DOI:10.19651/j.cnki.emt.2106215

基于注意力融合网络的 RGB-D 目标检测算法

朱书勤
(北京交通大学 软件学院 北京 100044)

摘 要：针对当前利用 RGB-D 图像进行目标检测出现的网络融合不充分和检测效率不高等问题，提出一种基于注意力机制的特征逐级融合网络结构。首先在基于 Yolo v3 的 Backbone 网络结构下，分别用标注好的 RGB-D 样本分别训练 RGB 和 Depth 网络，然后通过注意力模块增强两种特征，最后在网络中期逐层融合得到最终的特征权重。在具有挑战性的 NYU Depth v2 数据集上测试，得到本文方法的均值平均精度为 77.8%。通过对比实验得出，所提出的基于注意力机制的融合网络较同类算法性能有了明显提升。

关键词：目标检测；卷积神经网络；RGB-D 图像；注意力机制
中图分类号：TP75 **文献标识码**：A **国家标准学科分类代码**：520.2

RGB-D target detection algorithm based on attention fusion network

Zhu Shuqin
(School of Software Engineering, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Aiming at the problems of insufficient network fusion and low detection efficiency in current target detection using RGB-D images, a feature-level fusion network structure based on attention mechanism is proposed. First, under the backbone network structure based on Yolo v3, the RGB and Depth networks are trained separately with the labeled RGB-D samples, and then the two features are enhanced by the attention module, and finally the final feature weights are obtained by layer-by-layer fusion in the middle of the network. Tested on the challenging NYU Depth v2 data set, the average accuracy of the method in this paper is 77.8%. Through comparative experiments, it is concluded that the fusion network based on the attention mechanism proposed has significantly improved performance compared with similar algorithms.

Keywords: target detection; convolutional neural network(CNN); RGB-D image; attention mechanism

0 引 言

近年来，随着新型 RGB-D 传感器 Kinect、Xtion 等的出现，RGB-D 图像在图像识别领域的运用有了极速的发展。新型 RGB-D 传感器可以将图像的深度信息与彩色信息相融合，从而为解决计算机视觉基本问题提供新的思路。图像的深度信息可以在彩色信息的基础上，增加目标检测的额外信息，尤其适用于环境较为复杂的场景。同时图像的深度信息不随亮度、颜色等参量的变化而变化，提高了目标检测的稳定性。目前，将深度信息应用于目标识别算法已成为计算机视觉领域中的一个热门研究方向，被广泛应用于室内智能机器人的视觉识别与场景交互[1-3]。

文献[4]提出了一种基于 RGB-D 的显著性目标检测方法，将 RGB 图像高层特征和下层特征输入到卷积神经网络中，在卷积层根据最优权值对两个网络进行特征融合，在平均绝对误差和 F 测量上的结果优势较为突出。文献[5]提出的基于混合结构的 RGB-D 目标识别算法中，分别利用深层卷积神经网络(VGG-16)和改进 HONV 描述子的 Fisher Vector 编码对每一帧的 RGB 图像与深度图像提取特征表达，取得了比较优异的识别性能。文献[6]提出了一种新的双流卷积网络，将 RGB 图像和深度图像分别输入到两个结构相同的卷积网络中提取各自特征后，在卷积层根据最优权值对两个网络进行特征融合，最后通过全连接层得到输出。文献[7]提出的基于层次交替交互网络的 RGB-D 显著性目标识别算法中，通过分层交互模块(HAIM)不仅可以过滤图像深度信息中的干扰，还可以纯化深度以提高 RGB 效果，在提高检测精度的基础上同时提高了深度信息的质量。文献[8]提出基于多信息融合的 RGB 场景识别方法，通过多功能融合分类器为 RGB-D 数据描述最佳功能配置，并通过修正检测方法在保证检测精度的基础上提高了训练集的训练速度。

