基于 PADDLEPADDLE 的语义分割模型研究

计算机科学与技术（物联网方向）15-3 张涵

指导教师 孙钰

摘要

深度学习技术在当下十分火热，被广泛应用于人工智能领域。其中，卷积神经网络在图像识别领域具有很强的统治力，在全图或者局部的任务上都有很大的性能提升。早前的方法已经将卷积网络用于将每个像素标记为其封闭对象及区域的类别的语义分割，但这种方法存在缺点，而全卷积网络在传统卷积神经网络上做出了一些改进从而能够达到更好的效果。全卷积神经网络是端对端，像素对像素的训练，它在没有更进一步深入的机制的情况下就有了很好的效果。

基于全卷积神经网络的理念诞生了许许多多不同的模型，其中 Image Cascade Network（ICNet) 是一个基于 PSPNet 的实时语义分割网络，ICNet 的思路是使用高中低不同分辨率的图像作为网络模型的输入，不同分辨率图像输入的网络分支有着不同的运算复杂度的设计，最后将这几个分支输出的结果合并，这样能够结合高分辨率图像的精确推断和低分辨率图像输入分支的高运算速度， 兼顾了准确性和运算效率，使之能够达到高准确度的实时运算的效果。

关键词：深度学习，语义分割，全卷积神经网络

The research on semantic segmentation model based on PaddlePaddle

Computer Science and Technology(The Internet of Things)15-3 Zhang Han

Supervisor Sun Yu

Abstract

Deep learning technology has made many very constructive breakthroughs in the direction of artificial intelligence in the past decade and has developed rapidly. Among them, the convolutional neural network has a strong dominance in the field of semantic segmentation. The convolutional neural network not only improves the classification of the full graph, but also makes progress on the local tasks of structured output. Earlier methods have used convolutional networks to mark each pixel as a semantic segmentation of its closed objects and categories of regions, but this approach has drawbacks, and full convolutional networks have been made on traditional convolutional neural networks. Some improvements can achieve better results. The full convolutional neural network is end-to-end, pixel-to-pixel training, and it has a good effect without further in-depth mechanisms.

Based on the concept of a full convolutional neural network, many different models have been born. Image Cascade Network (ICNet) is a real-time semantic segmentation network based on PSPNet. The core idea of ICNet is to transform the input image into multiple different resolutions. Multi-resolution image input is calculated by sub-networks with different computational complexity, and then the results are combined to take advantage of the high processing efficiency of low-resolution images and the high inference quality of high-resolution images. The ICNet model is high in this way. A balance between the accuracy of the resolution image and the efficiency of the low complexity network enables it to achieve real-time results.

Key words: deep learning, semantic segmentation, full convolutional neural network

目录

1 绪论

1.1 毕业设计的背景

2006 年，Geoffrey Hinton 在《Science》发表了一篇文章[1]，提出了多隐层的神经网络有着优秀的学习能力，以及可通过逐层初始化降低深度神经网络在训练上的难度，自此之后，深度学习在工业界以及学术界开始大放光彩。最先在图像和语音领域中，深度学习得到了应用，从 2011 年开始，微软和谷歌使用深度学习的方法大幅提升了识别的正确率。2012 年，师从 Geoff Hinton 的 IIya Sutskever 和 Alex Krizhevsky 也使用了深度学习方法答大幅刷新了在 ImageNet 上的成绩，打败了 Google 团队。2012 年，谷歌首席架构师 Jeff Dean 和斯坦福大学教授 AndrewNg 主导的 GoogleBrain 项目，构建了一个有数十万处理器构成的深度神经网络，并应用到了语音、图像的识别上，取得了极佳的效果。此外，深度学习在搜索领域也备受青睐。如今，深度学习已经在图像、语音、大数据和自然语言处理等方面获得了全面且广泛的应用。深度学习技术使得计算机相关的各个领域都产生了巨大的变化和进步。

“卷积神经网络(CNN)的灵感来源于针对猫视觉皮层的研究，可以说是仿生学的杰作。卷积神经网络最初是为识别图像专门设计的一种神经网络，其主要特点在于局部感受域与权值共享。这些特点不但减少了神经网络的计算量，而且提高了其鲁棒性。这种网络结构极大的利用了图像在空间上相互关联的特性，对平移、比例缩放、倾斜或者共他形式的变形具有高度不变性”[2]。但是该方法也存在一些缺点。

1.2 毕业设计的研究目的和意义

在计算机视觉领域中，图像语义分割是主要课题之一，语义分割技术可以依据一些图像的特定性质把图像分割成多个具有不同含义的区域，使得计算机能够像人一样去理解图像内容，对一个场景做出分析，许多场景下都需要高效且精确的分割技术，以从影像中推理出相关的知识和语义，如自动驾驶、室内导航、医疗影像分析、虚拟现实等。其中，自动驾驶系统中，车载摄像头和激光雷达采集的图像经语义分割可以发现道路上的行人和车辆，以辅助驾驶和避让；在医疗影像分析中，语义分割可用于肿瘤图像分割和龋齿诊断等。如虚拟现实，自动驾驶，医学图像分析等。

1.3 国内外研究状况及研究成果

2014 年，Jonathan Long[3]等人提出了全卷积神经网络，全卷积网络(Fully Convolutional Networks)与传统的卷积神经网络的主要区别是把最后一层的全连接层换成了卷积层，对经过多层卷积和池化的缩小后图像进行反卷积，上采样恢复出一个与输入图片尺寸相同的 heatmap，输出每一个像素点对应的类别概率。通过这种方式，FCN 类的模型可以像素级别地进行预测，解决了图像的语义分割

