论文阅读笔记

论文

Image Segmentation Using Deep Learning: A Survey

摘要

图像分割是图像处理和计算机视觉中的关键主题，其应用包括场景理解，医学图像分析，机器人感知，视频监视，增强现实和图像压缩等。近来，由于深度学习模型在各种视觉应用中的成功，已经有大量旨在利用深度学习模型开发图像分割方法的工作。这篇论文提供了对文献的全面回顾，涵盖了语义和实例分割的各种开创性作品，包括全卷积像素标记网络，编码器-解码器体系结构，多尺度以及基于金字塔的方法，递归网络，视觉注意力模型和对抗环境中的生成模型，调查了这些深度学习模型的相似性，优势和挑战，研究了使用最广泛的数据集，并讨论了该领域有希望的未来研究方向。

这份调查论文的一些关键贡献可以总结如下：

- 这份调查涵盖了有关分割问题的当代文献，并概述了到2019年提出的100多种分割算法，分为10类。
- 对细分算法的各个方面进行全面的回顾和深入的分析，包括训练数据，网络架构的选择，损失函数，训练策略及其主要贡献。
- 概述了约20种流行的图像分割数据集，分为2D，2.5D（RGBD）和3D图像。
- 提供了按标准数据集划分的，用于细分领域的属性和性能的比较摘要。
- 基于深度学习的图像分割的一些潜在挑战和未来方向。

分类

根据深度学习的主要技术贡献将其分为以下几类：

- 1）完全卷积网络
- 2）图模型的卷积模型：
- 3）基于编码器-解码器的模型
- 4）基于多尺度和金字塔网络的模型
- 5）基于R-CNN的模型（例如分段）
- 6）扩展的卷积模型和DeepLab系列
- 7）基于递归神经网络的模型
- 8）基于注意力的模型
- 9）生成模型和对抗训练
- 10）具有主动轮廓模型的卷积模型
- 11）其他模型

流行的算法架构

CNN(卷积神经网络)

CNN是深度学习社区中最成功且使用最广泛的架构之一，尤其是对于计算机视觉任务而言。

CNN主要由三种类型的层组成：

1. 卷积层，其中对权重的核（或滤波器）进行卷积以提取特征；
2. 非线性层，它们在特征图上（通常是逐元素地）应用激活函数，以便能够通过网络对非线性函数进行建模；
3. 合并层，这些合并层用一些有关邻域的统计信息（平均值，最大值等）替换了特征图的一小部分邻域，并降低了空间分辨率。

每个单元都从上一层中较小的邻域（称为接收场）接收加权输入。通过堆叠图层以形成多分辨率金字塔，高层可以从越来越宽的接收场中学习特征。

CNN的主要计算优势在于，一层中的所有接收场均具有权重，因此与完全连接的神经网络相比，**参数数量明显减少（降维）**。

一些最著名的CNN架构包括：AlexNet ，VGGNet，ResNet，GoogLeNet，MobileNet和DenseNet。

Recurrent Neural Networks (RNNs) （循环神经网络） and the LSTM（长短期记忆人工神经网络）

RNN被广泛用于**处理顺序数据**，例如语音，文本，视频和时间序列，其中任何给定时间/位置的数据都取决于先前遇到的数据。

在每个时间戳上，模型都会收集当前时间的输入和上一步的隐藏状态，并输出目标值和新的隐藏状态。RNN通常在长序列方面存在问题，因为它们无法捕获许多实际应用中的长期依赖关系（尽管它们在这方面没有任何理论上的限制），并且经常会遇到梯度消失或爆炸的问题。

然而，一种称为长短期记忆的RNN旨在避免这些问题。 LSTM体系结构包括三个门（输入门，输出门，忘记门），三个门调节信息进出存储单元的信息流，该存储单元在任意时间间隔内存储值。

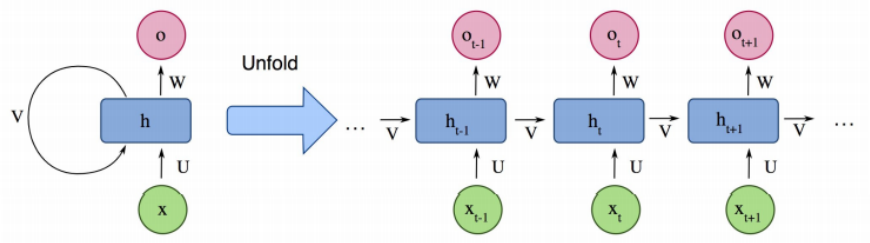


Fig. 3. Architecture of a simple recurrent neural network.

