度较低的问题，该文主要做了以下工作：

(1)在 YOLOv4-tiny 骨干网络中增加 VIT(Vision Transformer)[19]结构，有效改善了原网络提取深层图像特征能力欠缺的问题。

(2)根据 EfficientDet[20]的 Bi-FPN(Weighted Bi-directional Feature Pyramid Network)结构特点，构建自上而下的多尺度融合网络，在特征融合网络中加入自注意力机制，解决了原网络特征利用率不足的问题。

(3)使用 Ghost[21]模块替换传统卷积操作，降低计算冗余度，进一步优化网络体积规模，在保证轻量化的前提下有效提升了网络性能，降低了深度网络在嵌入式平台上的部署难度和成本。

2 YOLOv4-tiny

与 YOLOv4 不同，YOLOv4-tiny 简化了网络结构，降低了计算量，提升了推理速度，可有效应对实时处理的需求。YOLOv4-tiny 的网络模型如图 1 所示，主要由 Backbone、Neck 和 Predict 这 3 部分组成。

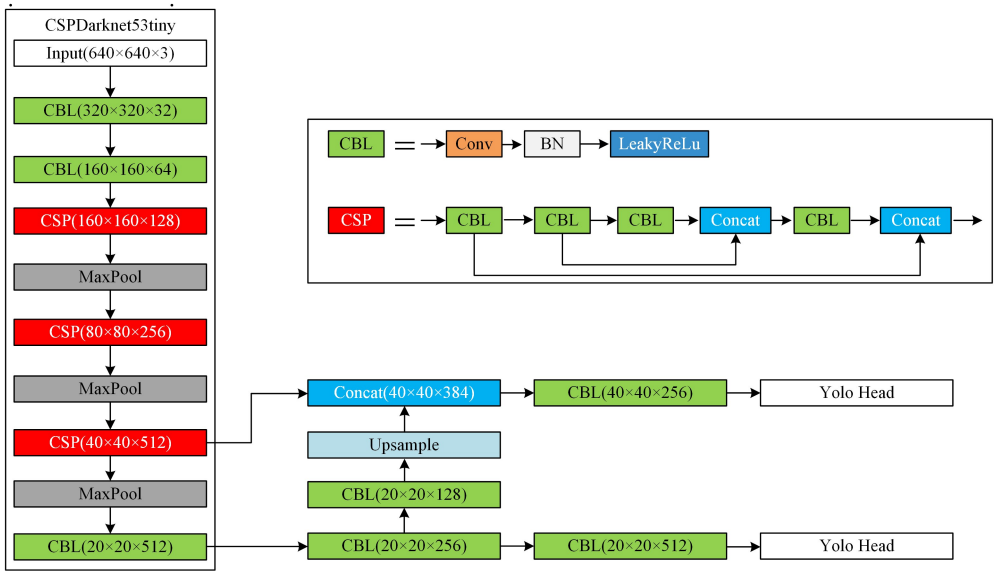

图 1 YOLOv4-tiny 的网络结构图

Fig. 1 Network structure diagram of YOLOv4-tiny

Backbone 部分采用 CSPDarknet53-Tiny 网络，Neck 部分采用特征金字塔网络 FPN，Predict 部分输出 2 个有效特征层 Yolo Head。YOLOv4-tiny 使用 CSPDarknet53tiny 作为骨干网络，由 CBL 层(Conv+BN+Leaky Relu)、CSP 层(Cross Stage Partial)和最大池化层(Max Pooling)构成。骨干网络前两个 CBL 层是由大小为 3 * 3、步长为 2 的卷积核、批标准化处理 BN(Batch normalization)和 Leaky Relu 激活函数组成，其主要作用是对特征图进行下采样操作。骨干网络末端 CBL 由大小为 3 * 3、步长为 1 的卷积核组成，有助于加强网络的特征学习能力。骨干网络中间堆叠了 3 次 CSP 和池化核大小为 2 * 2、步长为 2 的最大池化层，其作用是降低特征图的规模，从而减少网络的计算量。CSP 是由多个 CBL 组成的残差模块，可有效地降低网络的学习成本，增强学习能力，减少梯度消失和梯度爆炸所

带来的负面影响，增强网络的泛化能力。在骨干网络之后的 Neck 部分采用 FPN 网络，将 13 * 13 大小的特征层经过上采样操作与 26 * 26 大小的特征层进行融合，从而起到特征融合的效果。FPN 网络会输出两个不同大小的特征图，分别用于检测大目标与小目标，输出的特征图经过 Yolo Head 解码后，可得到每个锚框的位置信息和类别概率。