（semantic segmentation）问题。全卷积网络的优势之一在于可以接受任意大小的输入，而传统卷积神经网络由于全连接层的缘故无法做到这一点；在 CNN 中，为了对一个像素进行分类，该像素附近的一整个图像块需要被当成 CNN 的输入，用于网络的训练和预测，这样会带来存储方面的较大开销，并且由于图像块与图像块之间具有很多重复部分，会导致计算效率的下降，而 FCN 则很好地避免了上述这些问题。在全卷积神经网络这样一个对语义分割问题有着良好效果的模型提出之后，至今已陆陆续续诞生了许多不同种类基于 FCN 的模型和优化模型的新技术: 2015 年，Fisher Yu, Vladlen Koltun 等人提出了空洞卷积[4](Dilated Convolutions)；谷歌团队在 2014 年发布了 DeepLab[5]，至今已更新到 v4 版本；2018 年，卡耐基梅隆大学的 Xiaodan Liang 等人提出了 DSSPN[6]；除此以外还有 RefineNet[7]，剑桥大学的 SegNet[8]，U-net[9]等等。2016 年，香港中文大学和商汤组的 Hengshuang Zhao， Jianping Shi 等人提出了使用了金字塔池化模块(pyramid pooling module)的 PSPNet[10] (Pyramid Scene Parsing Network), 2018 年，该团队在 PSPNet 的基础上推出了一个能够兼顾准确率和运算效率的实时语义分割网络模型 ICNet[11] (Image Cascade Network)。这些网络模型对语义分割领域从不同方向进行了许多改进。

1.4 毕业设计的研究内容及方法

此次毕业设计的主要内容是构建实时语义分割网络模型 ICNet，并且在 cityscapes 数据集上进行训练，预测和评估。

研究具体分为准备阶段和实现阶段，在准备阶段，首先学习了 PYTHON 语言以及深度学习相关的原理及知识，并且重点阅读了计算机视觉方面关于图像语音分割方面的论文与源代码，熟悉了近年的发展状况以及各种网络模型的研究意义与特点以及具体实现方法，理解模型各层的原理及作用，为后期实现打好基础。并且学习了百度的开源深度学习框架 PaddlePaddle，阅读了运用了 PaddlePaddle 的实例代码及 Fluid API 的使用指南，了解其使用方法。

在后期实现阶段主要参考《ICNet for Real-Time Semantic Segmentation on High-Resolution Images》这篇论文，根据文章提供的设计思路及原理在 PaddlePaddle 平台上实现 ICNet 模型。

1.5 论文的构成

第一部分是绪论，包含：毕业设计的背景；毕业设计的研究背景和意义；国内外研究状况和研究成果；毕业设计的研究内容及方法。

第二部分是总体设计，包括环境及相关技术的介绍；卷积神经网络基本层；ICNet 网络架构；反向传播算法；参数优化的方法；防止过拟合。

第三部分是具体实现，包括数据集预处理，构建 ICNet 基础架构以及金字塔池化、学习率衰减等部分。

最后是结论、致谢以及参考文献等内容。

本文主题用 ICNet 网络模型对图像进行像素级别的分类预测。通过得出的实验数据验证 ICNet

进行实时语义分割的能力。

2 总体设计

2.1 环境及相关技术介绍

2.1.1 PaddlePaddle

PaddlePaddle 是百度研发的开源深度学习平台，有全面的官方支持的工业级应用模型，涵盖自然语言处理、计算机视觉、推荐引擎等多个领域，并开放多个领先的预训练中文模型。

PaddlePaddle 可提供深度学习并行技术的深度学习框架，拥有多端部署能力，支持服务器端、移动端等多种异构硬件设备的高速推理，预测性能有显著优势。目前 PaddlePaddle 已经实现了 API 的稳定和向后兼容，具有完善的中英双语使用文档。

2.1.2 Python

Python 是一种解释型计算机程序设计语言,支持面向对象的编程和结构化编程，Python 语法十分简洁，特点之一是使用缩进代替括号表示语句块。Python 具有丰富并且功能强大的库，并且是一种“胶水语言”，能够使由不同编程语言完成的功能模块结合在一起，并且是一种免费开源的语言，同时具有高可移植性和可扩展性。

2.2 基本层

2.2.1 概念

卷积神经网络（Convolutional Neural Network, CNN）是前馈神经元网络，其人工神经元可以覆盖区域的附近的一部分。 由数量众多的独立神经元构成每一层。相邻两层的神经元相互连接。 全卷积神经网络去掉了卷积神经网络的全连接层，目前包括 ICNet 在内的如今绝大多数语义分割模型都是全卷积神经网络，全卷积神经网络通常包括以下几个重要部分：卷积层，池化层，激活函数(ReLu)，上采样层。

2.2.2 卷积层

在卷积层中，我们使用一个特定大小的卷积核对图像进行卷积操作，这样相当于逐步选取整个图像的一小部分区域进行局部特征提取，因为图像中一个像素点周围的一片图像块有会有一定的相关性，把这些特征作为滤波器和整个图像做卷积运算，最终的到原始图像中的每一位置上的不同特征的激活值，卷积核的主要特性包括卷积核的大小，步长以及数量。通常卷积核的个数越多，能学习到的特征就越多，性能表现就越好。但是卷积核的数量和网络的复杂度若是过高，则会因为参数量太大导致过高的运算复杂度以及过拟合问题。所以卷积核的设计需要依据数据集和所用网络模型的特点来制定。