收稿日期：2021-03-29

注意力机制(attention mechanism)源于对人类视觉的研究。在计算机视觉领域,让机器像人类一样选择性地关注可见信息的重要部分,同时忽略其他无关信息,这种机制被称为注意力机制[9]。在遥感变化检测任务中,更多关注两幅影像发生变化的区域,因此加强变化区域的特征更有助于提高检测效率。近年来,很多研究者都在变化检测网络中添加不同类型的注意力模块,主要包含空间注意力和通道注意力,空间注意力的作用是增大了变化与未变化像素之间的距离差异值,通道注意力的作用是放大与地物变化相关的通道,抑制无关的通道。文献[10-12]都提出了不同的注意力机制用于增强图像中目标的显著特征,从而改善检测结果。

本文针对当前 RGB-D 网络融合不充分、检测效率不高的问题,提出一种基于注意力机制的融合网络来实现 RGB-D 目标检测。在 NYU Depth v2 数据集[13]上同时训练 RGB 和 Depth 网络,并在网络中间层将两者逐级融合,最后通过测试融合后网络的效果,对结果进行了比较评估和定量评价。

1　相关工作及方法

1.1　Backbone 网络结构

卷积神经网络是图像识别检测领域优秀的深度学习模型[14]。本文以当前流行的 Yolo v3[15]目标检测框架为基础,在原有的 Darknet53 网络上引入了注意力模块,分别训练 RGB 和 Depth 两类图像,在中间层将两者特征进行融合。

网络的训练过程可分为前向传播和后向传播两个阶段,在前向传播阶段,信息从输入层开始向前逐层传播,经过第 l 卷积层中第 j 个神经元输出的计算公式为:

$$a_j^l = f_c[w^l(\sum_{i\in M_j^l} a_i^{l-1} * k_{ij}^l) + b^l] \tag{1}$$

式中:a_i^{l-1} 为上一层的输出;f_c 为 ReLU 激活函数;k 为卷积核;M 为选择的输入特征图的集合;w^l 为卷积网络第 l 层的权重;b_l 为网络第 l 层的偏置;$*$ 代表卷积运算。后向传播的实质是根据损失函数的变化情况迭代调整网络的权重和偏置,从而获得最优网络参数。Yolo 算法中 loss 函数的定义为预测数据与标定数据之间的坐标定位误差、IoU 误差和分类误差三项的和。公式如下:

$$\begin{aligned} loss = {} & \lambda_{coord}\sum_{i=0}^{s^2}\sum_{j=1}^{B} I_{ij}^{obj}[(x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2] + \\ & \lambda_{coord}\sum_{i=0}^{s^2}\sum_{j=1}^{B} I_{ij}^{obj}[(\sqrt{w_i}-\sqrt{\hat{w}_i})^2+(\sqrt{h_i}-\sqrt{\hat{h}_i})^2] - \\ & \sum_{i=0}^{s^2}\sum_{j=1}^{B} I_{ij}^{obj}[\hat{C}_i^j\log(C_i^j)+(1-\hat{C}_i^j)\log(1-\hat{C}_i^j)] - \\ & \lambda_{noobj}\sum_{i=0}^{s^2}\sum_{j=1}^{B} I_{ij}^{noobj}[\hat{C}_i^j\log(C_i^j)+(1-\hat{C}_i^j)\log(1-\hat{C}_i^j)] - \\ & \sum_{i=0}^{s^2} I_{ij}^{obj}\sum_{c\in classes}[\hat{P}_i^j\log(P_i^j)+(1-\hat{P}_i^j)\log(1-\hat{P}_i^j)] \end{aligned} \tag{2}$$

式中:λ 表示权重,$\lambda_{coord}=5$,$\lambda_{noobj}=0.5$;x,y,w,h,c,p 为网络预测值;$\hat{x},\hat{y},\hat{z},\hat{h},\hat{c},\hat{p}$ 为标注值;I_i^{obj} 表示物体落入格子 i 中;I_{ij}^{obj} 和 I_{ij}^{noobj} 分别表示物体落入与未落入格子 i 的第 j 个边界框内。

1.2　注意力机制

在目标检测过程中,并不是所有高维特征都对目标判别有帮助,不相关的特征反而会使得检测更加困难,因此本文提出一种注意力融合模块用来增强有用信息,抑制无关信息。如图 1 所示,将输入的特征分别进行通道注意力和空间注意力操作后,将其通过逐元素相加的方式融合,得到加强后的特征图层,以更好地进行目标检测任务。

