摘要

在本文中，我们通过捕捉基于自我注意机制的丰富的上下文依赖来解决场景分割任务。与以往通过多尺度特征融合来捕捉上下文的工作不同，我们提出了一个双注意网络（DANet）来适应性地整合局部特征和它们的全局依赖关系。具体来说，我们在扩张的FCN之上附加了两种注意力模块，分别在空间和通道维度上模拟语义的相互依赖。位置注意模块通过所有位置的特征的加权和，选择性地聚集每个位置的特征。类似的特征会相互关联，而不考虑它们的距离。同时，通道注意模块通过整合所有通道图中的相关特征，选择性地强调相互依赖的通道图。我们将两个注意力模块的输出相加，进一步改善特征表示，这有助于获得更精确的分割结果。我们在三个具有挑战性的场景分割数据集，即Cityscapes、PASCAL Context和COCO Stuff数据集上实现了新的最先进的分割性能。特别是，在不使用粗略数据的情况下，在Cityscapes测试集上取得了81.5%的平均IoU得分。

1.介绍  
场景分割是一个具有挑战性的基本问题，其目标是将场景图像分割并解析为与语义类别相关的不同图像区域，包括东西（如天空、道路、草地）和离散物体（如人、汽车、自行车）。对这项任务的研究可以应用于潜在的应用，如自动驾驶、机器人感应和图像编辑。为了有效地完成场景分割的任务，我们需要区分一些混乱的类别，并考虑到具有不同外观的物体。例如，"田野 "和 "草地 "的区域往往是无法区分的，而 "汽车 "的对象可能经常受到尺度、遮挡和光照的影响。因此，有必要提高像素级识别的特征表示的鉴别能力。最近，人们提出了基于全卷积网络（FCN）[13]的最先进的方法来解决上述问题。一种方法是利用多尺度背景融合。例如，一些工作[3，4，29]通过结合不同的扩张卷积和池化操作产生的特征图来聚合多尺度语境。还有一些作品[15, 27]通过扩大具有分解结构的内核大小或在网络顶部引入有效的编码层来捕获更丰富的全局语境信息。此外，还提出了编码器-解码器结构[6，10，16]来融合中层和高层语义特征。虽然上下文融合有助于捕捉不同尺度的物体，但它不能利用全局视图中物体或东西之间的关系，这对场景分割也是至关重要的。

另一类方法采用递归神经网络来利用长距离的依赖关系，从而提高场景分割的准确性。基于二维LSTM网络的方法[1]被提出来捕捉标签上复杂的空间依赖关系。工作[18]建立了一个带有有向无环图的递归神经网络来捕捉局部特征上丰富的上下文依赖关系。然而，这些方法用递归神经网络隐含地捕捉全局关系，其有效性在很大程度上依赖于长期记忆的学习结果。为了解决上述问题，我们提出了一个新的框架，称为双注意网络（DANet），用于自然场景图像分割，如图2所示。它引入了一种自我注意机制，分别在空间和通道维度上捕捉特征的依赖性。具体来说，我们在扩张的FCN上面附加了两个平行的注意力模块。一个是位置注意模块，另一个是通道注意模块。对于位置注意模块，我们引入了自我注意机制来捕捉特征图中任意两个位置之间的空间依赖关系。对于某个位置的特征，它是通过对所有位置的特征进行加权求和来更新的，其中的权重是由相应的两个位置之间的特征相似性决定的。也就是说，任何两个具有相似特征的位置都可以促进相互改进，而不管它们在空间维度上的距离如何。对于通道注意模块，我们使用类似的自我注意机制来捕捉任何两个通道图之间的通道依赖性，并以所有通道图的加权和来更新每个通道图。最后，这两个注意力模块的输出被融合以进一步增强特征表示。应该指出，在处理复杂多样的场景时，我们的方法比以前的方法[4, 29]更加有效和灵活。以图中的街道场景为例。1为例。首先，第一行的一些 "人 "和 "红绿灯 "由于光线和视野的原因是不明显或不完整的物体。如果探索简单的上下文嵌入，来自被支配的突出物体（如汽车、建筑）的上下文会损害这些不明显的物体标签。相比之下，我们的注意力模型选择性地聚集了不显眼物体的相似特征，以突出它们的特征表征，避免了显眼物体的影响。第二，"车 "和 "人 "的尺度是多样的，识别这种多样的物体需要不同尺度的背景信息。也就是说，不同尺度的特征应该被同等对待，以代表相同的语义。我们的模型与注意力机制只是为了从全局角度自适应地整合任何尺度的相似特征，这在一定程度上可以解决上述问题。第三，我们明确地考虑了空间和通道的关系，因此场景理解可以从长距离的依赖关系中受益。

