2018.11.3

SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

Abstract

1.Introduction

2.Literature review

3.Architecture

3.1 Decoder Variants

3.2 training

3.3 analysis

4. benchmarking

4.1 Road Scene Segmentation

4.2 SUN RGB-D Indoor Scenes

5.discussion and future work

6.conclusion

Abstract

提出深度全卷积网络SegNet，用于语义像素分割。其核心可训练分割引擎包括编码器网络，相应的解码器网络，以及逐像素分类层。

编码器网络的体系结构与VGG16网络中的13个卷积层相同。解码器网络的作用是将低分辨率编码器特征图映射为用于像素分类的全输入分辨率特征图。

SegNet的新颖之处在于解码器上采样的方式。

我们将我们提出的架构与广泛采用的FCN以及众所周知的DeepLab-LargeFOV，DeconvNet架构进行比较。这种比较揭示了实现良好分割性能所涉及的内存与准确度之间的权衡。与其他架构相比，SegNet提供了良好的性能，具有竞争性的推理时间和最有效的推理内存

1. Introduction

语义分割应用广泛，尝试使用深度学习。有的方法是直接将用于类别预测的深度架构拿来进行像素标记，但是结果粗糙。主要是因为最大池化和下采样降低了特征图的分辨率。SegNet设计的动机源于将低分辨率特征映射到输入分辨率大小，从而进行像素分类的需求。映射必须产生对精确边界定位有用的特征。

SegNet由道路场景理解应用所推动的。应用需要能够建模外观（道路，建筑物），形状（汽车，行人）并理解不同类别（例如道路和人行道）之间的空间关系。引擎还必须具有基于其形状描绘小尺寸对象的能力。从计算的角度来看，在推断期间，网络必须在存储和计算时间方面都是有效的

编码器部分采用VGG16的架构，删去全连接层。解码器与编码器一一对应。

对于每个样本，存储池期间计算的最大位置的索引并将其传递给解码器。解码器通过使用存储的合并索引对特征图进行上采样。称为最大池化索引（max-pooling indices）。再使用可训练的解码器滤波器组对该上采样的映射进行卷积以重建输入图像。这个想法来源于一个用于无监督特征学习的结构。max-pooling indices好处：增强边界轮廓；减少参数从而能够进行端到端训练；该上采样形式可以很容易的被合并到任何编码-解码架构中

3.architecture

SegNet具有编码器网络和相应的解码器网络，后面是最终的像素分类层。编码器网络由13个卷积层组成，这些卷积层对应于VGG16网络中为物体分类设计的前13个卷积层。 因此，我们可以根据大型数据集分类训练的权重来初始化训练过程。 我们还可以丢弃完全连接的层，以便在最深的编码器输出处保留更高分辨率的特征图。显着减少了SegNet编码器网络中的参数数量（从134到14.7 M）（参见表6）。 每个编码器层具有相应的解码器层，因此解码器网络具有13层。 最终的解码器输出被馈送到多类soft-max分类器，以独立地为每个像素产生类概率。

编码器结构为卷积，batch normalization，ReLU函数，2×2最大池化。最大池化用于在输入图片的小空间移位中实现平移不变性。下采样可以产生大的输入图像的对于特征图每个像素的上下文信息。最大池化和下采样可以增加分类的鲁棒性和平移不变性，但是会丢失特征图的空间分辨率。边界细节极为重要，其损失对对于分割有害。因此，有必要在执行二次采样之前在编码器特征图中捕获和存储边界信息。本文的方法仅存储最大池化索引。即，为每个编码器特征图存储每个池化窗口中的最大特征值的位置。

解码器使用来自相应编码器特征图存储的最大池化索引对其输入特征图进行上采样。此步骤生成稀疏特征图。然后将这些特征图与可训练的解码器滤波器组进行卷积，以产生密集的特征图。之后是批量归一化。注意，对应于第一个编码器（最接近输入图像）的解码器将产生多通道特征图。最后的解码器输出端连接可训练的soft-max分类器，soft-max独立地对每个像素进行分类。