3　YOLO-TBG 算法

本文在 YOLOv4-tiny 网络的基础上，提出了 3 种改进的网络模型（YOLO-T，YOLO-TB，YOLO- TBG）。YOLO-T 在 YOLOv4-tiny 的基础上，在骨有自注意力机制和 CNN 捕捉卷积窗口内的局部信息不同，它具有全局性，利用注意力来捕获全局上下文信息之间的相关性，训练出的特征图上的每个点与其他各个点都会产生联系。本文通过在骨干网络中加入 VIT 结构来增加网络训练的深度，在第二个 CSP 节点后添加 VIT 的目的是为了使 FPN 的每个通道都能享有 VIT 所带来的性能提升。

在复杂场景下，检测目标的尺寸大小的不一，YOLOv4-tiny 网络模型只能输出两种大小的特征层，尺度的单一性很容易造成漏检。为了解决此类问题，我们提出了 YOLO-TB 网络模型。在 YOLO-T 的基础上进行进一步改进，借鉴 Bi-FPN 模型的特点，删除原有 FPN 只有单一输入通道的节点，使原有的两通道结构修改为三通道结构，添加上采样和下采样操作，对不同特征层进行特征融合，在每个融合节点处设置权重因子，使得网络融合具有注意力机制。改进后的 YOLO-TB 网络模型参数和计算量明显增加，为了在保留高精度的同时降低网络复杂度，我们使用 Ghost 模块替换传统卷积模块，最终得到 YOLO-TBG 网络，网络结构如图 2 所示。

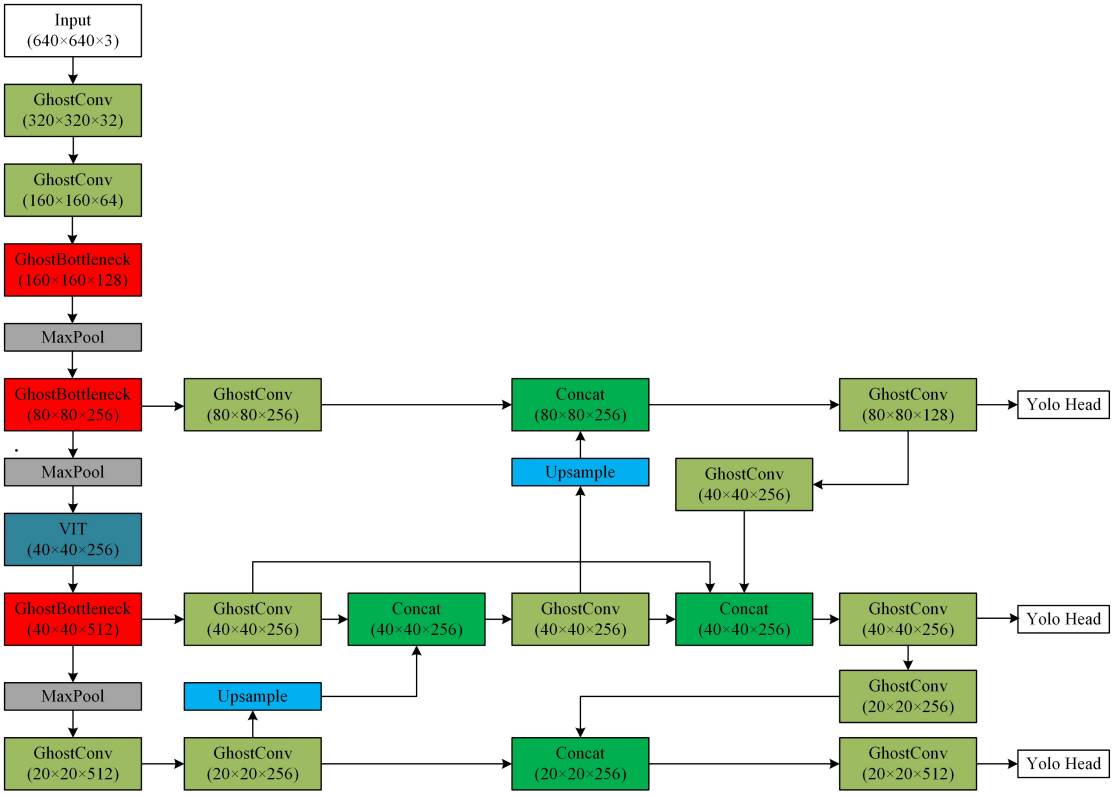

图 2　YOLO-TBG 网络结构图

Fig. 2　YOLO-TBG network structure diagram

3.1　自注意力机制模块 VIT

我们在 YOLOv4-tiny 网络的基础上添加了 VIT（Vision Transformer）结构，提出了 YOLO-T 网络模型。VIT 首先会将图像均分，然后对均分后的图像进行展平操作，将其处理为向量并添加一个可训练的类别信息，之后通过线性映射获得其位置编码信息，将图像信息与位置信息进行嵌入，利用 N 个 Transformer 编码器[22]对嵌入位置信息后的图像块进行处理，对最后一个编码器输出的特征向量进行变形操作，从而得到 VIT 处理后的特征矩阵，过程如图 3 所示。

