计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):147−155 [doi: 10.15888/j.cnki.csa.006716]

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

基于反卷积特征学习的图像语义分割算法①

郑　菲, 孟朝晖, 郭闯世

(河海大学 计算机与信息学院, 南京 211100)

摘　要: 随着深度学习的发展, 语义分割任务中许多复杂的问题得以解决, 为图像理解奠定了坚实的基础. 本文算法突出表现在两个方面, 其一是利用反卷积网络, 对卷积网络中不同深度的卷积层提取到的多尺度特征进行融合, 之后再次通过反卷积操作对融合后的特征图进行上采样, 将其放大到原图像的大小, 最后对每个像素进行语义类别的预测. 其二为了提升本文网络结构的性能, 提出一种新的数据处理方式, 批次中心化算法. 经过实验验证, 本文算法在 SIFT-Flow 数据集上语义分割的平均准确率达到 45.2%, 几何分割的准确率达到 96.8%, 在 PASCAL VOC2012 数据集上语义分割的平均准确率达到 73.5%.

关键词: 深度学习; 语义分割; 批次中心化; 多尺度特征; 反卷积网络

引用格式: 郑菲,孟朝晖,郭闯世.基于反卷积特征学习的图像语义分割算法.计算机系统应用,2019,28(1):147–155. http://www.c-s-a.org.cn/1003-3254/6716.html

Image Semantic Segmentation Algorithm Based on Deconvolution Feature Learning

ZHENG Fei, MENG Zhao-Hui, GUO Chuang-Shi

(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: With the development of deep learning, many complex problems in semantic segmentation tasks are solved, which lays a solid foundation for image understanding. The proposed algorithm highlights two aspects. Firstly, our algorithm fuses multi-scale features from different levels of deep convolutional network by using multi-level deconvolution network. Then our algorithm upsamples these feature maps by deconvolution, meanwhile zooms them up to the original image size to predict semantic categories pixel-to-pixel. The second one, we propose a new method for data processing which is batch centralization algorithm, in order to improve the performance of network structure in this study. Through experimental verification, the mean IoU of semantic segmentation on the SIFT-Flow dataset reaches 45.2%, and the accuracy of geometric segmentation reaches 96.8%. The mean IoU of semantic segmentation on the PASCAL VOC2012 dataset reaches 73.5%.

Key words: deep learning; semantic segmentation; batch centralization; multi-scale features; deconvolution network

图像语义分割是机器视觉中图像理解的重要一环, 其旨在通过一定的方法, 将图像中的每个像素分为不同的语义类别, 得到不同的分割区域, 实现从底层到高层的推理过程, 最终获得一幅具有像素语义标注的图像[1]. 准确的图像语义分割是实现诸多计算机视觉任务的基础, 如场景识别、场景理解和分析. 近年来, 随着深度神经网络被引入到图像语义分割任务中来, 该课题研究得到了快速发展. 应用在自动驾驶领域, 如定位道路、车体和行人, 获得物体轮廓信息等[2]; 在无人机领域, 如进行落地点检测、落地点场景识别等[3]; 还有在智能服务机器人、医学图像分析等领域中均取得了巨大的应用成果.

① 收稿时间: 2018-06-23; 修改时间: 2018-07-20; 采用时间: 2018-07-27; csa 在线出版时间: 2018-12-26

在深度卷积网络广泛应用到语义分割领域之前, 图像语义分割任务主要是根据图像自身的低阶视觉信息来进行分割, 比如图像的角点、边缘和色彩等. 彼时的算法研究有简单的像素级阈值法[4], 基于像素聚类的分割方法[5], 还有“图割法”的分割方法. 其中 Shi 等[6]提出的特征归一化分割算法是著名的“图割法”分割方法, 之后微软剑桥学院[7]提出的 Grab-cut 也是著名的交互式图像语义分割方法, 该方法利用图像中的纹理信息和边界信息, 尽可能减少了用户交互操作而得到比较好的前景与背景的分割结果. 在计算机视觉步入深度学习时代之后, 以往的算法在复杂困难的分割任务中所面临的难题, 如今也得到了很好的提升.