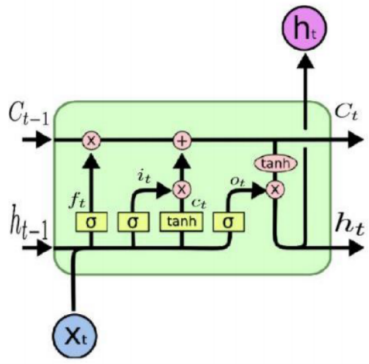


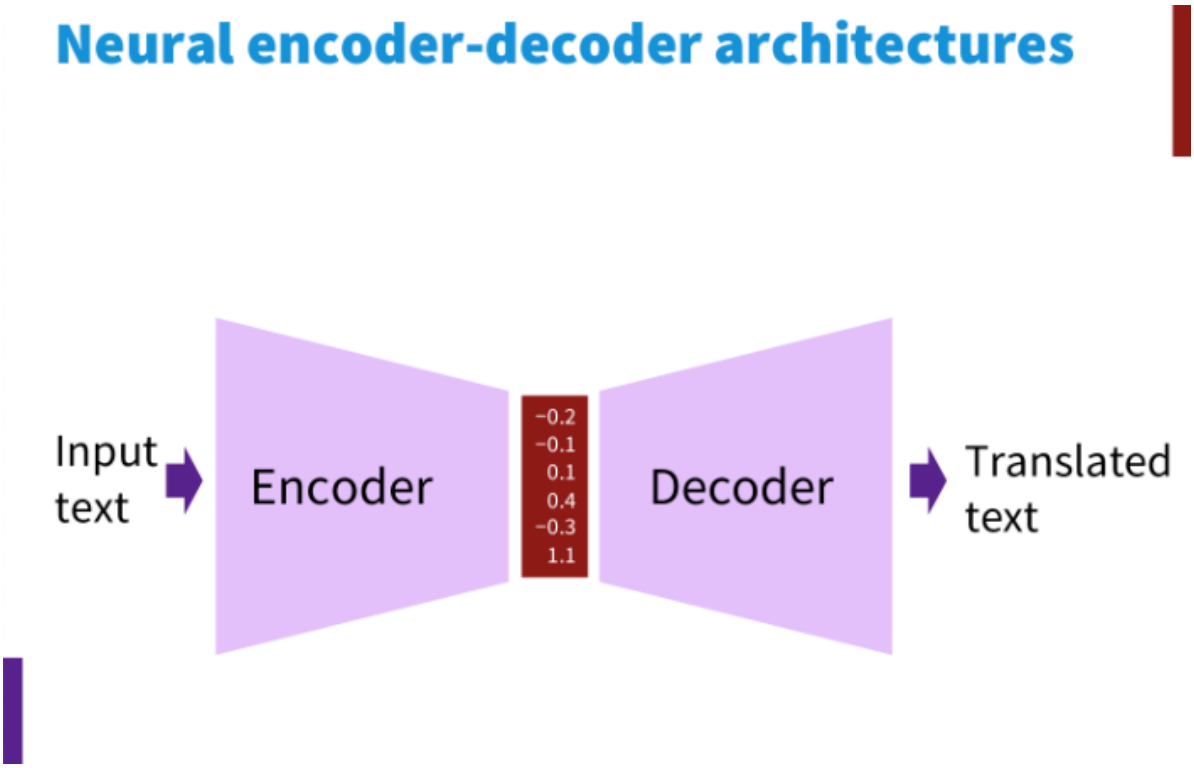
Fig. 4. Architecture of a standard LSTM module. Courtesy of Karpathy.