2.2.3 激活函数层

在神经网络中，激活函数一般运用在卷积层之后，激活函数是一个非线性函数，可以处理输入信号并转换成输出信号，将其送入下一层。常用的激活函数主要有：

(1) Sigmoid 函数

$$\sigma(x)=\frac{1}{1+e^{-x}} \tag{2.1}$$

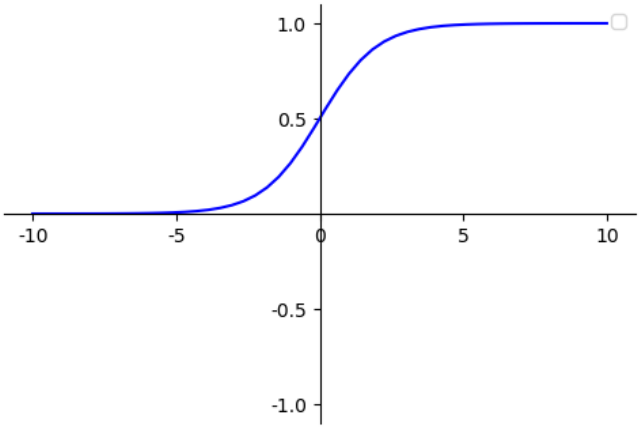

图 2. 1 Sigmoid 函数

Figure2.1 Sigmoid function

(2) Tanh 函数

$$\tanh(x)=\frac{\sinh(x)}{\cosh(x)}=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2.2}$$

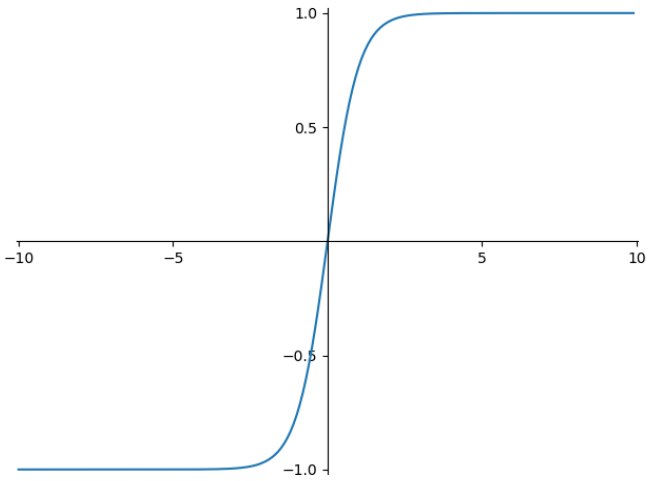

图 2.2 tanh 函数

Figure2.2 tanh function

(3)ReLu 函数

$$f(x)=\max(0,x) \tag{2.3}$$

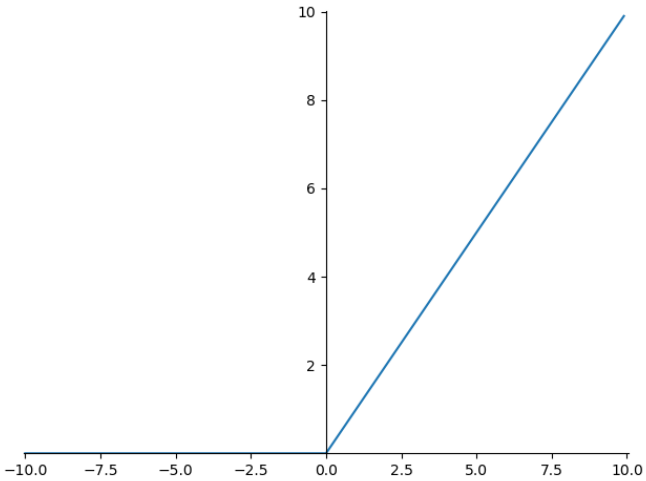

图 2.3 ReLu 函数

Figure2.3 ReLu function

在 ICNet 中，我们使用的是 ReLu 激活函数，Alex Krizhevsky 在一篇论文[11]中曾指出 ReLu 有着比传统的 sigmoid 函数和 tanh 函数更好的表现。Sigmoid 在函数在反向传播求误差梯度时求导需要较多运算量，而采用 Relu 激活函数则会降低对计算量的需求。并且，在深度网络中，当 sigmoid 函数和 tanh 函数反向传播时，梯度容易趋于消失，从而无法正常完成网络训练。此外，ReLu 将某些神经元的输出归零。这可以使得网络稀疏并减少参数相互依赖性，减轻过度拟合的问题。

2.2.4 池化层

在卷积层提取完图像特征之后，我们需要减少参数量，不然会带来很大的计算量以及过拟合问题，池化(pooling)操作可以在保留主要特征的同时减少参数量，并且提高系统鲁棒性。

常用的池化方式主要有平均池化、最大池化。

平均池化（Mean-pooling）是对区域内的点求平均值，特点是在特征提取时能够更好地保留图像背景信息。

最大池化(Max-pooling)是对区域内的点取最大值，特点是能更好地保留图像纹理信息。

2.2.5 反卷积层

语义分割的神经网络最终需要输出与原图尺寸大小相同的图进行预测，而输入图像经过双线性插值以及网络内各层的卷积和池化操作尺寸会变小，所以在网络的最高层需要对输出图像进行上采样回复到原有尺寸。双线性插值法或转置卷积、空洞卷积都可以达到上采样的目的，在本网络中使用转置卷积。

2.3 ICNet 网络模型的理论基础

2.3.1 影响运行速度的因素

在快速语义分割方面，速度是用来评判一个语音分割模型的重要标准。当一个高精确度的快速图像语义分割框架出现时，在视频上的语义分割将会大大受益。近年来 Yolo[12]和 SSD[13]的出现大大改善了它。可是在语义分割领域，对高判断速度模型的研究只能算是刚起步，已有的 SegNet[14]模型为了高速推断减少了了一些层来减少参数量；ENet[15]是一个规模很小的网络，这些方法提升了运算效率，但相比非实时的语义分割模型，准确性方面的表现并不理想。

