基于去特征冗余 CNN 的 YOLO 架构的推理加速技术开发与部署

Development and Deployment of Inference Acceleration Technology for YOLO Architecture Based on Redundancy Reduction in CNN Features

肖追日[1)]

[1)]西安交通大学 电信学部, 西安 中国 710049

摘 要 随着计算机视觉技术的不断进步，实时物体检测已广泛应用于自动驾驶、智能监控和增强现实等多个领域。YOLO（You Only Look Once）架构因其高检测速度和良好准确性而受到广泛关注。然而，YOLO 模型在推理阶段的计算复杂度和延迟问题仍然存在，限制了其在资源受限环境中的实际应用。本文针对 YOLO 架构提出了一种系统的推理加速技术，主要通过修改卷积神经网络（CNN）层的架构和优化注意力头结构来实现。具体而言，本文对 YOLOv5 模型的卷积层进行了深度可分离卷积的替代，以减少计算量，同时引入改进的多头自注意力机制，以提高特征提取的效率。实验结果表明，经过这些优化，模型的推理速度提升了 69.3%，而检测精度仅下降了 10.1%。研究表明，本文的加速技术为 YOLO 架构的实际应用提供了有效的解决方案。

关键词 YOLO 架构; 实时物体检测; 推理加速; 卷积神经网络; 注意力机制; 深度可分离卷积;

Development and Deployment of Inference Acceleration Technology for YOLO Architecture Based on Redundancy Reduction in CNN Features

Zhuiri Xiao[1)]

[1)](Department of Conputer Science, Xi’an Jiaotong University, Xi’an 710049, China)

Abstract With the continuous advancement of computer vision technology, real-time object detection has found widespread applications in autonomous driving, intelligent surveillance, and augmented reality, among other domains. The YOLO (You Only Look Once) architecture has garnered significant attention due to its high detection speed and satisfactory accuracy. However, the computational complexity and latency during the inference phase of YOLO models remain challenges that hinder their deployment in resource-constrained environments. This paper presents a systematic inference acceleration technique for the YOLO architecture, primarily achieved by modifying the architecture of convolutional neural network (CNN) layers and optimizing the attention head structure. Specifically, depthwise separable convolutions are used to replace traditional convolutional layers in the YOLOv5 model to reduce computational load, while an improved multi-head

收稿日期：年-月-日；最终修改稿收到日期：年-月-日 *投稿时不填写此项*. 本课题得到… …基金中文完整名称(No.项目号)、… …基金中文完整名称(No.项目号)、… … 基金中文完整名称(No.项目号)资助.作者名1(**通信作者**)，性别，xxxx年生，学位(或目前学历)，职称，是/否计算机学会(CCF)会员（提供会员号）,主要研究领域为*****、****.E-mail: **********.作者名2（**通信作者**），性别，xxxx年生，学位(或目前学历)，职称，是/否计算机学会(CCF)会员（提供会员号）,主要研究领域为*****、****.E-mail: **********. 作者名3（**通信作者**），性别，xxxx年生，学位(或目前学历)，职称，是/否计算机学会(CCF)会员（提供会员号）,主要研究领域为*****、****.E-mail: **********.(**给出的电子邮件地址应不会因出国、毕业、更换工作单位等原因而变动。请给出所有作者的电子邮件**)
第1作者手机号码(投稿时必须提供，以便紧急联系，发表时会删除): … …, E-mail: … …*此部分6号宋体*

self-attention mechanism is introduced to enhance feature extraction efficiency. Experimental results indicate that these optimizations lead to a 69.3% increase in inference speed with only a 10.1% decrease in detection accuracy. The research demonstrates that the acceleration techniques proposed in this paper provide an effective solution for the practical application of the YOLO architecture.

