基于 SSD-Mobilenet 模型的目标检测*

刘 颜 朱志宇 张 冰
（江苏科技大学电子信息学院 镇江 212003）

摘 要 为缩短在目标检测时模型的训练时间，加快网络的收敛速度，论文采取了一种卷积神经网络模型与迁移学习相结合的方法。SSD-Mobilenet 模型结合了 SSD 网络检测速度快和 Mobilenet 轻量型网络占用内存少的优点。首先，使用 COCO 数据集将 SSD-Mobilenet 模型进行预训练，得到模型参数、瓶颈描述因子，然后用 Pets 数据集重新训练网络的全连接层，运用迁移学习的思想，可以实现在较少数据集的情况下，通过短时间的训练模型即可收敛。实验结果显示总训练时长约为 11 个小时，检测准确率达到 74.5%。由此表明，采用 SSD-Mobilenet 模型与迁移学习相结合的方法可以缩短模型训练时间、加速模型的收敛速度且检准率比较高。

关键词 目标检测；卷积神经网络；SSD-Mobilenet；迁移学习
中图分类号 TP242.6 **DOI**：10. 3969/j. issn. 1672-9730. 2019. 10. 012

Target Detection Based on SSD-Mobilenet Model

LIU Yan ZHU Zhiyu ZHANG Bing
（College of Electronics and Information, Jiangsu University of Science and Technology, Zhenjiang 212003）

Abstract In order to shorten the training time of the model and speed up the rate of the convergence of the network during the target detection, this paper adopts a method combining convolutional neural network and transfer learning.The detection speed of the SSD network is fast and the mobilenet lightweight network takes up less memory. SSD-Mobilenet model combines the advantages of SSD network and Mobilenet network. Firstly, the SSD-Mobilenet model is pre-trained with the COCO dataset to obtain the parameters and bottleneck description factors of the model. Then the Pets dataset is used to retrain the full connection layer of the network. Using the idea of transfer learning, the model can converge in a short time with small datasets. The experimental results show that the total training time is about 11 hours, and the detection accuracy reaches 74.5%. This shows that the method combined with SSD-Mobilenet model and transfer learning can shorten the training time of the model, accelerate the convergence speed of the model and have a high detection accuracy.

Key Words target detection, convolutional neural networks, SSD-Mobilenet, transfer learning
Class Number TP242.6

1 引言

目标检测技术是计算机视觉领域中的重要组成部分，具有较高的研究价值。目标检测包括目标的定位和分类两个部分。随着计算机技术的发展和计算机视觉[1]原理的广泛应用，利用计算机图像处理技术实时跟踪目标的研究越来越热。目标的动态实时跟踪和定位在智能交通系统、智能监控系统、军事目标检测和手术器械定位等方面广泛应用[2]。2014年，RBG使用候选区域+卷积神经网络[3]设计了基于区域的卷积神经网络[4]框架，开启了基于深度学习目标检测热潮。R-CNN这个领域目前研究非常活跃，先后出现了 R-CNN、SPP-net、Fast R-CNN[5]、Faster R-CNN、R-FCN、YOLO[6]、SSD 等研究。

在深度学习研究中，结构简单、识别准确率高的模型是使用者一直追求的。Mobilenet [7]模型是 Google 提出的适合移动环境下的深度学习轻量级分类模型，延迟很低但可以保持一定的精度。SSD 结合了 YOLO 中的回归思想和 Faster R-CNN 中的

* **收稿日期**：2019年4月10日，**修回日期**：2019年5月23日
作者简介：刘颜，女，硕士，研究方向：深度学习、目标检测。朱志宇，男，博士，教授，研究方向：信号处理、智能控制。张冰，女，博士，教授，研究方向：信号处理、目标跟踪。

anchor 机制，不但有 YOLO 速度快的特性，而且可以像 Faster R-CNN 一样精准。SSD-Mobilenet 网络模型由于该模型结合了 Mobilenet 和 SSD 两种类型网络的优势，在识别精确度和性能消耗都有很好的表现。