Long 等[8]在 2014 年提出的全卷积神经网络 (Fully Convolution Network, FCN), 是深度学习语义分割工作的开山之作. FCN 是一个对整幅图像进行像素级密度预测[9](Pixelwise Dense Prediction) 的端对端网络模型, 输入任意尺寸的图像可直接得到相应的语义分割图, FCN 直接将传统卷积网络的全连接层替换为卷积层然后迁移到语义分割任务中, 通过跳转结构将深层的语义信息和浅层的位置信息进行融合以达到精确的分割效果. 然而 FCN 在池化过程中会造成信息丢失, 所以 Fishe 等[10]提出将 VGG 网络的最后两个池化层去掉, 然后将传统的卷积层替换成扩张卷积层 (Dilated Convolution), 这就需要不同扩张尺度的扩张卷积层来保证网络的感受野不受影响, 从而确保语义分割的准确度. Chen 等在语义分割工作中不断改进和实践, 至今提出了多种版本的方案. 最初提出的 DeepLab v1[11] 和 DeepLab v2[12]主要做了三个贡献, 首先用带孔卷积 (Atrous Convolution) 实现逐像素的密度预测, 其次提出带孔空间金字塔池化模型 (Atrous Spatial Pyramid Pooling, ASPP) 实现多尺度分割任务, 第三是利用深度卷积网络和概率图模型相结合准确定位物体边缘. DeepLab v3[13]改进了带孔空间金字塔池化模型, 通过带孔卷积级联获取多尺度的语义信息, 并且采用全连接条件随机场 (DenseCRF) 的后处理操作对预测结果进行优化. DeepLab v3+[14]为了融合多尺度语义信息, 引入编码器-解码器的架构, 并提出可任意控制编码器提取特征的分辨率, 通过带孔卷积来平衡精度与耗时.

经典卷积网络 (AlexNet[15], VGG[16], GoogLeNet[17] 等) 中不同的卷积层提取出来的特征包括了从浅层到深层的多尺度特征信息, 且特征图逐渐变小, 本文根据这一特性提出的算法主要思想包括两个方面, 其一通过反卷积网络将卷积网络提取到的不同尺寸的特征图放大到相同大小, 从而对多尺度特征信息进行融合; 其二再对这些特征图进行最后的反卷积将其放大到原图像的大小, 连接到 Softmax 分类器计算每个像素的损失函数, 基于整幅图像的损失函数反向训练网络. 并且本文提出一种新的数据处理方法, 批次中心化算法, 可以对输入数据同时进行激活和中心化操作, 有效提升网络的收敛速度和算法的平均准确率. 下面将阐述反卷积网络的概念, 以及批次中心化算法, 并且详细介绍本文算法和语义分割网络结构, 以及实验结果与分析.

1 反卷积网络

2010 年 Zeiler 等[18]提出利用反卷积网络 (Deconvolutional Networks) 无监督地学习图片的中低层特征, 操作的方向不再是从原图片到特征图, 而是从特征图到图片, 这里的反卷积就是卷积的前馈操作. 2011 年 Zeiler M D 等[19]又提出新的反卷积网络用来学习图片的中高层特征, 不同于以往的是该反卷积网络中加入了反池化 (up-pooling)和反卷积 (deconvolution). 2013 年 Zeiler 等[20]开始探究深度卷积网络良好性能背后的原理, 他们想知道深度卷积网络中每一个中间层的结果, 于是在每一层卷积后面接一个反卷积网络, 然后通过: 反池化——ReLU——反卷积的过程, 对卷积得到的特征图进行放大, 实现特征的可视化.

我们都知道池化起到下采样的作用, 最大池化操作可以得到上一层输出图中的最大激活值来帮助分类, 这一过程只保留了最大激活值而丢失了其余位置上的值, 因此池化是不可逆的. 但是反池化为了实现上采样, 可以做这样的近似操作, 通过记录最大池化过程中激活值的位置坐标, 然后在反池化的时候, 把池化过程中最大激活值所在位置上的值激活, 其他位置的值近似为 0. 如图 1 所示, 左边是池化过程, 右边是反池化过程.

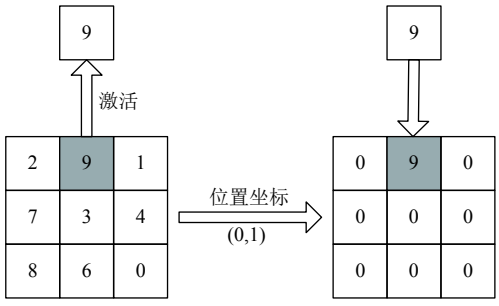

图 1 反池化示意图

卷积操作是将图片经过卷积后得到特征图, 而反卷积操作刚好相反. 现假设输入图片为 x, 特征图为 y, 卷积操作可表示为 $y=Cx$, 根据矩阵的运算性质可知, 反卷积的过程则是 $x=C^{\mathrm{T}}y$. 所以反卷积就是卷积操作在神经网络结构中正向和反向传播的相反过程, 其实还是理论意义上的卷积操作, 只是为了突出其特性而称作——反卷积. 采用卷积过程中转置后的卷积核对特征图进行反卷积, 并且由于其相反于卷积操作的特性还可以将特征图放大. 举例说明:若输入图片尺寸为 i、卷积核尺寸为 k、步长为 s、边缘扩充为 p、输出特征图尺寸为 o, 则卷积操作计算公式为:

$$o=\left\lfloor\frac{i+2p-k}{s}\right\rfloor+1 \tag{1}$$

现通过反卷积将特征图还原到原图像大小, 计算公式为:

$$i=s(o-1)+k-2p \tag{2}$$

反卷积示意图如图 2 所示.

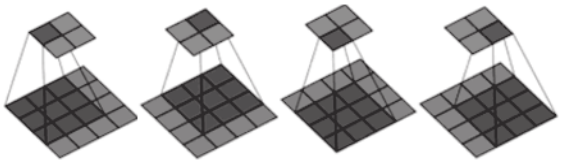

图 2 反卷积示意图

本文算法中将利用分层的反卷积网络对特征图进行上采样并学习特征, 最后将特征图上采样到原图像的像素空间, 以便逐像素计算损失反馈到网络中训练, 对图像的每一个像素进行类别预测.

2 批次中心化算法

在训练过程中网络的收敛速度跟输入数据是有关的, 如果一组训练样本具有较强的相关性, 则使训练网络很容易收敛. Sergey Ioffe 等[21]提出的批次归一化算法 (Batch Normalize, BN), 主要是对每一层输出的小批量数据进行归一化, 从而加快网络收敛. BN 算法将中心化和归一化合并完成, 然后输出采用 ReLU 激活函数.

受 BN 算法的启发, 本文提出一种新的数据处理方法, 批次中心化算法 (Batch Centralization, BC). BC 算法在训练时仍采用小批量数据 (Mini-batch) 处理的方式, 可以对输入数据同时进行激活和中心化操作. 针对一个有 m 个样本的小批量数据, 计算公式如公式 (3) 所示:

$$\begin{cases} y^{(k)}=f\left(\sum\limits_{j=1}^{n}w_j(x_j^{(k)}-\mu_j)\right),\ k=1,\cdots,m \\ \mu_j=\dfrac{1}{m}\sum\limits_{i=1}^{m}(x_j^{(i)}),\ j=1,\cdots,n \end{cases} \tag{3}$$

其中, μ_j 是一个输入样本的均值, 其下标表示上一层第 j 个神经元, $y^{(k)}$ 表示 BC 层处理后的输出, 上标表示第 k 个样本, w_j 则是上一层第 j 个神经元到该神经元的权重. $f(x)$ 是 Sigmoid 型激活函数, 用于对神经元输入的激活, 如公式 (4) 所示:

$$f(x)=\frac{1}{1+e^{-\alpha x}},\ \alpha>0 \tag{4}$$

由于一般在使用 Sigmoid 函数对输入值进行激活后, 在神经网络梯度反向传播时, 由于接近两端的数据梯度较小而容易造成梯度消失. 所以提出在 Sigmoid 函数中加入敏感性强度参数 $\alpha(\alpha>0)$, 函数图像如图 3 所示.

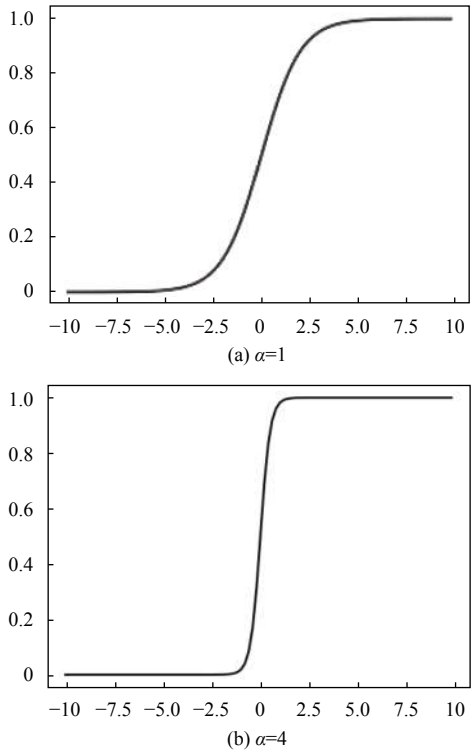

图 3 Sigmoid 函数

敏感性强度参数 α 对 Sigmoid 函数进行横向压缩, 由图 3 中不同敏感性强度参数的对比可知, α 值越大则中间部分的图像越陡, 梯度越大, 则对中间区域的值越

敏感. BC 算法的思想就是首先按照公式 (3) 对输入数据进行中心化处理, 然后通过一个变型 Sigmoid 函数 $f(x)$ 对中心化后的值激活. 使得有效数据尽可能集中在中心区域, 这就保证了中心化后的数据具有较大的梯度, 在一定程度上避免了在反向传播过程中梯度减小进而出现梯度消失的情况.