DeconvNet [52]和U-Net [15]与SegNet共享一个类似的架构，但有一些差异。 DeconvNet具有更大的参数化，需要更多的计算资源，并且更难以端到端训练（表6），主要是由于使用了完全连接的层（虽然是卷积方式）我们报告了几次与DeconvNet的比较。论文第4节。

与SegNet相比，U-Net [15]（针对医学成像社区提出）不重用池化索引，而是将整个特征图（以更多内存为代价）传输到相应的解码器并将它们连接到上采样（通过反卷积）解码器特征图。与VGG网络架构一样，U-Net中没有conv5和max-pool五块。另一方面，SegNet使用来自VGG网的所有预先训练的卷积层权重作为预训练权重。

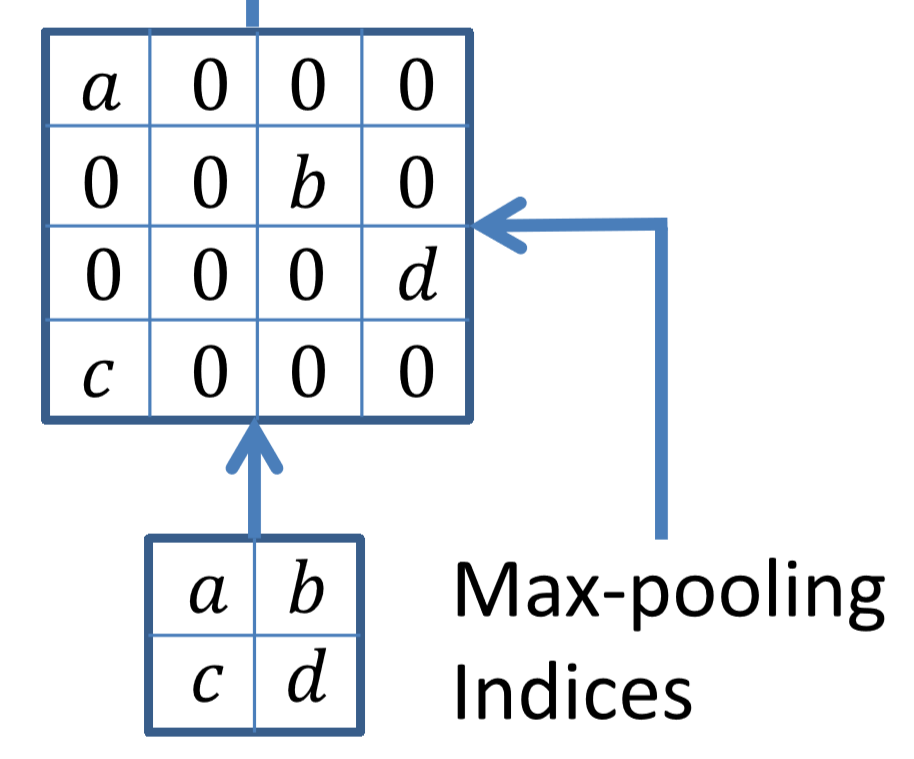
3.1 Decoder Variants

将SegNet与其他编码-解码结构进行对比

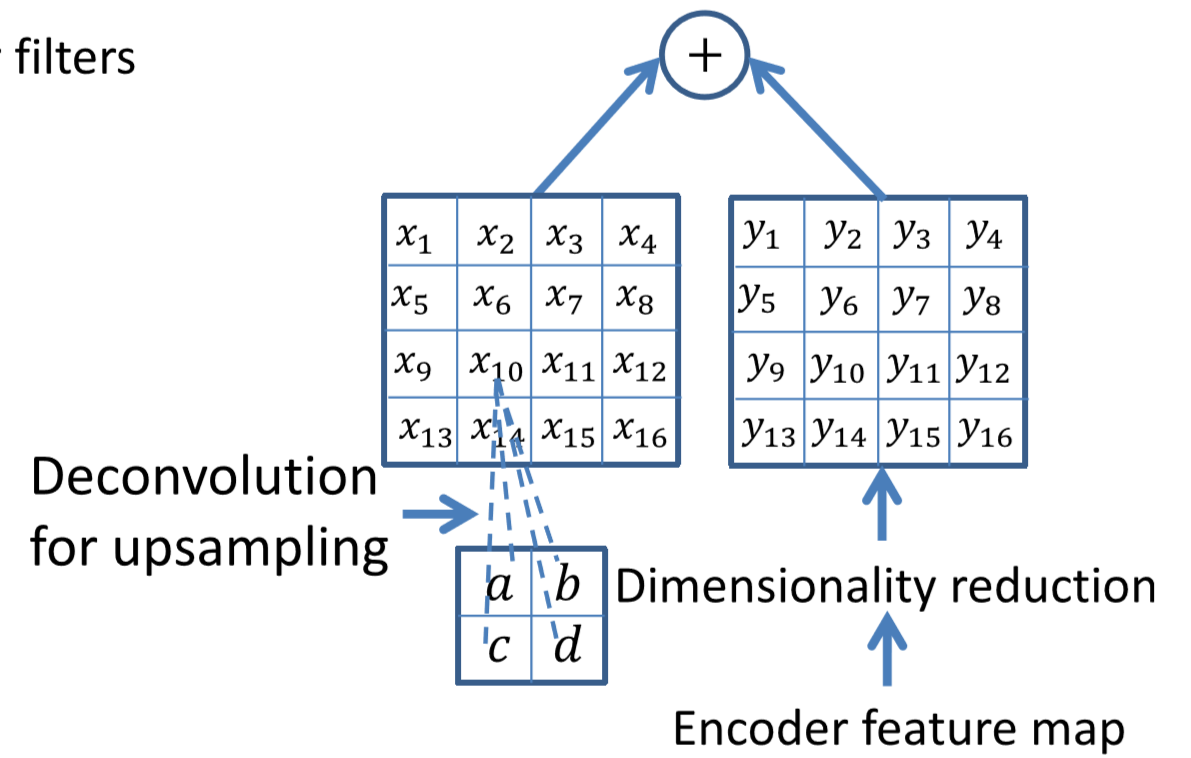
为了分析SegNet并将其性能与FCN（解码器变体）进行比较，我们使用较小版本的SegNet，称为SegNet-Basic。具有四个编码器和四个解码器。 SegNet-Basic中的所有编码器执行最大池化和二次采样，并且相应的解码器使用接收的最大池索引对其输入进行上采样。在编码器和解码器网络中的每个卷积层之后使用批量归一化。在卷积之后没有使用偏差，并且在解码器网络中不存在ReLU非线性。此外，选择在所有编码器和解码器层上的7×7的恒定内核大小以提供用于平滑标记的宽的上下文。在106×106像素的输入图像中。