2　基于 SSD—Mobilenet 模型的目标检测

2.1　Mobilenet 卷积神经网络

Mobilenet 使用深度可分离卷积来构建轻量级深度神经网络。深度可分离卷积将标准卷积分解成为如图 1 所示的深度卷积和一个如图 2 所示 1×1 的逐点卷积。如图 3 所示，假设输入特征图有 M 个通道，输出特征图的通道数 N，卷积核大小为 $D_K \times D_K$。将深度卷积与标准卷积部分的计算量相比：

$$\frac{D_K \times D_K \times D_F \times D_F \times M + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_K^2} \quad (1)$$

可见，Mobilenet 网络的卷积方式和标准卷积方式相比大大减少了计算量。Mobilenet 卷积神经网络模型结构为表 1 所示。Mobilenet 总共 28 层，由输入层，13 层卷积层，以及平均池化层与全连接层组成。

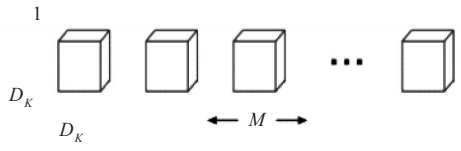


图 1　深度卷积

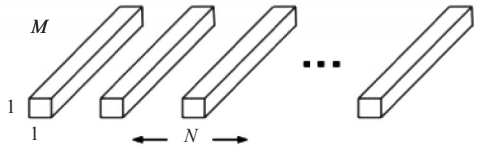


图 2　逐点卷积

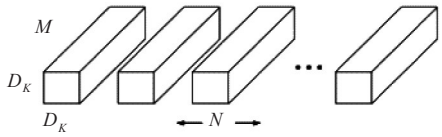


图 3　标准卷积

2.2　SSD 卷积神经网络

SSD，全称 Single Shot MultiBox Detector，是 Wei Liu 在 ECCV 2016 提出的目标检测算法[8]。SSD 网络结构如图 4 所示，算法的主网络结构是 VGG16，将两个全连接层改成卷积层再增加 4 个卷积层构造网络结构。SSD 在计算损失函数的时候，用到了两项的和，包括定位损失函数与回归损失函数。总的损失函数：

$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c) + \alpha L_{loc}(x,l,g)) \quad (2)$$

N 是匹配的 default boxes 的个数，x 表示匹配了的框是否属于类别 p，取值 $\{0,1\}$；g 是真实值（ground truth box）；c 是指所框选目标属于类别 p 的置信度（confidence）。

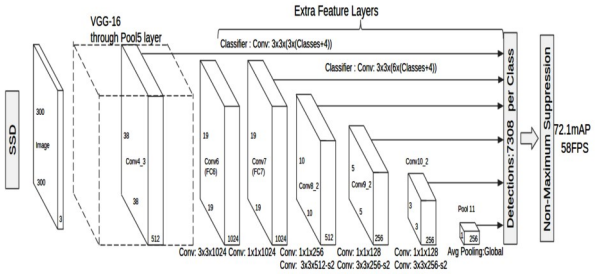


图 4　SSD 网络模型

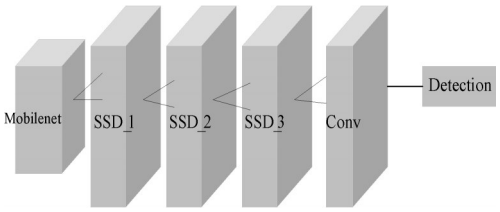


图 5　SSD-Mobilenet 卷积神经网络结构

2.3　SSD-Mobilenet 网络模型

SSD-Mobilenet 网络结构类似于 VGG_SSD 网络结构，在 conv13 卷积层后增加 8 个卷积层，其中 6 个卷积层用于检测目标。SSD-Mobilenet 网络模型用 SSD 模型作为基础模型，结合 Mobilenet 使用参数少，降低计算量的特点。其网络结构如图 5 所示。由于该模型结合了 Mobilenet 和 SSD 两种类型网络的优势，在保证良好精确度的基础上，使用小规模参数网络，减少计算量，降低资源消耗，缩短了训练时间，改善模型性能。

