图像语义分割综述

魏来

摘要 图像的语义分割是计算机视觉中重要的基本问题之一，其目标是对图像的每个像素点进行分类，将图像分割为若干个视觉上有意义的或感兴趣的区域, 以利于后续的图像分析和视觉理解。近年来，以神经网络为基础的深度学习技术的发展和应用，极大地推动了语义分割的发展。
关键词 图像语义分割；深度学习；神经网络

目录

1 问题介绍

图像分类、物体检测和图像语义分割是计算机视觉的三大核心研究问题。图像语义分割任务极具挑战性。图像语义分割 (Image Semantic Segmentation) 融合了传统的图像分割和目标识别两个任务，将图像分割成一组具有一定语义含义的块，并识别出每个分割块的类别，最终得到一幅具有逐像素语义标注的图像。目前, 图像语义分割是计算机视觉和模式识别领域非常活跃的研究方向，并在很多领域具有广泛的应用价值。

Figure 1: 语义分割示例

然而, 图像语义分割任务是一个非常具有挑战性的问题, 同文献 [1] 中对物体识别任务的总结类似, 其难点主要体现在以下几个方面:

1) 物体层次: 同一物体, 由于光照、视角、距离的不同, 拍摄出的图像也会有很大不同;
2) 类别层次: 类别层面上所面临的难点主要来自两个方面, 即类内物体之间的相异性和类间物体之间的相似性;
3) 背景层次: 通常干净的背景有助于实现图像的语义分割, 但实际场景中的背景往往是错综复杂的, 这种复杂性也大大提升了图像语义分割的难度;

图像语义分割问题本质上是分类问题，是对输入的图像的像素进行分类的问题。下面从这个角度进行介绍:

1.1 分类对象与可分的类

图像语义分割的分类对象一般为像素，算法对每个像素进行分类。

在哪些类的集合上训练算法是设计算法的关键问题，许多算法使用固定的类的集合，有些甚至只工作在两个类别问题上，比如前景和背景 [2]，或者街道和非街道 [5]。

也有些无监督学习的算法完全不区分类别，以及有些分割算法能够识别它们不知道的类。例如文献 [6] 中加入了空白类作为算法无法识别的图像中的类。

1.2 输入数据

有些算法对输入数据的形式有一定要求，不同的数据形式也包含不同的信息:

- **灰度图像与彩色图像：** 灰度图像在许多方面有广泛应用，而彩色图像则包含更多信息，使用范围也更广。

- **是否包含深度信息：** 在机器人、自动驾驶等领域可能会用到深度图像。
- **单一图像与多张图像：** 单一图像分割是最广泛的图像分割方式，在多张图像分割方面也有一些尝试。有些工作可以对一组图像进行联合处理，从而找到图像中相同的某个语义对象。

1.3 操作模式

分类器的操作模式可以分为主动式 [7] 或被动式，主动式指机器人可以移动物体来进行分割图像，被动式指接收不能被影响的图像。在被动式算法中，有些是完全自动的，有些是在交互模式中的。例如 [8] 某系统可以通过用户点击来粗略标记背景，再由算法寻找精确的分割边界。

1.4 评价指标

评价指标对于系统的设计是非常关键的。语义分割系统的用户都希望得到一个正确的结果，**准确率**是最广泛使用的评价方式。但是对不同的分割算法进行比较时，也有一些其他的质量评价方式。

1. **准确率**表示了语义分割完成的正确性。然而这一指标也有许多的表现形式，一种给读者直观感受的方式就是画图例如上图 图1。这种方式只能用于解释说明或特定的问题。对于总体的正确率，有几个单位可以定义。

   一个有效的对比方法是比较各算法的像素准确率，即被正确分类的像素占总像素的比例。但是该方法有两个缺点：

   1）对于不平衡类，即某大片区域为一类，这样的情况此方法无法正确反映准确性。

   2）对于人工粗标记的图像，算法可能将那片区域再进行分类，该方式也不适用。

   文献 [14] 中引入以下三个参数来避免问题 1)：平均正确率（mean accuracy)、平均交叠率（mean intersection over union）和频率加权交叠率（frequency weighted intersection over union)。一些文章中也给出了其它方法来避免以上两个问题如加入空白类、F 参数等。(具体定义略)

2. **速度**由于需求场景不同，我们很难对于每幅图像给出一个限制时间，而且这个处理时间和硬件甚至数据情况有很大关系。

   但是，即使使用不同的硬件，算法的处理时间仍然是非常关键的，这可以使读者估计在固定时间限制内优化算法的可能性。同时，算法的吞吐量也是非常重要的。

3. **稳定性**当输入的图像有轻微变化（如模糊或转换角度）时，算法的结果不应当有太大的偏差。

4. **内存占用**算法的内存占用和可运行的硬件有很大关系。

2 传统的图像分割算法

2.1 阈值法

阈值分割法 [9] 是图像分割领域最基础的方法之一，原理是根据图像中像素的颜色或灰度值的不同，对图像进行分割。这一方法的关键在于阈值的选取。最常用的阈值方法是基于灰度直方图的方法，如最大类间方差法 (OTSU)[5]、最小误差法、最大熵法等。该方法的优点是计算简单、运算效率较高、速度快。但该方法对语义的识别效果一般，缺点也非常明显，当图像中像素的灰度较接近或颜色差别不大时，出错概率较高。这些缺点在后续方法中可以得到解决。

2.2 边缘检测法

基于边缘检测的方法主要是通过检测出区域的边缘来进行分割，利用区域之间特征的不一致性，首先检测图像中的边缘点，然后按一定策略连接成闭合的曲线，从而构成分割区域。图像中的边缘通常是灰度、颜色或纹理等性质不连续的地方。对于边缘的检测，经常需要借助边缘检测算子来进行，其中常用的边缘检测算子包括：Roberts 算子、Laplace 算子、Prewitt 算子、Sobel 算子、Rosonfeld 算子、Kirsch 算子以及 Canny 算子等。