Transformer 编码器要求输入是 token（向量）序列，即二维矩阵[num_token，token_dim]，num_token 对应的是向量的个数，token_dim 对应每个向量的长度。对于图像数据而言，其数据格式是三维矩阵，需要先将数据变换为二维矩阵形式。例如输入特征图 $X\in R^{H\times W\times C}$，$H$ 和 W 表示图像的高度和宽度，$\boldsymbol{C}$ 表示特征图的通道数，其为三维矩阵，需要对特征图进行分割并将其展平为二维矩阵，分割后的图像表示 $X\in R^{N\times(P^2\cdot C)}$，$N=HW/P^2$，$N$ 表示分割后的特征图的个数，P 表示为特征图的分辨率。利用可训练的矩阵 $\boldsymbol{E}$ 对分割后的图像 X 进行线性映射，映射为长度为 D 的向量，并连接可训练的类别信息 $X_{\text{class}}\in R^{1\times D}$。将连接类别信息后的向量通过线性映射，得到位置编码 E_{pos}，它表示被切割后的图像块在原图像中的空间位置信息，将两矩阵进行融合得到矩阵 $Z_0\in R^{(N+1)\times D}$，如式(1)所示。

$$Z_0=[X_{\text{class}};X_1E;X_2E;\cdots;X_NE]+E_{pos},E\in R^{(P^2\cdot C)\times D},\quad E_{pos}\in R^{(N+1)\times D} \tag{1}$$

将得到的Z_0送入 VIT 的编码器(Encoder)中,编码器是由层归一化(Layer Norm,LN)、多头注意力机制(Multiheaded Self-Attention,MSA)以及多层感知器(Multilayer Perceptron,MLP)组成。如图 4 所示,编码器分为两层,分别为多头注意力机制和多层感知器,每一层在输入前都会做一次层归一化操作,对输入和输出做残差连接,其中多层感知器 MLP 包含两层线性映射层和两层 Dropout 层[23]以及 GELU 激活函数[24]。

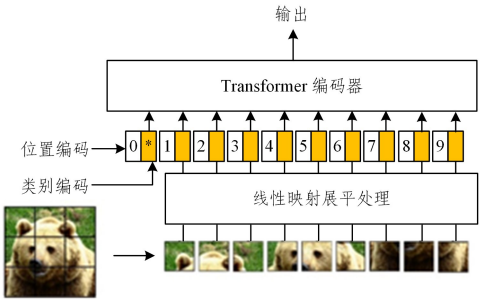

图 3 VIT 结构图

Fig. 3 VIT structure diagram

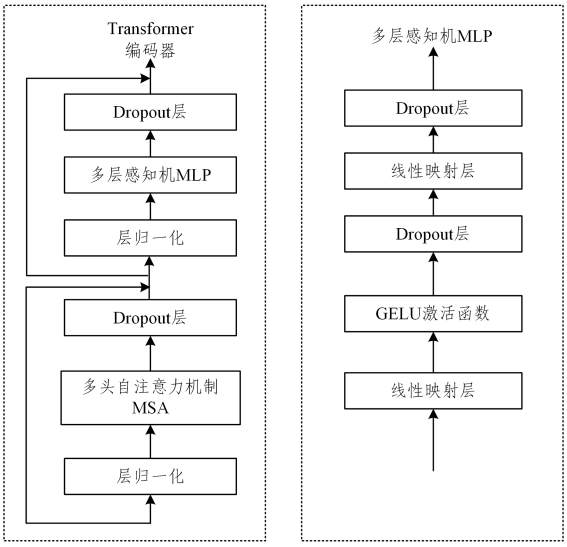

图 4 Transformer 编码器和多层感知机结构图

Fig. 4 Diagram of Transformer encoder and multilayer perceptron structure

对于第 L 层 transformer 编码器,其输入记为Z_{l-1},输出的结果记为Z_l,将其变形为输入 VIT 之前的图像尺寸,变形后的特征图即 VIT 最终提取到的图像特征,如式(2)、式(3)所示。

$$Z_l'=MSA(LN(Z_{l-1}))+Z_{l-1},l=1,\cdots,L \tag{2}$$

$$Z_l=MLP(LN(Z_l'))+Z,l=1,\cdots,L \tag{3}$$

MSA 处理Z_0之前,需经过一次层归一化处理,之后使用 3 个可训练矩阵$\boldsymbol{W}_q$,$\boldsymbol{W}_k$,$\boldsymbol{W}_v$对Z_0进行处理,得到对应的 q,k,v 并将其分为 h 份,h 为设置的头的数量;然后对得到的 Q 与 K 进行点积操作,并将点击结果除以$\sqrt{d}$,(其中,d 为 Q,K,V 的维度);最后将 softmax 函数与 V 相乘,即可得到每个头对应的输出特征,如式(4)所示:

$$\begin{aligned}head\,d_i&=\mathrm{Attention}(ZW_q,ZW_k,ZW_v)\\&=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right)V,W_q,W_k,W_v\in R^{D\times D}\end{aligned} \tag{4}$$