2.4　迁移学习的概念

迁移学习的目的是将从一个环境中学习到的知识运用到新的环境中。神经网络所学习到的知识主要体现于在特定任务上所训练得到的权重参数上，因而迁移学习的本质即为权重的迁移。传统的机器学习要求训练数据与测试数据相同分布，而且无法获取足够多的有效样本，需要重新标注大量的样本数据以满足需求。使用迁移学习的方法可以解决以上两个问题，允许将学习到的已有知识运用到新的且仅有少量样本数据的领域内。这意味着对于一个已经训练过的现有模型，并重新训练它最后一层或几层网络，不需要太长的训练时间便可以得到效果较好的网络模型[9~15]。

2.5　基于SSD-Mobilenet模型的迁移学习方法

经过预训练的SSD-Mobilenet网络模型可以网络结构与参数的分离,只要网络结构不变,就可以使用之前已经训练好的参数初始化网络。在实验中采用了COCO2014数据集对模型进行预训练,这个数据集图像包括80类目标,包括人类、交通工具、动物、家用电器、电子产品等12个大类。用于重新训练全连接层的参数的数据集使用的是牛津大学视觉几何组-宠物数据集(Visual Geometry Group-Pets Dataset),这个数据集包含37种不同种类的宠物图片,本文使用了其中shiba_inu与pomeranian这两类图片。

Type/Stride	Filter Shape	Input Size
Conv/s2	3×3×3×32	224×224×3
Conv dw/s1	3×3×32 dw	112×112×112×32
Conv/s1	1×1×32×64	112×112×112×32
Conv dw/s2	3×3×64 dw	112×112×112×64
Conv/s1	1×1×64×128	56×56×64
Conv dw/s1	3×3×128 dw	56×56×128
Conv/s1	1×1×128×128	56×56×128
Conv dw/s2	3×3×128 dw	56×56×128
Conv/s1	1×1×128×256	28×28×128
Conv dw/s1	3×3×256 dw	28×28×256
Conv/s1	1×1×256×256	28×28×256
Conv dw/s2	3×3×256 dw	28×28×256
Conv/s1	1×1×256×512	14×14×256
5× Conv dw/s 1 Conv/s 1	3×3×512 dw 1×1×51×512	14×14×512 14×14×512
Conv dw/s2	3×3×512 dw	14×14×512
Conv/s1	1×1×512×1024	7×7×512
Conv dw/s2	3×3×1024 dw	7×7×1024
Conv/s1	1×1×1024×1024	7×7×1024
Avg Pool/s1	Pool 7×7	7×7×1024
FC/s1	1024×1000	1×1×1024
Softmax/s1	Classifier	1×1×1000

图6　Mobilenet卷积神经网络模型结构

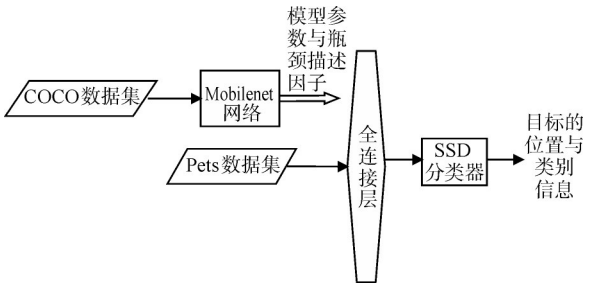


图7　检测流程图

目标检测流程图如图7所示,首先Mobilenet网络用COCO数据集进行预训练得到模型参数、瓶颈描述因子和初始特征,然后目标域Pets数据集重新确定全连接层参数,最后将训练好的Mobilenet模型与融合特征传入SSD网络,即可实验对图片中目标的定位与分类。