我们的主要贡献可以概括为以下几点。

- 我们提出了一个具有自我注意机制的新型双注意网络（DANet），以增强特征表征对场景分割的判别能力。

- 提出了一个位置注意模块来学习特征的空间相互依赖性，并设计了一个通道注意模块来模拟通道相互依赖性。它通过对局部特征的丰富的上下文依赖关系进行建模，大大改善了分割结果。

- 我们在三个流行的基准上取得了最先进的结果，包括Cityscapes数据集[5]，PASCAL Context数据集[14]和COCO Stuff数据集[2]。

3.DANet

在这一节中，我们首先介绍了我们网络的一般框架，然后介绍了两个注意力模块，它们分别在空间和通道维度上捕捉长距离的背景信息。最后，我们描述了如何将它们聚集在一起进行进一步的细化。

3.1. 概述

给定一张场景分割的图片，东西或物体在比例、光照和视图上是不同的。由于卷积操作会导致局部接受场，与具有相同标签的像素对应的特征可能有一些差异。这些差异会带来类内的不一致，影响识别的准确性。为了解决这个问题，我们通过在注意力机制下建立特征间的关联来探索全局性的背景信息。我们的方法可以自适应地聚合长范围的上下文信息，从而改善场景分割的特征表示。

如图所示。如图2所示，我们设计了两种类型的注意力模块，在扩张的残差网络产生的局部特征上绘制全局背景，从而获得更好的像素级预测的特征表示。

我们采用预训练的残差网络，以扩张策略[3]为骨干。值得注意的是，我们去掉了下采样操作，在最后两个ResNet块中采用了扩张卷积，从而将最终的特征图大小扩大到输入图像的1/8。它保留了更多的细节，而没有增加额外的参数。然后，来自扩张的残差网络的特征将被送入两个平行的注意力模块。

以图3中上部的空间注意模块为例，我们首先应用一个卷积层来获得降维的特征。

然后，我们将这些特征送入位置注意模块，并通过以下三个步骤生成新的空间长距离背景信息的特征。

第一步是生成一个空间注意力矩阵S，该矩阵对特征的任何两个像素之间的空间关系进行建模。

第二步，我们在注意力矩阵和原始特征之间进行矩阵乘法。

第三，我们对上述相乘的结果矩阵和原始特征进行元素相加的操作，以获得反映长距离语境的最终表征E。

同时，通道维度上的长程上下文信息被通道关注模块所捕获。

除了第一步，捕捉通道关系的过程与位置注意模块相似，在这一步中，通道注意矩阵是以通道维度计算的。最后，我们将两个注意模块的输出汇总，以获得更好的特征表示，用于像素级预测。