边缘检测算法比较适合边缘灰度值过渡比较显著且噪声较小的简单图像的分割。对于边缘比较复杂以及存在较强噪声的图像，则面临抗噪性和检测精度的矛盾。若提高检测精度，则噪声产生的伪边缘会导致不合理的轮廓：若提高抗噪性，则会产生轮廓漏检和位置偏差。

因此，这一方法常常被用作图像分割的预处理算法。先对图像进行大致检测，然后用后续方法再对图像进行处理。

2.3 主动轮廓模型

主动轮廓模型 [10] 又称为 Snake 模型，是 Kass 等人于 1987 年提出的，活动轮廓即定义在图像域的曲线或者曲面，在与自身几何特性相关的内力以及图像数据相关的外力共同作用下，以最小化能量函数的形式向边界运动。经过二十多年的发展，活动轮廓模型已经在边缘检测、图像分割以及运动跟踪中得到了广泛的应用。按照曲线的表达方式的不同，活动轮廓模型大致可以分为两大类：参数活动轮廓模型和几何活动轮廓模型。

2.4 分水岭算法

分水岭算法是以数学形态学作为基础的一种区域分割方法。其基本思想是将梯度图像看成是假想的地形表面，每个像素的梯度值表示该点的海拔高度。原图中的平坦区域梯度较小，构成盆地，边界处梯度较大构成分割盆地的山脊。分水岭算法模拟水的渗入过程，假设水从最低洼的地方渗入，随着水位上升，较小的山脊被淹没，而在较高的山脊上筑起水坝，防止两区域合并。当水位达到最高山脊时，算法结束，每一个孤立的积水盆地构成一个分割区域。由于受到

图像噪声和目标区域内部的细节信息等因素影响，使用分水岭算法通常会产生过分割现象，分水岭算法一般是作为一种预分割方法，与其它分割方法结合使用，以提高算法的效率或精度。

2.5 区域生长法

区域生长方法 [11] 也是一种常用的区域分割技术，其基本思路是首先定义一个生长准则，然后在每个分割区域内寻找一个种子像素，通过对图像进行扫描，依次在种子点周围邻域内寻找满足生长准则的像素并将其合并到种子所在的区域，然后再检查该区域的全部相邻点，并把满足生长准则的点合并到该区域，不断重复该过程直到找不到满足条件的像素为止。该方法的关键在于种子点的位置、生长准则和生长顺序。

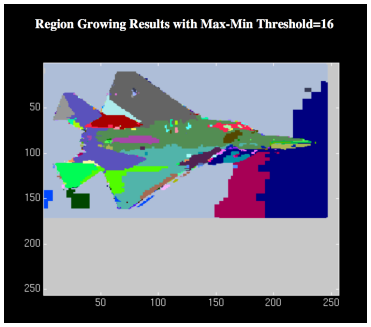


Figure 2: 区域生长法

2.6 随机决策森林

在机器学习中，随机森林 [12] 是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。Leo Breiman 和 Adele Cutler 发展出推论出随机森林的算法。而”Random Forests” 是他们的商标。这个术语是 1995 年由贝尔实验室的 Tin Kam Ho 所提出的随机决策森林（random decision forests）而来的。这个方法则是结合 Breimans 的”Bootstrap aggregating” 想法和 Ho 的”random subspace method” 以建造决策树的集合。

2.7 基于图论的图像分割

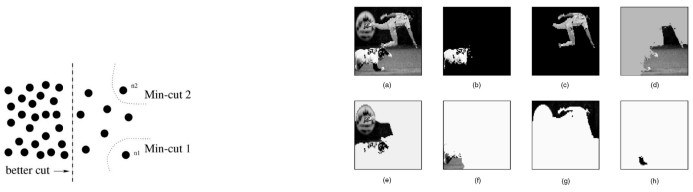


Figure 3: 基于图论的算法

基于图论的图像分割技术是近年来图像分割领域的一个新的研究热点。其基本思想是将图像映射为带权无向图，把像素视作节点，节点之间的边的权重对

应于两个像素间的不相似性度量，割的容量对应能量函数。运用最大流 / 最小流算法对图进行切割，得到的最小割对应于待提取的目标边界。该方法具有快速、鲁棒、全局最优、抗噪性强、可扩展性好等优点。基于图论的方法本质上将图像分割问题转化为最优化问题，是一种点对聚类方法，其最优分割基本原则就是使划分成的两个子图 (区域) 内部相似度最大，而子图之间的相似度最小。

2.8 马尔可夫随机场与条件随机场

马尔可夫随机场（Markov Random Fields）是无向的概率图论模型，是计算机视觉中广泛使用的一种模型。思路是对像素赋予一个随机的值，然后通过概率的方式进行计算和分类。

条件随机场模型（Conditional Random Fields）是 Lafferty 于 2001 年，在最大熵模型和隐马尔科夫模型的基础上，提出的一种判别式概率无向图学习模型，是一种用于标注和切分有序数据的条件概率模型。[13]

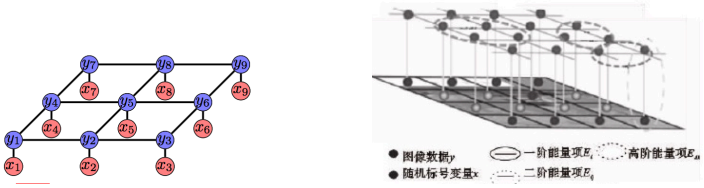


Figure 4: CRF

前 DL 时代的语义分割工作多是根据图像像素自身的低阶视觉信息（Low-level visual cues）来进行图像分割。由于这样的方法没有算法训练阶段，因此往往计算复杂度不高，但是在较困难的分割任务上（如果不提供人为的辅助信息），其分割效果并不能令人满意。一些传统算法常被用作图像预处理或后处理，与神经网络配合使用。