3　实验过程及结果分析

3.1　训练环境

Tensorflow是Google的可移植机器学习和神经网络库,对Python具有良好的支持,它还有用于显示和理解描述其计算的数据流图Tensorboard。本实验的运行平台为Windows 10,开发语言为Python,使用的深度学习框架为Tensorflow,本实验的运行环境如图8所示。

配置名称	版本
Python	3.5.2
Anaconda	4.2.0 (64bit)
Tensorflow-gpu	1.6.0
Jupyter Notebook	4.2.1
Cuda	9.0.0
Cudann	7.0.5
显卡	NVIDIA GeForce 920M

图8　实验运行环境表

filename	width	height	class	xmin	ymin	xmax	ymax
pomeranian_1.jpg	494	500	pomerianian	1	42	482	427
pomeranian_10.jpg	500	333	pomerianian	191	46	488	264
pomeranian_11.jpg	500	332	pomerianian	145	49	488	328
pomeranian_12.jpg	500	375	pomerianian	148	4	494	369
pomeranian_13.jpg	500	375	pomerianian	144	48	479	247
pomeranian_14.jpg	500	375	pomerianian	11	4	484	341

图9　部分标注信息表格文件示例

3.2　数据集的准备

本文在Pets数据集中选取了博美犬和柴田犬这两类的图片各200张,其中160张作为训练集,30张作为测试集,10张用于检测训练好的模型。首先标注图像使用的是labellmg工具,样本被标注好后,生成与样本一一对应的xml文件。然后,调用xml_to_csv.py脚本生成表格,如图9所示,图的第一列filename是图像的名称;第二列和第三列代表了图片的大小;第五列到第八列代表在目标检测框的坐标;class的值设为pomerianian,代表博美犬,若为shiba_inu代表柴田犬。接下来,调用调用generate_tfrecord.py把对应的csv格式转换成Tensorflow能识别的tfrecord格式。

3.3　开始训练与结果分析

本文的训练过程是在GPU上训练。如图10(a)所示每一步的运行时间大约为0.9s。如图11(a)所示,随着训练过程的进展,总损失会降低到可能的最小值,当总损失不会因更多迭代而下降可以认定到训练已经完成。总共训练了40K步,总损失值约为3.2。

训练完成后,保存训练好的模型就可以开始做一些检测了。在test_images文件夹里放了每个分类各10张图片。这20张图片的平均检测准确率达到74.5%。如图12,最高的准确率达到99%。如图13所示,这两张检测结果图出现了同一个检测对象有多个检测框的情况,这个问题存在的原因可能是在打标签时由于每张图片中狗的外貌和姿态以及背景的差别,打框时有的框出了目标对象身体

的全部，有的只框出来目标对象身体的一部分，这使得网络接收到的信息存在着差别。如图 11 所示，检测中还存在分类错误的情况，将博美犬分到柴田犬类，出现这种情况的原因不能完全归于模型的不够优秀，事实上，图片中的博美犬确实和柴田犬有着极大的相似。