这种小尺寸的SegNet-Basic允许我们探索许多不同的变体（解码器）并在合理的时间内训练它们。类似地，我们创建了FCN-Basic，这是我们分析的FCN的可比版本，它与SegNet-Basic共享相同的编码器网络，但使用了所有解码器中使用的FCN解码技术，即使用转置卷积。

SegNet（也是SegNet-Basic）使用的解码技术，其中在上采样步骤中不涉及学习。然而，上采样的图与可训练的多通道解码器滤波器卷积以使其稀疏输入致密化。每个解码器滤波器具有与上采样特征图的数量相同的通道数。较小的变体是解码器滤波器是单通道的变体，即它们仅卷绕它们相应的上采样特征映射。此变体（SegNetBasic-SingleChannelDecoder）显着减少了可训练参数和推理时间的数量。

图3右侧是FCN（也是FCN-Basic）解码技术。 FCN模型的重要设计元素是编码器特征图的降维步骤。这压缩了编码器特征图，然后在相应的解码器中使用。编码器特征图（例如64个通道）的维度减小是通过将它们与1×1×64×K可训练滤波器进行卷积来执行的，其中K是类的数量。压缩的K通道最终编码器层特征图是解码器网络的输入。在该网络的解码器中，通过使用固定或可训练的多通道上采样kernel的反卷积来执行上采样。我们将kernel大小设置为8×8。注意，相比之下，SegNet使用可训练解码器滤波器的多通道卷积在上采样之后执行，致密化特征图。 FCN中的上采样特征映射具有K个通道。然后将其逐个元素地添加到相应的分辨率编码器特征图以产生输出解码器特征图。使用双线性插值权重来初始化上采样kernel。