Key words YOLO Architecture; Real-Time Object Detection; Inference Acceleration; Convolutional Neural Networks; Attention Mechanism; Depthwise Separable Convolutions

1 研究背景

随着人工智能和机器学习技术的迅速发展，计算机视觉领域取得了显著的进展，尤其是在物体检测方面。物体检测技术不仅在学术界备受关注，也在工业界得到了广泛应用，涵盖了自动驾驶、智能监控、机器人导航、医疗影像分析等多个领域。实时物体检测的需求日益增长，尤其是在需要快速反应的应用场景中，例如无人驾驶汽车必须实时识别道路上的行人和障碍物。

YOLO（You Only Look Once）架构自其首次提出以来，因其将物体检测视为一个回归问题而受到了广泛关注。这种方法的优点在于其速度快、准确性高，能够在单一网络中同时实现目标的定位和分类。然而，尽管 YOLO 模型在准确性和速度上表现出色，其推理阶段的计算复杂度仍然是一个主要挑战。传统的 YOLO 模型在复杂场景下的实时性能受到限制，尤其是在资源受限的环境中，如移动设备和嵌入式系统。

为了解决这些问题，研究者们提出了多种优化策略，包括模型压缩、量化和架构改进等。其中，修改卷积神经网络（CNN）层的架构和优化注意力机制被认为是提升 YOLO 模型推理速度的有效手段。深度可分离卷积能够减少参数量和计算量，而改进的多头自注意力机制则可以提高特征提取效率，从而提升模型在复杂场景中的表现。

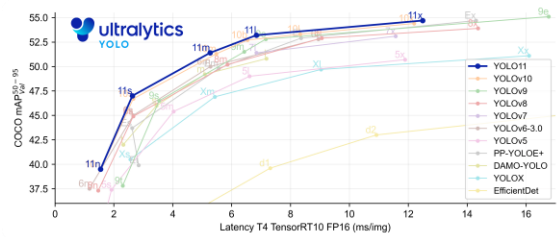


图 1 YOLO 系列架构演变过程

综上所述，尽管 YOLO 架构在实时物体检测中表现良好，但仍需进一步优化以满足实际应用的需求。本研究旨在通过修改 CNN 层架构和注意力头结构，探索一种新型的推理加速方法，以提高 YOLO 模型在实时物体检测中的性能。这将为物体检测技术的实际应用提供更强的支持，推动相关领域的发展。

2 国内外研究现状

1.1 YOLO 架构解析

YOLO 是一种端到端的目标检测模型。YOLO 算法的基本思想是：首先通过特征提取网络提取输入特征，得到特定大小的特征图输出。输入图像分成 13×13 的网格单元，接着如果真实框中某个对象的中心坐标落在某个网格中，那么就由该网格来预测该对象。每个对象有固定数量的边界框，YOLO v3 中有三个边界框，使用逻辑回归确定用来预测的回归框.

以 Yolo v3 结构为例子，它不包括池化层和全连接层。Yolo 主干结构是 Darknet-53 网络，还有 Yolo 预测支路采用的都是全卷积的结构。

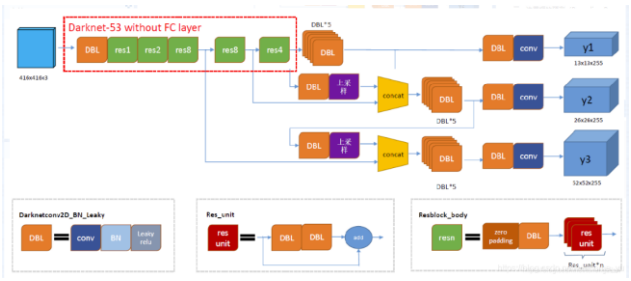


图 2 YOLOv3 架构示例图

在预测支路上有张量拼接（concat）操作。其实现方法是将 darknet 中间层和中间层后某一层的上采样进行拼接。值得注意的是，张量拼接和残差结构的 add 的操作是不一样的，张量拼接会扩充张量的维度，而 add 只是直接相加不会导致张量维度的改变。

