计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(10):226−231 [doi: 10.15888/j.cnki.csa.006600]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于迁移学习的商品图像检测方法①

胡正委, 朱　明

(中国科学技术大学 信息科学技术学院, 合肥 230031)
通讯作者: 胡正委, E-mail: zhengwei_hu@163.com

摘　要: 近年来, 对象识别方法被应用到多个领域. 如人脸检测, 车辆检测. 然而模型训练所需要的边框标定需要很大的工作量. 本文通过基于迁移学习的方法, 将物体检测任务迁移到商品检测, 且不需要边框标定. 本文在分类层和边框回归层之间建立关系层, 来学习两种任务之间的关联. 本文建立了一个商品数据集, 并提出了一种深度学习训练方法, 解决了可旋转物体的检测问题. 基于 Faster RCNN 框架, 本文提出一种候选选择方法, 可以在无边框标定情况下训练商品分类. 本文提出的商品检测方法不需要边框标定, 而且很容易训练并应用到其它数据集.

关键词: 物体检测; 迁移学习; 关系层; 深度学习训练方法; 边框标定

引用格式: 胡正委,朱明.基于迁移学习的商品图像检测方法.计算机系统应用,2018,27(10):226–231. http://www.c-s-a.org.cn/1003-3254/6600.html

Product Image Detection Based on Transfer Learning

HU Zheng-Wei, ZHU Ming

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230031, China)

Abstract: In recent years, object detection is transferred to other fields, for example, face and vehicle detection. However, the bounding-box labeling is a huge resources cost work. This study solves the problem that transfer object detection task to other domain dataset without bounding-box label. A relationship layer is built to learn the relationship between classification and regression task. In addition, we construct a product dataset, on which rotatable object detection is solved using our training method. A proposal selecting method is proposed for training classification based on faster RCNN framework without bounding-box label. We propose a object detection method without bounding-box annotation. The method is easy to transfer to other datasets and training.

Key words: object detection; transfer learning; relationship layer; training method; bounding-box label

对象检测方法, 从图像分类和定位扩展而来. 近年来受到许多研究者的关注, 如人脸识别, 车辆检测等。对象检测是针对图像中包含多个物体时, 对其进行识别和定位. 而图像分类和定位针对图像中仅包含单个物体的情况. 在深度学习广泛应用之前, 对象检测效果最好的方法是可变形组件模型[1]. 而近年来最成功的方法主要包括: 两级方法和单级方法. 其中单级方法有 YOLO[2]和 SSD[3]. 两级方法包括基于区域的快速卷积神经网络 (Faster RCNN)[4]和其扩展方法. 其中类似 Faster RCNN 的两级方法具有较高的准确率.

通过迁移学习, 卷积神经网络可以被应用在很多领域. 可以理解为, 通过某一任务学习到的知识可以被迁移到另外一个任务[5]. 迁移学习不仅节约了数据训练的时间, 也能有效防止深度学习网络的过拟合. 本文使用 Mask RCNN[6]网络来进行商品检测. 在 ImageNet 数据集训练的全卷积网络 (FCN), 被用来解决分割问题[7].

① 基金项目: 中科院先导专项课题 (XDA06011203)
Foundation item: Strategic Priority Program of Chinese Academy of Sciences (XDA06011203)
收稿时间: 2018-03-15; 修改时间: 2018-04-18; 采用时间: 2018-04-20; csa 在线出版时间: 2018-09-28

在 ImageNet 上训练的 VGG[8]模型被用来进行人脸识别. 所有这些方法, 都需要在新的数据上重新调整模型. 但是, 在实际场景中, 数据的标定是非常难的工作, 需要消耗很大的财力物力和人力[9]. 对于物体检测任务的迁移学习方法, 研究者需要进行边框标定来训练边框回归任务.

