1. 介绍

Deeplabv3是一种先进的语义图像分割算法，它通过引入扩张卷积来解决深度卷积神经网络（DCNNs）在处理图像时遇到的分辨率降低问题。这种技术通过在卷积核中插入间隔（即“洞”），允许网络在不增加参数的情况下捕获更高分辨率的特征，从而增强了对细节的感知能力。

在语义分割任务中，Deeplabv3面临着两个主要挑战。首先是由于连续的池化或卷积操作导致分辨率降低，这使得DCNNs倾向于学习越来越抽象的特征，而忽视了图像的细节。为了解决这一问题，Deeplabv3采用了扩张卷积，这种方法在语义图像分割领域已被证明是有效的。通过使用扩张卷积，网络能够在内部计算出与输入图像相同分辨率的特征图，而无需学习额外的参数。

第二个挑战是多尺度对象的存在。为了应对这一挑战，Deeplabv3提出了多种策略。首先，通过将DCNN应用于图像金字塔，可以在不同的尺度上提取特征，使得不同大小的对象在相应的特征图上更加突出。其次，采用编码器-解码器结构，利用编码器捕获的多尺度特征，并在解码器部分恢复图像的空间分辨率。第三，通过在原始网络的基础上串联额外的模块，来捕获更广泛的空间信息。最后，Deeplabv3还引入了扩张空间金字塔池化（Atrous Spatial Pyramid Pooling, ASPP）模块，该模块能够在多个尺度上获取卷积特征，并通过融合全局上下文的图像级特征来进一步提升性能。

DeepLabv3的设计还包括了串联或并行的多尺度扩张卷积模块，这使得网络能够更有效地捕获不同尺度的上下文信息。实验结果表明，Deeplabv3在PASCAL VOC 2012语义图像分割基准测试中取得了优异的性能，证明了其在图像分割领域的先进性和实用性。通过这些创新的设计和策略，Deeplabv3不仅提高了分割的精度，还增强了网络对不同尺度和复杂场景的适应能力。

1. 相关工作

已有的研究表明全局特征和上下文交互对于正确分类像素进行语义分割至关重要。在本项目中探讨了四种类型的全卷积网络（FCNs），这些网络通过利用上下文信息来执行语义分割任务。

图像金字塔是一种常用的方法，它使用相同的模型处理不同尺度的输入图像。小尺度输入能够编码长距离的上下文信息，而大尺度输入则保留了小物体的细节。例如，Farabet等人通过拉普拉斯金字塔变换输入图像，将每个尺度的输入提供给深度卷积神经网络（DCNN）并合并所有尺度的特征图。其他研究者则采用从粗糙到精细的策略依次应用多尺度输入，或者直接将输入调整为几个尺度并融合所有尺度的特征。然而，这种方法的主要缺点是，对于更大的或更深的DCNN来说，由于GPU内存的限制，它通常只在推理阶段应用。

编码器-解码器模型由两部分组成：编码器使得特征图的空间维度逐渐减小，以便于捕获更长距离的信息；解码器使得物体细节和空间维度逐渐恢复。例如，一些研究使用反卷积来学习低分辨率特征响应的上采样。SegNet重用了编码器中的池化索引，并学习额外的卷积层来密集化特征响应。U-Net在编码器特征和相应的解码器激活之间添加了跳跃连接。ReﬁneNet和其他一些研究展示了基于编码器-解码器结构的模型在多个语义分割基准测试中的有效性。

上下文模块模型包含串联的额外模块以编码长距离上下文。一种有效的方法就是将DenseCRF整合到DCNNs中。此外，一些研究提出联合训练CRF和DCNN组件，而其他研究在DCNNs的信念图之上采用几个额外的卷积层来捕获上下文信息。最近，一些研究提出学习一种通用且稀疏的高维卷积，或者结合高斯条件随机场和DCNNs进行语义分割。