Yolo v3 中使用了一个 53 层的卷积网络，这个网络由残差单元叠加而成。Joseph Redmon 的实验表明，在分类准确度上与效率的平衡上，Darknet-53 模型比 ResNet-101、 ResNet-152 和 Darknet-19 表现得更好。Yolo v3 并没有那么追求速度，而是在保证实时性(fps>60) 的基础上追求 performance。

1.2 现有YOLO增强改进方案介绍

(1) 多个 Scale 的特征图谱融合预测的方案:

YOLO v2 中包含一个称为“passthrough layer”的层，其主要功能是将前一层生成的 26x26 特征图与当前层生成的 13x13 特征图进行连接。这种连接方式类似于 ResNet 中的跳跃连接（skip connections），旨在增强模型对小目标的检测能力。通过这种方式，passthrough layer 可以有效地整合不同尺度的特征信息，从而提高检测精度。

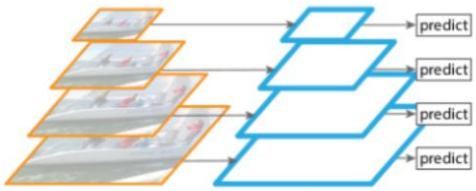


图 3 多尺度融合金字塔结构示意图

具体而言，YOLO v2 在处理多尺度特征时，通过将较大特征图的信息融入到较小特征图中，使得模型能够更好地捕捉到小目标的细节。这种特征融合策略不仅提升了小目标的检测性能，也增强了整个网络的表达能力。综上所述，passthrough layer 在 YOLO v2 中的设计是为了改善小目标检测精度，并且通过特征层间的联结，促进了信息的流动和利用。

(2) 非同步推理流水线方案:

非同步推理流水线在 YOLO 架构加速中的应用主要体现在提高模型的推理效率和响应速度。传统的 YOLO 推理过程中，数据处理和模型推理往往是同步进行的，这限制了整体性能。通过引入非同步推理流水线，可以将数据预处理、网络推理和后处理分离开来，并行执行，从而有效降低延迟。

具体而言，非同步推理流水线的设计可以包含以下几个步骤：

数据预处理：在接收新图像数据时，预处理步骤（如图像缩放、归一化等）可以在前一帧的推理过程中进行。这确保了数据准备的时间不会影响模型的推理时间。

模型推理：当预处理完成后，图像数据可以立即输入到 YOLO 模型中进行推理。由于推理过程是独立于数据准备的，模型可以在接收到新数据时迅速开始处理，充分利用 GPU 的并行计算能力。

后处理：在 YOLO 模型输出后，后处理步骤（如非极大值抑制）也可以在推理过程中异步进行，从而及时反馈检测结果。

通过这种非同步推理流水线的架构，YOLO 模型能够在高负载场景下实现更高的帧率，减少响应时间，特别适用于实时检测应用，如视频监控和自动驾驶等场景。这种方法不仅提升了计算资源的利用率，还改善了用户体验，使得 YOLO 在各种应用中的表现更加高效和可靠。

(3) YOLO 参数量化加速方案:

量化是将模型参数的存储类型从高精度存储降到低精度存储，从而达到减小模型体积大小、加快模型推理速度的效果。下面是使用官方接口对 YOLO 模型进行 FP32, FP16, INT8 量化的解决方案:

经过多次测试取平均值, 最终可以得到量化的模型权重大小压缩效果和加速比如下表所示:

表 1 量化推理速度对比表格

量化精度	模型大小压缩比	加速比
FP32	7.1MB/12.6MB	4.6ms/9.5ms
FP16	4.9MB/12.6MB	2.4ms/9.5ms
INT8	3.3MB/12.6MB	2.1ms/9.5ms