本文提出的方法，可以不需要边框标定来进行物体的检测. 我们基于 Faster RCNN 框架提出一种不需边框的标定的分类任务训练方法. 并通过在分类任务和边框回归任务之间添加关系层, 来学习二者之间的关联, 并在 COCO[10]数据集上进行关系层的训练. 在商品数据集的分类任务训练完成之后, 可以通过训练完成的关系层直接推断边框回归, 并无需进行边框回归的训练. 除此之外, 我们构建了一个商品数据集, 并且解决了可旋转商品的检测问题.

1 基于 Faster RCNN 的商品分类

Faster RCNN 是一个两级物体检测方法, 包括区域候选网络 (RPN) 和头网络 (Network Head), 如图 1 所示. 其中区域候选网络为头网络生成区域候选集和特征. 分类层和回归层是特征提取之后的两个并行的分支. 通常利用 Faster RCNN 进行物体检测的迁移学习, 至少需要对分类层的全连接层 (FC layer) 和回归层的全连接层进行在训练. 如果需要, 前面的卷积层也要训练. 但是, 在实际应用中对物体边框的标定需要很大的工作量. 本文提出一种方法, 可以学习分类层与回归层之间关系, 从而根据分类层可以直接推导出回归层, 而且不需要边框标定数据的训练.

我们在公开的 COCO 数据集上进行预训练工作,其中包含 80 种常见物体类别. 并具有 2 500 000 个标签示例和 328 000 张图像. 我们还在本地数据集进行了验证, 具体描述在第 3 章实验部分.

和传统的机器学习一样, 训练数据和测试数据应该服从相同的分布. 若将关联层从 COCO 数据集迁移到本地商品数据集上，就需要 COCO 数据的分类层参数和商品数据的分类层参数满足相同分布. 即分类层应该在 COCO 数据集和本地商品数据集联合训练.

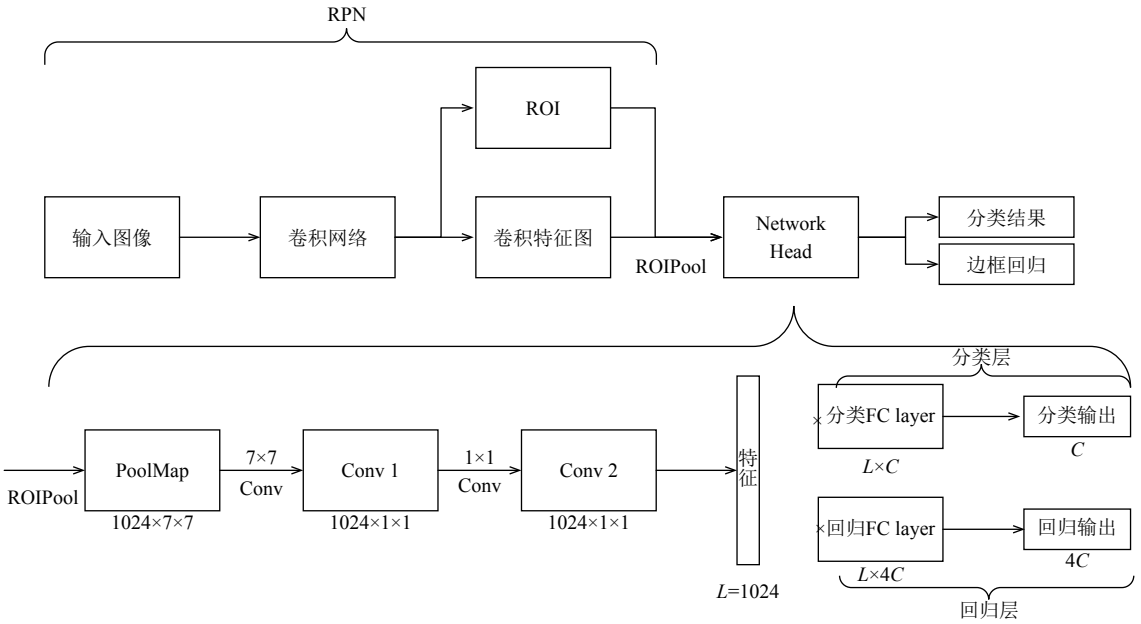

图 1 Faster RCNN 框架

1.1 候选选择方法

Faster RCNN 的分类层是根据真实的边框选择的 ROI(兴趣候选) 区域来训练的. 训练得到的 RPN 网络用来判断一个候选区域是物体还是背景, 是一个二分类模块.