空间金字塔池化模型采用空间金字塔池化来捕获不同范围的上下文。例如，ParseNet利用图像级特征来获取全局上下文信息。Deeplabv2提出了扩张空间金字塔池化（ASPP），其中并行的扩张卷积层采用不同的速率捕获多尺度信息。金字塔场景解析网络（PSP）在几个网格尺度上执行空间池化，并在多个语义分割基准测试中展示了出色的性能。此外，还有一些研究基于LSTM来聚合全局上下文。

1. 方法

3.1在本节中介绍扩张卷积在提取用于语义分割的密集特征中的应用。扩张卷积是一种强大的工具，它可以在不牺牲分辨率的情况下提取特征，这对于精确的语义分割至关重要。

深度卷积神经网络（DCNNs）在语义分割任务中表现出了显著的有效性。然而，这些网络中连续的池化和步距的组合显著降低了结果特征图的空间分辨率。通常，在最新的DCNNs中，每个方向上的空间分辨率会降低32倍。

扩张卷积是一种最初为未衰减小波变换的高效计算而开发的技术，并已在DCNNs中得到应用。扩张卷积通过在输入特征图上应用一种特殊的卷积操作，允许以更高的分辨率提取特征。具体来说，扩张卷积通过在卷积核权重之间插入间隔（即“洞”），从而增加了卷积核的视野，而不增加参数数量。

扩张卷积的一个关键优势是它允许模型显式控制全卷积网络中特征响应的密度。在DCNNs中，输出步长表示输入图像空间分辨率与最终输出分辨率的比率。例如，对于用于图像分类任务的DCNNs，最终的特征响应通常是输入图像尺寸的1/32，因此输出步长为32。如果希望在DCNNs中将计算特征响应的空间密度加倍，即输出步长为16，可以通过将最后一个降低分辨率的池化或卷积层的步长设置为1来避免特征衰减。然后，所有后续的卷积层都被替换为具有特定速率的扩张卷积层，从而允许网络提取更密集的特征响应，而无需学习任何额外的参数。

通过这种方式，扩张卷积不仅提高了特征提取的分辨率，而且保持了参数数量的效率。这种特性使得扩张卷积成为语义分割任务中提取密集特征的理想选择。通过串联或并行部署扩张卷积模块可以进一步提升网络的性能，更好地捕获图像中的上下文信息，从而实现更准确的语义分割。

3.2 串联扩张卷积模块的设计。首先复制了几个最后的ResNet块并将其串联排列。这些块中包含三次3×3卷积，除了最后一个块中的卷积之外，其余卷积都包含步长2，这种设计的动机在于，通过引入步长可以在更深的网络层中更容易地捕获长距离的信息，应用扩张卷积，其扩张因子由所需的输出步长值确定，其中输出步长设置为16。

在Deeplabv3模型中尝试了将ResNet块串联到block7，即在block4的基础上额外复制block5、block6、block7。如果不采用扩张卷积，输出步长将达到256。

通过这种设计可以在不同的网络层中捕获不同尺度的特征，从而更好地处理多尺度的语义分割问题。这种方法不仅提高了模型对长距离信息的捕获能力，而且通过调整扩张卷积的速率，有效地控制了特征响应的密度，避免了连续步长带来的信息衰减问题。这种多网格方法的应用，使模型能够更加灵活地适应不同的分割任务，提高了语义分割的准确性和鲁棒性。

3.3 在ASPP中引入了批量归一化技术，这有助于稳定和加速模型的训练过程。ASPP通过具有不同扩张速率的卷积有效地捕获了多尺度信息。然而，随着采样速率的增大，有效卷积核的权重的数量会逐渐减少。在极端情况下，当扩张速率值接近特征图大小时，3×3卷积核不再能够捕获整个图像的上下文，而是退化为一个简单的1×1卷积核，因为只有中心的卷积核权重是有效的。

为了解决这个问题，并为模型合并全局上下文信息，采用了图像级特征的策略。具体来说是在模型的最后一个特征图上应用全局平均池化，然后将得到的图像级特征提供给一个带有256个卷积核和批量归一化的1×1卷积。接着，将这些特征通过双线性上采样到所需的空间维度。