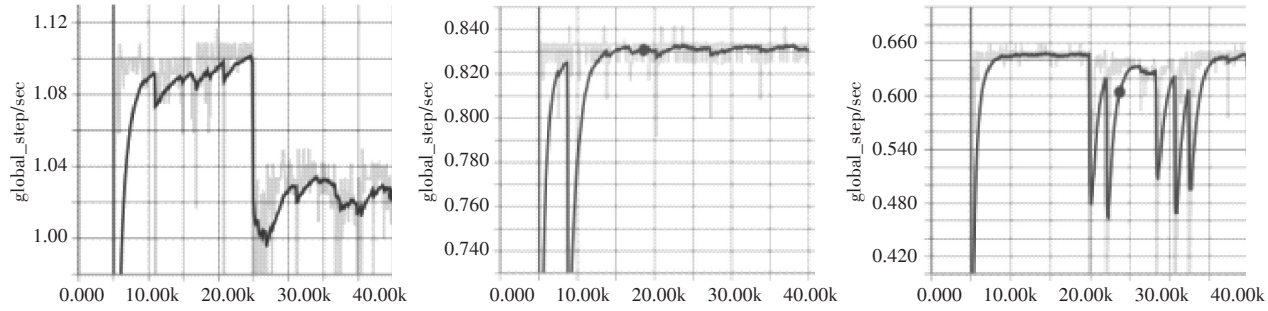


图 10　训练步数与时间关系图

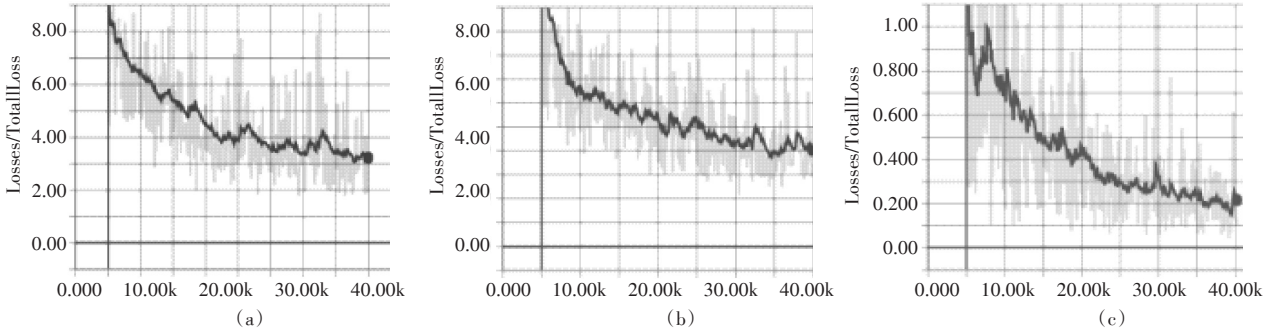
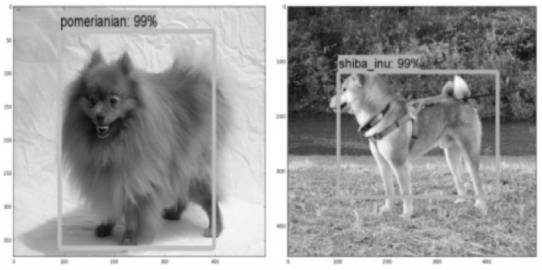
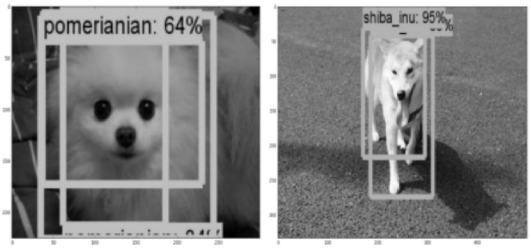