通过利用 COCO 数据集训练 Faster RCNN 之后,我们发现其中的 RPN 网络可以在我们的本地数据集上用来生成候选. 当输入图片只包含一个物体时, 最大概率的候选框基本上包含真实的物体. 所以在没有边框标定的情况下, 我们用这个候选框作为真实的候

选框来训练分类层. 即使选出来的候选框含有噪声, 也不影响分类层的训练[11,12].

我们构建的商品数据集较小, 但是在应用样本扩展方法之后, 边框的标定将会是一个巨大的工作量. 所以候选选择方法非常重要.

1.2 训练分类层

我们发现, 由于我们构建的商品数据集的特殊性.其中的商品对象是可旋转的, 这点和 COCO 数据集不同. 只训练分类层的全连接层不能使网络正确的识别图像.

为了使整个网络具有旋转不变性, 我们利用样本扩展之后的数据训练了全连接层之前的卷积层. 样本扩展方法包括随机旋转、平移、放缩. 训练之后的网络可以正确的分类出几乎所有的图片.

1.3 联合训练

为了确保两个数据集的分类层参数服从同一分布，需要对两个数据集进行分类层的联合训练. 因此, 首先在本地商品数据集上训练网络的分类全连接层和之前的两个卷积层. 然后保持卷积层参数不变, 在 COCO 数据集和本地商品数据集上同时训练分类全连接层.

2 关系层学习

在完成分类全连接层的训练之后, 需要对分类层和回归层之间的关系层进行训练.

通常的想法是在分类输出之后添加一层全连接层,其输出作为回归输出. 但是分类层的输出是特征降维的结果, 损失了一部分信息. 而且不同数据集的类别数会有不同的输出维度. 所以不能直接在分类输出后直接添加全连接层.

所以我们在分类层的参数后添加一个全连接层来生成回归层的参数.

如图 2 所示, 分类层和回归层之前的特征维度为 $L(1024)$. 分类类别数为 C, 回归层的输出维度为 $4C$. 所以 Faster RCNN 的分类层参数和回归层参数大小分别为 $L\times C$ 和 $L\times 4C$. 公式 (1) 中w_{cls}表示分类层参数, 参数大小为 $L\times C$, 公式 1 中w_{reg}是回归层参数, 参数大小为 $L\times 4C$. 我们的目的是通过 $L\times C$ 大小的分类层参数生成回归层参数. 其中 L 是常量, C 是数据集中类别数目, 是可变的. 关系层模型可以定义为:

$$w_{reg} = r(w_{cls}^{\mathrm{T}} w_{rel}) \tag{1}$$

我们构建了名为关系层的全连接层, 在公式 (1) 中用w_{rel}表示, 其参数大小为 $L\times 4L$. 其接受输入为 $C\times L$ 的数据, 生成 $C\times 4L$ 大小的输出. 因此, 分类层的参数需要变换为 $C\times L$ 大小. C 可以看做输入样本数量, L 是输入特征维度, $4L$ 是输出特征维度. 其输出需要通过r函数变为 $L\times 4C$ 大小. 即使类别数 C 改变, 关系层的网络结构也不需要改变. 公式 (2) 定义了回归结果:

