智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向车规级芯片的对象检测模型优化方法

宫大汉， 于龙龙， 陈辉， 杨帆， 骆沛， 丁贵广

引用本文:
宫大汉, 于龙龙, 陈辉, 等. 面向车规级芯片的对象检测模型优化方法[J]. 智能系统学报, 2021, 16(5): 900–907.
GONG Dahan, YU Longlong, CHEN Hui, et al. Object detection model optimization method for car–level chips[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(5): 900–907.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202107057

吴文俊人工智能科技进步奖一等奖

成果名称：开放环境自适应视觉感知计算关键技术与应用
获 奖 人：丁贵广、颜成钢、郭雨晨、张志伟、敖乃翔、殷俊、全书学、刘霁、何宇巍、陈辉、何涛、项刘宇、丁霄汉、朱树磊、师文喜
完成单位：清华大学、杭州电子科技大学、北京达佳互联信息技术有限公司、新疆联海创智信息科技有限公司、浙江大华技术股份有限公司、OPPO 广东移动通信有限公司

丁贵广

清华大学软件学院副院长，北京信息科学与技术国家研究中心副主任。特别研究员，博士生导师。

主要从事视觉感知计算、机器学习、数据检索等方面的研究，重点关注视觉感知计算和弱监督学习理论与方法的研究，面向端侧和边缘侧计算能力有限、功耗限制高等场景，研究深度视觉模型压缩技术，研发端侧视觉计算系统和平台。先后主持基金委重点项目、重点研发项目、国家973、863等项目数十项。发表高水平学术论文近百篇，获授权发明专利数十项，相关成果成功应用于公安部二十三局、快手、新疆联海创智、数码视讯等单位，曾获国家科技进步二等奖、中国电子学会技术发明一等奖。

团队简介

项目团队由清华大学、杭州电子科技大学、北京达佳互联信息技术有限公司、新疆联海创智信息科技有限公司、浙江大华技术股份有限公司、OPPO 广东移动通信有限公司六家单位组成。项目团队围绕“开放环境自适应视觉感知计算”开展了长期的产学研合作，在基础理论、关键技术、实际系统、商业化产品等多个方面刻苦攻关，形成了一系列突破性成果，发表高水平学术论文六十余篇，授权发明专利三十余项、软件著作权十余项，在多项国际权威技术竞赛中名列前茅，形成了广泛的国内外影响力。团队成果在公共安全、互联网监管、智能设备等重要领域得到广泛应用，取得了显著的经济与社会效益。

DOI: 10.11992/tis.202107057

面向车规级芯片的对象检测模型优化方法

宫大汉[1,2], 于龙龙[3], 陈辉[2,4], 杨帆[1,2], 骆沛[5], 丁贵广[1,2]

(1. 清华大学 软件学院, 北京 100084; 2. 清华大学 北京信息科学与技术国家研究中心, 北京 100084; 3. 涿溪脑与智能研究所, 浙江 杭州 311121; 4. 清华大学 自动化系, 北京 100084; 5. 禾多科技 (北京) 有限公司, 北京 100102)

摘 要: 卷积神经网络复杂的网络结构使得模型计算复杂度高, 限制了其在自动驾驶等实际终端场景中的应用。针对终端场景下的计算资源受限的问题, 本文从轻量化深度模型设计和车规级芯片模型部署验证两方面进行研究。针对深度模型计算效率和检测精度的矛盾, 本文设计了基于中心卷积的轻量化对象检测模型, 实现功耗低且精度高的模型性能。进一步, 本文基于量化感知训练的模型加速部署方法在车规级芯片上开展了系统级部署验证, 在车规级芯片 tda4 上成功实现了高效的对象检测模型, 在自动驾驶场景中取得了良好的性能。

关键词: 人工智能; 计算机视觉; 对象检测; 终端设备; 车规级芯片; 卷积神经网络; 模型加速; 模型量化

中图分类号: TP18 **文献标志码:** A **文章编号:** 1673−4785(2021)05−0900−08

中文引用格式: 宫大汉, 于龙龙, 陈辉, 等. 面向车规级芯片的对象检测模型优化方法 [J]. 智能系统学报, 2021, 16(5): 900–907.

英文引用格式: GONG Dahan, YU Longlong, CHEN Hui, et al. Object detection model optimization method for car-level chips[J]. CAAI transactions on intelligent systems, 2021, 16(5): 900–907.

Object detection model optimization method for car-level chips

GONG Dahan[1,2], YU Longlong[3], CHEN Hui[2,4], YANG Fan[1,2], LUO Pei[5], DING Guiguang[1,2]

(1. School of Software, Tsinghua University, Beijing 100084, China; 2. BNRist Tsinghua University, Beijing 100084, China; 3. Zhuoxi Institute of Brain and Intelligence, Hangzhou 311121, China; 4. Department of Automation, Tsinghua University, Beijing 100084, China; 5. HoloMatic Technology (Beijing) Co., Ltd, Beijing 100102, China)

Abstract: Convolutional neural networks have achieved great success in visual perception tasks. Its complex network structure makes the model computationally complex, which limits its application in actual terminal scenarios such as autonomous driving. Aiming at the problem of limited computing resources in terminal scenarios, in this paper, we conduct research from two aspects: lightweight deep model design and the model deployment and verification on car-level chips. As for the contradiction between the calculation efficiency of deep models and the detection accuracy, we design a lightweight object detection model based on the center-convolution, enjoying low power consumption and high accuracy model performance. Furthermore, based on the method of quantization aware training, we carried out system-level deployment and verification on car-level chips. We successfully implemented a high-efficiency object detection model on the car-level chips, i.e. tda4, and achieved good performance in autonomous driving scenarios.