通过这种设计，Deeplabv3模型不仅能够有效地捕获不同尺度的特征，而且还能够整合全局上下文信息，从而提高了语义分割的准确性和鲁棒性。这种改进的ASPP模块为语义分割模型提供了一个强大的工具，使其能够更好地处理各种复杂场景和不同尺度的物体。

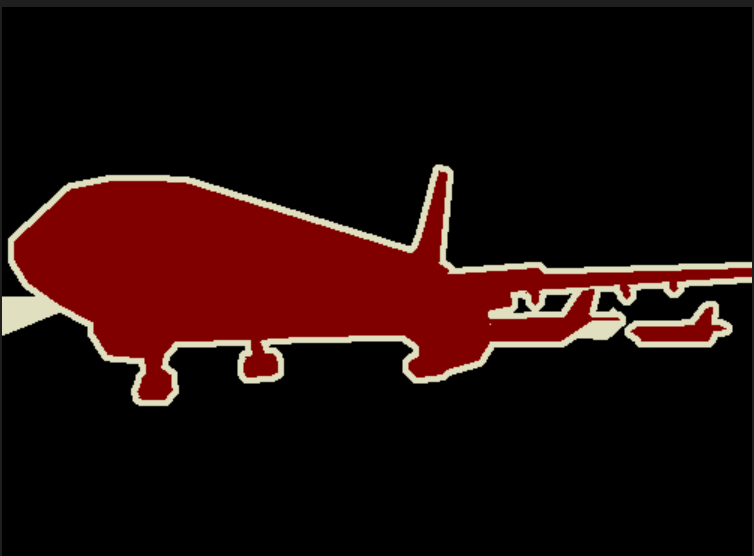
1. 实验

4.1 数据集

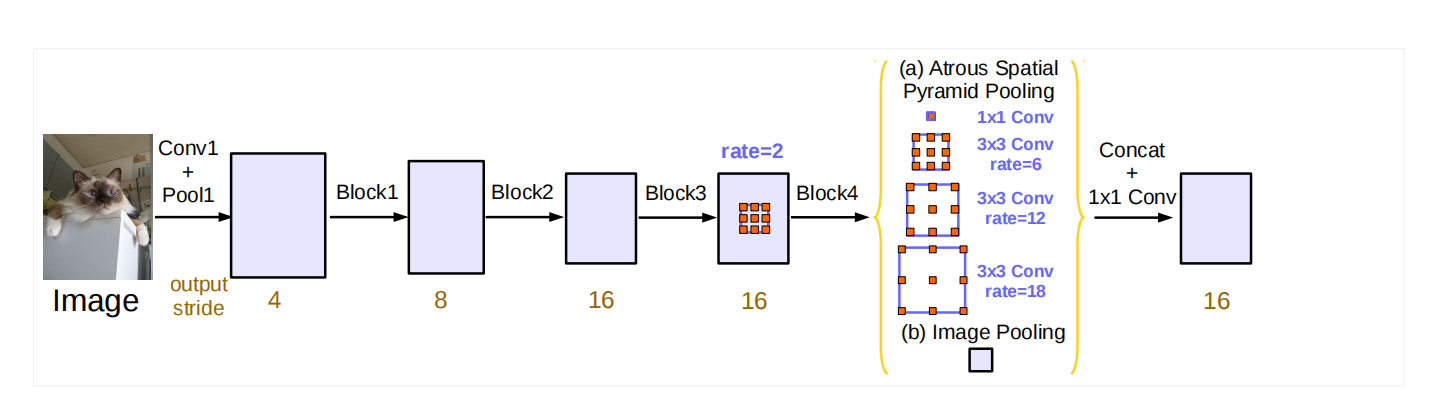
PASCAL VOC挑战赛 （The PASCAL Visual Object Classes ）是一个世界级的计算机视觉挑战赛，PASCAL全称是Pattern Analysis, Statical Modeling and Computational Learning，是一个由欧盟资助的网络组织。PASCAL VOC挑战赛主要包括以下几类：图像分类(Object Classification)，目标检测(Object Detection)，目标分割(Object Segmentation)，行为识别(Action Classification) 等。