影响运行速度的最重要的因素是图像分辨率，分析表明，与全分辨率图像相比，半分辨率图像只花费四分之一的时间。一种简单的方法是利用低分辨率图像作为输入，使用比例为 1/2 和 1/4 的下采样图像输入到基于 FCN 的框架，如 PSPNet。获得了判断结果之后，我们再将其上采样恢复到原始尺寸。

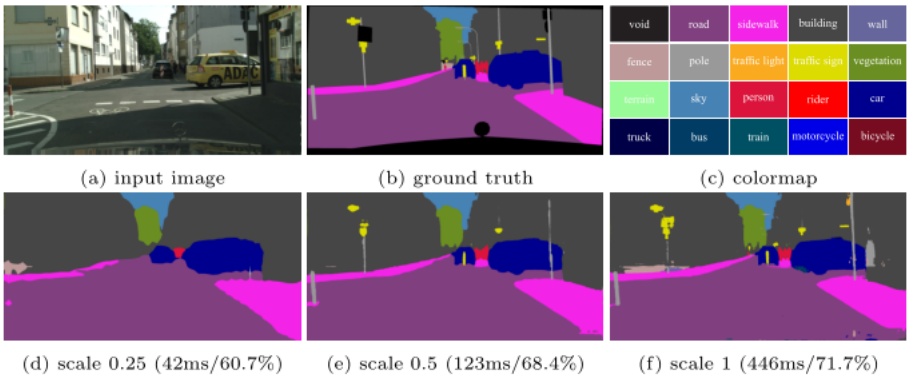

图 2.4 PSPNet 在 cityscapes 验证集上做出的预测

Figure2.4 prediction of PSPNet50 on the validation set of Cityscapes.

如图 2.4 所示，下采样尺度为 0.25 时候耗费时间较少，但分类结果不够精细，对于细节的分辨和分类不够理想；尺度为 0.5 时，相对能够分类出更多的物体，但较远处细小的交通物体和人仍无法被较好地分类，而且需要时间太多，无法达到实时运算。

除了直接对输入图像进行降采样以外，另一个直接的选择是在推理过程中按较大的比例缩小功能图。FCN 对其进行了 32 倍取样，DeepLab 进行了 8 倍下采样。使用 1：8,1：16 和 1:32 的下采样比测试 PSPNet50 结果如图 2.5 所示。较小的特征映射可以以推断质量的下降为代价产生更快的预测。丢失的信息同样包含在低层次的信息中。另外，考虑到即使使用 1:32 比例的最小特征图，系统在推断中仍然需要大约 131ms，与实时分割的目标仍然存在差距。

表 2. 1 选择下采样因子 8, 16 和 32 时在 PSPNet50 上耗费的总时间

Table2.1 Total time spent on PSPNet50 when choosing downsampling factors 8, 16 and 32

Downsample Size	8	16	32
mIOU(%)	71.7	70.2	67.1
Time(ms)	446	177	131

此外还可采用模型压缩方案,对于每个卷积核，计算权重 weights 的 L1 的和，之后对 L1 和进行降序排序，只留下前几个有意义的值。试验对比如图 2.6 所示。

表 2. 2 不同核权值的性能和运行速度对比

Table2.2 Model compression with kernel keeping rates 1, 0.5 and 0.25.

Kernel Keeping Rates	1	0.5	0.25
mIOU(%)	71.7	67.9	59.4
Time(ms)	446	170	72

虽然只有部分的核权值得到了保留，但耗时较长而且精度下降很多，比如当 Rates 为 0.25 时，仍然需要长达 72ms 的计算时间。

2.3.2 模型架构

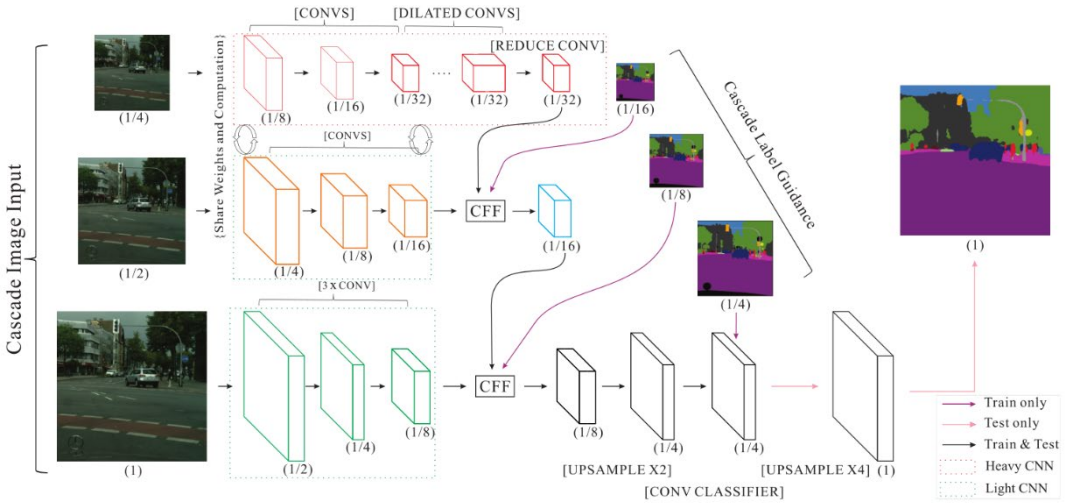

图 2.5 ICNet 的网络结构

Figure2.5 Network architecture of ICNet

上述的几种方法均无法达到较好的效果，采用下采样输入图像虽然可以减小运行时间,但预测结果会丢失很多细节信息,直接输入图像,又需要很多运行时间。而如图 2.7 所示结构的 ICNet 模型中，低分辨率图像输入的网络分支采用了高复杂度的设计，高分辨率图像输入的网络分支采用低复杂度