Keywords: artificial intelligence; computer vision; object detection; terminal equipment; car-level chip; convolutional neural network; model acceleration; model quantization

卷积神经网络(convolutional neural network, CNN)在图像识别、对象检测等视觉感知任务上取得了巨大的成功。由于其优秀的性能, 使得 CNN 已经成为一种标准的智能结构, 在智能手机、可穿戴设备、IoT 终端设备、自动驾驶等智能应用中扮演着重要的角色。然而, CNN 参数量巨大、计算复杂度高的特点限制了它在实际场景中的大范围落地应用。比如在自动驾驶领域, 如果 CNN 模型部署到云端, 那么网络传输的稳定性将决定系

收稿日期: 2021−07−27.

基金项目: 国家自然科学基金项目(U1936202, 61925107); 中国博士后科学基金创新人才计划项目(BX2021161).

通信作者: 丁贵广. E-mail: dinggg@tsinghua.edu.cn.

统是否能及时应对不断变化的外部环境，进而影响整体自动驾驶性能。而受限于网络传输的不稳定性，这种云端计算终端响应的方式在实际自动驾驶系统中难以广泛应用。所以，基于车载终端设备实现深度模型的部署成为一种越来越流行的解决方案。

然而，面向车载终端设备的深度模型部署存在着计算资源和模型性能的矛盾。具体来说，深度模型复杂度越高，通常它的效果越好，但其计算速率越差。考虑到终端设备上缺少高性能计算单元（如 GPU）的支持，使得复杂度高的深度模型无法满足计算效率的要求。而如果采用复杂度低的小模型，又面临模型性能不佳的问题。因此，如何在计算资源受限条件下实现高精度深度模型的部署对自动驾驶至关重要。

本文面向自动驾驶场景，研究车规级芯片上的对象检测[1-2]模型高效计算方法。本文使用先进的对象检测模型 CenterNet[3] 作为实验模型。这种模型通过预测对象中心点的位置和包围盒的偏移量，进而确定对象的整体位置。通常使用主流深度学习模型，比如 ResNet[4] 作为模型的主干网络。主干网络可以将输入的视觉图像均分为网格区域，然后 CenterNet 判断每个网格点是否是某个对象的中心。考虑到车载系统缺乏高性能计算单元 GPU，本文采用轻量型深度模型 ResNet18 作为 CenterNet 的主干网络。

ResNet18 通过堆叠 3×3 卷积，使得在一个 3×3 的卷积窗口中，卷积核的中心点对应的区域可以不断“吸收”周围区域的信息来增强自身的特征表达。这种方式的缺点在于 3×3 卷积将中心区域的特征和周围区域的特征同等对待，容易降低中心区域的特性在特征中的表达，加大了模型混淆中心区域和其他区域的可能性，不利于 CenterNet 对关键点的定位。

为了解决这个问题，本文提出一个基于中心卷积的对象检测模型。具体来说，本文为 3×3 卷积增加一个 1×1 卷积旁路，这种旁路对应 3×3 卷积核的中心区域，可以充分学习中心区域的隐藏特性。本文对 3×3 卷积和 1×1 卷积的输出进行加和融合，并将这种结构命名为中心卷积。所设计的中心卷积可以独立学习中心区域的特性和周围区域的关联信息，有效增强中心区域的特征表达，促进关键点的预测。3×3 卷积和 1×1 卷积都是线性变换，因此在模型部署推理时，可以很方便地将 1×1 卷积旁路的参数融合到 3×3 卷积核的中心点上，恢复成标准 3×3 卷积，大大降低了推理时的模型大小。相比于标准的 3×3 卷积，所提出的中心卷积不增加推理复杂度，但是具有更强的特征学习能力，使得模型检测效果更佳。

本文用中心卷积替换了 ResNet18 中的 3×3 卷积，构建了基于中心卷积的 CenterNet 模型，并在实际的车规级计算芯片上进行了模型部署验证。尽管 ResNet18 的计算复杂度不高，但为了充分利用车载系统的计算资源，希望尽可能在保持精度的情况下，提升 CenterNet 的计算效率。为此，本文采用一套基于量化的部署流程：首先使用量化感知的训练方式对给定的 CenterNet 进行重训练，得到 int8 数值精度的 CenterNet 模型，然后调用底层开放接口，将模型部署在芯片上。通过这种量化感知训练，可以得到低比特的模型，减少了模型的大小，并且保持了足够高的模型精度。而在芯片线上推理过程中，模型以低比特 int8 进行运算，相比于 float32 运算，处理速度更快。

综上所述，本文提出了一种中心卷积来替代 ResNet18 中的 3×3 卷积，构建了基于中心卷积的 CenterNet 对象检测模型；进一步，采用基于量化的模型部署方法，实现在车规级芯片上的模型效果验证。