目标分割任务重，Pascal VOC数据集中主要包含20个目标类别，分别是Vehicles中的Aeroplane、Bicycle、Boat、Bus、Car、Motorbike、Train，Household中的Bottle、Chair、Dining table、Potted plant、Sofa、TV/Monitor，Animals中的Bird、Cat、Cow、Dog、Horse、Sheep和Other中的Person类。

语义分割任务中，训练验证的具体划分是，训练集1464张，验证集1449张。从文件夹中可以观察，两个相同前缀的图像分别表示原图和mask图，例如：

4.2 实现细节



在代码中，主要使用了论文中的该网络框架进行了原文的复现，训练过程中的一些参数如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Batch size | Epoch | Lr | Momentum | Weight-decay |
| 4 | 30 | 0.0001 | 0.9 | 1e-4 |

数据准备：首先，需要准备和预处理训练数据集Pascal VOC2012。这通常包括图像的收集、标注（像素级别的标签）、归一化处理以及数据增强等步骤，具体使用了torchvision库中的transforms模块对数据进行了resize、totensor和normalize操作。

网络构建：接着，构建Deeplabv3的网络结构。这包括编码器-解码器架构、扩张卷积层、扩张空间金字塔池化（ASPP）模块以及用于恢复空间分辨率的上采样模块。

损失函数定义：定义损失函数以评估模型的预测与真实标签之间的差异，在本次实验中使用了交叉熵损失作为损失函数。

优化器选择：使用随机梯度下降法，用于在训练过程中调整网络的权重。

训练过程：开始训练过程，其中包括前向传播、损失计算、反向传播和权重更新。在每次迭代中，模型会处理一批图像（batch size=4），并根据损失函数的结果更新网络权重。

评估与调整：在训练过程中，定期对模型进行评估，以监控其在验证集上的性能。

模型保存：在训练过程中，定期保存模型的权重，以便在训练中断或完成时能够恢复到最佳状态。

推理与应用：训练完成后，使用训练好的模型对新的图像进行推理，以进行语义分割。这通常涉及将图像输入到模型中，并得到每个像素的类别标签。

4.3 评价指标

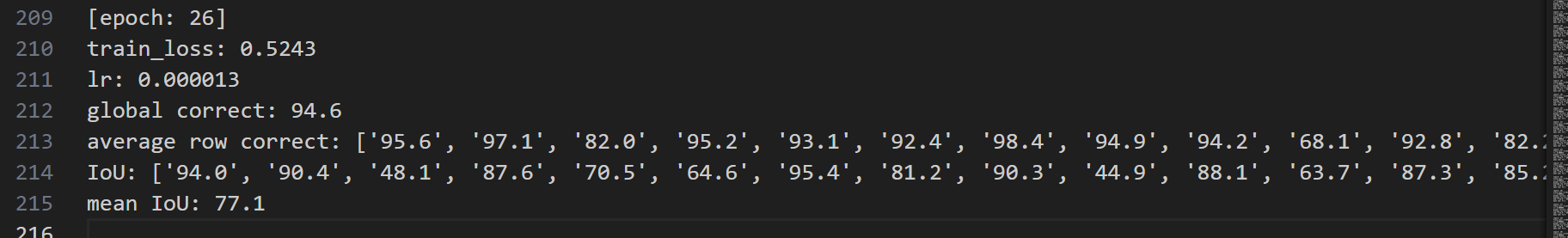
准确率（Pixel Accuracy）：准确率是最简单的评价指标之一，它衡量的是模型正确分类的像素数占总像素数的比例。具体来说，对于一幅图像，如果模型预测的类别与真实标注的类别一致，则认为该像素分类正确。然后，将所有像素中分类正确的像素数除以总像素数，得到准确率。

交并比（IoU）：交并比是一种更为复杂的评价指标，它衡量的是模型预测的分割区域与真实标注的分割区域之间的重叠程度。IoU对于评价模型的定位精度非常有效，特别是在对象边缘的精确度上。