Encoder-Decoder and Auto-Encoder Models 编码-译码架构

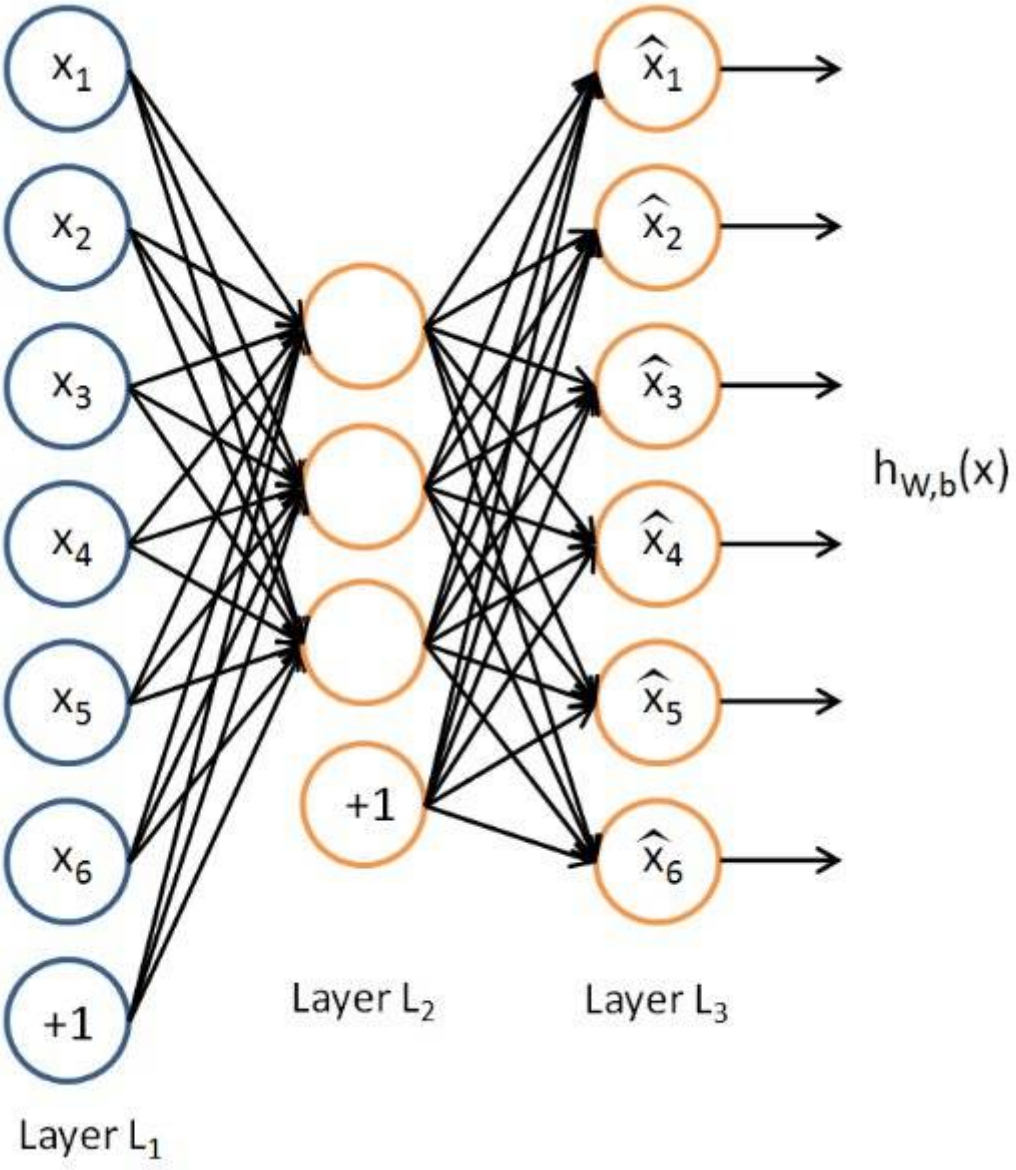
Encoder-Decoder（编码-解码）是深度学习中非常常见的一个模型框架，一个encoder是一个接收输入，输出特征向量的**网络**（FC, CNN, RNN, etc）。这些特征向量其实就是输入的特征和信息的另一种表示。