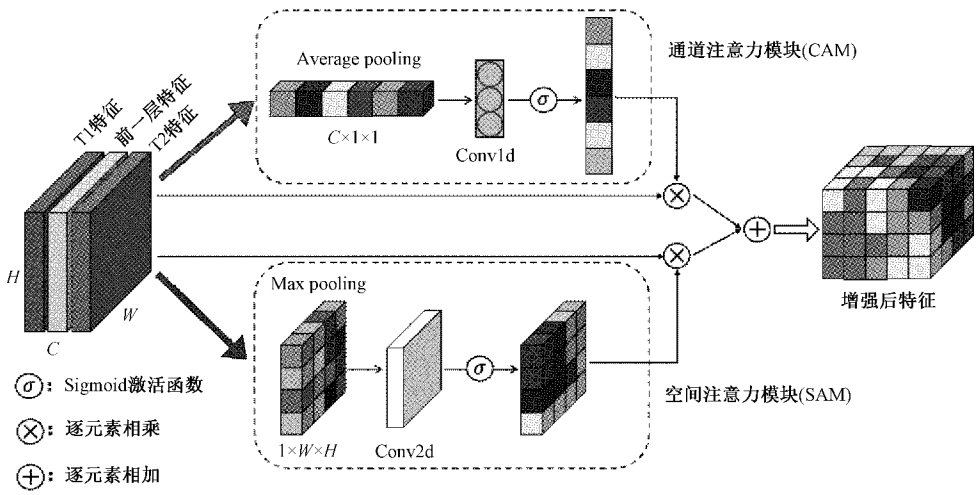


图 1　注意力融合模块内部结构

1.3　双通道网络融合方法

对 RGB-D 图像进行联合检测时,主要通过卷积神经网络对 RGB 和 Depth 图像的信息进行融合。当前对于 RGB-D 信息融合较为常见的简单模式为早期和后期融合[16],而在

全连接层之前的中期融合会取得更好的识别效果。本文在文献[17]的基础上提出中期逐层融合的方法(如图 2 所示)，将 Depth 网络每一个卷积层的输出以不同的权重融合到相应 RGB 网络的特征图层中，融合之前分别添加各自的注意力模块用来增强各自特征，之后作为下一层的输入，进入 RGB 网络的全连接层中识别场景的物体类别。

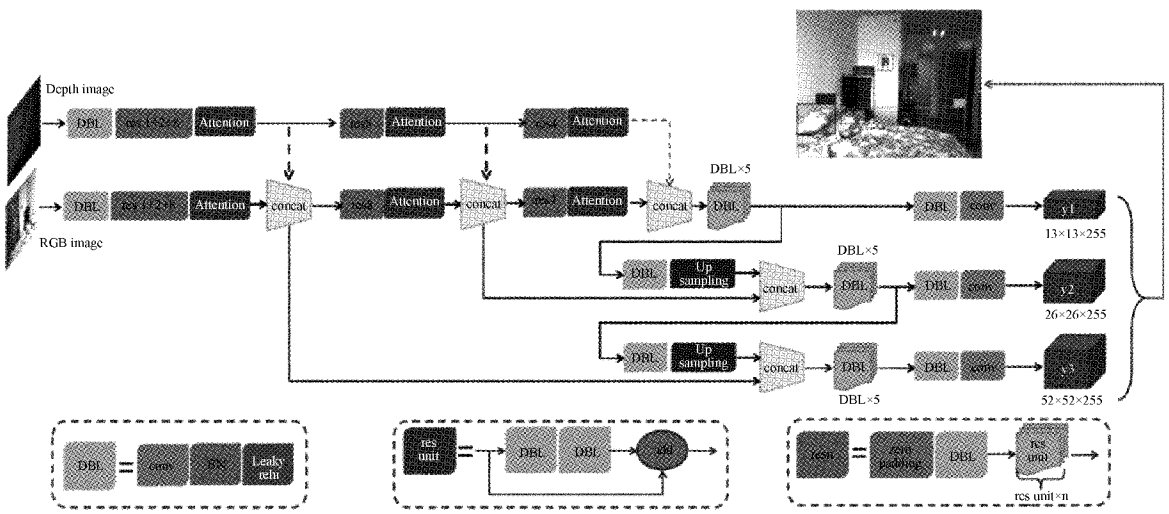


图 2　RGB-D 目标检测网络模型结构

设网络中第 l 层中第 i 个神经元为 a_i^l，M_r 和 M_d 分别是从 RGB 网络和 Depth 网络中选择的输入特征图的集合，根据式(1)可得到融合层的第 j 个神经元计算公式：