的设计，通过这种方式在高分辨率图像的准确性和低复杂度网络的效率之间获得了平衡。

低分辨率所提出的框架如图所示。缩放 1 的输入全分辨率图像在缩放 1/2 和 1/4 的缩放后生成两个较低分辨率的图像。这些级联图像分三路输入我们的 ICNet。对于最低分辨率输入，它通过顶级分支，这是一个基于 FCN 的 PSPNet 架构。由于输入尺寸只有原始尺寸的 1/4，因此卷积层相应地缩小了 1/8 的比例特征贴图，并且产生了原始空间尺寸的 1/32。然后使用几个膨胀的卷积层来放大感受野而不下采样空间尺寸，输出具有原始尺寸 1/32 尺寸的特征图。

对于 1/2 尺寸的中分辨率图像，它在第二个分支中处理。通过数个卷积层的处理后，输出特征图的大小为原始图的 1/16。为了将 1/16 大小的特征图与顶部分支中的 1/32 大小的特征图融合，ICNet 使用了一个将在之后讨论的级联特征融合（CFF）单元。这种融合产生了具有原始分辨率 1/16 的组合特征映射。另外，中低分辨率输入的两个子网的卷积参数可以共享，从而节省了计算量并减少了参数数量。

对于高分辨率图像，与第二个分支中的操作类似，它由几个卷积层处理，降采样率为 8，得到 1 / 8 尺寸的特征图。由于中分辨率图像的网络分支中已经很好地学习到了语义细节，因此我们可以对处理高分辨率输入时的卷积层数量进行减少。

这里我们只使用三个卷积层，每个卷积层的内核大小为 3×3，步长 2 将分辨率减小到原始输入的 1/8。与分支 2 中描述的融合类似，我们使用 CFF 单元把当前分支输出的图像与先前 CFF 单元的输出合并成一个特征图。最后，我们得到原始图像尺寸的 1/8。

对于每个分辨率的输出特征图，我们首先对其进行两倍的上采样。为了使学习过程更好，我们使用级联标记引导策略。使用 1/16,1/8 和 1/4 的 ground truth 标签指导低，中和高分辨率输入的学习阶段。在测试阶段放弃低和中指导操作，只留下高分辨率分支。这种级联标签大大降低了计算量且保证了结果的准确性。

该级联标签使得梯度优化更加平滑以便于训练迭代。随着每个分支学习能力的增强，最终的预测图没有被任何一个单一分支所支配。同时，在测试期间放弃引导部分也是效率方面的一种收益。

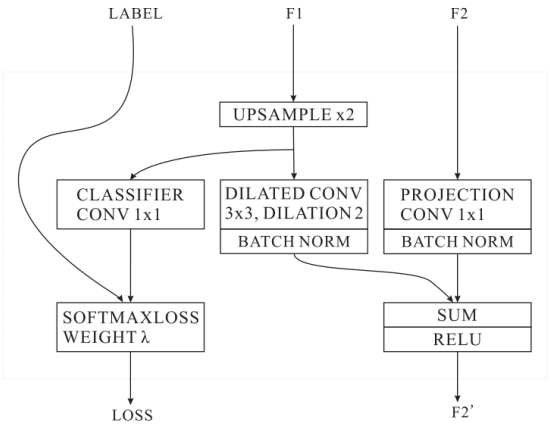

图 2.6 级联特征融合单元

Figure2.6 Cascade feature fusion.

为了结合来自不同分辨率图像的级联特征，我们最终提出了如图 2.8 所示的级联特征融合（CFF）单元。该单元的输入包含三个分量：分辨率为 H1×Wi×C1 和 H2×W2×C2 的两个特征图 F1 和 F2，真实标签分辨率为 H2×W2×1，F2 的尺寸是 F1 的尺寸的 2 倍。上采样应用于使 F1 尺寸与 F2 相同。然后应用一个尺寸为 3×3 和膨胀 1 的扩张卷积层来改善上采样特征。

这种扩张卷积可以结合来自多个相邻像素的特征信息，相比之下，直接上采样则无法做到这一点。相比于原始特征的反卷积操作，扩张卷积只需要较小的卷积核。在我们的实现中，对于特征 F2，利用具有内核大小 1×1 的投影卷积层以与特征 F1 的输出相同的尺寸来投影它。

然后使用两个批处理标准化层来标准化这两个特征。接着是具有'SUM'操作和 ReLU 层的元素层，我们得到融合特征 F，其具有与 F2 相同的分辨率。为了加强对 F1 的学习，我们对上采样的 F1 使用辅助标签指导。如图中的辅助损失权重设置为 0.4。

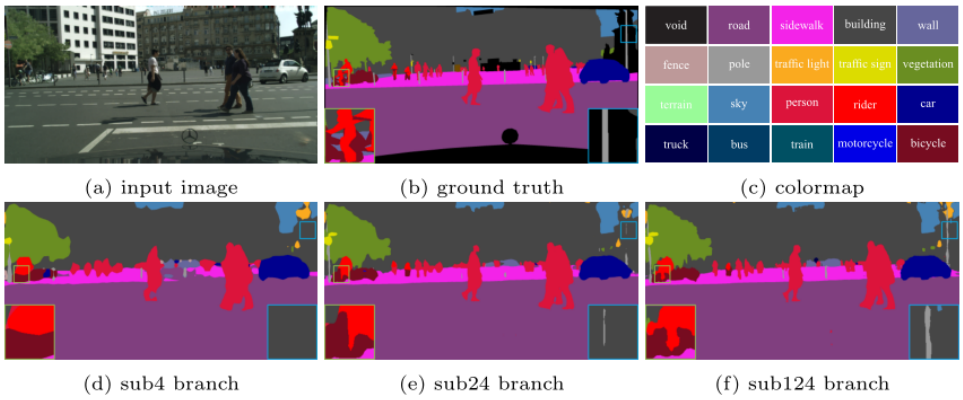

图 2.7 Cityscapes 数据集上 ICNet 各个分支上的预测

Figure2.7 prediction of ICNet in each branch on Cityscapes dataset

同时，在 ICNet 中，最深层的网络结构最多只适用于低分辨率数据，能够有效地提取大多数语义信息。因此，即使超过 50 层，推理操作和内存消耗也不大，分别为 18ms 和 0.6GB。