3 反卷积特征学习的语义分割

3.1 多尺度特征融合

我们都知道对于一个深层卷积网络, 其最初一两层学习到的基本上是颜色、边缘等低层特征; 再往后就开始学习稍微复杂的特征, 比如纹理、线条等这些比较有区别性的特征; 更深层次的网络学习的特征就更加完整, 具有明显的辨别性特征, 比如物体的轮廓以及显著的位置信息.

图像中的语义信息和物体位置信息是以非线性金字塔的形式进行编码的, 然后作为图像的浅层和深层特征. 语义分割任务中, 图像的语义信息和图像中物体的位置信息存在一种关系: 全局语义信息解决的是“是什么”的问题, 而局部位置信息解决的是“在哪里”的问题. 如图 4 左边的特征金字塔所示. 想要对图像进行准确的语义分割, “在哪里”和“是什么”显然是问题的关键, 如何将二者结合也是该领域一直在探索的话题. 所以在语义分割网络中, 不仅要提取图像的浅层特征, 还要提取深层特征, 将这些多尺度特征进行融合进而对图像的每个像素准确分类.

本文算法提出, 将浅层卷积、深层卷积以及更深层卷积提取到的多尺度特征分别通过相应的反卷积进行融合, 如图 4 所示. 这里之所以要使用反卷积, 是因为卷积和池化过程中会使特征图变小, 并且卷积层数越多特征图越小, 为了方便融合需要将不同尺度的特征图统一上采样到相同尺寸, 还要保证尽可能少的丢失特征信息.

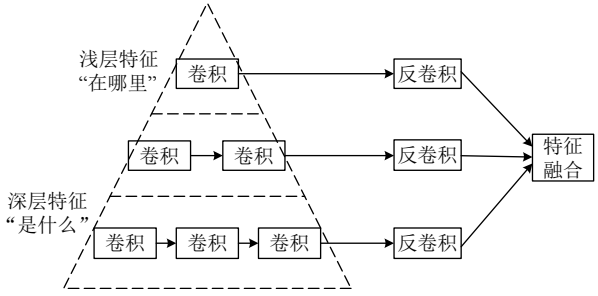

图 4 多尺度特征融合

3.2 语义分割网络结构

在 3.1 小节中详细介绍了本文算法是如何进行多尺度特征融合的, 下面介绍整个反卷积特征学习网络的结构, 阐明该网络结构是如何实现图像语义分割的.

目前的深度卷积网络, 如 AlexNet[15]、GoogLeNet[17]、VGG[16]以及 ResNet[22]等, 通过不断加深的卷积网络结构能够更好地学习图像中的特征, 根据每年 ILSVRC 挑战赛的结果可知, 网络结构层次越深往往分类的准确率越高. 这就说明卷积网络学习图像特征的本领已经非常强大了, 那么如何将这些用于分类的卷积网络的本领发挥到图像语义分割任务中来, 就需要充分理解语义分割任务和分类任务的区别.

以往的分类识别任务是对图像进行特征提取, 将特征学习的结果送入分类器中, 最后得到标签 TOP5 的概率得分以此判断图像所属类别. 而在语义分割任务中需要通过特征学习来预测每个像素所属的类别, 得到一副具有像素语义标注的图像. 因此语义分割任务中需要计算每个像素的损失函数, 对图像进行逐像素的密度预测. 而在卷积网络提取特征的过程中, 卷积和池化操作改变了原图像的尺寸, 为了完成逐像素预测就要将特征图放大到原图像的像素空间.

Long 等[8]提出的全卷积神经网络, 是在第 7 层卷积之后添加一层 1×1 卷积核的、通道数为 21 的卷积来预测特征图每个像素位置上粗糙的分类得分, 后面紧跟一个反卷积层用来将这个粗糙的预测得分双线性上采样到原图像的像素密度预测. 显然这种只通过一层反卷积就简单粗暴地将最后的特征图上采样到原图像的像素空间, 丢失了许多特征信息, 这对最后的像素级密度预测非常不利. 而本文算法提出将每一步卷积和池化的过程都逐步反向操作来实现上采样, 逐步还原特征图中的信息, 减少特征信息的丢失.