1　相关工作

对象检测领域的研究热点经历了由基于锚点框（anchor-based）检测模型到无锚点框（anchor-free）检测模型。其中，以 CornerNet[5] 为代表的基于关键点预测的对象检测模型的出现引起了研究者的广泛兴趣，anchor-free 的方法渐渐兴起。CornerNet 首次提出预测包围框的一对角点来检测目标，并使用角池化操作来更好地定位包围框的角点。之后，Zhou 等[6] 提出 ExtremeNet 来预测对象的极值点和中心点进而实现目标检测的目的，后面又进一步提出 CenterNet 只预测中心点来检测目标。Liu 等[7] 改进了 CenterNet，提出了 TTFNet,使用高斯核编码来引入更多的回归样本，极大地加快了模型的训练速度。而 Zhou 等[8] 又将中心点预测的思想引入到两阶段（two-stage）检测模型中，获得了显著的性能。

需要指出的是，本文重点关注深度模型在车规级芯片上的高效推理。具体来说，本文希望在有限的计算资源条件下实现高性能和高效率的对象检测模型。考虑到 CenterNet 具有简单易扩展的优点，所以本文采用 CenterNet 为实验模型进行探索，而不采用近期对 CenterNet 的改进工作进行实验，如 CenterNet2[8]、TTFNet[7]。

2　基于中心卷积的对象检测模型

2.1　中心卷积

本文的中心卷积关注对 3×3 卷积的改进。首先，为了更好地描述中心卷积，将卷积运算描述为：给定输入特征图为$\boldsymbol{M}\in\mathbf{R}^{U\times V\times C}$，卷积输出特征图为$\boldsymbol{O}\in\mathrm{R}^{U\times V\times D}$，那么对于一个卷积核大小为$H\times W$且有$\left(\left\lfloor\frac{H}{2}\right\rfloor,\left\lfloor\frac{W}{2}\right\rfloor\right)$填充的卷积，可以用$\boldsymbol{F}\in\mathbf{R}^{H\times W\times C\times D}$表示它的参数，且满足：

$$\boldsymbol{O}_{i,j,k}=\boldsymbol{M}^{i,j}*\boldsymbol{F}^{k},k\in[1,D] \tag{1}$$

式中：$*$表示矩阵元素相乘；$\boldsymbol{M}^{i,j}$是从$\boldsymbol{M}$中以(i,j)为中心截取的一个大小为$H\times W\times C$的矩阵；$\boldsymbol{F}^{k}$表示第 k 个大小为$H\times W\times C$的矩阵。3×3 卷积和 1×1 卷积都可以用式 (1) 来描述。

图 1 展示了 3×3 卷积运算。卷积以 3×3 为滑动窗口对输入的特征图$\boldsymbol{M}$进行处理，处理时，将中心区域(i,j)信息和窗口内的区域信息进行融合，输出的特征可作为中心区域(i,j)的更高层次的特征表示。可以看到，这种方式直接对窗口内的区域特征进行融合，没有考虑到每个区域的不同，这样就容易减弱中心区域的特性信息在更高层次特征中的表达，使得在 CenterNet 中对关键点的预测效果不佳。

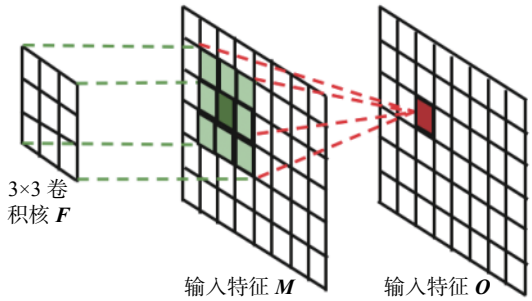
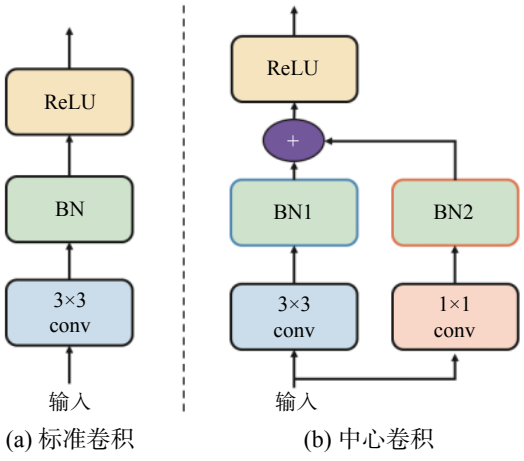


图 1　3×3 卷积运算

Fig. 1　3×3 convolution operation

为了解决上述问题，本文提出中心卷积（center-convolution）来增强中心区域的信息。中心卷积被设计为双分支的架构，其中一个分支是标准的 3×3 卷积后接一个批规范层（batch normalization, BN）[9]，用于融合领域的信息，另外一个分支是一个 1×1 的卷积旁路，同样后接一个批规范层，用于增强中心区域自身的信息。需要注意的是，两个分支学习到的信息是不一样的，所以学习到的特征空间中的分布不一样。因此，双分支的两个 BN 层的参数是不共享的。在 BN 层将不同信息进行规范化约束后，对两个分支进行加和操作得到最终的中心区域的特征。图 2 展示了标准的 3×3 卷积和所提出的中心卷积的差异。