在计算机视觉步入深度学习时代之后，语义分割同样也进入了全新的发展阶段，以全卷积神经网络（Fully convolutional networks，FCN）[14] 为代表的一系列基于卷积神经网络“训练”的语义分割方法相继提出，屡屡刷新图像语义分割精度。

3 基于深度学习的分割算法

3.1 神经网络的历史

模拟神经网络的原创文章发表于 1943 年，两位作者是麦卡洛可（McCulloch）和皮茨（Pitts)。他们合作的成果就是神经网络的第一篇文章：“A Logical Calculus of Ideas Immanent in Nervous Activity”，发表在《数学生物物理期刊》上。这篇文章也成了控制论的思想源泉之一。

1949 年，神经心理学家 Hebb 出版《行为组织学》(Organization of Behavior)，该书提出了被后人称为“Hebb 规则”的学习机制。后来的各种无监督机器学习算法或多或少都是 Hebb 规则的变种。

1957 年，康奈尔大学的实验心理学家弗兰克 · 罗森布拉特在一台 IBM-704 计算机上模拟实现了一种他发明的叫作“感知机”（Perceptron）的神经网络模型。这个模型可以完成一些简单的视觉处理任务。罗森布拉特在理论上证明了单层神经网络在处理线性可分的模式识别问题时，可以收敛，并以此为基础，做了若干“感知机”有学习能力的实验。罗森布拉特在 1962 年出版了《神经动力学

原理：感知机和大脑机制的理论》（Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms）。

明斯基是人工智能的奠基人之一，是达特茅斯会议的组织者。明斯基和麻省理工学院的佩普特合作出版了《感知机：计算几何学》（Perceptrons: An Introduction to Computational Geometry)。在书中，明斯基和佩普特证明单层神经网络不能解决 XOR（异或）问题。

1974 年，哈佛的一篇博士论文证明了在神经网络多加一层，并且利用“后向传播”（Back-propagation）学习方法，可以解决 XOR 问题。这篇论文的作者是沃波斯（Werbos)。沃波斯这篇文章刚发表时并没引起多少重视，那时正是神经网络研究的低谷。

1982 年，在加州理工担任生物物理教授的霍普菲尔德提出了一种新的神经网络，可以解决一大类模式识别问题，还可以给出一类组合优化问题的近似解。这种神经网络模型后被称为霍普菲尔德网络。1984 年，霍普菲尔德用模拟集成电路实现了自己提出的模型。霍普菲尔德模型的提出振奋了神经网络领域。一些科学家受到鼓励开始了“连接主义”（Connectionism）运动。

连接主义运动的成果之一就是那本著名的被称为 PDP（Parallel and Distributed Processing) 的文集。1993 年，美国电气电子工程师学会 IEEE 开始出版《神经网络会刊》。

神经网络由一层一层的神经元构成。层数越多，就越深，所谓深度学习就是用很多层神经元构成的神经网络达到机器学习的功能。辛顿（Hinton）就是“深度学习”的作者，他 2006 年的一篇文章开辟了这个新领域。

2012 年，斯坦福大学人工智能实验室主任 Andrew Ng（吴恩达）和谷歌合作建造了一个当时最大的神经网络，参数多达十七亿的神经网络。后来 Ng 在斯坦福大学建了个更大的神经网络，参数高达一百一十二亿。人脑的神经连接有一百万万亿个。从计算能力上说，如果这个人工神经网络要是能接近大脑，每个人工神经元必须能达到一万个大脑神经元的功能。这个神经网络会用到大量的图形处理芯片 GPU，GPU 是模拟神经网络的完美硬件，因为每个 GPU 芯片内都有大量的小核心。这和神经网络的大规模并行性天然相似。硬件的进步让以往不可能的成为可能。

3.2 逻辑回归与各类神经网络

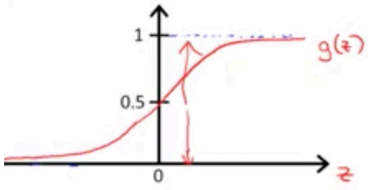
3.2.1 逻辑回归

逻辑回归的基本思路是将 $h_\theta(x) = \theta^T x$ 替换为 $Sigmoid\ Function$，即逻辑回归函数，将 $\theta^T x$ 的值转化为 0-1 之间的数，从而进行分类。$SigmoidFunction$ 为：

$$g(x) = 1/1 + e^{-z}$$

图像为：

Figure 5: Sigmoid Function

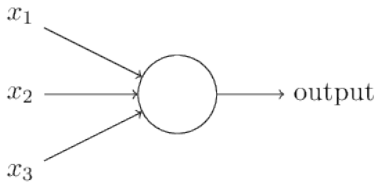


再依据此值，$h_\Theta(x)$ 的输出确认为 0 或 1，即输出为正类或负类。相应地，$Cost\ Function$ 变为：

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log h_\theta(x^{(i)}) + (1-y^{(i)}) log(1-h_\theta(x^{(i)})) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