图 11　训练步数与总损失关系图



(a)博美犬分类　　(b)柴田犬分类

图 12　准确率 99%



(a)博美犬分类　　(b)柴田犬分类

图 13　一个目标有多个检测框图

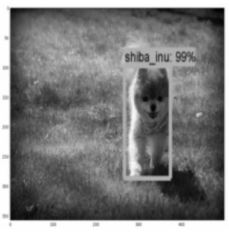


图 14　分类错误

3.4　几种模型的效果对比

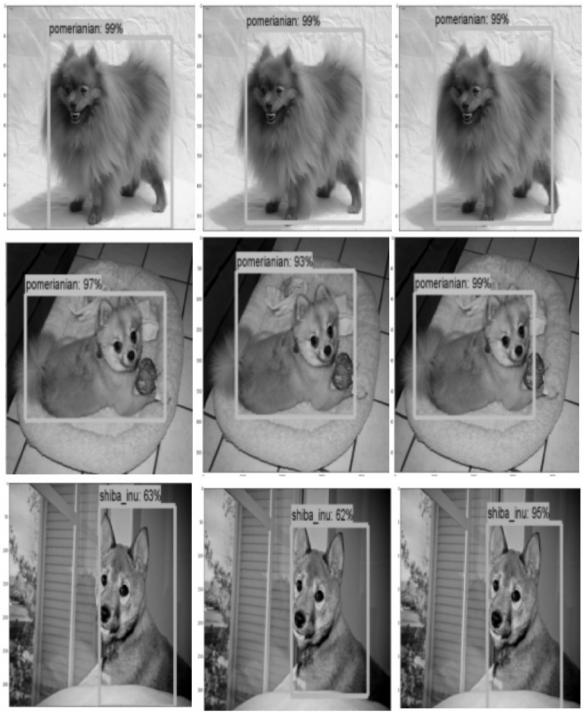


图 15　同一张图片使用三种模型做目标检测

本实验还训练了 SSD_inception 网络与 faster_rcnn 网络进行实验结果的对比。训练步数与时间的关系图如图 10(b)、10(c)，所示，总损失与时间的关系图如图 11(b)、11(c)所示。同一张图片用三

种模型检测效果如图15所示,从左至右分别使用的模型为SSD_Mobilenet, SSD_inception, faster_rcnn。

	step	loss	sec/step	test time/s	accuracy
SSD-Mobilenet	40000	3.2	0.9	62	74.5
SSD_inception	35000	4.1	1.2	75	76.2
faster_rcnn	35000	0.2	1.6	210	77.5

图16 三种模型性能参数比较

4 结语

SSD-Mobilenet是一种新型目标识别网络模型,它既结合了轻量型Mobilenet网络节省存储空间以及低能耗的优点,又有SSD网络同时保持高效率与高准确率的特点。运用迁移学习的思维方式,在COCO数据集进行过预训练的SSD-Mobilenet模型再用Pets数据集进行网络参数微调,使得网络经过较短的训练时间便可以收敛具有很好的检测目标的能力。在这次的实验中出现的对一个目标有多个检测框和由于两类目标的相似性分类错误的问题,在今后的实验中需要进一步探究,力求得到更加性能优秀的网络。

参考文献

[1] 郑南宁. 计算机视觉与模式识别[M]. 北京:国防工业出版社,1998.

[2] Ciresan D C, Meier U, Masci J, et al. High-Performance Neural Networks for Visual Object Classification [R]. Switzerland: Dalle Molle Institute for Artificial Intelligence, 2011.

[3] LeCun Y, Bengio Y. Convolutional Networks for Images, Speech, and Time Series[J]. The Handbook of Brain Theory and Neural Networks, 1995, 3361(10):255-258.

[4] LeCun Y, Bottou L, Bengio Y, et al. Gradient-Based Learning Applied to Document Recognitio [J]. IEEE, 1998, 86(11): 2278-2324.

[5] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[6] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016:779-788.

[7] Hinton G E, Salakhutdinov R R. Reducing the Dimensionality of Data with Neural Networks [J]. Science, 2006, 313(5786): 504-507.

[8] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot Multibox Detector [C]//Computer Vision-ECCV 2016. Berlin: Springer-Verlag, 2016:21-37.

[9] Tahmoresnezhad J, Hashemi S. Visual Domain Adaptation via Transfer Feature Learning[J]. Knowledge & Information Systems, 2016, 50(2):1-21.

[10] Mahmud M M H. On universal Transfer Learning [J]. The Oretical Ccomputer Science, 2009, 410 (19): 1826-1846.

[11] Oyen D, Lane T. Transfer learning for Bayesian Discovery of Multiple Bayesian Networks [J]. Knowledge and Information Systems,2015,43(1):1-28.

[12] Yang L, Zhang J. Automatic Transfer Learning for Short Text Mining[J]. Eurasip Journal on Wireless Communications & Networking, 2017, 2017(1):42.

[13] Wei F, Zhang J, Yan C, et al. FSFP: Transfer Learning From Long Texts to the Short[J]. Applied Mathematics & Information Sciences, 2014, 8(4):2033-2040.

[14] 庄福振, 罗平, 何清,等. 迁移学习研究进展[J]. 软件学报, 2015, 26(1):26-39.

[15] Pan S J, Yang Q. A Survey on Transfer Learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10):1345-1359.