将每个头得到的特征向量进行融合,并通过一个训练矩阵W^o,将所得的特征图重塑为输入 VIT 之前的尺寸,即可得到 VIT 输出后的特征图,如式(5)所示:

$$MultiHead(Q,K,V)=Concat(head\,d_1,\cdots,head\,d_h)W^o,\ W^o\in R^{D\times D} \tag{5}$$

YOLO-T 的骨干网络中拥有 VIT 模块,利用 VIT 模块中的多头自注意力机制,弥补了一般的卷积操作在特征挖掘不充分的缺陷,使得网络在特征挖掘方面表现得更加优秀。

3.2 改进的 Bi-FPN

YOLO-T 网络虽然在特征挖掘上性能有所提升,但在经过骨干网络提取特征后,依然只能输出 40 * 40,20 * 20 两个尺度的特征图,在复杂场景下,检测目标的尺寸通常是大小不一的,面对这种应用场景,网络很容易造成漏检,满足不了现实需求。为了解决上述问题,我们将改进的 Bi-FPN(Bidirectional Feature Pyramid Network)添加到 YOLO-T 网络中,提出了 YOLO-TB 网络模型。在骨干网络第二个 CSP 结构引出一条通道,将原本的两通道改为三通道,这 3 个通道的尺度大小为 80 * 80,40 * 40,20 * 20,分别检测大、中、小目标,极大程度地避免了原网络出现漏检的现象。

Bi-FPN 是一种加权双向(自顶向下+自低向上)特征金字塔网络,它允许简单和快速的多尺度特征融合。与 YOLOv4 网络所使用的 PANet(Path Aggregation Network)结构相比,PANet 是在 FPN(Feature Pyramid Networks)的基础上增加了一条自顶向上的通道,融合了顶层向底层和底层向顶层的双向特征融合,参数量较大。而 Bi-FPN 不像 PANet 那样只有一次的顶层向底层和底层向顶层的特征融合,Bi-FPN 进行了多次的这样的特征融合,并且进行了多次的堆叠,在融合后的特征的基础上还进行了原始对应层的特征的叠加。同时在结构上 Bi-FPN 删除了只有一条输入边或输出边的节点,简化了双向网络。其次如果原始输入节点和输出节点处于同一层,则在原始输入节点和输出节点之间添加一条额外的边,目的在于在不增加太多成本的情况下融合更多的特性。图 5 是 PANet 和 Bi-FPN 的结构图。

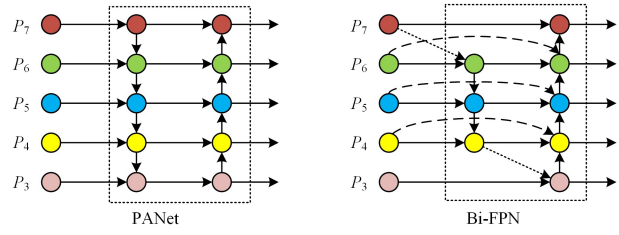

图 5 PANet 和 Bi-FPN 的结构图

Fig. 5 Structure diagram of PANet and Bi-FPN

以往的特征融合方式是直接在特征图通道上堆叠不同尺寸特征,并未考虑不同的输入特征图对最后输出特征的贡献程度是不同的。对于此问题,Bi-FPN 引入一个可训练的权重因子,来调节各个不同输入对输出的特征图贡献程度。并且 Bi-FPN 还采用了快速归一化的融合策略(Fast Normalized Fusion),相对基于 softmax 方式的融合策略(softmax-based Fusion),此方法训练稳定,训练速度提高了近 30%,且两者的效果相当。快速归一化的融合策略如式(6)所示:

$$O=\sum_i\frac{w_i}{\varepsilon+\sum_j w_j}\cdot I_i \tag{6}$$

其中，O 表示输出的特征图，i 和 j 表示输入特征融合节点的特征图数，$i=j$，I_i 为输入的特征图；ε 为防止分母为 0 的常数，值为 0.0001；w_i 和 w_j 表示输入节点的特征图权重，权重的初始值范围为 0 到 1 之间的随机数。在网络训练阶段，每一次训练时权重都将通过 ReLU(Rectified Linear Unit)激活函数来确保数值大于 0。在多次训练后，每个特征融合节点输入的特征图将获得当前检测算法学习到的最优权重。