$$result_{reg} = fv \cdot w_{reg} \tag{2}$$

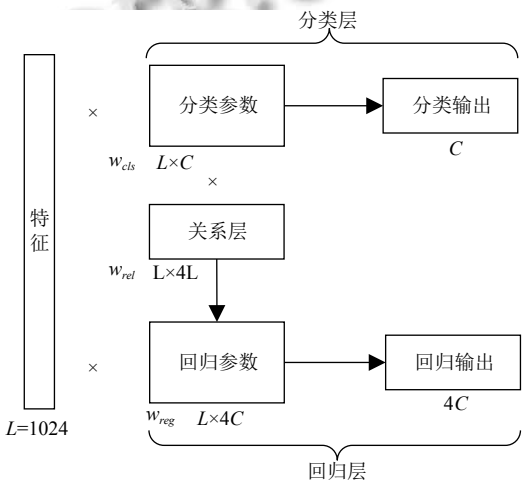

图 2 关系层模型

$N\times 4C$ 大小的边框回归结果是通过直接将 $L\times 4C$ 大小的参数w_{reg}与 $N\times L$ 大小的特征fv相乘. 其中 N 表示 ROI 区域的数量. 我们只需要在 COCO 数据集上进行关系层的训练.

3 实验分析

3.1 数据集

本文在构建的商品数据集和已经公开的货架商品数据集[13]上验证了提出了方法. 货架商品数据集是关于香烟包装的货架. 本文提出的数据集如图 3 所示. 利用本文的方法. 我们不需要在商品数据集上训练边框回归. 所以本文构建的商品数据集训练图像仅包含类别信息. 数据集中共包含 3200 张训练图像和 1600 张测试图像，共计 40 个商品类别. 训练图像通过 2 个摄像头在 4 个不同的视角下拍摄的, 测试图像通过另一个摄像头在同样的 4 个视角下拍摄的. 每张图像只有一个商品对象, 用来进行分类任务的训练. COCO 数据

集共 80 个类别, 并包括非常多的图片以及边框和类别标注.

本文构建的商品数据集和 COCO 数据集主要的区别在于, 本文商品数据集中的物体时可旋转的, 且训练数据远远少于 COCO 数据集. 这无疑增加了分类任务的难度.

图 3 本文构建的商品数据集训练图片示例

3.2 分类和检测结果

在实验中, 我们使用 Faster RCNN 的扩展版 Mask RCNN, 其除 Faster RCNN 方法外利用了特征金字塔网络 (FPN) 和兴趣区域对其 (ROIAlign) 方法[6]. 实验中不需要进行分割, 因此我们移除了 Mask RCNN 中的分割分支, 只使用其分类和回归分支, 并用本文提出的关系层将二者关联起来, 并学习二者的关系. 而且回归层参数不需训练, 只利用分类层参数和关系层参数即可推导出来. 实验中，Mask RCNN 使用 Python 语言并基于 Keras 和 Tensorflow[14]深度学习框架实现.硬件使用 Intel Core i7-7800x CPU 与 2 个 NVIDIA TITANX GPU 上进行, 每个 GPU 拥有 12 GB 显存. GPU 用过将内存中的数据读入显存进行运算, 由于 GPU 具有很高的数据带宽和众多可以并行计算处理单元, 在矩阵数值计算方面的速度远远由于 CPU, 因此研究者们更青睐使用 GPU 进行深度神经网络的计算任务。 初始的学习率为 0.001, 并在训练时手动调整. 动量参数 Momentum 为 0.9.

因为训练图片较少, 所以我们进行了样本增强. 生成了数百万张图片, 并利用了提出的候选选择方法来进行分类层的训练. 图 4 为部分训练图片和其对应的被选择的候选边框. 图中的黑色边界是由于 Mask RCNN 方法中的零填充 (Zero Padding) 导致的.

图 4 候选选择方法的结果

为了保证 Mask RCNN 网络能够具有识别商品数据的旋转不变性, 我们首先在本地的商品数据集上训练其分类层. 然后对 COCO 数据集和本地商品数据集进行分类层的联合训练, 在测试数据上达到了 86.6% 的准确率. 因为训练图片中有些视角没有出现在测试图片中, 所以准确率相对较高. 然后只在 COCO 数据集上训练关系层, 迁移到本地商品数据集上不需再训练. 如图 5 所示, 其中虚线表示前 4 个候选区域, 实线表示回归之后的边框位置, 商品对象候选区域的回归边框趋向于同一位置, 可见关系层可以有效地迁移到商品数据上而无需边框标注数据的再训练.