分支 2 中有 17 个卷积层，分支 3 中只有 3 个卷积层。通过这些少量的控制层来处理中值和高分辨率的输入数据，来自分支 2 和 3 的推理操作总数受到很好的约束。与此同时，分支 2 中的卷积层与第一分支共享权重和计算。因此，使用两个分支仅花费 6ms 来构建融合特征映射。

分支 3 层次更少。虽然分辨率很高，但分支三的推断时间仅为 9 毫秒。通过所有这三个分支，ICNet 能够成为一个非常高效和内存友好的模型。

2.4 反向传播算法

反向传播（Backpropagation）是一种用于通过梯度下降等优化方法学习人工神经网络的方法。该方法计算网络中所有权重的损失函数的梯度，之后逐渐将参数向最优点方向优化。

反向传播需要每个输入值的预期已知输出来计算损失函数梯度，每一次迭代的梯度可以用链式法则来计算。并且反向传播需要节点的激活函数可微。

构成反向传播的两个阶段分别是激励传播和权重更新。

激励传播为第一阶段，网络中每次迭代的传播环节包括两步：首先对于前向传播，将训练输入信号发送到网络以获得激励响应。之后在.反向传播中，将所产生的激励响应输入对应的目标输出求差值，得到隐藏层和输出层的响应误差。

第二阶段是权重更新，首先将输入激励乘上响应误差以获得权重梯度，之后将此梯度乘上一个比例并取反后加到权重上。误差扩大的方向被梯度方向指明，因此在更新权重时对其取反可以减少权重引起的误差。

通常这两个阶段在训练中会反复迭代，直到网络的输出结果达到期望目标为止。

2.5 参数优化的方法

在学习过程中，我们需要选择优化器用来更新和计算影响模型训练和模型输出的网络参数，使其不断接近最优值，从而最小化损失函数。常用的优化器方法有批量梯度下降法(Batch gradient descent,BGD)，随机梯度下降算法(Stochastic gradient descent,SGD)，以及引入了动量(Momentum)的 SGD 等等。

和 BGD 的一次用所有数据计算梯度相比，SGD 每次更新时对每个样本进行梯度更新，在较大的数据集中，样本相似的可能性比较高，如此一来 BGD 在计算梯度时会出现冗余的情况，而 SGD 一次只进行一次更新，这样一来就没有冗余，速度也会更快，而且还能够新增样本。

$$\theta = \theta - \eta \cdot \nabla_{\theta} J(\theta; x^{(i)}; y^{(i)}) \tag{2.4}$$

SGD 是通过每个样本来迭代更新一次，若样本量足够大，参数更新到最优解的时候可能只需

要样本中的一部分。然而 BGD 一次迭代就要用到数十万规模大小的训练样本，而且一次迭代无法达到最优。缺点是 SGD 的噪音较 BGD 要多，SGD 以高方差进行快速更新，目标函数会出现抖动，使得 SGD 并不是每一次迭代都在向最优值方向更新参数。所以虽然训练速度快，但是准确度会下降，并不一定是全局最优。然而另一方面，也正是因为计算的抖动，可以让梯度的计算跳出局部最优值，最终到达全局的最优点。

由于 SGD 在某些接近局部最优的情况下会不断震荡困在该点附近，有一个优化方案可以改善该问题，就是引入动量(Momentum), Momentum 模拟物理里动量的概念，积累之前的动量来替代真正的梯度。

$$\begin{aligned} v_t &= \gamma v_{t-1} + \eta \nabla_\theta J(\theta) \\ \theta &= \theta - v_t \end{aligned} \tag{2.5}$$

通过添加一个衰减因子到历史更新向量，并加上当前的更新向量。当梯度保持相同方向时，动量因子加速参数更新；而梯度方向改变时，动量因子能降低梯度的更新速度。这样一来让 SGD 速度更快，并且抑制震荡。

在本文的 ICNet 模型中，就是使用了 Momentum 优化器进行参数优化。

2.6 防止过拟合

当一个网络对于已有的数据拟合得过于好的时候，很可能会使模型失去一定的泛化能力，从而在对未知数据进行预测的时候导致准确性低下。在对模型进行训练时，有可能遇到训练数据量不足，或者在对模型进行过度训练时，常常会导致模型的过拟合。为了解决过拟合问题，可以采用正则化方法，做法是在损失函数后面添加一个额外项，常用的额外项一般有两种，L1 正则化和 L2 正则化。L1 正则化和 L2 正则化可以看作是损失函数的惩罚项。所谓“惩罚”是指限制损失函数中的某些参数。

L1 正则化是指权值向量 ω 中各个元素的绝对值之和，公式 2.3 为带 L1 正则化的损失函数。

$$J = J_0 + \alpha \sum_w |w| \tag{2.6}$$

L2 正则化是指权值向量 ω 中各个元素的平方和然后再求平方根，公式 2.4 为带 L2 正则化的损失函数。

$$J = J_0 + \alpha \sum_w w^2 \tag{2.7}$$

本文的 ICNet 模型使用了 L2 正则化，正则化可以惩罚过大的权重，更倾向于学习小权重，以此增强网络模型鲁棒性。

3 具体实现

3.1 cityscapes 数据集预处理

使用 cityscapes 数据集提供的预处理脚本中的 createTrainIdLabelImgs.py 在标注数据集 gtFine 中生成用于训练的 ID 标签的 png 文件。