训练的过程即为对此函数进行优化达到最小值。
此结构的示意图为：

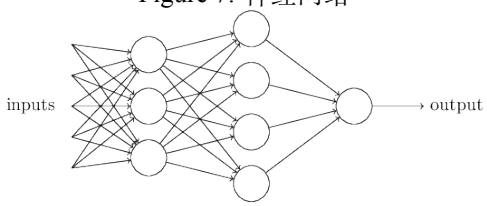
Figure 6: 单个神经元



3.2.2 神经网络（多层感知器）与深度神经网络

把这些单个神经元结构组织在一起，便形成了神经网络。下图是一个三层神经网络结构：

Figure 7: 神经网络



为什么要用神经网络?——因为很多实际问题中，feature 数目过大，参数数量巨大，比如计算机视觉的问题。

上图中最左边的原始输入信息称之为输入层，最右边的神经元称之为输出层(上图中输出层只有一个神经元)，中间的叫隐藏层。输入数据得到输出数据进行预测的过程被称为前向传播（Frontpropagation)。层数越多，参数及参数之间的关系函数就越多，可以实现的运算就越复杂。深度神经网络系统中的层数比较多，达到 8-10 层（普通神经网络的层数通常 3-4 层)。

神经网络的 $Cost\ Function$ 为：

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} log h_\Theta(x^{(i)})_k + (1-y_k^{(i)}) log(1-h_\theta(x^{(i)}))_k + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_1}\sum_{j=1}^{s_1+1}(\Theta_{ji}^{(l)})^2$$

可以看出，神经网络的 $Cost\ Function$ 非常复杂，要优化起来也不容易。神经网络的参数优化算法为后向传播（Backpropagation)。训练优化过程为：首先为各参数设置不对称的值，然后通过前向传播对训练集计算偏差，后向传播计算偏导优化参数，然后不断利用 backprop 对参数进行优化，直到收敛到最小值或达到较好的正确率。

随着神经网络层数的加深，优化函数越来越容易陷入局部最优解，并且这个“陷阱”越来越偏离真正的全局最优。利用有限数据训练的深层网络，性能还不如较浅层网络。同时，随着网络层数增加，“梯度消失”现象更加严重。

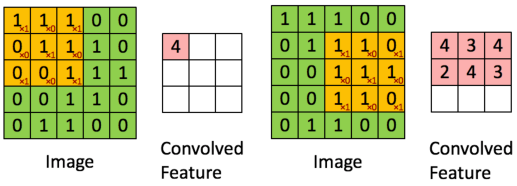
2006 年，Hinton 利用预训练方法缓解了局部最优解问题，将隐含层推动到了 7 层 [16]，神经网络真正意义上有了“深度”，由此揭开了深度学习的热潮。这里的“深度”并没有固定的定义——在语音识别中 4 层网络就能够被认为是“较深的”，而在图像识别中 20 层以上的网络屡见不鲜。为了克服梯度消失，ReLU、maxout 等传输函数代替了 sigmoid，形成了如今 DNN 的基本形式。

3.2.3 卷积神经网络（CNN）

全连接 DNN 的结构里下层神经元和所有上层神经元都能够形成连接，带来的潜在问题是参数数量的膨胀。假设输入的是一幅像素为 1K*1K 的图像，隐含层有 1M 个节点，光这一层就有 10^{12} 个权重需要训练，这不仅容易过拟合，而且极容易陷入局部最优。另外，图像中有固有的局部模式（比如轮廓、边界，人的眼睛、鼻子、嘴等）可以利用，显然应该将图像处理中的概念和神经网络技术相结合。对于 CNN（Convolutional Neural Networks）来说，并不是所有上下层神经元都能直接相连，而是通过“卷积核”作为中介。同一个卷积核在所有图像内是共享的，图像通过卷积操作后仍然保留原先的位置关系。

“卷积”的本质意义就是“加权叠加”，在对图像处理进行卷积时，根据卷积核的大小，输入和输出之间也会有规模上的差异。如图：

Figure 8: CNN



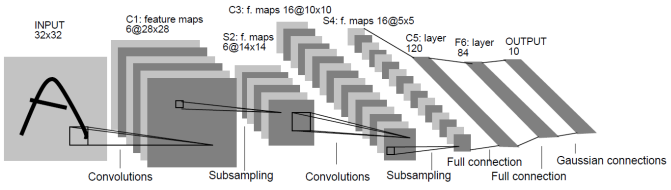
上图左边 5*5 的方块视为图像输入，黄色移动的 3*3 以及里面的数字（*1/*0）是卷积核，该卷积核按照步长为 1 的顺序依次从原始输入的左上角一直移动计算叠加到右下角，卷积核一共移动 9 次。