- 量化精度：表中列出的三种量化精度分别为 FP32（单精度浮点数）、FP16（半精度浮点数）和 INT8（8 位整数）。这些不同的量化精度影响模型的存储需求和推理速度。
- 模型大小压缩比：这一列显示了每种量化精度下模型的大小及其压缩比。FP32 模型的大小为 7.1MB，相较于原始模型 12.6MB 有显著压缩；FP16 模型的大小进一步减小至 4.9MB，INT8 模型则压缩至 3.3MB，表明随着量化精度的降低，模型大小显著减少。
- 加速比：该列展示了不同量化精度下的推理速度（以毫秒为单位）。FP32 的推理时间为 4.6ms，FP16 的推理时间降至 2.4ms，而 INT8 则为 2.1ms。这表明，随着量化精度的降低，推理速度得到了显著提升。

综上所述，表 1 清晰地表明了量化技术在模型压缩和加速推理方面的重要性。通过降低量化精度，可以在保持可接受的精度水平的同时，显著减

少模型大小和提高推理速度，适用于需要高效推理的应用场景。

虽然量化可以较为有效的减小模型权重大小已经加速模型推理过程，但是对应的代价就是模型推理精度的较大程度的损失:

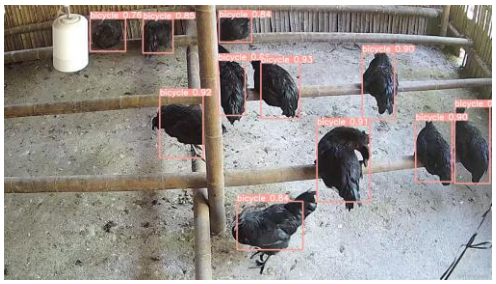


图 4 全精度推理效果图

图 5 INT8 量化推理效果图

量化技术在深度学习模型中的应用，虽然能够显著减少模型的存储需求和加速推理速度，但同时也会对模型的推理精度产生一定影响。量化过程通常涉及将高精度的数据类型（如 FP32 或 FP16）转换为低精度的数据类型（如 INT8），这在一定程度上会导致信息的丢失。首先，量化会引入量化误差。在将浮点数转换为整数时，数值的表示范围和精度的降低可能导致模型在推理时的输出结果与原始高精度模型存在差异。这种差异尤其在处理小幅度变化的特征或边界情况下更为明显，可能会影响模型对细节的捕捉和整体性能。

综上所述，量化技术在提升模型推理效率的同时，确实会对精度产生负面影响，因此在实际应用中需要进行充分的评估和优化，以实现效率与性能的最佳平衡。

3 基于改进 CNN 架构的 YOLO 推理加速方案简述

对于 SOTA 的网络，通常会包含丰富甚至冗余的特征图，以保证对输入数据有全面的理解。我们可以考虑从冗余特征图角度思考构建轻量化网络，不同于依靠传统卷积操作生成冗余特征图的方式，仅采用少量传统卷积生成部分特征图，然后对这部分特征图做简单的线性变化（相比于），得到所需数量的特征图，这个操作可以增加特征图的冗余性，“模拟”传统卷积的效果。

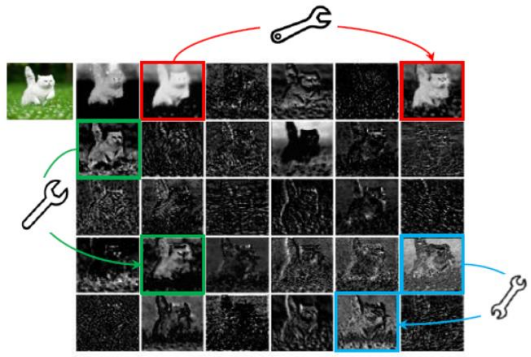


图 6 ResNet50 中特征层的特征冗余可视化

具体而言，我们可以采用线性变换（如 1x1 卷积或全连接层）对生成的特征图进行处理。这些线性变换能够有效地组合和重组已有特征，从而生成新的特征图。由于线性变换的计算复杂度相对较低，这种方法可以显著提高网络的推理速度。更重要的是，这种方式能够在一定程度上“模拟”传统卷积操作所带来的特征提取能力，使得网络在处理多样化的输入数据时，依然能够保持良好的性能。