$$a_j^l = f_{concat}\{[w_r^l \sum_{i\in M_{rj}^l}(a_i^{l-1} * k_{ij}^l)+b_r^l],[w_d^m \sum_{i\in M_{dj}^m}(a_i^{m-1} * k_{ij}^m)+b_d^m]\} \tag{3}$$

式中：w_r^l 和 b_r^l 分别为 RGB 网络第 l 卷积层的权重和偏置；w_d^m 和 b_d^m 分别为 Depth 网络第 m 卷积层的权重和偏置；f_{concat} 为特征级联操作。

2　实验研究

2.1　实验环境及数据预处理

实验采用深度学习框架 pytorch 进行模型训练、特征融合与物体识别，在搭载 Titan X 显卡和 CUDA 9.0 GPU 驱动上运行。工作站 CPU 为 Inter(R) Xeon(R) CPU E5-2695 v2@2.4 GHz，内存为 192 G。实验使用 NYU Depth v2 数据集，其中包含 26 种室内场景类型，1 000 多个类别和 1 449 张补全深度的 RGB-D 图像。

为了在 NYU Depth v2 数据集上实现目标识别任务，重新用检测框手动标记原始数据，制作训练样本。为了降低时间成本，课题组共标记了 500 张 RGB-D 图像中的 10 个类别，并且另选 80 张图像作为测试集。

2.2　训练过程及可视化

实验采用批量标准化操作(BN)，每次迭代 64 张图片进行训练，共迭代 30 200 次。训练阶段采用动量项为 0.9 的随机梯度下降，权值的初始学习率为 0.001，衰减系数设为 0.000 5。分别单独训练 RGB 和 Depth 网络参数，训练过程将 loss，IoU 和 Recall 三项指标可视化(如图 3 所示)，用来评价模型的性能，其中 IoU 和 Recall 计算公式如下：

$$IoU = \frac{\text{area}(BB_{dt} \cap BB_{gt})}{\text{area}(BB_{dt} \cup BB_{gt})} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

其中，area 表示面积；BB_{gt} 为训练的参考标准框(ground truth box)；BB_{dt} 为检测边界框(detection truth box)；TP(true positive)表示被判定为正样本，事实上也是正样本的个数；FN(false negative)表示被判定为负样本，但事实上是正样本的个数。

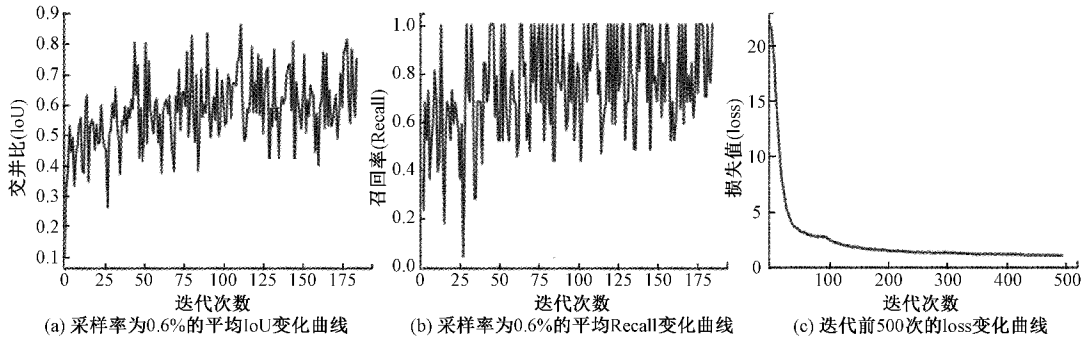
为了改善可视化效果，图 3 中 loss 曲线的采样率为 10%，IoU 和 Recall 曲线采样率均为 0.6%，在迭代 30 200 次后 IoU 曲线值最终平均稳定在 0.75，Recall 曲线值最终平均稳定在 0.89。图 3(c)所示为迭代前 500 次的 loss 变化曲线，可以看出在训练前 100 个迭代周期内 loss 值迅速下降，之后变化极为缓慢，最终稳定在 0.32。