(a) 标准卷积　　(b) 中心卷积

图 2　标准卷积和中心卷积

Fig. 2　Standard convolution and center-convolution

中心卷积尽管因为引入了旁路分支而增加了模型的复杂度，但实际上，所引入的 1×1 卷积以及 BN 层可以融合进 3×3 卷积中，因此，形式上就等价于标准的 3×3 卷积。具体来说，给定中心卷积，其中的 3×3 卷积设为$\boldsymbol{F}^{3\times3}\in\mathbf{R}^{3\times3\times C\times D}$，1×1 卷积设为$\boldsymbol{F}^{1\times1}\in\mathbf{R}^{1\times1\times C\times D}$，对于给定的输入特征中的一个 3×3 的分块$\boldsymbol{M}\in\mathbf{R}^{3\times3\times C}$，卷积层的输出$\boldsymbol{O}\in\mathbf{R}^{D}$为

$$\boldsymbol{O}_{k}=\sum_{i,j}^{3,3}\boldsymbol{F}_{(i,j,:,k)}^{3\times3}*\boldsymbol{M}_{(i,j,:)}+\boldsymbol{F}_{(0,0,:,k)}^{1\times1}*\boldsymbol{M}_{(1,1,:)} \tag{2}$$

式 (2) 可以很容易地将$\boldsymbol{F}^{1\times1}$融合到$\boldsymbol{F}^{3\times3}$中的 (1,1) 的张量上，得到新的3×3卷积$\boldsymbol{F}^{\mathrm{new}}\in\mathbf{R}^{3\times3\times C\times D}$：

$$\boldsymbol{F}^{\mathrm{new}}=\begin{cases}\boldsymbol{F}_{(i,j,:,:))}^{3\times3}, & i\neq1,j\neq1\\ \boldsymbol{F}_{(1,1,:,:)}^{3\times3}+\boldsymbol{F}_{(0,0,:,:)}^{1\times1}, & i=1,j=1\end{cases} \tag{3}$$

图 3 展示了中心卷积的融合过程。在模型训练时，本文将中心卷积设计为图 2 所示的双分支的结构，这样可以利用冗余的 1×1 旁路分支来增强模型对图像显著性特征的学习，提高网络的学习性能，而在推理阶段，利用图 3 所示的分支融合机制，可以很好地将冗余参数融合进主干 3×3 分支中，融合后的计算等价于标准的 3×3 卷积，不会提高模型的推理复杂度。

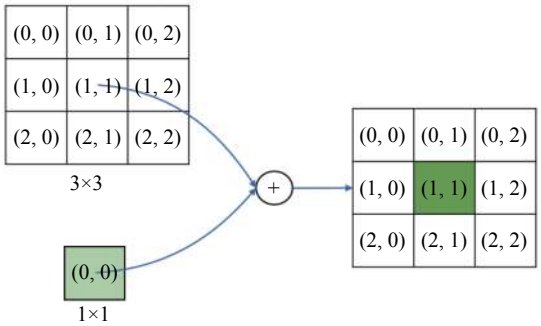


图 3　将 1×1 卷积融合到 3×3 卷积中

Fig. 3　Merging 1×1 into 3×3

2.2　基于中心卷积的 CenterNet 检测模型

本文选择 CenterNet 来构建检测模型。Cen-

terNet 抛弃了传统的基于锚点框（anchor）[10] 的对象检测方法，而是通过预测目标对象的中心点位置和大小来预测对象的包围框。

CenterNet 的结构包含两个部分。对于给定的一张图像$\boldsymbol{I}\in\mathbf{R}^{W\times H\times 3}$，首先使用主干网络对图像进行特征抽取，得到一个14×14的特征图，接着，使用反卷积扩大特征图分辨率，得到一个热力图张量$\widehat{\boldsymbol{Y}}\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$，其中 R 是下采样率，C 是类别数目。如果$\widehat{\boldsymbol{Y}}_{x,y,c}=1$表示点$(x,y)$是检测到的关键点，反之则是背景。

使用本文提出的中心卷积来构建 ResNet18 模型，并在 ImageNet[11] 上进行预训练。为了方便描述，本文将得到的模型命名为 C-ResNet18。并基于 C-ResNet18 卷积神经网络设计主干网络构建 CenterNet 对象检测模型，得到的模型称为 CR-CenterNet。为了对比，本文也基于 ResNet18 构建了 CenterNet，得到的模型称为 R-CenterNet。对比效果详见实验部分。

3　基于量化的深度模型部署

考虑到在自动驾驶车辆上，通常部署一些专用的终端芯片和设备来执行相关算法和模型。计算资源受限是该类设备的一大缺陷，这给复杂度高的智能模型的应用带来了巨大的挑战。面向终端设备的深度模型加速技术能够显著降低深度模型的复杂度，有利于深度模型在终端设备上的部署。

本文采用基于量化的模型加速方法，以最大化车载系统底层计算模块的计算效率。深度模型在训练过程中使用浮点精度来表示参数和数据的数值，从硬件原理来说，整数运算比相同位数的浮点运算更快且更省电，如果将深度模型的计算全部转化为整数运算，势必带来极大的加速效果。基于量化的模型加速方法就是通过将浮点（float32）精度数值量化到短型整数（int8）精度数值，实现了模型的高效运算。