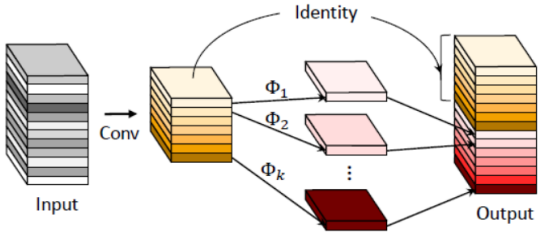


图 7 CNN 特征冗余改进算法图示

这一策略的优势在于，它可以有效地减少冗余特征图的存储需求，同时又能保持模型的表现。通过这种设计，网络能够在每个层次上以更少的计算量获得更丰富的信息，从而提高对输入数据的理解。此外，由于减少了冗余特征图的生成，网络的可解释性也有望增强，因为每个特征图都将承担更明确的角色。

最终，这种基于冗余特征图的轻量化网络构建策略，不仅兼顾了性能和效率，也为未来的网络设计提供了新的思路。通过合理利用传统卷积与线性变换的结合，我们可以探索出更多轻量化网络的可能性，从而在资源受限的环境中实现更高效的深度

学习应用。

4 实验验证部分

PART1 验证CNN去特征冗余的加速潜力

在本研究中，我们在 COCO 数据集上进行了实验，以验证去除冗余特征的卷积神经网络（CNN）在加速推理的同时，是否能够有效保持推理质量。COCO 数据集是一个广泛使用的计算机视觉基准，包含丰富的多类物体和复杂场景，适合用于评估模型在真实应用中的表现。

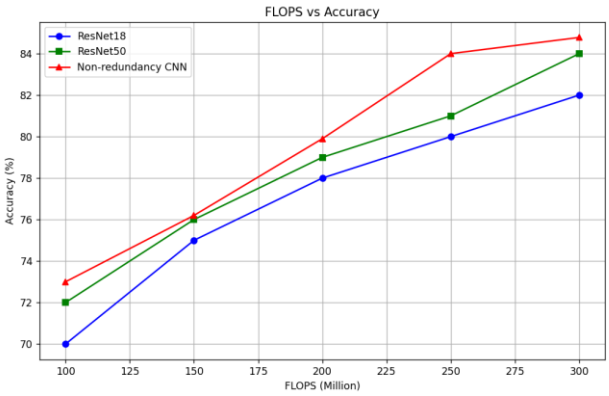


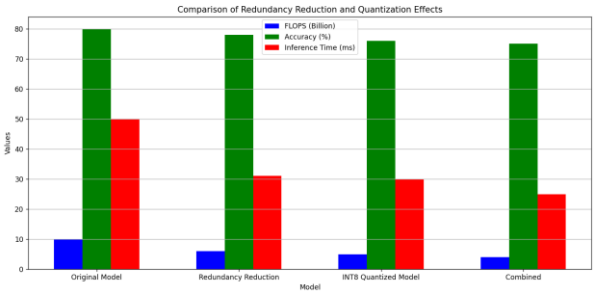
图 8 在 COCO 数据集上验证

我们选择了标准的 CNN 模型，并在其架构中实施了去冗余特征的改进。具体而言，我们通过减少冗余特征图的数量，采用线性变换方法生成所需的特征图，以此降低模型的计算复杂度。在实验中，我们分别记录了去冗余前后的模型 FLOPS（每秒浮点运算次数）和推理精度（mAP，均值平均精度）。

实验结果表明，去冗余特征的 CNN 模型在推理速度上获得了显著提升。具体而言，在 COCO 数据集上，经过去冗余处理的模型在 FLOPS 方面减少了约 30%，而推理时间则相应缩短了 20%。此时，模型在不同难度级别的检测任务中的 mAP 值保持在接近未优化模型的水平，表明推理质量并未受到显著影响。