3.2 生成训练表

编写了一个脚本 generatelist.py 脚本用来遍历数据集，获取图片及标注的文件名及地址写入 train.list 和 val.list，用于之后网络的训练和验证。

3.3 ICNet 框架

ICNet 网络的核心架构：三个不同分辨率输入所对应的不同子网以及连接彼此的级联特征融合单元(CCF)，在 icnet.py 中完成，主要代码如下：

```
def icnet(data, num_classes, input_shape):
    image_sub1 = data       #原图
    image_sub2 = interp(data, out_shape=input_shape * 0.5)     #原图缩小 1/2


    s_convs = shared_convs(image_sub2)              #1/2 图进行
    sub4_out = sub_net_4(s_convs, input_shape)
    sub2_out = sub_net_2(s_convs)                #sub2 与 sub4 共享卷积参数
    sub1_out = sub_net_1(image_sub1)
    sub24_out = CCF24(sub2_out, sub4_out, input_shape) #sub2_out 与 sub4_out 通过 CCF 融合成 sub24_out
    sub124_out = CCF124(sub1_out, sub24_out, input_shape)   #sub24_out 与 sub1_out 通过 CCF 融合成 sub124
    #输出
    conv6_cls = conv(
        sub124_out, 1, 1, num_classes, 1, 1, biased=True, name="conv6_cls")
    sub4_out = conv(
        sub4_out, 1, 1, num_classes, 1, 1, biased=True, name="sub4_out")
    sub24_out = conv(
        sub24_out, 1, 1, num_classes, 1, 1, biased=True, name="sub24_out")
```

```
    return sub4_out, sub24_out, conv6_cls
```

3.4 金字塔池化

子网 sub4 中需要用到金字塔池化，对输入图像进四次不同尺度的平均池化再利用双线性插值将其上采样到输入尺寸，最后合成结果，以提取图像的多尺度特征。

3.5 学习率自适应衰减

在模型的训练过程中，学习率是一个很重要的参数，若学习率过低会增加训练时长，过高则会导致难以收敛至最优值，为了使网络能够更好地收敛，在 ICNet 的训练过程中使用了多项式衰减用于使学习率逐渐下降，使得能够保障训练速度的同时，避免出现接近最优值附近时反复波动的状况。

4 结论分析

4.1 结论分析的算法

均值 IOU（Mean Intersection-Over-Union）是语义图像分割中的常用的评价指标之一，它首先计算每个语义类的 IOU，然后计算类之间的平均值。定义如下:

$$IOU=true\_positive/(true\_positive+false\_positive+false\_negative) \quad (4.1)$$

在一个 confusion 矩阵中累积得到预测值，然后从中计算均值 mIOU。

4.2 结论分析的过程

在 cityscapes 数据集的预处理中，createTrainIdLabelImgs.py 调用了 label.py，其中定义了不同类与 ID、色彩之间的对应关系，我们使用了其中的 19 类，有 19 个标签和色彩与之对应，让他们的对应关系如下所示：

```
#           name                    id    trainId        color
    Label(  'road'                ,  7 ,       0 ,  (128, 64,128) ),
    Label(  'sidewalk'            ,  8 ,       1 ,  (244, 35,232) ),
    Label(  'building'            , 11 ,       2 ,  ( 70, 70, 70) ),
    Label(  'wall'                , 12 ,       3 ,  (102,102,156) ),
    Label(  'fence'               , 13 ,       4 ,  (190,153,153) ),
    Label(  'pole'                , 17 ,       5 ,  (153,153,153) ),
    Label(  'traffic light'       , 19 ,       6 ,  (250,170, 30) ),
    Label(  'traffic sign'        , 20 ,       7 ,  (220,220,  0) ),
    Label(  'vegetation'          , 21 ,       8 ,  (107,142, 35) ),
    Label(  'terrain'             , 22 ,       9 ,  (152,251,152) ),
    Label(  'sky'                 , 23 ,      10 ,  ( 70,130,180) ),
    Label(  'person'              , 24 ,      11 ,  (220, 20, 60) ),
    Label(  'rider'               , 25 ,      12 ,  (255,  0,  0) ),
    Label(  'car'                 , 26 ,      13 ,  (  0,  0,142) ),
    Label(  'truck'               , 27 ,      14 ,  (  0,  0, 70) ),
```