decoder同样也是一个**网络**（通常与编码器相同的网络结构，但方向相反），它从编码器获取特征向量，并输出与实际输入或预期输出最近似的结果。准确的说，Encoder-Decoder并不是一个具体的模型，而是一类框架。Encoder和Decoder部分可以是任意的文字，语音，图像，视频数据，模型可以采用CNN，RNN，BiRNN、LSTM、GRU等等。所以基于Encoder-Decoder，我们可以设计出各种各样的应用算法。

这些模型在图像到图像转换问题以及NLP中的序列模型中非常流行。此处的输出可以是图像的增强版本（例如，在图像去模糊或超分辨率中），也可以是分割图。



自动编码器神经网络是一种无监督机器学习算法、有三层的神经网络：输入层、隐藏层（编码层）和解码层。该网络的目的是重构其输入，使其隐藏层学习到该输入的良好表征。其应用了反向传播，可将目标值设置成与输入值相等。自动编码器属于无监督预训练网络（Unsupervised Pretained Networks）的一种。



Generative Adversarial Networks (GANs) 对抗网络

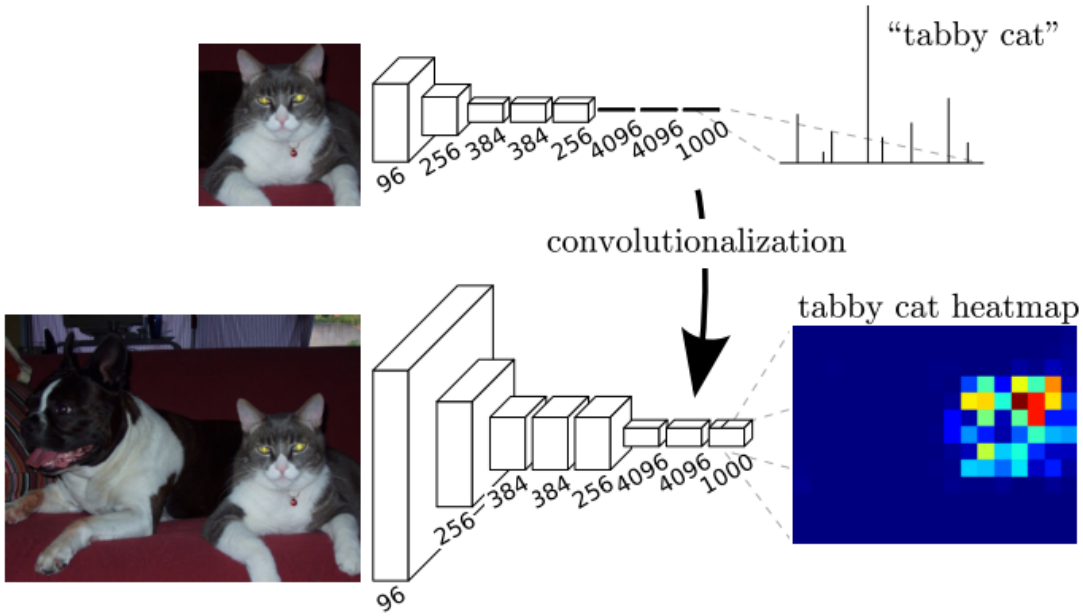
包括生成器和判别器两个部分。生成器接收随机变量并生成“假”样本，判别器则用于判断输入的样本是真实的还是合成的。两者通过相互对抗来获得彼此性能的提升。判别器所作的其实就是一个二分类任务，我们可以计算他的损失并进行反向传播求出梯度，从而进行参数更新。

总体来说， GANs简单的想法就是用两个模型，一个生成模型，一个判别模型。判别模型用于判断一个给定的图片是不是真实的图片（从数据集里获取的图片），生成模型的任务是去创造一个看起来像真的图片一样的图片。而在开始的时候这两个模型都是没有经过训练的，这两个模型一起对抗训练，生成模型产生一张图片去欺骗判别模型，然后判别模型去判断这张图片是真是假，最终在这两个模型训练的过程中，两个模型的能力越来越强，最终达到稳态。