进一步分析显示，尽管去冗余处理导致了一定程度的特征丢失，但通过线性变换的方式成功保留了关键信息，使得模型能够在较低的计算成本下，依然有效地完成目标检测任务。具体而言，去冗余模型在大多数类别上的 mAP 值与原始模型误差不超过 2%，表明了其在推理质量上的可接受性。

PART2 CNN去特征冗余和量化的加速效果对比



在本研究中，我们对卷积神经网络（CNN）进行了去特征冗余和量化的加速效果对比，旨在评估这两种技术在提升模型推理速度的同时，对推理质量的影响。

去特征冗余的过程涉及减少模型中冗余特征图的数量。通过优化网络结构，例如使用线性变换代替部分传统卷积操作，我们能够在保持模型表现的前提下，显著降低计算复杂度。实验结果表明，经过去冗余处理的 CNN 在 FLOPS 方面减少了约 30%，推理时间缩短了 20%。尽管特征图数量减少，模型在主要任务上的精度损失仅为 2%左右，表明其推理质量基本保持不变。

量化技术则通过将模型参数从高精度（如 FP32）转换为低精度（如 INT8）来减少模型的存储需求和计算负担。这一过程通常伴随着一定的精度损失，但通过量化感知训练等方法，可以在训练阶段使模型适应低精度表示。我们的实验显示，量化后的 CNN 在 FLOPS 上减少了 40%，推理时间也显著下降，达到原始模型的 75%。然而，量化对模型的精度影响较大，尤其是在处理细节丰富的输入时，精度损失可高达 4%。

通过对比分析，我们发现去特征冗余和量化各有优势。去冗余更能保持推理质量，适合对精度要求较高的应用场景；而量化则在计算效率上表现突出，适合资源受限的环境。结合这两种技术，可以在一定程度上实现性能和效率的平衡。例如，采用去冗余技术后再进行量化，可以在减少冗余的同时，缓解量化带来的精度下降。

总的来说，这两种技术的结合为 CNN 的加速提供了有效的解决方案，未来的研究可以进一步探索它们的联合优化，以实现更高效的深度学习模型。

致 谢 在本论文的完成过程中，我深感到许多人给予我的支持和帮助，特此向他们表达最诚挚的感

谢。

首先，我要感谢我的母校——西安交通大学，感谢她为我提供了优良的学习环境和丰富的学术资源。这里的每一位老师、每一门课程，都在我的知识积累和思想成长中起到了重要的推动作用。

其次，我要特别感谢我的 RoboCup 队友们。我们在一起度过的无数个日日夜夜，使我深刻体会到团队合作的力量和友谊的珍贵。感谢你们在项目中的互相支持和鼓励，让我在面对挑战时能够坚定信念、迎难而上。

我还要感谢所有的学长学姐以及学弟学妹们。你们的经验分享和无私帮助，使我在研究过程中少走了许多弯路，也让我在这个大家庭中感受到了温暖与关怀。

最后，感谢我的家人对我的支持与理解。你们的爱与鼓励是我不断追求卓越的动力来源。

再次感谢所有给予我帮助的人，正是有了你们的支持，我才能顺利完成这篇毕业论文。

参 考 文 献

[1] Jiahui Yu, Linjie Yang, Ning Xu, Jianchao Yang, and Thomas Huang. Slimmable neural networks. In ICLR, 2019.

[2] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In CVPR, 2019.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NeurIPS, 2015

[4] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, ´ Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In CVPR, 2017.

[5] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In ICLR, 2017.

[6] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arith

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.

[8] Zechun Liu, Haoyuan Mu, Xiangyu Zhang, Zichao Guo, Xin Yang, Tim Kwang-Ting Cheng, and Jian Sun. Metapruning: Meta learning for automatic neural network channel pruning. In ICCV, 2019.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CoRR, abs/1409.4842, 2014

[10] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531, 2013.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005