九次的位置对应到右侧的 3*3 的相应格内，格中的数字便是卷积值，（此处是卷积核所覆盖的面积内元素相乘再累加的结果）。

9 次移动计算完毕后，右侧 3*3 的新矩阵为此次卷积层的计算结果。

卷积层之间的卷积传输的示意图如下：

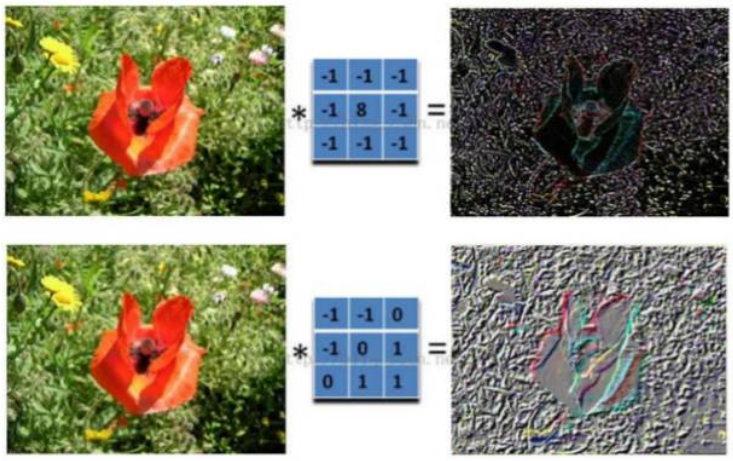
Figure 9: 卷积神经网络



在实际计算过程中，输入是一张原始图片和滤波器 filter（一组固定的权重，也就是上面我们说的卷积核对应的实际意义）做内积后得到新的二维数据。

不同的滤波器 filter 会得到不同的输出数据，比如轮廓、颜色深浅，如果想提取图像的不同特征，需要用不同的滤波器 filter 提取想要的关于图像的特定信息。

Figure 10: 不同卷积核的处理结果不同



上图为一个卷积层中的卷积处理过程，注意上下两次卷积核内容是不同的，所以得到两种处理结果。

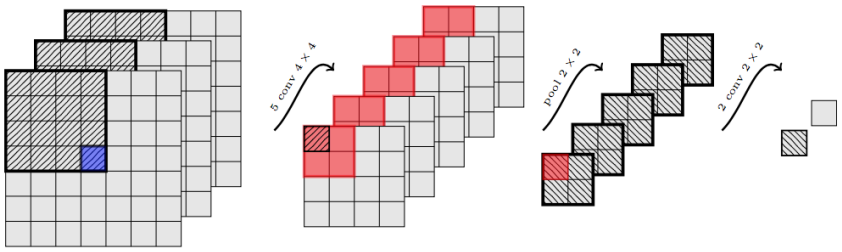
等号右边的新的二维信息在 CNN 网络中，会作为下一个卷积层的输入，即在下一个卷积层计算时，右边的图像会作为输入的原始图像。

CNN 中还有 max-pooling 等操作进一步提高鲁棒性。

我们注意到，对于图像，如果没有卷积操作，学习的参数量是灾难级的。CNN 之所以用于图像识别，正是由于 CNN 模型限制了参数的个数并挖掘了局部结构的这个特点。顺着同样的思路，利用语音语谱结构中的局部信息，CNN 照样能应用在语音识别中。在普通的全连接网络或 CNN 中，每层神经元的信号只能向上一层传播，样本的处理在各个时刻独立，因此又被成为前向神经网络 (Feed-forward Neural Networks)。

3.2.4 循环神经网络（RNN）

Figure 11: RNN [17]



全连接的 DNN 还存在着另一个问题——无法对时间序列上的变化进行建模。然而，样本出现的时间顺序对于自然语言处理、语音识别、手写体识别等应用

非常重要。对了适应这种需求，就出现了另一种神经网络结构——循环神经网络 RNN(Recurrent convolutional Neural Networks)。而在 RNN 中，神经元的输出可以在下一个时间戳直接作用到自身，即第 i 层神经元在 m 时刻的输入，除了（i−1）层神经元在该时刻的输出外，还包括其自身在（m−1）时刻的输出。

RNN 可以看成一个在时间上传递的神经网络，它的深度是时间的长度！正如我们上面所说，“梯度消失”现象又要出现了，只不过这次发生在时间轴上。对于 t 时刻来说，它产生的梯度在时间轴上向历史传播几层之后就消失了，根本就无法影响太遥远的过去。因此，之前说“所有历史”共同作用只是理想的情况，在实际中，这种影响也就只能维持若干个时间戳。为了解决时间上的梯度消失，机器学习领域发展出了长短时记忆单元 LSTM，通过门的开关实现时间上记忆功能，并防止梯度消失。

事实上，不论是哪种网络，他们在实际应用中常常都混合着使用，比如 CNN 和 RNN 在上层输出之前往往会接上全连接层，很难说某个网络到底属于哪个类别。不难想象随着深度学习热度的延续，更灵活的组合方式、更多的网络结构将被发展出来。尽管看起来千变万化，但研究者们的出发点肯定都是为了解决特定的问题。

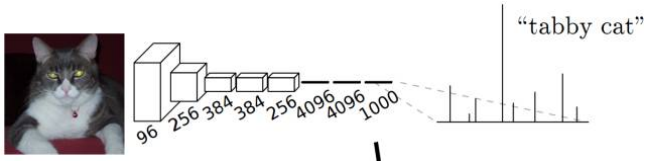
3.3 神经网络在语义分割中的应用

在此前使用的图像识别算法中，主流的技术是卷积神经网络算法（Convolutional Neural Networks），即 CNN。

在 2015 年的 CVPR（国际计算机视觉与模式识别会议）上发表了一篇论文 [14]，提出了 FCN 即全卷积神经网络（Fully Convolutional Networks）。

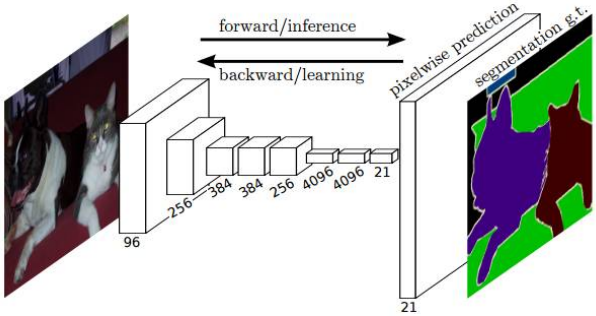
CNN 的输入是图像，输出是一个结果，或者说是一个值，一个概率值。

Figure 12: 图像级理解



FCN 提出所追求的是，输入是一张图片是，输出也是一张图片，学习像素到像素的映射。

Figure 13: 像素级理解



“全卷积”体现在网络中的后三层。CNN 的后三层都是一维的向量，计算方式不再采用卷积，所以丢失了二维信息，而 FCN 网络中，将这三层全部转化为 1*1 的卷积核所对应等同向量长度的多通道卷积层，使后三层也全部采用卷积计算，整个模型中，全部都是卷积层，没有向量，所以称为“全卷积”。

FCN 将第 6 层和 7 层分别从 4096 长度的向量转化为 4096 通道的卷积层，第 8 层则是 21 通道的卷积层。之所以第 8 层从 1000 缩减到 21，是因为 FCN 使用的识别库是 PASCAL VOC，在 PASCAL VOC 中有 20 种物体分类，另外一个 background 分类。