基于深度学习的的图像分割算法模型

FCN Fully Convolutional Networks 全卷积网络

FCN将传统CNN中的全连接层转化成一个个的卷积层。如下图所示，在传统的CNN结构中，前5层是卷积层，第6层和第7层分别是一个长度为4096的一维向量，第8层是长度为1000的一维向量，分别对应1000个类别的概率。FCN将这3层表示为卷积层，卷积核的大小(通道数，宽，高)分别为（4096,1,1）、（4096,1,1）、（1000,1,1）。所有的层都是卷积层，故称为全卷积网络。



FCN有两大明显的优点：一是可以接受任意大小的输入图像，而不用要求所有的训练图像和测试图像具有同样的尺寸。二是更加高效，因为避免了由于使用像素块而带来的重复存储和计算卷积的问题。FCN的缺点也比较明显：一是得到的结果还是不够精细。进行8倍上采样虽然比32倍的效果好了很多，但是上采样的结果还是比较模糊和平滑，对图像中的细节不敏感。二是对各个像素进行分类，没有充分考虑像素与像素之间的关系，忽略了在通常的基于像素分类的分割方法中使用的空间规整（spatial regularization）步骤，缺乏空间一致性。

Convolutional Models With Graphical Models

FCN忽略了可能有用的场景级语义上下文。 为了集成更多上下文，几种方法将概率图形模型（例如条件随机场（CRF）和马尔可夫随机场（MRF））纳入DL结构。

Chen等提出了一种基于CNN和全连接的CRF语义分割算法，为了克服深层CNN的不良定位特性，他们将最终CNN层的响应与完全连接的CRF相结合。 他们表明，与以前的方法相比，他们的模型能够以更高的准确率定位分割边界。

Encoder-Decoder Based Models

一般情况

Noh等发表了有关基于反卷积（也称为转置卷积）*（我们想要建立在一个矩阵中的1个值和另外一个矩阵中的多个个值的关系,这就是像在进行卷积的逆向操作,这就是转置卷积的核心思想）*的语义分割的早期论文。 他们的模型由两部分组成，一个是使用VGG16的卷积层作为编码器，另一个解码器是将特征向量作为输入并生成像素级类别概率图的反卷积网络。 反卷积网络由反卷积层和反池化层组成，这些层识别逐个像素的类标签并预测分割掩码。

在另一个被称为SegNet的架构中，Badrinarayanan等人提出了一种用于图像分割的卷积编码器-解码器架构，SegNet的核心可训练分段引擎包括一个编码器网络（在拓扑上与VGG16网络中的13个卷积层相同），以及一个相应的解码器网络，其后是按像素分类层。 SegNet的主要新颖之处在于解码器对其较低分辨率的输入特征图进行升采样。具体来说，它使用在相应编码器的最大池化步骤中计算出的合并索引来执行非线性上采样，从而无需学习上采样。然后，将（稀疏的）上采样图与可训练的滤波器进行卷积以生成密集的特征图。

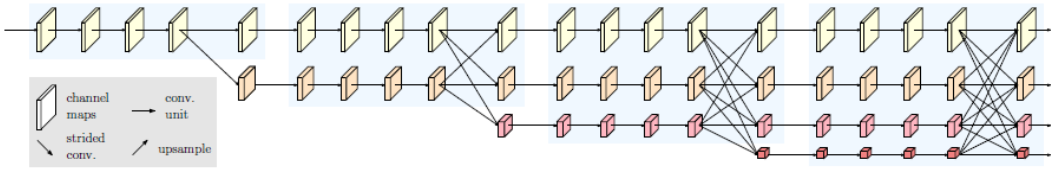


Fig. 13. Illustrating the HRNet architecture. It consists of parallel high-to-low resolution convolution streams with repeated information exchange across multi-resolution steams. There are four stages. The 1st stage consists of high-resolution convolutions. The 2nd (3rd, 4th) stage repeats two-resolution (three-resolution, four-resolution) blocks. From [119].