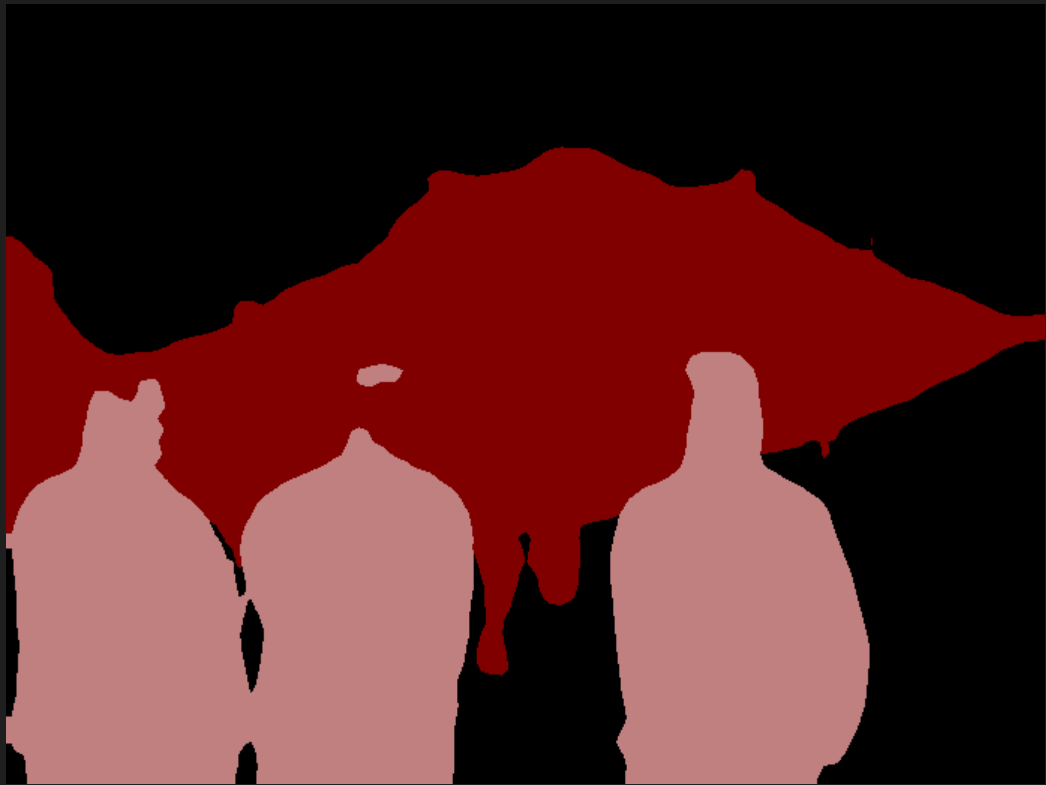
IoU的计算基于预测的分割区域（预测的像素集合）和真实的分割区域（标注的像素集合）。首先计算两者的交集（同时被预测和标注为某一类别的像素集合），然后计算并集（被预测或标注为某一类别的像素集合），最后，IoU是交集与并集的比值。

4.4 实验效果

经过30轮的训练后，发现准确率最高是94.6%，平均IOU最高是77.1%，运行日志截图如下：



同时，使用第26个epoch的模型对一张图片进行测试，查看其分割效果，结果如下图所示，左图为原图，右侧图是分割之后的图，可以观察出，第一行的三个人被分割成了浅粉色，飞机被分割成了深红色，证明我们的模型训练有效，不仅能够在类别上进行区分，还能够较为准确的分割出目标的边缘和轮廓。观察第二行图片，模型可以将人和杯子分割出来，杯子呈现紫色。





5、结论

本项目使用了Deeplabv3模型完成了语义分割的任务，Deeplabv3采用上采样卷积核的扩张卷积来提取密集的特征图，并捕获上下文信息。具体来说，为了编码多尺度信息，使用扩张空间金字塔池化模块通过在多个采样参数和有效视野下增强了图像级特征，通过训练后，在Pascal VOC数据集上达到了94.6%的准确率和77.1%的交并比，实现了语义分割的效果。