CNN 的识别是图像级的识别，也就是从图像到结果，而 FCN 的识别是像素级的识别，对输入图像的每一个像素在输出上都有对应的判断标注，标明这个像素最可能是属于一个什么物体/类别。

在实际的图像语义分割测试时，输入是一个 H*W*3 的三通道彩色图像，而输出是一个 H*W 的矩阵。

这就可以简单看做每一个像素所携带的信息是多维的，比如颜色，就分为 3 层，分别对应 R、G、B 三个值。

所以在进行卷积的时候，每一个通道都是要独立计算的，计算完之后再叠加，得到最终卷积层结果。

如果卷积核移动的步长为 1，那么卷积是按照像素排列去挨个计算的，计算量可想而知会有多么庞大。但是在实际中，相邻的像素往往都是一类，按照像素依次计算就显得冗余，所以在卷积之后会对输出进行一次**池化**（pooling）处理。

池化简单来说就是将输入图像切块，大部分时候我们选择不重叠的区域，假如池化的分割块大小为 h*h，分割的步长为 j，那么一般 h=j，如果需要重叠，只需要 h>j 即可。

对完整图像切分，再取切分区域中所有值的均值或最大值作为代表该区域的新值，放入池化后的二维信息图中。得到的新图就是池化结果。

在 CNN 和 FCN 的网络模型中，每一个卷积层，都包含了 [卷积 + 池化] 处理，这就是传说中的“下采样”，但这样处理之后的结果是：图像的像素信息变小了，每一层的像素信息都是前一层的 1/2 大小，到第五层的时候，图像大小为原始图像的 1/32。

在 CNN 算法里，这并没有什么要紧的，因为 CNN 最终只输出一个结果：“这个图是什么”，但是 FCN 不同，FCN 是像素级别的识别，也就是输入有多少像素，输出就要多少像素，像素之间完全映射，并且在输出图像上有信息标注，指明每一个像素可能是什么物体/类别。

所以就必须对这 1/32 的图像进行还原。

这里用**“反卷积”**，对第 5 层进行反卷积，可以将图像扩充至原来的大小（近似原始大小），一般会大一点，但是会裁剪掉，会变大的原理略复杂，这里先不提。——这个“反卷积”称为**“上采样”**（和下采样对应）。

较浅的卷积层（靠前的）的感受域比较小，学习感知细节部分的能力强，较深的隐藏层 (靠后的)，感受域相对较大，适合学习较为整体的、相对更宏观一些的特征。

所以在较深的卷积层上进行反卷积还原，会丢失很多细节特征。

于是在反卷积步骤时，考虑采用一部分较浅层的反卷积信息辅助叠加，更好地优化分割结果的精度：

该论文方法（FCN）的结果如图15

可以看出，FCN 能够一定程度上预测出对象的大致轮廓，但是在边缘的准确性比较差。这主要是因为网络还是存在一定的下采样，并且网络是全卷积，所以对位置不是很敏感。Deeplab 的一个想法 [18] 是使用传统方法中的图模型 (CRF) 来做后处理（post-processing），优化当前神经网络得出的结果。使用 CRF 将算法的能量函数加入了一个二元项，将像素之间的语义联系/关系考虑了进去。

Figure 14: 浅层叠加反卷积示意

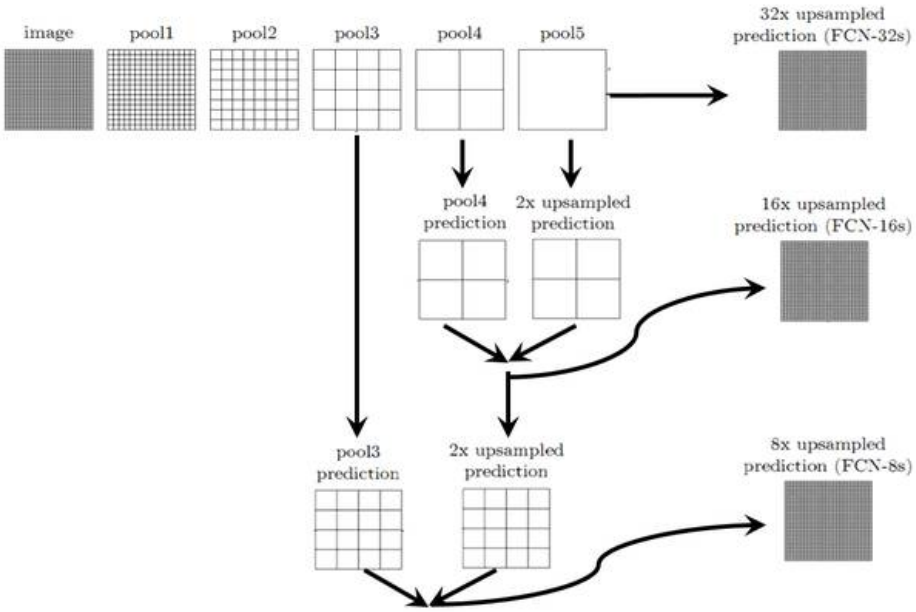
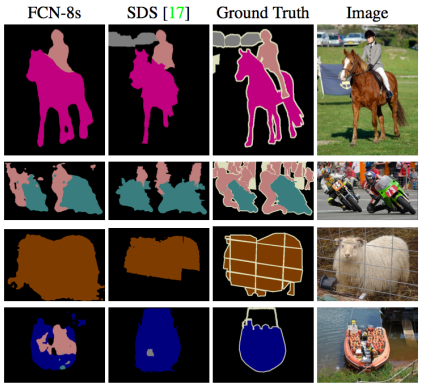