医疗图像

U-Net和V-Net 是两个众所周知的此类体系结构，现在也已在医疗领域之外使用。

Ronneberger等提出了用于分割显微镜图像的U-Net。他们的网络和训练策略依靠数据增强来更有效地从可用的带注释的图像中学习。

Multi-Scale and Pyramid Network Based Models

多尺度分析是图像处理中的一个相当古老的想法，已被部署在各种神经网络体系结构中。 这种最突出的模型之一是Lin等人提出的特征金字塔网络（FPN）。它主要是为目标检测而开发的，但后来也应用于分割。

R-CNN Based Models (for Instance Segmentation)

区域卷积网络（R-CNN）及其系列Fast R-CNN，Faster R-CNN，Mask-RCNN在目标检测应用中被证明是成功的。 R-CNN的某些扩展已被广泛用于解决实例分割问题。 即同时执行对象检测和语义分割的任务。 特别是，为目标检测而提出的Faster R-CNN结构（使用区域建议网络（RPN）生成候选框。RPN提取感兴趣区域（RoI），RoIPool层从这些建议框中计算特征，以推断出边界框坐标和对象的类别。

Dilated Convolutional Models and DeepLab Family

扩张卷积（空洞卷积）为卷积层引入了另一个参数，即扩张率。 空洞卷积在实时分割领域中很流行。

Recurrent Neural Network Based Models

使用RNN，可以将像素链接在一起并进行顺序处理以建模全局上下文并改善语义分割。 但是，挑战之一是图像的自然2D结构。

Visin等提出了一种基于RNN的语义分割模型ReSeg。

在另一项工作中，Byeon等 使用长短期记忆（LSTM）网络开发了场景图像的像素级分割和分类。 他们研究了自然场景图像的二维（2D）LSTM网络，同时考虑了标签的复杂空间依赖性。 在这项工作中，分类，分割和上下文集成都由2D LSTM网络执行，从而允许在单个模型中学习纹理和空间模型参数。

Attention-Based Models

Chen等提出了一种注意力机制，可以学习在每个像素位置轻柔地加权多尺度特征。 他们采用了强大的语义分割模型，并结合多尺度图像和注意力模型对其进行了训练。 注意机制的性能优于平均池化和最大池化，它使模型能够评估不同位置和比例下特征的重要性。

Generative Models and Adversarial Training

Luc等提出了一种对抗训练的语义分割方法。 他们训练了一个卷积语义分割网络，以及一个对抗网络，该网络将真实分割图与由分割网络生成的图区分开来。 他们表明，对抗训练方法可以提高Stanford Background和PASCAL VOC 2012数据集的准确性。

CNN Models With Active Contour Models

对FCN和活动轮廓模型（ACM）之间的协同作用的探索[7]最近引起了研究兴趣。一种方法是根据ACM原理制定新的损失函数。例如，受全球能源公式的启发，Chen等人提出了一个监督丢失层，该层在训练FCN的过程中结合了预测面罩的面积和大小信息，并解决了心脏MRI中的心室分割问题。

图像数据集

PASCAL VOC： 是计算机视觉中最流行的数据集之一，其带注释的图像可用于5个任务-分类，分割，检测，动作识别和person layout。该数据集分为训练和验证两套，分别具有1,464和1,449张图像。

PASCAL Context： 是PASCAL VOC 2010检测挑战的扩展，它包含所有训练图像的逐像素标签。 它包含400多个类。此数据集的许多对象类别太稀疏，因此，通常会选择59个常见类别的子集来使用。

NYU-D V2： 由Microsoft Kinect的RGB和深度相机记录的各种室内场景的视频序列组成。 它包括来自3个城市的450多个场景中的1,449张密集标记的RGB和深度图像对。