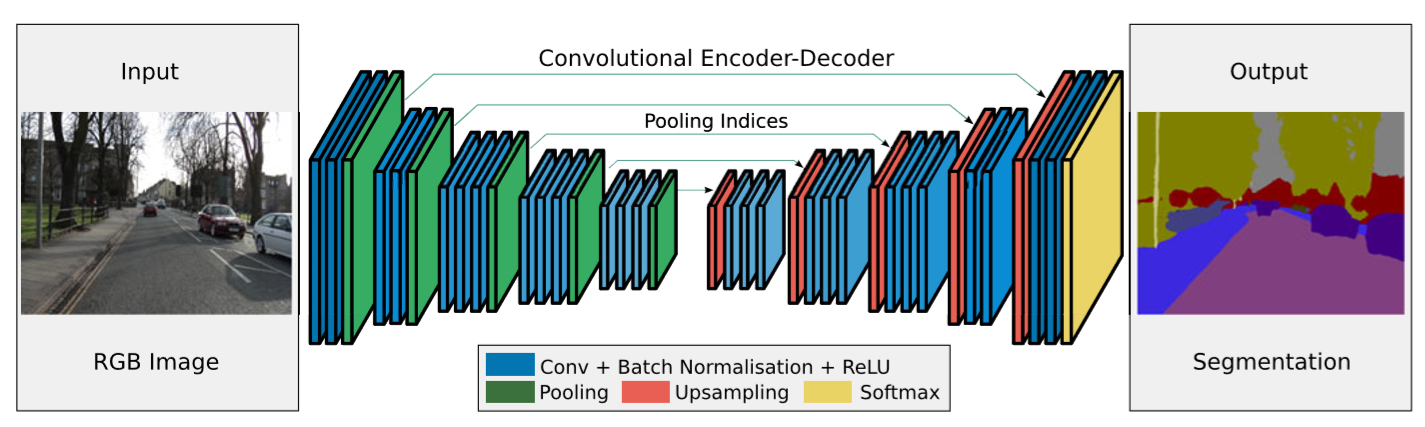


FCN解码器模型需要在推理期间存储编码器特征图。 这对嵌入式应用来说可能是存储密集型的。 另一方面，SegNet对于池化索引需要几乎可忽略不计的存储成本。 我们还可以创建FCNBasic模型的变体，该模型丢弃编码器特征图添加步骤，并且仅学习上采样内核（FCN-Basic-NoAddition）。