2.3　实验结果评价比较

1)定性评价

根据本章提出的 RGB-D 特征逐级融合策略，在经过训练的融合网络中对 NYU Depth v2 数据集的 80 张测试图像进行目标检测。为了检验 RGB-D 网络检测性能的优劣，又单独在原始 Yolo v3 架构上训练了 RGB 和 Depth 网络，并将 3 种检测结果做了定性比较，如图 4 所示。

实验选取了 5 张不同场景的图像，图 4 中第 1 列为 RGB 网络的检测结果，中间列为 Depth 网络的检测结果，第 3 列为本章提出的 RGB-D 逐级融合网络的检测结果。可以看出，RGB-D 图像的检测结果最优，有效降低了误检和漏检的概率，例如在图 4(c)中的 people 和 desk，图 4(l)



(a) 采样率为0.6%的平均IoU变化曲线　(b) 采样率为0.6%的平均Recall变化曲线　(c) 迭代前500次的loss变化曲线

图 3　网络训练过程各项指标可视化

(a) 场景1中RGB网络检测结果　(b) 场景1中Depth网络检测结果　(c) 场景1中RGB-D融合网络检测结果

(d) 场景2中RGB网络检测结果　(e) 场景2中Depth网络检测结果　(f) 场景2中RGB-D融合网络检测结果

(g) 场景3中RGB网络检测结果　(h) 场景3中Depth网络检测结果　(i) 场景3中RGB-D融合网络检测结果

(j) 场景4中RGB网络检测结果　(k) 场景4中Depth网络检测结果　(l) 场景4中RGB-D融合网络检测结果

(m) 场景5中RGB网络检测结果　(n) 场景5中Depth网络检测结果　(o) 场景5中RGB-D融合网络检测结果

图 4　NYU Depth v2 数据集中 5 个场景的 RGB-D 目标检测结果

中的 bed，图 4(o)中的 chair 等。在 RGB 图像的检测结果中，对于光线昏暗和表面纹理模糊的物体容易出现漏检情况(如图 4(d)中 tv 和 cabinet)；在 Depth 图像的检测结果中，对于边缘轮廓被遮挡或视线远处的物体，由于其深度信息较为模糊，容易被漏检或误检(如图 4(b)中的 people 被漏检，图 4(n)远处的 chair 被误检等)。在 RGB-D 图像的检测结果中可以看出，RGB-D 特征逐级融合网络能够互补彩色和深度两种信息的各自特征，实现检测结果准确率和召回率的有效提升。

2)定量分析

为了验证本文方法的识别效果，除了与利用实验中训练好的 RGB 和 Depth 网络单独识别的结果对比之外，还与文献[5]和[6]中的方法在不同类别的平均精度(AP)上做了对比，结果如图 5 所示。可以看出本文提出的 RGB-D 融合方法较基于 RGB 和 Depth 单独图像的平均精度有所提高，同时也优于文献[5]和[6]所提方法。表 1 所示为所有类别的均值平均精度(mAP)和检测时间，可以看出本文方法的 mAP 在 NYU Depth v2 数据集上达到 77.8%，高于同类方法的测试结果，但由于网络结构较为复杂，检测时间也相对较长。

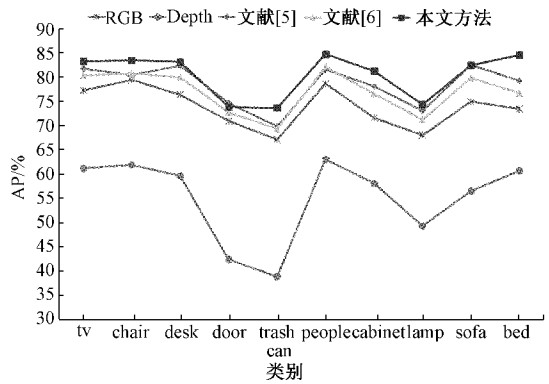


图 5　不同方法的检测结果对比

表 1　不同检测方法的 mAP 与时间比较

	RGB	Depth	文献[5]	文献[6]	本文方法
mAP/%	69.9	50.1	73.8	73.0	77.8
时间/ms	41.6	39.7	452.0	326.4	526.7