Label('bus' , 28 , 15 , (0, 60,100)),

Label('train' , 31 , 16 , (0, 80,100)),

Label('motorcycle' , 32 , 17 , (0, 0,230)),

Label('bicycle' , 33 , 18 , (119, 11, 32)),

在评估程序中，我们计算模型在验证集上的平均交并比(mIOU)来判断模型的准确度

最终在 cityscapes 验证集上得到的 mean IOU 为 67.0%，证明了 ICNet 网络的高准确性。

模型输入、输出的图片以及 ground truth 文件的对比样例如图 4.1、4.2、4.3 所示。

图 4.1 输入图片

Figure4.1 Input image

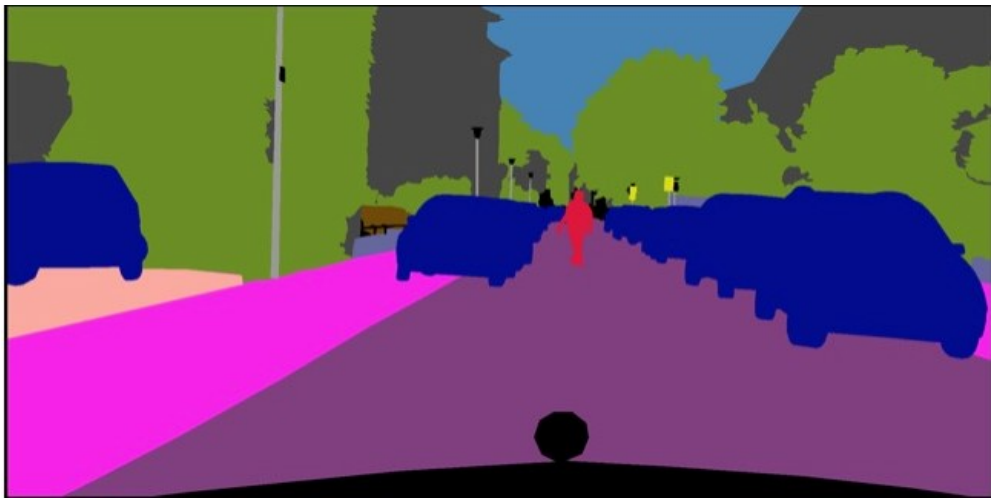

图 4.2 对应的 Ground Truth 图片

Figure4.2 The ground truth image that corresponding to the input image

图 4.3 模型输出的图片

Figure4.3 Output image

5 总结与展望

5.1 工作总结

本文使用实时语义分割模型 ICNet 在 cityscapes 数据集上进行了训练及预测。现有的大多数语义分割模型虽然在准确率上达到了很好的效果，但是算力消耗非常大，对每张图片的预测需要花费较长时间，无法达到实时预测。而 ICNet 模型采用了级联特征融合单元将不同尺度的特征图融合，还使用了金字塔池化和空洞卷积等操作进一步学习图像的多种尺度上的特征，使得模型拥有较高准确度；并且 ICNet 对不同分辨率的输入图像进行不同复杂度的处理，降低了计算复杂度，使该模型能够进行实时运算。ICNet 拥有较理想的判断准确度，且其实时性使得该模型可应用于车载自动驾驶系统，使车载计算机能够辨别载具，行人，道路，建筑，标志牌等，辅助计算机做出避让、转向等决策。

5.2 工作展望

语义分割问题是计算机视觉领域的一个大问题，ICNet 在实时性方面做出了较大突破，但由于工作时间和能力问题，在一些方面仍然拥有不足和提升空间。

1） 观察 ICNet 模型的预测结果可发现，模型在对于远处细小标志牌和行人的判断上仍有提升空间，可以尝试适当调整各子网层数及参数量以优化预测结果。
2） 对于自动驾驶系统来说，如果能根据连续图像判断出行人及载具的运动方向可以更好地辅助计算机做出更精准的决策，ICNet 模型可以对图像进行逐像素分割，但无法判断物体运动方向，可以在这方面做出一些进步使得 ICNet 更适合于自动驾驶系统。

致谢

历时大半年，终于完成了这篇论文，与此同时，大学四年的本科生涯也即将告一段落，为了完成这一之前从未接触过的领域的课题，我查阅了许多的深度学习相关的资料文献和论文，学习了python这门编程语言和百度的开源深度学习平台PaddlePaddle的使用方法，也花了很多时间去调试、修改代码，在这期间遇到了许多意料之中和意料之外的困难，但是老师和同学们都给了我极大的帮助，帮助我克服了这些困难。首先要感谢我的毕设指导老师孙钰老师，孙钰老师的悉心指导使我们收获良多，并且提供了许多相关资源帮助我们学习以及克服过程中的难点，锻炼了我们的专业实力，使我们在面对未来的工作或学习生活中，有了更多信心和能力。其次，我也要感谢过去四年为我们传授知识的所有信息学院的老师，没有你们为我们打下的知识基础我们将难以完成这项任务。同时还要感谢我的父母一直以来对我的信任和支持，正是因为他们对我的关照，我才能更专注于毕业设计相关的工作。更应该感谢的，还有在此领域做出贡献的无数学者们，这些先驱者的璀璨成果，无时无刻不在指引着后来者前进的道路。

这次毕业论文的完成过程使我收获了许多，除了知识，更让我明白了自己未来的道路，以及人与人之间互相帮助的精神的可贵。最后感谢我的母校北京林业大学给我们提供的良好的学术氛围、和睦的人际关系，以及一切美好的回忆。

参考文献

[1] Hinton G E, Osindero S, The Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7):1527-1554

[2] 王若辰. 基于深度学习的目标检测与分割算法研究[D]. [出版地不详]: 北京工业大学, 2016

[3] Jonathan Long, Evan Shelhamer, Trevor Darrell: Fully convolutional networks for semantic segmentation[C]. In CVPR 2015

[4] Fisher Yu, Vladlen Koltun: Multi-Scale Context Aggregation by Dilated Convolutions[C]. In ICLR 2016

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille: Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[C]. arXiv:1412.7062, 2014

[6] Xiaodan Liang, Hongfei Zhou, Eric Xing: Dynamic-structured Semantic Propagation Network[C]. In CVPR 2018

[7] Guosheng Lin, Anton Milan, Chunhua Shen, Ian Reid: RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation[C]. arXiv:1611.06612, 2016

[8] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation[C]. arXiv:1511.00561, 2015

[9] Olaf Ronneberger, Philipp Fischer, Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation[C]. In MICCAI 2015

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia: Pyramid Scene Parsing Network[C]. In CVPR 2017

[11] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, Jiaya Jia: ICNet for Real-Time Semantic Segmentation on High-Resolution Images[C]. In ECCV 2018

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi: You Only Look Once: Unified, Real-Time Object Detection[C]. arXiv:1506.02640

[13] Jisoo Jeong, Hyojin Park, Nojun Kwak: Enhancement of SSD by concatenating feature maps for object detection[C]. arXiv:1705.09587

[14] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks[C]. In NIPS'2012

[15] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, Eugenio Culurciello: ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation[C]. arXiv:1606.02147, 2016