我们可以在每一层解码器上添加64个编码器特征图，以创建更加内存密集的SegNet变体（SegNet-Basic-EncoderAddition）。

另一个和更多内存密集型FCN-Basic变体（FCN-Basic-NoDimReduction）是没有为编码器特征图执行降维。

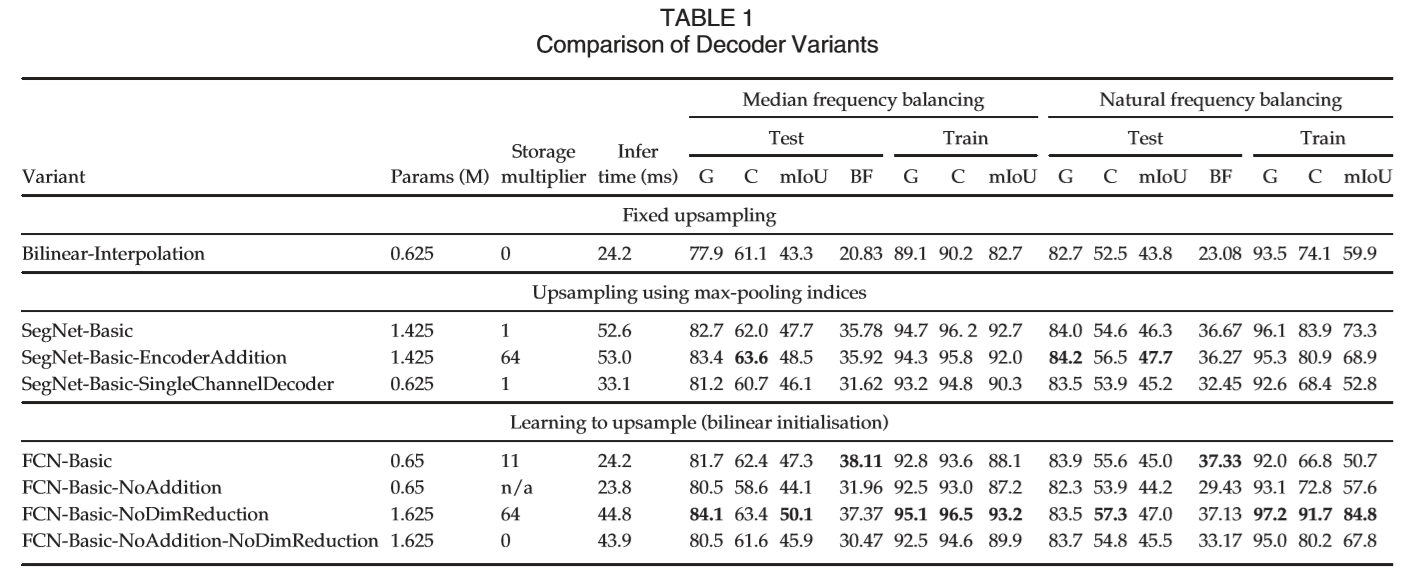


3.2training

367次训练和233次测试RGB图像（白天和黄昏场景），分辨率为360c×480。挑战是划分11个类别，如道路，建筑，汽车，行人，标志，杆子，人行道等。我们对RGB输入执行局部对比度归一化(local contrast normalization)。

使用随机梯度下降，固定学习率为0.1，动量为0.9。我们选择在验证数据集上执行最高的模型。

使用交叉熵损失(cross-entropy loss)作为训练网络的目标函数



使用三种常用的性能测量：全局准确度（G）测量在数据集中正确分类的像素百分比，

类平均准确度（C）是所有类上的预测准确度的平均值

mIoU平均交叉比，mIoU度量是比类平均准确度更严格的度量

语义轮廓测量semantic contour measure (BF).

到基于双线性插值的上采样，没有任何学习，基于所有精度测量表现最差。

其他的方法都比其好，表明了在分割中学习解码器的重要性

比较SegNet-Basic和FCN-Basic时，我们发现在所有测量精度的测试中，两者都表现得同样出色。

区别在于SegNet在推理期间使用较少的内存，因为它仅存储最大池化索引。

FCN-Basic使用降维在每个解码器层中具有较少的特征图。这减少了解码器网络中的卷积数，因此FCN-Basic在推理（正向传递）期间更快。

SegNet-Basic中的解码器网络使其整体上比FCN-Basic更大。这使得它具有更高的灵活性，因此在相同的迭代次数下实现了比FCN-Basic更高的训练精度。

总的来说，当推理时间内存受到限制但是在某种程度上可以妥协推理时间时，我们看到SegNet-Basic优于FCN-Basic。

SegNet-Basic在解码器方面与FCN-Basic-NoAddition最相似。

SegNet-Basic的性能优越，部分原因在于其较大的解码器尺寸。

与FCN-Basic相比，FCN-Basic-NoAddition的准确性也较低。 这表明捕获编码器特征图中存在的信息以获得更好的性能至关重要。

FCN-Basic-NoAddition-NoDimReduction模型的大小略大于SegNet-Basic，因为最终的编码器特征映射没有被压缩以匹配类别K的数量。

该FCN变体的性能在测试中比SegNet-Basic差。 这表明使用较大的解码器是不够的，捕获编码器特征图信息以便更好地学习也很重要，特别是细粒度的轮廓信息（注意BF测量中的下降）。 在这里，有趣的是，与大型模型（如FCN-Basic-NoDimReduction）相比，SegNet-Basic具有更具竞争力的训练精度。