原始的 Bi-FPN 是将特征图连续 5 次进行下采样操作，在每次下采样后将得到的特征图输入到特征融合网络进行融合。我们删除 Bi-FPN 中的 P7 和 P6 这两个通道，只对 P3，P4，P5 这 3 个通道进行特征融合，删除 P5 层输入与输出的残差连接，同时在多个融合节点添加注意力机制。此注意力机制是将每个输入融合节点的特征图乘上一个可训练的归一化权重因子 w，修改后的 Bi-FPN 结构如图 6 所示。

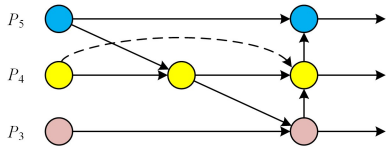

图 6　修改后的 Bi-FPN 结构图

Fig. 6　Modified Bi-FPN structure diagram

P4 通道的融合过程如式(7)、式(8)所示：

$$P_4^{td}=Conv\left(\frac{w_1\cdot P_4^{in}+w_2\cdot Resize(P_5^{in})}{w_1+w_2+\varepsilon}\right) \tag{7}$$

$$P_4^{out}=Conv\left(\frac{w_1\cdot P_4^{in}+w_2'\cdot P_4^{td}+w_3'\cdot Resize(P_3^{out})}{w_1'+w_2'+w_3'+\varepsilon}\right) \tag{8}$$

其中，P_4^{td} 是 P4 通道的中间特征，P_5^{in} 是输入进 P5 通道的特征，P_4^{out} 是输出 P4 通道的特征，P_3^{out} 是输出 P3 通道的特征。Conv 是卷积操作，Resize 是上采样或者下采样操作。P_5^{in} 经过上采样操作与 P_4^{in} 进行融合，再经过卷积操作得到 P_4^{td}。P_3^{out} 经过下采样操作与 P_4^{in} 和 P_4^{td} 进行融合，然后经过卷积操作之后得到 P_4^{out}。融合点处各个特征图所乘的权重因子 w 的初始值是 0 到 1 的一个可训练的随机数。ε 的数值为 0.0001，是为了防止分母为 0 而设置的。式(7)和式(8)的分子中加号代表不同特征图的堆叠操作。修改后的 Bi-FPN 结构将分别输出 3 种不同规格的检测头，分别是 20 * 20，40 * 40，80 * 80 这 3 个检测头，最后再用 YOLOv4-tiny 的解码算法生成最后的预测框。

改进后的 YOLO-TB 网络模型与原网络相比在检测不同大小目标上表现得更加优异，极大地避免了漏检问题。

3.3　Ghost 模块

改进后的 YOLO-TB 网络模型的参数和计算量明显增加，为在保留高精度的同时降低网络复杂度，我们在 YOLO-TB 网络模型的基础上使用 Ghost 模块替换传统卷积模块，提出了 YOLO-TBG 网络。

传统卷积操作输出的特征图包含丰富的甚至是冗余的图像特征，旨在达到对原始图像特征全面提取的目的，显然，该过程中会存在大量的冗余计算。针对减少传统卷积操作冗余计算的问题，华为诺亚方舟实验室提出 Ghost 轻量级卷积模块以及在小型卷积网络上使用的 Ghost Bottleneck 的方法。在进行传统卷积操作时，给定输入特征图 $X\in R^{c\times h\times w}$，其中 c 为特征图的通道数，h 和 w 分别为特征图的高和宽。传统卷积操作得到特征图的过程如式(9)所示：

$$Y=X*f+b \tag{9}$$

其中，$*$ 表示卷积操作，$f\in R^{c\times k\times k\times n}$ 表示卷积核，c 和 k 分别表示卷积核的通道数和尺寸，n 表示卷积核的数量，b 表示偏置项，$Y\in R^{h'\times w'\times n'}$ 为输出的特征图。根据以上信息，可以计算出传统卷积操作所需的浮点运算规模为 $n\times h'\times w'\times c\times k\times k$。

为减少计算冗余，Ghost 模块先用普通卷积操作获得通道数为 m 的特征图，将得到的特征图通过简单的线性运算生成通道数为 s 的特征图，然后将两次得到的特征图进行叠加得到最终的特征图，其中 $n=m\times s$，目的是为了确保普通卷积操作与 Ghost 模块输出的特征图大小一致。Ghost 模块的具体操作如式(10)、式(11)所示：

$$Y'=X*f' \tag{10}$$

$$y_{ij}=\Phi_{i,j}(y_i'),\forall i=1,\cdots,m,\forall j=1,\cdots,s \tag{11}$$

式(10)中，特征图 X 通过卷积核 $f'\in R^{c*k*k*m}$ 得到通道数为 m 的特征图 $Y'\in R^{h'*w'*m}$，然后将 Y' 上每一个通道的特征图经过 s 个线性操作生成与映射数量对等的特征图，接着将每个通道输出的特征图进行堆叠，对 Y' 进行叠加，输出最终的特征图。式(11)表示第 i 通道的特征图 y_i' 通过第 j 个线性操作 $\varphi_{i,j}$ 得到特征图 y_{ij}。在第 s 个线性运算时，会进行恒等映射，用于保留原始特征。图 7 是 Ghost 模块示意图。