然后我们在含有多个物体的图片上验证我们的方

法. 图 6 表示在本文构建的商品数据集和香烟包装数据集上经过非极大值抑制 (NMS) 之后的商品对象检测结果. 可以看出, 很少有物体被漏检或者误检. 我们在没有进行边框回归训练的情况下验证了我们的方法, 并与 Varol[13]所提出的方法进行了比较, 如表 1 所示. 本文对 mAP[15]指标也进行了计算. 从结果可以看出, 本文提出的方法的检测效果超过其他方法. 而且本文提出的数据集难度高于香烟包装商品数据集. 主要因为我们的数据集具有物体可旋转、多个视角和对象遮挡等情况.

图 5 候选区域和对应的边框回归结果

(a) 本文数据集结果 (b) 烟草包装数据集结果

图 6 非极大值抑制之后的检测结果

表 1 在两种数据集上的方法对比

数据集	香烟包装数据集			本文构建数据集
方法	Varol[13]且有人工约束	Varol[13]无人工约束	本文方法	本文方法
召回率 (%)	94	89	97.2	78.1
准确率 (%)	81	69	99.9	71.0
mAP(%)	-	-	88.0	61.3

4 结论与展望

本文提出了一种无需边框标定的分类训练方法,并且解决了可旋转物体的分类问题. 此外, 本文构建了一个关系层模型来学习分类层与回归层之间的关系, 来

实现商品检测定位, 并迁移到跨域的商品数据集上. 本文的方法很容易训练并迁移到其它数据集. 未来, 我们将致力于研究如何提高物体检测的精度和可解释性, 并将其扩展到分割领域并应用到自动化零售系统中.

参考文献

1 Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA, 2008: 1–8. [doi: 10.1109/CVPR.2008.4587597]

2 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA, 2016: 779–788. [doi: 10.1109/CVPR.2016.91]

3 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. In: Leibe B, Matas J, Sebe N, et al, eds. Computer Vision – ECCV 2016. Cham: Springer, 2016: 21–37. [doi: 10.1007/978-3-319- 46448-0_2]

4 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

5 Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359. [doi: 10.1109/TKDE.2009.191]

6 He KM, Gkioxari G, Dollar P, *et al*. Mask R-CNN. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2980–2988. [doi: 10.1109/ICCV.2017.322]

7 Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640–651. [doi: 10.1109/TPAMI.2016.2572683]

8 Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition. Proceedings of the British Machine Vision Conference, 2015: 1–12.

9 Dai JF, He KM, Sun J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. Proceedings of 2005 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1635–1643. [doi: 10.1109/ICCV.2015.191]

10 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. In: Fleet D, Pajdla T, Schiele B, *et al*, eds. Computer Vision – ECCV 2014. Cham: Springer, 2014: 740–755. [doi: 10.1007/978-3-319-10602-1_48]

11 Natarajan N, Dhillon IS, Ravikumar PK, *et al*. Learning with noisy labels. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 1196–1204.

12 Bekker AJ, Goldberger J. Training deep neural-networks based on unreliable labels. Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China, 2016: 2682–2686. [doi: 10.1109/ICASSP.2016.7472164]

13 Varol G, Kuzu RS. Toward retail product recognition on grocery shelves. Proceedings of Sixth International Conference on Graphic and Image Processing. Beijing, China, 2015: 944309. [doi: 10.1117/12.2179127]

14 Abadi M, Barham P, Chen JM, *et al*. TensorFlow: A system for large-scale machine learning. arXiv:1605.08695, 2016.

15 George M, Floerkemeier C. Recognizing products: A per-exemplar multi-label image classification approach. In: Fleet D, Pajdla T, Schiele B, *et al*, eds. Computer Vision – ECCV 2014. Cham: Springer, 2014: 440–455. [doi: 10.1007/978-3-319-10605-2_29]