Figure 15: FCN 结果



	pixel acc.	mean acc.	mean IU	f.w. IU	geom. acc.
Liu *et al.* [25]	76.7	-	-	-	-
Tighe *et al.* [36]	-	-	-	-	90.8
Tighe *et al.* [37] 1	75.6	41.1	-	-	-
Tighe *et al.* [37] 2	78.6	39.2	-	-	-
Farabet *et al.* [9] 1	72.3	50.8	-	-	-
Farabet *et al.* [9] 2	78.5	29.6	-	-	-
Pinheiro *et al.* [31]	77.7	29.8	-	-	-
FCN-16s	**85.2**	**51.7**	39.5	76.1	**94.3**

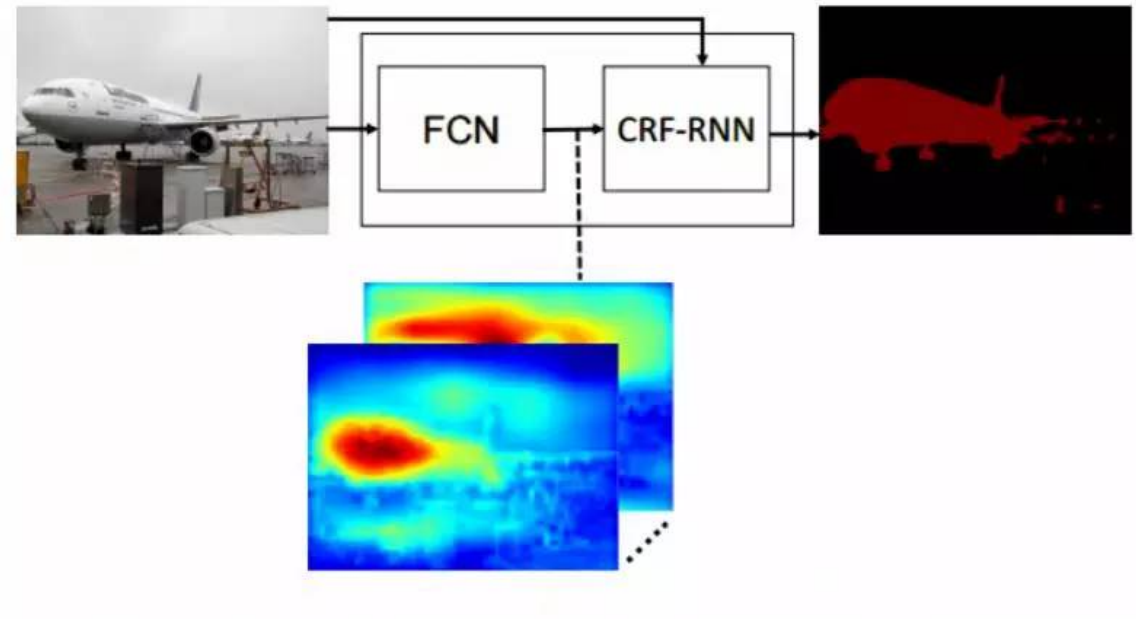
除了这个突破点之外，还修改了 FCN 中的网络结构，提出了一种带”hole” 的卷积。

使用“疏松的卷积核”来处理卷积，可以达到在不增加计算量的情况下增加感受域，弥补不进行池化处理后的精度问题。人为加大了卷积核内部元素之间的距离。

在 ICLR2016 的一篇文章 [19] 给这种滤波器提出了一个更文雅的名字”Dilated Filter”。

从图像语义分割引入深度学习方法之后，以上述的几篇文章为主，大致确立了一个通用框架：

Figure 16: 图像语义分割深度学习方法框架



即前端使用 FCN 进行特征粗提取，后端使用 CRF/MRF 优化前端的输出，有些论文还把 CRF 优化过程拆解成迭代过程接到一个神经网络后面，当成是一个可用学习的模组，最后得到分割图。

此外，还有一些方法可用于解决语义分割问题，如最近兴起的**生成对抗网络（Generative Adversarial Networks）**及其他监督、无监督学习方法。

4 常用数据集

当某种图像语义分割算法被提出时，需要采用一个或多个数据集来验证算法的有效性，深度学习兴起之后, 数据库变得更加重要。基于同样的深度模型，数据量的增加通常可以有效提升图像语义分割的性能。本节对计算机视觉中语义分割方面的主流数据库进行概述。

4.1 PASCAL VOC

在 2005 到 2012 年之间，PASCAL VOC（Pattern Analysis, Statistical modelling and ComputAtional Learning, Visual Object Classes）每年都会发布用于图像分类、物体检测或语义分割的数据库，并在相应的数据库上开展算法竞赛。2007 年以来，数据库的类别数据被固定为 20，主要涉及了日常生活中常见的物体，包括：aeroplane、bicycle、bird、boat、bottle、bus、car、cat、chair、dining table、dog、horse、motobike、person、pot ted plant、sheep、sofa、train 和 tv/monitor。

目前用于评测语义分割算法的图片通常来自 PASCAL VOC 2012，原始的竞赛数据库提供了 1464 张具有标注信息的训练图片。2014 年, 文献 [15] 重新标注了近一万幅图像，将训练样本数据提升到 10582。PASCAL VOC 属于多标签数据库，每张图片中包含了一个或多个物体。物体的尺度变化很大、图片的背景复杂，且同一图片内的不同物体之间往往存在遮挡现象，因此图像语义分割的难度较大。PASCAL VOC 是目前最知名的评估图像语义分割算法的数据库，对语义分割算法提出了很大的挑战，也极大地推动了语义分割研究的发展。