假设要量化的浮点数是x，量化后的整数是y，两者的换算公式为：$x=S(y-Z)$，$y=\mathrm{round}\left(\frac{x}{S}+Z\right)$。其中，$S$ 表示 x 和 y 之间的比例关系，Z 表示量化后的整数 0 点，对应量化前的 0 点。它们的计算为

$$S=\frac{x_{\max}-x_{\min}}{y_{\max}-y_{\min}},\quad Z=\mathrm{round}\left(y_{\max}-\frac{x_{\max}}{S}\right)\tag{4}$$

式中：$x_{\max}$和$x_{\min}$分别表示 x 的最大值和最小值；$y_{\max}$和$y_{\min}$分别表示 y 的最大值和最小值；round 函数表示四舍五入运算。

基于上述的数值量化方法，可以很容易地将预训练好的深度模型进行参数量化。但是实验表明，这种后量化的方式存在量化误差累积的缺陷，使得量化后的模型性能产生了极大的损失。目前的很多深度学习框架，比如 pytorch[12] 和 TensorFlow[13] 都使用了一种量化感知的方法，将参数量化的过程融合进模型训练过程中，让网络参数能够适应量化带来的信息损失。本文采用同样的方法对深度模型进行处理。具体来说，将量化算法集成为一个模块，并串联在卷积参数层的后面参与特征计算，流程如图 4 所示。因为量化操作里面的 round 函数不是可导的，所以在反向传播的时候无法将梯度准确地传给前面层的参数。为了解决这个问题，本文重构了量化层的前向和后向操作，在前向时按照量化操作正常进行，反向时跳过量化层，直接把卷积层的梯度回传到量化前的卷积参数中来。因为卷积层的梯度是经过量化操作的，因此可以模拟量化误差，把误差的梯度回传到原来的参数上，使得原来的参数去自适应地感知量化产生的误差。本文的实验结果表明，量化感知训练的方式可以避免模型的性能下降，而其计算效率成倍提升。

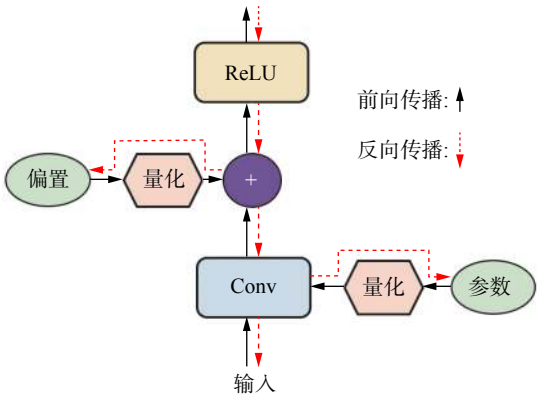


图 4　带量化层的卷积操作

Fig. 4　Convolution with the quantization layer

4　面向车规级芯片的检测验证系统

本文在车规级计算芯片上搭建了对象检测系统。为了达到这个目标，本文首先针对真实的自动驾驶场景收集了一批数据，并采用人工的方式进行数据标注。本文一共收集了两批数据：泊车数据和公路数据。泊车数据在室外泊车场景采集，共有 7 848 张 RGB 图像，每张图像标注了 2D 的包围框和物体接地点位置。如图 5 所示，红框是车辆的包围框，彩色点是车辆的接地点。公路数据是在公路场景采集的，由 22 213 张 RGB 图像组成，标注了 3D 的包围框。

(a) 泊车数据

(b) 公路数据

图 5　泊车数据和公路数据

Fig. 5　Parking data and road data

成本是智能算法在自动驾驶场景面临的一个重要的问题。自动驾驶场景复杂广泛，如果使用全监督的方式进行数据标注，标注成本将不可估量。其次，对象的 3D 信息是保障自动驾驶安全可行的一种重要数据，但其采集设备代价昂贵，不适合车辆量产。对此，本文设计了一套从 2D 图片到 3D 空间推断的对象检测流程。具体来说，本文用泊车数据中的 2D 包围框和接地点训练了所提出的基于中心卷积的对象检测模型 CR-CenterNet，然后利用输出的 2D 框和接地点信息，逆投影到世界坐标系中，得到对象的 3D 位置信息，即输入是 2D 图像，输出是 3D 位置信息（这里用鸟瞰图表示，如图 6 所示）。在训练阶段，CR-CenterNet 只在泊车数据上训练，测试时只在公路数据上进行测试。实验发现，尽管模型在训练过程中没有感知到公路数据上的对象信息，但是在公路数据上仍然具有较好的性能，说明模型具有较好的泛化能力。

图 6　自动驾驶的检测任务

Fig. 6　A detection task toward the autopilot

芯片的计算效率是限制深度模型在车载系统上应用的一个关键因素。本文采用常用的 tda4 芯片来搭建对象检测算法验证系统。首先将训练好的 CR-CenterNet 模型进行量化（如第 3 节内容所述），然后注入到 tda4 芯片上进行运算。整体的开发流程如图 7 所示。