下面介绍本文反卷积特征学习的语义分割网络, 示意图如图 5 所示. 第 2 节所介绍的批次中心化算法 (BC), 是对数据同时进行中心化和激活操作, 其作用就相当于经典分类卷积网络中的 ReLU 激活层和归一化层. 本文算法将卷积网络中的激活层和归一化层替换为 BC 层, 其可行性和优越性将在第 4 节实验中进行验证.

本文的语义分割网络输入的原图像尺寸为 H×W, 其后融合了第三层卷积后的特征图、第四层卷积后的特征图和最后一层卷积的特征图, 实现了将图像中浅层到深层的多尺度特征进行融合. 图 5 中 Conv1-11-96

表示第一层 11×11 的卷积核、通道数 96 的卷积操作, Maxpool1-3-96 表示第一层 3×3 的窗口、通道数为 96 的最大池化. 而我们都知道卷积和池化操作会改变图像的尺寸, 尤其在本文算法中需要融合不同尺寸的特征图, 为了方便特征融合, 可以先分别将每一层的特征图经过反卷积层上采样到相同尺寸, 再进行融合.

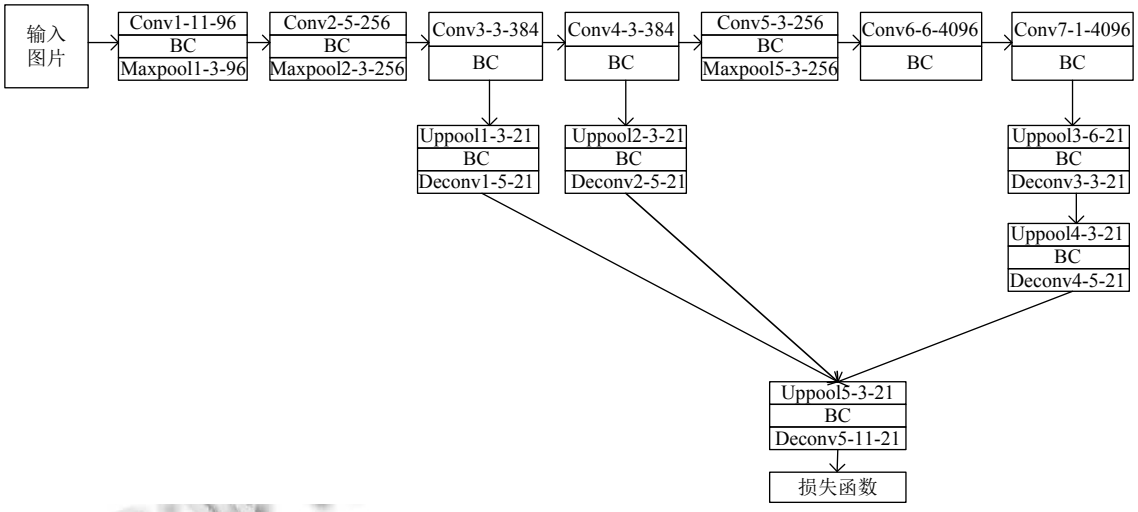

图 5 网络结构

这里反卷积网络采用反池化-反激活-反卷积的操作顺序. 本文中的反池化操作采用文献[19,20]中提出的方法. 对于反激活, 在卷积网络中激活函数是用于保证每层输出的激活值都是正数, 那么对于反卷积过程我们依然需要保证每层特征图为正值, 因此这个反激活其实跟激活没有区别, 依然采用我们之前的 BC 层. 而反卷积操作中的所有参数都要参与训练, 以达到特征学习的目的. 图 5 中 Uppool1-3-21 表示第一层 3×3 的窗口、通道数 21 的反池化操作, Deconv1-5-21 表示第一层 5×5 的卷积核、通道数 21 的反卷积操作. 本文卷积-反卷积的网络结构是根据编码器-解码器原理设计的, 所以算法中添加的反卷积网络是根据前面特征提取过程中卷积和池化操作来设计的, 这里可以理解为“镜像原理”, 前面一组卷积和池化对应于后面的一组反池化和反卷积, 逐渐将特征图放大到原图大小. 这样一来可以更大程度上减少特征信息的损失, 二来反卷积网络也要进行特征学习, 帮助更精确地完成分割.