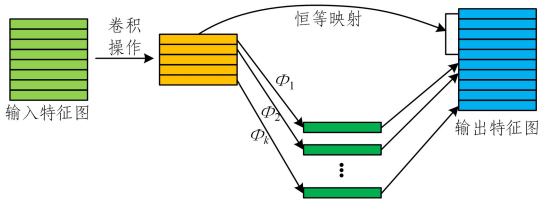

图 7　Ghost 模块示意图

Fig. 7　Ghost module diagram

Ghost 模块中线性操作采用 $d*d$ 大小的卷积核，传统卷积与 Ghost 模块的浮点计算量之比如式(12)、式(13)所示：

$$r_s=\frac{n*h'*w'*c*k*k}{\frac{n}{s}*h'*w'*c*k*k+(s-1)*\frac{n}{s}*h'*w'*d*d} \tag{12}$$

参数量之比为：

$$r_s=\frac{n*c*k*k}{\frac{n}{s}*c*k*k+(s-1)*\frac{n}{s}*d*d} \tag{13}$$

当 d 与 k 近似，且 $s\ll c$ 时，计算量与参数量之比可约等于 s。

Ghost bottleneck 分为步长为 1 和步长为 2 的两种结构，Yolov4tiny 骨干网络中的 CSP 是步长为 1 的结构。Ghost bottleneck 步长为 1 的结构采用两个 Ghost 模块相连，并在每个模块后增加一个批量归一化层，两个 Ghost 中间添加 Relu 激活函数，在输入和输出部位添加残差结构，图 8 给出了 Ghost bottleneck 结构。为减少网络计算量以及模型参数，并且提升网络速度，我们使用步长为 1 的 Ghost bottleneck 替换 yolov4-tiny 网络中的 CSP 结构。

在网络规模不变的情况下，改进后的 YOLO-TBG 相比 YOLOv4-tiny 原网络检测精度取得了明显提升，可同时满足边缘计算及实时系统对深度网络轻量化和准确度的要求。

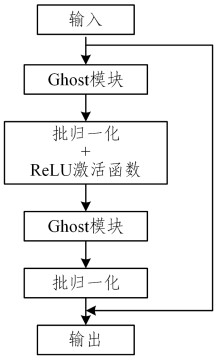

图 8 Ghost bottleneck 结构图

Fig. 8 Ghost bottleneck structure diagram

4 实验结果与分析

4.1 实验数据以及环境配置介绍

本文实验所采用的数据集是微软公司发布的 MS COCO2017 数据集，其中训练集有 118 287 张，验证集有 5 000 张，共计 123 287 张，包含 80 个种类，所有图片已经人工标注好真实目标框的位置、大小以及种类信息。为方便进行测试，本实验采用的测试集为 COCO 数据集的验证集。

实验所使用的操作系统为 Linux 系统，GPU 采用 GeForce RTX 2080 Ti 的 11GB 显卡，CPU 为 Intel(R) Xeon(R) E5-2620 V4 2.10 GHz，编译软件为 Pycharm。实验中设置总训练轮次为 100，批量大小为 8，初始学习率为 0.01，衰退率为 0.0005，输入的图片尺寸大小统一设置为 640 * 640。实验将会选取各个模型训练后得到的损失最小的权重文件进行对比，对比项为各个模型 IOU 阈值为 50 的平均精度 AP_{50}、不同目标尺寸下的平均精度、模型参数量、检测速度 FPS。

4.2 消融实验及结果分析

本文进行了消融实验来验证的各种改进方法对 YOLOv4-tiny 原模型的优化效果，实验对比了 YOLOv4-tiny 原网络以及改进后得到的 YOLO-T、YOLO-TB 和 YOLO-TBG 网络，实验结果如表 1、表 2 所列。

表 1 改进的 YOLO 网络模型与其他网络的性能对比

Table 1 Performance comparison between the improved YOLO network model and other networks

方法	mAP_{50}/%	召回率/%	参数量	FPS
YOLOv3	64.1	62.0	61 922 845	56.8
YOLOv3-tiny	34.6	36.7	8 849 182	200.0
YOLOv4-tiny	37.2	40.7	6 056 606	158.7
YOLO-T	42.1	43.6	6 500 638	102.0
YOLO-TB	51.1	52.6	10 517 521	64.1
YOLO-TBG	48.6	49.6	5 980 801	72.4