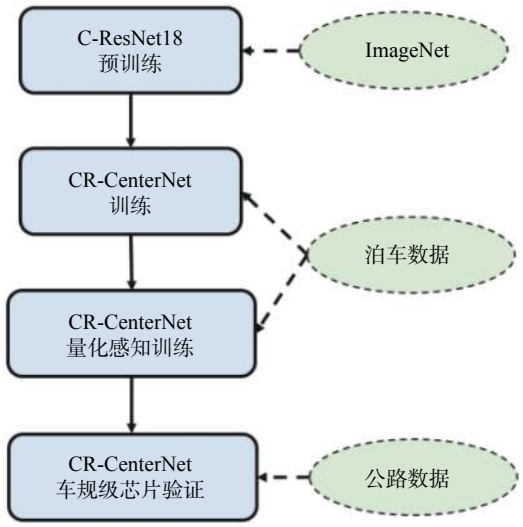


图 7　面向车规级芯片的检测系统构建

Fig. 7　Constructing the detection system on car-level chips

5　实验

5.1　中心卷积

在本节中，将验证所提出的中心卷积的有效性。ImageNet 是目前国际上主流的大型图像分类评测集，在人工智能的发展史上具有举足轻重的地位。AlexNet[14]、VGGNet[15]、GoogleNet[16]、ResNet[4] 和 DenseNet[17] 的成功都离不开 ImageNet，ImageNet 也已经成为计算机视觉领域的标准数据集。研究者通常会使用 ImageNet 来验证新提出的模型的有效性，然后在下游任务上进行验证，如对象检测[12,18]、实例分割[19-21]、行人重识别[22-24] 等。

因此，本文在 ImageNet 上对 C-ResNet18 进行了预训练，并在图像分类任务上展示其有效性。在训练时，本文采用批训练（batch）的方式来训练 C-ResNet18，每批次采样 256 张图片，并训练 120 轮。模型的训练学习率设置为 0.1，采用简单的数据增广策略，如随机裁剪和随机水平翻转。

为了展示所提出的中心卷积的有效性，本文把 ResNet18 作为基线模型，并和 C-ResNet18 进行对比。对比结果如表 1 所示。

表 1　在 ImageNet 上的分类准确率

Table 1　Classification accuracy on ImageNet　%

模型	top-1准确率	top-5准确率
ResNet18-torchvision	69.76	89.08
ResNet18-Ours	70.75	89.84
C-ResNet18	**71.59**	**90.32**

为了对比公平，这里展示了由 pytorch 官方提供的 ResNet18 结果（ResNet18-torchvision）和本文复现的 ResNet18 的结果（ResNet18-Ours）。从表 1 的实验结果可以看出，相比于 ResNet18-torchvision，本文复现的 ResNet18 获得了更高的分类准确率，原因是本文采用更好的训练技巧。而本文所提出的 C-ResNet 优于 ResNet18-torchvision 和 ResNet18-Ours，其中，相比于 ResNet18-torchvision，C-ResNet18 在 top-1 准确率上提高了 1.83%，在 top-5 准确率上提高了 1.24%；相比于 ResNet18-Ours，C-ResNet18 在 top-1 准确率上提高了 0.84%，在 top-5 准确率上提高了 0.48%。和 ResNet18-Ours 的性能对比，可以看出所提出的中心卷积对模型性能的增益效果显著，很好地证明了所提出的中心卷积的有效性。

5.2　CR-CenterNet

本节验证所提出的中心卷积在本文所构建的面向自动驾驶真实场景中的对象检测任务的应用效果。本文使用所构建的基于中心卷积的对象检测模型（CR-CenterNet，见 1.2 节）在泊车数据上进行模型训练，实现接地点的预测。尽管泊车数据上提供了 2D 框的标注数据，但本文致力于研究从 2D 数据中推断出对象的 3D 空间位置。因此，本文构建 CR-CenterNet 时只预测跟对象有关的接地点位置，然后使用逆投影算法[25] 推断对象的 3D 空间位置信息，并可视化真实场景的鸟瞰图。

考虑到对象检测模型的复杂度较高，因此，本文训练 CR-CenterNet 时采用分布式训练方式。具体来说，本文使用 3 张 2 080Ti 的英伟达显卡训练模型，每张卡在一个批次内训练 3 张图片，即批次大小是 9。本文设置 CR-CenterNet 的学习率为 3e-5。整体训练进行了 300 轮，并保存最后一轮的结果进行模型评测。

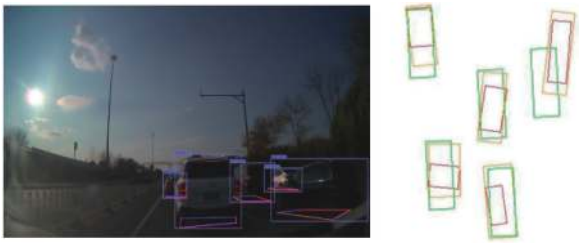
为了验证 CR-CenterNet 的性能，本文也基于 ResNet18 训练了 CenterNet 模型（R-CenterNet）。在模型评测时，本文计算鸟瞰图模式下，算法推断的包围框和人工标注的包围框的交并比，将交并比大于 0.5 的预测当作是正确的预测。表 2 展示了基线模型 R-CenterNet 和本文的 CR-CenterNet 的整体性能对比。