融合后再接一层反卷积网络, 将特征图还原到原图大小, 这里设计成 21 通道是对应数据集 PASCAL VOC 的 21 类语义标签. 最后将 21 张特征图中每个对应位置的像素值送入 Softmax 层进行语义类别的预测, 利用预测结果和标签计算最终的损失函数. 如公式 (5) 所示:

$$J(x_{h,w};\theta)= -\sum_{i=1}^{m}\sum_{k=1}^{K}1\{y_{h,w}^{(i)}=\hat{y}_k\}*\log(P_{h,w}(x^{(i)}|\theta))+\frac{\lambda}{2}\|\theta\|^2 \quad (5)$$

上式表示图像中一个像素点的损失计算. 其中, $1\{\cdot\}$是示性函数. 针对一个有 m 个样本的小批量数据集, $x=\{x(1),x(2),\cdots,x(m)\}$, 原图像经过上述语义分割网络的处理, 得到对图像中每个像素的预测, 结果为 $P_{h,w}(x^{(i)}|\theta)$, 下标 h, w 表示图像中 (h, w) 的位置, θ 表示网络中的参数向量. y_k 表示语义标签的第 k 类. λ 表示正则项的超参数.

由公式 (5) 可以推出最终整幅图像的损失计算如公式 (6) 所示:

$$J(\theta)=\sum_{h=1}^{H}\sum_{w=1}^{W}J(x_{h,w};\theta) \quad (6)$$

4 实验

本文实验目的主要有两个, 其一是验证所提出的批次中心化 (BC) 在本文算法中的可行性和优越性. 其二主要评估本文算法在 SIFT Flow 和 PASCAL VOC2012 两个数据集上的性能表现, 以及与其他算法的对比实验结果.

4.1 数据集介绍

本节实验所使用到的数据集有 PASCAL VOC2012 和 SIFT-Flow.

(1) Pascal VOC 2012

PASCAL VOC 挑战赛是视觉对象的分类识别和检测的一个基准测试, 提供了检测算法和学习性能的标准图像注释数据集和标准的评估系统. PASCAL VOC 数据集包括 20 个对象.

所有的标注图片都有物体检测需要的标签, 但只有部分图片有图像分割的标签. 对于物体检测任务, VOC2012 的训练集、验证集和测试集包含 VOC2008-VOC2011 的所有对应图片; 对于图像分割任务, VOC2012 的训练集和验证集包含 VOC2007-VOC2011 的所有图片, 测试集只包含 VOC2008-VOC2011. 对于分割任务的标签有两个部分, 其一是语义分割 (Semantic Segmentation) 标注出每一个像素的类别, 如图 6(b) 所示; 其二是实例分割 (Instance Segmentation) 标注出每一个像素属于哪一个对象如图 6(c) 所示.

(a) 原图 (b) 语义分割 (c) 实例分割

图 6 PASCAL VOC 数据集

本文所研究的语义分割属于第一种类别分割, 通过本文算法提出的网络对每个像素进行类别预测.

(2) SIFT-Flow 数据集

SIFT-Flow 数据集中包含 2688 张图片, 是针对森林、城市街景、道路、建筑物等室外场景, 其中包含语义分割和几何分割两种标签. 语义分割的标签 0 表示背景、1~33 表示 33 种类别, 几何分割的标签–1 表示背景、1–3 分别表示天空、水平和垂直. 如图 7 所示, 该数据集中对应原图像的每个像素在语义标签中是用数字表示的, 28 表示天空, 6 表示建筑, 要通过 Matlab 将标签图显示出来.

4.2 批次中心化算法实验

本实验首先讨论批次中心化算法 (BC) 中, Sigmoid 函数的敏感性强度参数 α 的设置对最终准确率的影响. 选取 α=1 开始, 以 0.25 的步长进行多次实验, 在 PASCAL VOC2012 数据集上的结果如图 8 所示.

(a) 原图 (b) 语义分割标签

图 7 SIFT-Flow 数据集

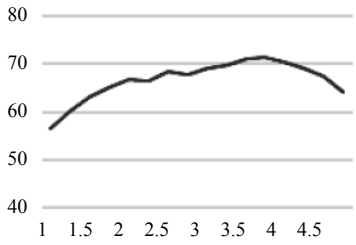

图 8 α 取值对准确率的影响

由图 8 可知, α=3.75 时本文算法的准确率最优, 因此下面所有的实验中 α 取值均为 3.75.

下面验证批量中心化 (BC) 算法在本文网络模型中的性能. 在 SIFT-Flow 和 PASCAL VOC2012 两个数据集上, 首先将网络结构中的激活层和归一化层采用 ReLU 和 BN 进行训练, 再将网络结构中的 ReLU 和 BN 用一层 BC 来代替进行训练. 最终测试结果如表 1 所示.