表 2 原网络与改进网络对不同尺寸目标的检测精度

Table 2 Detection accuracy of the original network and the improved network for objects of different sizes

方法	mAP_{small}/%	mAP_{median}/%	mAP_{large}/%
YOLOv4-tiny	11.2	26.4	23.1
YOLO-T	13.0	30.2	27.5
YOLO-TB	16.9	36.0	41.2
YOLO-TBG	15.4	32.9	38.4

从实验结果可以看出，YOLOv3-tiny 与 YOLOv4-tiny 具有相似的特点，即网络的参数量较少，模型结构较为简单，检测速度较快，但精度较低，提取较深层次的特征能力较弱。使用 VIT 对网络骨干进行改造得到 YOLO-T，网络参数相比 YOLOv4-tiny 原网络略有增加，FPS 降低了 56.7，但 mAP_{50} 提升了 4.9%，且对不同目标尺寸下的精度都有所增加，由此证明了 VIT 结构的有效性。YOLO-TB 的 mAP_{50} 为 51.1%，相比 YOLO-T 的 mAP_{50} 提升了 9%，且不同目标尺寸下的平均精度都有明显增加，这表明 Bi-FPN 改进后的 YOLO-TB 能够提取更深层次的特征，且自上而下的多尺度融合结构能够提高模型对图片特征的利用率，但模型参数量明显增加，且 FPS 下降了 37.9。为了压缩网络模型并降低网络计算量，使用 Ghost 卷积模块对 YOLO-TB 进行改进得到 YOLO-TBG，YOLO-TBG 相比 YOLO-TB 检测精度 mAP_{50} 下降了 2.5%，召回率下降了 3%，但网络参数量从原来的 10 517 521 下降到了 5 980 801，FPS 有所提升，有效降低了模型的部署成本，YOLO-TBG 综合性能比 YOLO-TB 具有明显优势。以上模型 FPS 降低是因为 mAP 的提高会增加检测目标框的数量，且增加的网络模块造成了额外的计算量，增加了时间成本，但该检测速度仍符合实际检测场景的实时性要求。

此外，为了更加直观地说明改进后的模型与 YOLOv4-tiny 之间的区别，本文选取了测试集中的 3 张不同场景的图片进行目标检测实验对比，从左到右依次是：YOLOv3-tiny、YOLOv4-tiny 检测结果、YOLO-TBG 检测结果。

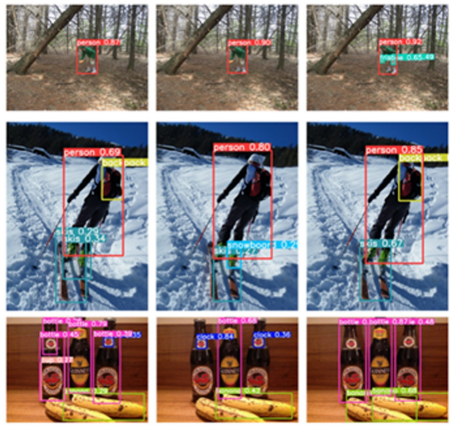

图 9 YOLOv3-tiny，YOLOv4-tiny，YOLO-TBG 的检测结果

Fig. 9 Detection results of YOLOv3-tiny, YOLOv4-tiny and YOLO-TBG

从图 9 中可以看到，YOLOv3-tiny 与 YOLOv4-tiny 的检测结果相差不大，存在着目标框偏移以及漏检的问题。改进的 YOLOv4-tiny(YOLO-TBG)则弥补了以上网络的缺陷，能够很好地检测出小目标，也能够提取大目标特征信息，能够有效降低原网络漏检以及误检率，提高目标框的置信度，并且在多个目标堆叠的情况下也能取得较好的检测出结果，不发生目标框偏移的情况。

结束语 本文算法是基于现实中工业领域方向的项目需求所提出的，为满足算法模型轻量化、检测精度高且容易部署的项目需求所做出的算法改进，目前本文算法已经被应用到现实工业领域项目中。由于在实际项目中，甲方已有的相关工作更加支持使用 YOLOv4 系列的算法，因此本文为了贴合实际项目要求，进行了基于 YOLOv4 系列算法改进的工作，以及相关的实验对照。

本文提出的 YOLO-TBG 是在 YOLOv4-tiny 的骨干网络

中添加 VIT 模块，通过利用 VIT 模块中的多头自注意力机制，来补足一般的卷积操作特征挖掘不充分的缺点，在少量增加网络参数的情况下提高网络的检测精度。简化 Bi-FPN 结构，适当增加网络中的节点，同时引入注意力机制，提高了不同层之间特征融合的鲁棒性，能够有效提高多个尺度的目标检测精度。使用 Ghost 模块替换传统卷积操作，减少了模型冗余计算量，有效压缩了网络体积，提升了网络运行速度，降低了算法的部署成本。实验结果表明，相比原 YOLOv4-tiny 网络在 COCO 数据集中的表现，YOLO-TBG 网络的检测准确率提高了 11.4%，可有效弥补原网络精确度不足的缺陷，且检测速度仍能满足工业需求。