表 2　R-CenterNet 和 CR-CenterNet 的性能对比
Table 2　Performance comparison between R-CenterNet and CR-CenterNet　%

模型	精度
R-CenterNet	36.6
CR-CenterNet	**42.5**

从表 2 可以看出，CR-CenterNet 可以比 R-CenterNet 获得 5.9% 的提升，进一步证明了所提出的中心卷积在这个任务上的有效性。

图 8 展示了算法的对象检测效果，包括 CR-CenterNet 预测的接地点，以及逆投影后的鸟瞰图效果。图 8（a）中也展示了标注的包围框，图 8（b）中绿色的是标注的鸟瞰图矩形框，红色的是使用 R-CenterNet 得到矩形框，橙色的是使用 CR-CenterNet 得到的矩形框。可以看到，即使因为遮挡问题而无法从图像中直接看出接地点，所提出的 CR-CenterNet 也可以有效推断出对象的接地点；相比于 R-CenterNet，CR-CenterNet 对对象的 3D 空间位置推断（鸟瞰图）更加准确，也侧面说明了所提出的 CR-CenterNet 的有效性和优越性。



(a) 接地点　　(b) 鸟瞰图

图 8　对象检测效果可视化
Fig. 8　Visualization of detected objects

5.3　面向车规级芯片的模型验证系统

Tda4 芯片是一款由世界第三大半导体制造商德州仪器（TI）推出的面向新一代智能驾驶应用的车规级芯片，具有性能强、成本低、功耗低、安全性较高等优势，因此被许多汽车厂商和一级供应商选为计算平台。本文采用基于量化的方法成功将所提出的基于中心卷积的对象检测模型部署到该款车规级芯片 tda4 上。为了展示模型在 tda4 上的推理性能，在不同的计算平台上部署了本文提出的 CR-CenterNet，并测试了模型处理单张图片的时间开销。表 3 展示了各个平台上的时间开销对比情况。

表 3　不同平台上 CR-CenterNet 的时间开销
Table 3　Time consumption of CR-CenterNet in various platforms.

平台	配置	是否量化	时间开销/ms
服务器CPU	Xeon Gold 5 118	否	776
服务器GPU	2080Ti	否	**18.2**
笔记本CPU	i5-11300H	否	1 090
笔记本GPU	mx450	否	130
tidl模拟器	i5-11300H	否	>10 000
tda4	C7x DSP	是	**64**

本文选择了 6 种不同的计算平台，包括服务器端 CPU 和 GPU、笔记本 CPU 和 GPU、芯片模拟器和车规级芯片 tda4，对比结果如表 3 所示。可以看到，在服务器端，无量化版的 CR-CenterNet 在 CPU 上达到 776 ms 的时间开销，而量化版的 CR-CenterNet 在 tda4 上可以提升超 10 倍，时间开销下降到 64 ms。由此可见，量化版的深度模型可以在计算性能更受限的车规级芯片上取得比服务端高性能 CPU 更优的计算速率。

此外，本文所采取的量化方法在训练过程中进行，可以抵抗由于量化所导致的模型精度的骤降。表 4 展示了 CR-CenterNet 在量化前后的检测性能的对比，可以观察到，经过量化感知训练的模型在检测效果上跟量化前的模型差异不大（仅下降了 1.4%）。

表 4　量化对模型效果的影响

Table 4　Impact of the quantification method on the detection performance　%

量化	检测精度
无量化	42.5
量化	**41.1**

从上述分析可以看到，基于量化的部署方法可以提升模型的计算效率且保持模型的精度，满足了车规级芯片的计算需求，因此，本文将整套算法集成到支持 tda4 芯片的开发板上，形成一套面向车规级芯片的检测模型验证系统。图 9 展示了利用 tda4 进行计算得到的检测效果。可以看到，对于输入的 RGB 图片，所构建的车规级芯片验证系统能够准确地给出 3D 空间位置信息（鸟瞰图）。

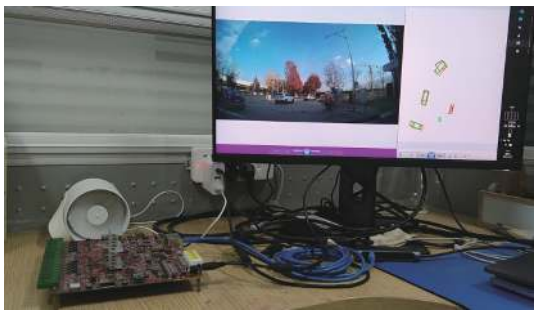


图 9　面向车规级芯片的对象检测模型验证系统

Fig. 9　Object detection model verification system for car-level chips

6　结束语

本文针对自动驾驶场景下智能模型计算效率要求高和终端设备计算资源受限的矛盾，提出了基于中心卷积的轻量化卷积神经网络和基于量化的深度模型部署方法。所提出的中心卷积在训练时为标准 3×3 卷积引入了 1×1 卷积旁路，可以增强模型对视觉信息的学习，而在推理时，可以方便地将旁路融合进 3×3 卷积中，减少了计算量且保持了和原来模型一样的性能。所采用的量化模型部署方法可以降低模型的大小，在保持量化前模型精度的情况下成倍提升模型的计算效率。基于轻量化中心卷积结构和量化技术，本文成功将深度对象检测模型在车规级芯片 tda4 上部署，在自动驾驶场景上取得了良好的检测性能。未来，有望集成到真实车辆驾驶系统中，在真实自动驾驶场景下发挥更大的作用。

参考文献：

[1] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of Annual Conference on Neural Information Processing Systems 2015. Montreal, Canada, 2015: 91−99.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 779−788.