表 1 中给出了是否使用 BC 算法对本文语义分割算法最终的平均准确率, 明显得到使用 BC 算法的准确率更高, 网络结构的性能更好.

表 1 BC 算法验证结果

数据集	SIFT-Flow	PASCAL VOC2012
本文算法 (ReLU+BN)	41.8	70.3
本文算法 (BC)	45.2	73.5

4.3 语义分割实验

本节主要评估本文算法在 SIFT-Flow 和 PASCAL VOC2012 两个数据集上的性能表现, 以及与其他算法的对比实验结果.

首先在数据集 SIFT-Flow 上进行对比实验, 因为

该数据集包含了语义分割和几何分割两种标签, 所以我们在这两种任务上独立训练本文的模型. 将整个数据集分为训练集、验证集和测试集, 训练集包含 1612 张图片, 验证集包含 806 张图片, 测试集包含 270 张图片. 使用 Caffe 框架实现图 5 的网络结构, 整个网络是端对端的结构, 在数据集 SIFT-Flow 中图片尺寸均是 256×256, 输入到网络中进行训练. 训练中使用了小批量随机梯度下降 (Mini-Batch Gradient Descent, MBGD) 算法, 调整学习率为 1×10^{-1}, 正则项的超参数 λ 设置为 0.0005, 冲量 (momentum) 设置为 0.99, batch 大小设置为 5, 迭代训练 150 000 次. 在相同数据集上和 Liu 等[23]的算法、Tighe 等[24,25]的算法、Farabet 等[26]的算法以及 Long 等[8]提出的 FCN-16s 进行对比, 对最终的结果进行分析, 如表 2 所示.

表 2 SIFT-Flow 上实验结果对比

算法	Pixel acc.	Mean acc.	Mean IoU	Geom acc.
Liu 等[23]	76.7	-	-	-
Tighe 等[24,25]	78.6	39.2	-	90.8
Farabet 等[26]	785	29.6	-	-
FCN-16s[8]	85.2	51.7	39.5	94.3
本文算法	88.3	53.5	45.2	96.8

SIFT-Flow 上的语义分割 (中间部分) 和几何分割 (右边部分) 的实验结果是独立的. 其中 Pixel acc.指语义分割的像素准确率, Mean acc.指平均准确率, Mean IoU 指的是不同语义类别识别准确度的平均值, Geom acc.指几何分割的像素准确率. 可以看出本文算法在语义分割上的像素准确率达到 88.3%, 几何分割上的像素准确率达到 96.8%, 表现较好.

下面在 PASCAL VOC2012 数据集上进行实验, 选取 1747 张训练集、874 张验证集以及 1165 张测试集. 使用 Caffe 框架实现图 5 的网络结构, 先将所有图片统一缩放到 500×500, 输入到网络中进行训练. 训练中仍采用小批量随机梯度下降算法, 学习率、超参数、冲量、batch 大小以及迭代次数分别设置为 1×10^{-2}、0.0005、0.9、5、150 000. 并且和 FCN-8s[8]、DeepLab v1[11]、DeconvNet[27]三种语义分割网络进行对比. 在 PASCAL VOC2012 数据集上各种类别的准确率和平均准确率的结果对比分析如表 3 所示.

表 3 中列出了 PASCAL VOC2012 中 20 种语义类别和背景 (bkg), 观察得到本文算法在大部分语义类别上的准确率表现较好, 并且平均准确率达到 73.5% 相比其他算法也有一定的提高. 现选取 FCN-8s、DeepLab v1 和本文算法在部分测试集上的语义分割图作对比, 如图 9 所示, 其中最右侧是语义分割标签图. 可见本文算法分割结果更好, 物体边缘分割更明显.

表 3 PASCAL VOC2012 上实验结果对比

算法	FCN-8s	DeepLab v1	DeconvNet	本文算法
bkg	91.2	**93.1**	92.7	92.8
areo	76.8	84.4	85.9	**88.3**
bike	34.2	54.4	42.6	**55.6**
bird	68.9	81.5	78.9	**83.3**
boat	49.4	63.6	62.5	**65.9**
bottle	60.3	65.9	66.6	**69.4**
bus	75.3	85.1	87.4	**88.7**
car	74.7	79.1	77.8	**82.3**
cat	77.6	83.4	79.5	**85.2**
chair	21.4	**30.7**	26.3	29.5
cow	62.5	74.1	73.4	**76.1**
table	46.8	59.8	60.2	**63.4**
dog	71.8	79.0	70.8	**81.1**
horse	63.9	76.1	76.5	**78.9**
mbk	76.5	**83.2**	79.6	82.6
person	73.9	**80.8**	77.7	79.6
plant	45.2	**59.7**	58.2	59.3
sheep	72.4	**82.2**	77.4	81.5
sofa	37.4	50.4	52.9	**54.4**
train	70.9	73.1	75.2	**77.1**
tv	55.1	63.7	59.8	**68.3**
mean	62.2	71.6	69.6	**73.5**