FCN-BasicNoAddition和SegNet-Basic-SingleChannelDecoder之间的另一个有趣的比较

使用max-pooling indices进行采样和更大的解码器可以获得更好的性能。 这也证明SegNet是一种很好的分割架构，特别是当需要在存储成本，精度与推理时间之间找到折衷方案时。

在最好的情况下，当内存和推理时间都不受约束时，较大的模型（如FCN-Basic-NoDimReduction和SegNet-EncoderAddition）都比其他变体更准确。 特别是，在FCN-Basic模型中丢弃维数降低导致具有高BF分数的FCN-Basic变体中的最佳性能。 这再一次强调了分割架构中存储和准确度之间的权衡。

总结：

1. 当编码器特征图被完整储存的时候，效果最好
2. 当推理内存受到限制时，压缩编码器特征图（降维或者最大池化索引）并储存，再放入一个合适的解码器中使用，可以提高性能
3. 更大的解码器表现更好
4. benchmarking

两种分类。

道路场景分割，用于自动驾驶

室内场景分割，用于AR增强现实

对比了三个网络

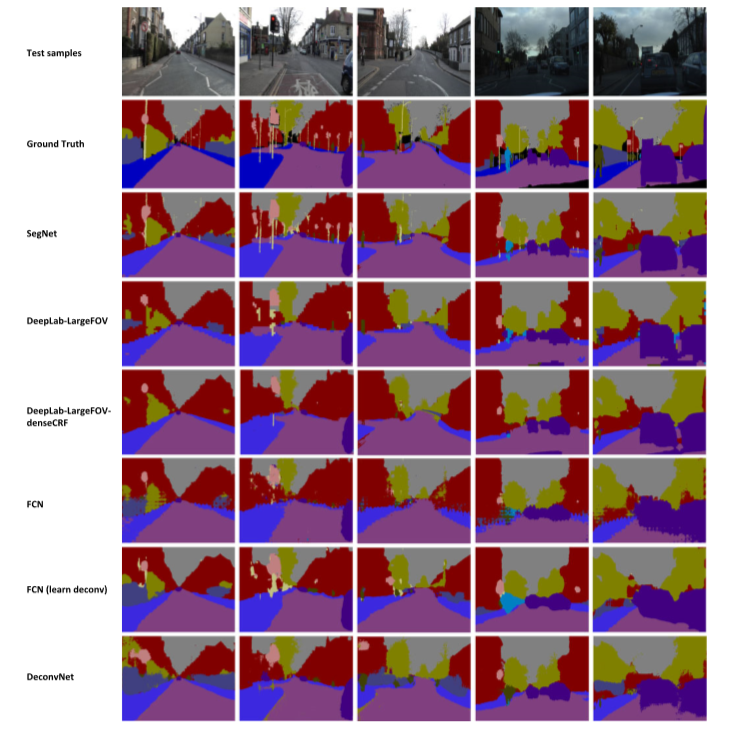
FCN , DeepLab-LargFOV and DeconvNet

Road Scene Segmentation

与一些较大型号相比，SegNet显示出卓越的性能，够很好地分割大类和小类。 DeepLab-LargeFOV是最有效的模型，尽管较小的类丢失，但CRF后处理可以产生有竞争力的结果。

具有学习反褶积的FCN明显优于固定双线性上采样。

DeconvNet是最大的模型，也是最难以训练的。 它的预测不会保留小类。



我们提出了SegNet，用于语义分割的一个深度卷积网络结构。 SegNet背后的主要动机是需要设计一个有效的道路和室内场景理解架构，这在内存和计算时间方面都很有效。我们分析了SegNet并将其与其他重要变体进行了比较，以揭示设计分段架构所涉及的实际权衡，特别是训练时间，内存与准确性。

那些存储编码器网络特征的架构完全表现最佳，但在推理期间消耗更多内存。另一方面，SegNet更有效，因为它只存储特征映射的最大池索引，并在其解码器网络中使用它们以实现良好的性能。