[3] ZHOU Xingyi, WANG Dequan, KRÄHENBÜHL P. Objects as points[J]. arXiv:1904.07850, 2019.

[4] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770−778.

[5] LAW H, DENG Jia. Cornernet: detecting objects as paired keypoints[C]//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany, 2018: 734−750.

[6] ZHOU Xingyi, ZHUO Jiacheng, KRÄHENBÜHL P. Bottom-up object detection by grouping extreme and center points[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 850−859.

[7] LIU Zili, ZHENG Tu, XU Guodong, et al. Training-time-friendly network for real-time object detection[C]//Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence. New York, USA, 2020: 11685−11692.

[8] ZHOU Xingyi, KOLTUN V, KRÄHENBÜHL P. Probabilistic two-stage detection[J]. arXiv:2103.07461, 2021.

[9] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//Proceedings of the 32nd International Confer-en-

ce on Machine Learning. Lille, France, 2015: 448−456.

[10] 洪文亮. 基于改进的 Faster R-CNN 的目标检测系统的研究 [D]. 长春: 吉林大学, 2019.
HONG Wenliang. Research on systems of object detection based improved Faster R-CNN[D]. Changchun: Jilin University, 2019.

[11] DENG Jia, DONG Wei, SOCHER R, et al. Imagenet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 248−255.

[12] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high-performance deep learning library [C]//Proceedings of Annual Conference on Neural Information Processing Systems 2019. Vancouver, Canada, 2019: 8026−8037.

[13] ABADI M, BARHAM P, CHEN Jianmin, et al. TensorFlow: a system for large-scale machine learning[C]//Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. Savannah, USA, 2016: 265−283.

[14] KRIZHEVSKY ALEX, SUTSKEVER ILYA, HINTON GEOFFREY E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. Lake Tahoe, USA, 2012, 25: 1097−1105.

[15] KAREN SIMONYAN, ANDREW ZISSERMAN. Very deep convolutional networks for large-scale image re-cognition[EB/OL]. arXiv preprint, 2014. https://arxiv.org/abs/1409.1556.

[16] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1−9.

[17] HUANG Gao, LIU Zhang, LAURENS VAN DER MAATEN, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii, USA, 2017: 4700−4708.

[18] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2980−2988.

[19] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2261−2269.

[20] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Inter-vention. Munich, Germany, 2015: 234−241.

[21] 林开颜, 吴军辉, 徐立鸿. 彩色图像分割方法综述 [J]. 中国图象图形学报 (A 辑), 2005, 10(1): 1−10.
LIN Kaiyan, WU Junhui, XU Lihong. Survey of color image segmentation[J]//Journal of image and araphics (A), 2005, 10(1): 1−10.

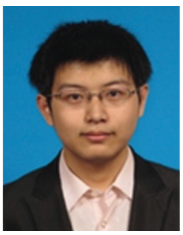
[22] ZHENG Liang, YANG Yi, HAUPTMANN A G. Person re-identification: past, present and future[J]. arXiv: 1610.02984, 2016.

[23] LUO Hao, GU Youzhi, LIAO Xingyu, et al. Bag of tricks and a strong baseline for deep person re-identification[C]//Proceedings of 2019 IEEE/CVF Conference on Com-puter Vision and Pattern Recognition Workshops. Long Beach, USA, 2019: 1487−1495.

[24] 宋婉茹, 赵晴晴, 陈昌红, 等. 行人重识别研究综述 [J]. 智能系统学报, 2017, 12(6): 770–780.
SONG Wanru, ZHAO Qingqing, CHEN Changhong, et al. Survey of person reidentification[J]. CAAI transactions on intelligent systems, 2017, 12(6): 770–780.

[25] 刘瑞芝, 孙士杰, 王菽裕, 等. 基于三维垂直逆投影面 的枚举车速检测算法 [J]. 电子设计工程, 2016, 24(14): 165–167.
LIU Ruizhi, SUN Shijie, WANG Shuyu, et al. Enumerated vehicle speed detection algorithm based on three-dimensional vertical back projection surface[J]. Electronic design engineering, 2016, 24(14): 165–167.

作者简介：



宫大汉，博士研究生，主要研究方向为轻量化深度模型结构设计和边缘设备智能推理引擎构建。

于龙龙，算法工程师，主要方向为嵌入式智能设备开发和模型部署。

丁贵广，副教授，博士，主要研究方向为多媒体信息处理、计算机视觉感知。获国家科技进步二等奖 1 项、人工智能学会科技进步奖一等奖 1 项、中国电子学会技术发明一等奖 1 项。主持和参与重点项目、重点研发项目等国家级项目数十项。发表学术论文近百篇。