本文对轻量化网络 YOLOv4-tiny 进行了一系列改进，虽然提升了网络的检测精度，但检测速度上相比原网络有所降低，根据实验数据分析可以得到：VIT 对网络检测精度有着明显提升的作用，但也加重了网络的计算量，降低了网络的检测速度。因此，在之后的研究中会针对 VIT 的变种结构进行研究，达到在保持精度的同时降低模块的冗余计算量。网络使用 Bi-FPN 结构后检测精度提升显著，且 Ghost 降低了提升所带来的负面影响，但改进后的网络在检测速度上仍有改善空间，在之后的研究中可以研究如何对网络进行模型剪枝，设计更为轻量化的网络。

参 考 文 献

[1] BOCHKOVSKIY A，WANG C Y，LIAO H. YOLOv4：Optimal Speed and Accuracy of Object Detection[J/OL]. arXiv：2004.10934，2020.

[2] REDMON J，FARHADI A. YOLOv3：An Incremental Improvement[J]. arXiv e-prints，2018.

[3] REDMON J，FARHADI A. YOLO9000：Better，Faster，Stronger[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE，2017：6517-6525.

[4] 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.

[5] LIU W，ANGUELOV D，ERHAN D，et al. Ssd：Single shot multibox detector[C]// Computer Vision－ECCV 2016：14th European Conference，Amsterdam，The Netherlands. Springer International Publishing，2016：21-37.

[6] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，39(6)：1137-1149.

[7] HE K，GKIOXARI G，P DOLLÁR，et al. Mask R-CNN][C]// IEEE Transactions on Pattern Analysis & Machine Intelligence. IEEE，2017.

[8] LIU S，QI L，QIN H，et al. Path Aggregation Network for Instance Segmentation][C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2018.

[9] WANG C Y，BOCHKOVSKIY A，LIAO H Y M. Scaled-yolov4：Scaling cross stage partial network[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：13029-13038.

[10] WANG B，LE H X，LI W J，et al. Improved mask detection algorithm of YOLO lightweight network[J]. Computer Engineering and Applications，2021，57(8)：62-69.

[11] ZHANG X，ZHANG Y Q，HE B，et al. Research on remote sensing image aircraft target detection technology based on YOLOv4-tiny[J]. Optical Technology，2021，47(3)：344-351.

[12] ZHANG X，WAN T，WU Z，et al. Real-time detector design for small targets based on bi-channel feature fusion mechanism[J]. Applied Intelligence，2022，52(3)：2775-2784.

[13] LU D，MA W Q. Gesture recognition based on improved YOLOv4-tiny algorithm[J]. Journal of Electronics and Information，2021，43(11)：3257-3265.

[14] TIAN Y，MAO W，YUAN S，et al. A Decision Support System for Power Components Based on Improved YOLOv4-Tiny[J]. Scientific Programming，2021，2021：1-11.

[15] LIN Y，CAI R，LIN P，et al. A detection approach for bundled log ends using K-median clustering and improved YOLOv4-Tiny network[J]. Computers and Electronics in Agriculture，2022，194：106700.

[16] WANG L，ZHOU K，CHU A，et al. An improved light-weight traffic sign recognition algorithm based on YOLOv4-tiny[J]. IEEE Access，2021，9：124963-124971.

[17] HUI T，XU Y L，JARHINBEK R. Detail texture detection based on Yolov4-tiny combined with attention mechanism and bicubic interpolation[J]. IET Image Processing，2021，15(12)：2736-2748.

[18] GUO C，LV X，ZHANG Y，et al. Improved YOLOv4-tiny network for real-time electronic component detection[J]. Scientific Reports，2021，11(1)：22744.

[19] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：Transformers for image recognition at scale[J]. arXiv：2010.11929，2020.

[20] TAN M，PANG R，LEQ V. EfficientDet：Scalable and Efficient Object Detection[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2020.

[21] HAN K，WANG Y，TIAN Q，et al. GhostNet：More Features From Cheap Operations[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). IEEE，2020.

[22] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is All You Need[J]. arXiv：1706.03762，2017.

[23] HINTON G E，SRIVASTAVA N，KRIZHEVSKY A，et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science，2012，3(4)：212-23.

[24] HENDRYCKS D，GIMPEL K. Gaussian Error Linear Units(GELUs)[J]. arXiv：1606.08415，2016.

DOU Zhi，born in 1983，male，Ph.D，associate professor. His main research interests include image processing，pattern recognition.