图 9 语义分割对比实验结果

5 总结与展望

本文提出了一种新的图像语义分割算法, 首先融合卷积网络中的多尺度特征信息, 再通过反卷积网络

将融合后的特征图放大到原图像大小, 最后对图像的每个像素点进行密度预测. 本文提出的批次中心化算法在网络中表现也较好, 在 SIFT-Flow 数据集上语义分割的平均准确率达到 45.2%, 几何分割的准确率达到 96.8%, 在 PASCAL 数据集上的平均准确率达到 73.5%. 当前语义分割任务的实时性是一个亟待解决的问题, 本文语义分割网络需要进一步优化和提升, 在保证分割准确率的同时提升速度.

参考文献

1 刘丹, 刘学军, 王美珍. 一种多尺度 CNN 的图像语义分割算法. 遥感信息, 2017, 32(1): 57–64. [doi: 10.3969/j.issn.1000-3177.2017.01.011]

2 魏云超, 赵耀. 基于 DCNN 的图像语义分割综述. 北京交通大学学报, 2016, 40(4): 82–91. [doi: 10.11860/j.issn.1673-0291.2016.04.013]

3 熊志勇, 张国丰, 王江晴. 基于多尺度特征提取的图像语义分割. 中南民族大学学报 (自然科学版), 2017, 36(3): 118–124. [doi: 10.3969/j.issn.1672-4321.2017.03.025]

4 Mardia KV, Hainsworth TJ. A spatial thresholding method for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1988, 10(6): 919–927. [doi: 10.1109/34.9113]

5 Giannakeas N, Karvelis PS, Exarchos TP, *et al*. Segmentation of microarray images using pixel classification-comparison with clustering-based methods. Computers in Biology and Medicine, 2013, 43(6): 705–716. [doi: 10.1016/j.compbiomed.2013.03.003]

6 Shi JB, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888–905. [doi: 10.1109/34.868688]

7 Rother C, Kolmogorov V, Blake A. "GrabCut": Interactive foreground extraction using iterated graph cuts. ACM SIGGRAPH. Los Angeles, CA, USA. 2004. 309–314.

8 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]

9 Pinheiro PO, Collobert R. Recurrent convolutional neural networks for scene labeling. Proceedings of the 31st International Conference on Machine Learning. Beijing, China. 2014. 82-90.

10 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. International Conference on Learning Representations. SanJuan, Puerto Rico. 2016.

11 Chen LC, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected CRFs. Computer Science, 2014(4): 357–361.

12 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

13 Chen LC, Papandreou G, Schroff F, *et al*. Rethinking atrous convolution for semantic image segmentation. arXiv: 1706.05587, 2017.

14 Chen LC, Zhu YK, Papandreou G, *et al*. Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv: 1802.02611, 2018.

15 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097–1105.

16 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.

17 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1–9.

18 Zeiler MD, Krishnan D, Taylor GW, *et al*. Deconvolutional networks. Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 2528–2535.

19 Zeiler MD, Taylor GW, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning. Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain. 2011. 2018–2025.

20 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Proceedings of 2014 European Conference on Computer Vision. Springer. Cham. 2014. 818–833.

21 Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France. 2015. 448–456.

22 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on

Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.

23 Liu C, Yuen J, Torralba A. SIFT flow: Dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(5): 978–994. [doi: 10.1109/TPAMI.2010.147]

24 Tighe J, Lazebnik S. SuperParsing: Scalable nonparametric image parsing with superpixels. In: Daniilidis K, Maragos P, Paragios N, eds. Computer Vision —ECCV 2010. Berlin, Heidelberg. Springer. 2010. 352–365.

25 Tighe J, Lazebnik S. Finding things: Image parsing with regions and per-exemplar detectors. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA. 2013. 3001–3008.

26 Farabet C, Couprie C, Najman L, *et al*. Learning hierarchical features for scene labeling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1915–1929. [doi: 10.1109/TPAMI.2012.231]

27 Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1520–1528.