3　结　　论

本文设计了一种基于注意力机制的双通道融合网络，用于 RGB-D 图像的目标检测任务。在 Yolo v3 框架的基础上，先分别训练 RGB 和 Depth 两种卷积神经网络，利用注意力机制增强两者各自的特征，在网络中期将该两种特征逐级融合，通过特征互补最终实现对 RGB-D 图像中目标的有效识别。

通过对实验结果的对比分析，发现本文提出的方法可以在 NYU Depth v2 数据集实现 RGB-D 物体的有效检测，并且与文中同类型的两种方法相比，检测结果的均值平均精度(mAP)明显提高。但是由于中期逐级融合的方法复杂度较高，计算量较大使得检测速度较慢。接下来的工作将重点围绕压缩模型结构、提高检测速度方面展开研究。

参考文献

[1]　李润顺. 基于深度学习的三维目标识别算法研究[D]. 大连：大连理工大学，2017.

[2]　孙宁嘉，于纪言，王晓鸣. 适用于复杂场景的多目标跟踪算法[J]. 仪器仪表学报，2019，40(3)：129-140.

[3]　张培培，王昭，王菲. 基于深度学习的图像目标检测算法研究[J]. 国外电子测量技术，2020，39(8)：42-47.

[4]　HUANG R，XING Y，WANG Z. RGB-D salient object detection by a CNN with multiple layers fusion[J]. IEEE Signal Processing Letters，2019，26(4)：552-556.

[5]　李威. 基于特征学习的 RGB-D 目标识别算法研究[D]. 武汉：华中科技大学，2016.

[6]　刘帆，刘鹏远，张峻宁，等. 基于双流卷积神经网络的 RGB-D 图像联合检测[J]. 激光与光电子学进展，2018，55(2)：380-388.

[7]　LI G，LIU Z，CHEN M，et al. Hierarchical alternate interaction network for RGB-D salient object detection[J]. IEEE Transactions on Image Processing，2021，30：3528-3542.

[8]　GONG W，ZHANG B，LI X. An efficient RGB-D scene recognition method based on multi-information fusion[J]. IEEE Access，2020，8：212351-212360.

[9]　林羽晨，张金艺，秦政，等. 融合双重注意力机制的复合头部动作识别[J]. 电子测量技术，2020，43(11)：85-90.

[10]　WANG Q，WU B，ZHU P，et al. ECA-Net：Efficient channel attention for deep convolutional neural networks [C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition，2020，DOI：10.1109/CVPR42600.2020.01155.

[11]　QIN Z，ZHANG P，WU F，et al. FcaNet：Frequency channel attention networks[J]. ArXiv Preprint，2020，ArXiv：2012.11879.

[12]　LI X，WANG W，HU X，et al. Selective kernel networks[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition，2020：510-519.

[13]　SILBERMAN N，HOIEM D，KOHLI P，et al. Indoor segmentation and support inference from RGBD images [C]. European Conference on Computer Vision. Springer，Berlin，Heidelberg，2012，DOI：org/10.1007/978-3-642-33715-4_54.

[14]　LU H T，ZHANG Q C. Overview of application of depth convolutional neural network in computer vision[J]. Data Acquisition and Processing，2016，31(1)：1-17.

[15] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. ArXiv E-prints, 2018, ArXiv:1804.02767.

[16] TU S Q, XUE Y J, LIANG Y, et al. RGB-D image classification methods[J]. Laser & Optoelectronics Progress, 2016, 53(6): 060003.

[17] HAZIRBAS C, MA L, DOMOKOS C, et al. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture[C]. Asian Conference on Computer Vision. Springer, Cham, 2016, DOI: 10.1007/978-3-319-54181-5_14.

作者简介

朱书勤,学士学位在读,主要研究方向为人工智能与图像识别等。

E-mail:zsqcool3@163.com