4.2 MSRCv2

微软研究推出的包含 591 张带有像素级别标注的照片。有 21 个类：aeroplane, bike, bird, boat, body, book, building, car, cat, chair, cow, dog, face, flower, grass, road, sheep, sign, sky, tree, water。并且，该数据集还带有空类别标记，为了给那些不属于以上 21 类的像素或是靠近语义对象边界的像素进行分类。

4.3 CITYSCAPES

该数据库主要应用于都市街道场景的语义分割，通常采用的语义类别包含 19 种，即 road、side-walk、building、wal、fence、pole、traficlight、trafic sign、vegetation、terain、sky、person、rider、car、truck、bus、train、motorcycle 和 bicycle。该数据库中用于训练和校验的精细标注的图片数量为 3475，同时也包含了 2 万张粗糙的标记图片。该数据库面临的难点主要包括两点。一方面，街道场景的背景通常比较复杂且大量远端的物体尺度较小。另一方面，图像的尺度很大 (2048×1024)，而该数据库的应用场景通常又要求语义分割算法能够快速预测出区域的语义内容. 该数据的提出将会极大地推动自动驾驶或辅助驾驶等相应产业的发展。

4.4 MSCOCO

该数据库较 PASCALVOC2012 更为复杂，被标注的物体类别数目达到 80 个。同时，用于训练的图像样本数量也远超 PASCALVOC2012(约 82783vs。10582)。目前，该数据库的图片通常被用来对 DCNN 进行预训练，从而进一步提升 DCNN 在某个特定语义分割任务上的性能。

Table 1: 常用数据集对比

数据库	类别数目	训练集数目	校验集数目	测试集数目
PASCAL VOC 2012	20	1464	1449	1452
PASCAL VOC 2012+	20	10582	1449	1452
PASCAL-CONTEXT	59	4998	5105	—
PASCAL-PERSON-PART	6	1716	—	1817
MS-COCO2014	80	82783	40504	40775

5 总结

深度学习极大地提升了许多传统问题的解决效果，改变了许多问题的研究思路。也把许多问题的模式归结于网络的设计和训练上。越来越多的问题可以使用深度学习的方法解决。

图像语义分割引入深度学习后，问题的解决效果迎来了一次突破。而且在科研人员的不断研究和尝试中还在继续发展。通过最近的查阅文献和思考，我认为还有以下几点问题值得关注研究：

1）**基本模型结构的创新** FCN 的提出是对于网络结构的创新，带给了我们一个新的思路，也激发了人们创意和热情。因此基本结构的改进和创新能够另辟蹊径，为解决问题提供新的思路和方法。

2）**模型能力的训练提升** 当前，各项竞赛中有许多排名靠前的算法的基本思路基本相似，而差异在于模型训练程度的不同。避免过拟合的同时进行更加完全的网络训练能够更好地提升算法效果。

3）**训练时间的优化** 当前的深度学习模型复杂程度高，隐层多，参数与参数之间的函数关系多，导致训练时间过长。因此更快地进行训练，如通过并行计算进行训练或硬件性能的提升，让以前难以实现的算法成为可能。

4）**深度学习与其它方法的结合** 神经网络算法是黑盒算法，可解释性差，容易出现过拟合。将深度学习与其它方法进行结合，从传统方法或其它方法的角度看问题，或许可以带来一些启示。

参考文献

[1] 黄凯奇, 任伟强, 谭铁牛. 图像物体分类与检测算法综述 [J]. 计算机学报, 2014,37(6):1225-1240.

[2] J. Reynolds and K. Murphy, "Figure-ground segmentation using a hierarchical conditional random field," in Computer and Robot Vision, 2007. CRV ' 07. Fourth Canadian Conference on, May 2007, pp. 175–182.

[3] 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，1673-0291(2016)04-0082-10

[4] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmenta- tion," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3241–3248.

[5] S. Bittel, V. Kaiser, M. Teichmann, and M. Thoma, "Pixel-wise segmentation of street with neural networks," arXiv preprint arXiv:1511.00513, 2015.

[6] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller, "Multi-class segmentation with relative location prior," International Journal of Computer Vision, vol. 80, no. 3, pp. 300–316, Apr. 2008.

[7] D. Schiebener, J. Schill, and T. Asfour, "Discovery, segmentation and reactive grasping of unknown objects." in Humanoids, 2012, pp. 71–77.

[8] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," ACM Transactions on Graphics (TOG), vol. 23, no. 3, pp. 309–314, 2004.

[9] 刘文萍, 吴立德 1 图像分割中阈值选取方法比较研究 [J]1 模式识别与人工智能,1997 ,10 (3) :27422771

[10] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," International journal of computer vision, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[11] Levine, M & Shaheen, M., (1981) "A Modular Computer Vision System for Image segmentations" , IEEE PAMI, Vol. 3, No. 5, pp. 540-554.

[12] T. K. Ho, "Random decision forests," in Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on, vol. 1. IEEE, 1995, pp. 278–282.

[13] 毛凌，解梅基于高阶 CRF 模型的图像语义分割计算机应用研究，2013，11

[14] Jonathan Long, Evan Shelhamer and Trevor Darrell. "Fully Convolutional Networks for Semantic Segmentation." IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[15] HARIHARANB,ARBELAEZP,BOURDEVL, etal. "Semantic contours from inverse detectors" [C].IEEEIn- ternationalConferenceon Computer Vision,2011:991 -998.

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

[17] P. H. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene parsing," arXiv preprint arXiv:1306.2795, 2013.

[18] Chen, Liang-Chieh, et al. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs." arXiv preprint arXiv:1412.7062(2014). Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, Liang-Chieh Chen and George Papandreou and Iasonas Kokkinos and Kevin Murphy and Alan L Yuille, ICLR 2015.

[19] Yu, Fisher, and Vladlen Koltun. "Multi-Scale Context Aggregation by Dilated Convolutions." arXiv preprint arXiv:1511.07122 (2015).