ScanNet： 是RGB-D视频数据集，在1,500多次扫描中包含250万个视图，并以3D相机，表面重建和实例级别语义分割进行注释。

性能评价

理想情况下，应该从多个方面评估模型，例如定量**精度，速度（推断时间）和存储要求（内存占用）**。但是，到目前为止，大多数研究工作都集中在**评估模型准确性的指标**上。

pixel accuracy：像素精度，正确分类的像素比率除以像素总数。

$$\mathrm{PA} = \frac{\sum_{i=0}^{K} p_{ii}}{\sum_{i=0}^{K} \sum_{j=0}^{K} p_{ij}},$$

Pii代表正确识别的像素数目，pij表示原本是i类被识别为j类的数目。

Mean Pixel Accuracy：平均像素精度，pa的拓展，其中以每个类的方式计算正确像素的比率，然后在类的总数上求平均值。

Intersection over Union (IoU) or the Jaccard Index：语义细分中最常用的指标之一。它定义为预测的分割图和实际分割之间的交集面积，除以预测的分割图和实际分割之间的并集面积

$$\mathrm{IoU} = J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

Mean-IoU：另一个流行的指标，它定义为所有类别的平均IoU。

Precision / Recall / F1 score：是用于报告许多经典图像分割模型准确性的流行指标。可以为每个类别以及整体级别定义精度和召回率。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

Dice coefficient：是另一种流行的图像分割指标（在医学图像分析中更常用），可以将其定义为预测图和真实图的重叠区域的两倍，再除以两幅图像中的像素总数。 Dice系数与IoU非常相似。

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}.$$

挑战机遇

1. 数据集的需求

   已经创建了几个大型图像数据集用于语义分割和实例分割。然而，仍然需要更具挑战性的数据集以及用于不同种类图像的数据集。对于静止图像，具有大量对象和重叠对象的数据集将非常有价值。这可以使训练模型更适合于处理密集的对象场景，以及对象之间的大量重叠，这在现实世界的场景中很常见。随着3D图像分割的日益普及，尤其是在医学图像分析中，也非常需要大型3D图像数据集。这些数据集比其较低维度的数据集更难创建。现有的可用3D图像分割数据集通常不够大，有些是合成的，因此更大，更具挑战性的3D图像数据集可能非常有价值。

2. 可解释的深度模型

   尽管基于DL的模型在具有挑战性的基准上取得了可喜的性能，但有关这些模型的问题仍然存在。例如，深度模型究竟要学习什么？我们应该如何解释这些模型学到的特征？能在给定的数据集上达到一定分割精度的最小神经架构是什么？尽管可以使用一些技术来可视化这些模型的学习卷积核，但仍缺乏对这些模型的基本行为/动力学的具体研究。

3. 半监督和无监督学习

   这些技术对图像分割特别有价值，因为在许多应用领域，尤其是在医学图像分析中，收集用于分割问题的标记样本是有难度的。转移学习方法是在大量带标签的样本（可能来自公共数据集）上训练通用图像分割模型，然后在某些特定目标应用程序的几个样本上对该模型进行微调。自我监督学习是另一个有希望的方向。借助自我监督学习，可以捕获图像中的许多细节，从而以更少的训练样本来训练分割模型。基于强化学习的模型也可能是另一个潜在的未来方向，因为在图像领域少有人关注。

4. 各种实时应用模型

   在许多应用中，精度是最重要的因素。但是，在某些应用中，至关重要的是要具有能够以接近实时或至少接近普通相机帧速率（至少每秒25帧）运行的分割模型。这对于例如部署在自动驾驶汽车中的计算机视觉系统很有用。当前的大多数模型都远没有达到这个帧速率。例如，FCN8大约需要100毫秒来处理低分辨率图像。基于扩张卷积的模型有助于在某种程度上提高分割模型的速度，但仍有很大的改进空间。

5. 3D点云分割

   许多工作专注于2D图像分割，但是处理3D点云分割的工作却很少。但是，对点云分段的兴趣日益增长，在3D建模，自动驾驶汽车，机器人技术，建筑建模等方面具有广泛的应用。处理3D无序和非结构化数据（例如点云）带来了一些挑战。例如，尚不清楚在点云上应用CNN和其他经典深度学习架构的最佳方法。基于图的深度模型可能是探索点云分割的潜在领域，