3.2. 位置注意模块

鉴别性特征表示对于场景理解至关重要，这可以通过捕获长距离的上下文信息来获得。然而，许多工作[15, 29]表明，由传统的FCNs产生的局部特征可能会导致对物体和东西的错误分类。为了在局部特征上建立丰富的上下文关系模型，我们引入了一个位置注意模块。位置注意模块将更广泛的上下文信息编码到局部特征中，从而提高了它们的表示能力。接下来，我们详细说明了自适应聚合空间背景的过程。如图3(A)所示，给定一个局部特征A∈RC×H×W，我们首先将其输入卷积层，分别生成两个新的特征图B和C，其中{B，C}∈RC×H×W。然后我们将它们重塑为RC×N，其中N=H×W是像素的数量。之后，我们在C和B的转置之间进行矩阵乘法，并应用softmax层来计算空间注意图S∈RN×N。

公式1

其中sji衡量第i个位置对第j个位置的影响。两个位置的特征表示越相似，就越有助于它们之间的关联性。同时，我们将特征A送入卷积层，生成一个新的特征图D∈RC×H×W，并将其重塑为RC×N。然后，我们在D和S的转置之间进行矩阵乘法，将结果重塑为RC×H×W。最后，我们将其乘以一个比例参数α，并与特征A进行元素相加运算，得到最终输出E∈RC×H×W，如下所示

公式2

其中α初始化为0，并逐渐学习分配更多的权重[28]。从公式2中可以推断出，每个位置的结果特征E是所有位置的特征和原始特征的加权和。因此，它有一个全局的语境观，并根据空间注意图选择性地聚合语境。相似的语义特征实现了相互增益，从而提高了类内紧凑性和语义一致性。

3.3. 通道注意模块

高层特征的每个通道图都可以被看作是一个特定的类反应，不同的语义反应是相互关联的。通过利用通道图之间的相互依存关系，我们可以强调相互依存的特征图，改善特定语义的特征表示。因此，我们建立了一个通道注意模块来明确地模拟通道之间的相互依赖关系图3(B)说明了通道注意模块的结构。与位置注意模块不同，我们直接从原始特征A∈RC×H×W计算出通道注意图X∈RC×C。具体来说，我们将A重塑为RC×N，然后在A和A的转置之间进行矩阵乘法。最后，我们应用softmax层来获得通道注意图X∈RC×C。

公式3  
其中xji衡量第i个通道对第j个通道的影响。此外，我们在X和A的转置之间进行矩阵乘法，将其结果重塑为RC×H×W。然后，我们将结果乘以一个比例参数β，并与A进行元素相加运算，得到最终输出E∈RC×H×W。  
公式4  
方程4显示，每个通道的最终特征是所有通道的特征和原始特征的加权和，它模拟了特征图之间的长程语义依赖关系。它有助于提高特征的可辨别性。  
值得注意的是，在计算两个通道的关系图之前，我们没有采用卷积层来嵌入特征，因为它可以保持不同通道图之间的关系。此外，与最近的工作[27]不同的是，我们利用所有相应位置的空间信息来建立通道相关性模型，通过全局汇集或编码层来探索通道关系。

3.4. 注意力模块嵌入网络

为了充分利用长距离的上下文信息，我们将这两个注意力模块的特征聚合起来。具体来说，我们通过卷积层转换两个注意力模块的输出，并进行元素求和来完成特征融合。最后，再通过卷积层来生成最终的预测图。我们没有采用级联操作，因为它需要更多的GPU内存。我们注意到，我们的注意力模块很简单，可以直接插入现有的FCN管道中。它们没有增加太多的参数，但却能有效地加强特征表示。

5. 结论

在本文中，我们提出了一个用于场景分割的双注意网络（DANet），它利用自我注意机制自适应地整合了局部语义特征。具体来说，我们引入了一个位置注意模块和一个通道注意模块来分别捕捉空间和通道维度上的全局依赖性。消融实验表明，双注意模块能有效地捕捉长距离的上下文信息，并给出更精确的分割结果。我们的注意力网络在四个场景分割数据集上持续取得了出色的表现，即Cityscapes、Pascal VOC 2012、Pascal Context和COCO Stuff。此外，降低计算复杂性和增强模型的鲁棒性也很重要，这将在未来的工作中进行研究。