计算机视觉课程报告

Dual Attention Network for Scene Segmentation

梁垒
21821198
计算机科学与技术学院
516604842@qq.com

摘要（Abstract）

本文中，使用基于自注意力机制捕获丰富的语境关联来解决场景分割问题。与以往通过多尺度特征融合捕获语境的研究不同，本文提出了一种双重注意力网络（DANet）来自适应地将局部特征与其全局依赖关系相结合。具体来说，本文在传统的基于空洞卷积的 FCN 上添加了两种注意力模块，分别对空间维度和通道维度的语义相互关联进行建模。位置注意力模块通过所有位置的特征加权总和选择性地聚集每个位置的特征。无论距离远近，相似的特征都会相互关联。同时，通道注意力模块通过整合所有通道图中的相关特征，有选择地强调相互关联的通道图。本文将两个注意力模块的输出相加，以进一步改进特征表示，这有助于获得更精确的分割结果。本文在三个具有挑战性的场景分割数据集（Cityscapes、PASCAL Context 和 COCO Stuff）上取得了当前最佳分割性能。特别是，在不使用粗略数据的情况下，在 Cityscapes 测试集的平均 IoU 分数达到了 81.5 %。

1. Introduction

语义分割具有广泛的应用范围，从场景理解，推断对象之间的支持关系到自主驾驶.依靠低级别视觉线索的早期方法已经被流行的机器学习算法所取代.特别的，深度学习后来在手写数字识别、语音、整图分类以及图片中的检测上都取得了成功.现在图像分割领域也对这个方法很感兴趣等.然而，近来的很多方法的都尽力直接采用设计来图像分类的方法进行语义分割.结果虽然令人鼓舞，但是比较粗糙.这主要是因为max-pooling和sub-sampling减少了特征图的分辨率。

1.1. 动机

场景分割是一个基本且具有挑战性的问题，其目标是将场景图像分割和解析到与语义类别相关联的不同图像区域，包括填充物（例如天空，道路，草地）和离散物体（例如人，汽车，自行车）。 该任务的研究可应用于潜在的应用，例如自动驾驶，机器人传感和图像编辑。 为了有效地完成场景分割任务，我们需要区分一些令人困惑的类别，并将不同外观的对象分解。 例如，“田地”和“草地”的区域通常是难以区分的，并且通常难以区分的“汽车”可能具有不同的尺度，遮挡和照明。因此，有必要提高像素级识别特征表示的辨别能力

1.2. 相关工作

通常情况下分析图像使用CNN网络，通过一系列卷积核依次扫过图像的每一个区域，通过不同的卷积核来发现图像不同位置上的一些图形特征，这就构成了图像低层的特征选择过程，之后在此基础上继续使用卷积核来将基层特征进行组合，进而形成高层的特征，最后再将结果向量进行分类，这种模型受启发于生物神经中的感受野，并取得了很好的效果。

但是这类模型同样有他的局限性，例如在选取底层特征时卷积核较为固定，其鲁棒性较差，卷积核的大小在扫描图象时也是固定的，会受到图像的规模影响，再者，大量的特征需要大量的卷积核来训练，会导致参数增加，大幅增加了计算量，最后，底层的特征在高层特征分析时往往会被转化为有或没有来考虑，这就导致底层特征的一些细节在向后传播时被忽略掉，这往往会导致信息的丢失甚至导致模型的误判，所以不能将底层特征与高层特征相结合也是传统CNN 的一个缺陷。

对于场景分割任务，一种方法是利用多标签文本融合。 例如，一些作品通过组合由不同的扩张卷积和池化操作生成的特征图来聚合多尺度上下文。有些方法通过使用分解结构扩大核大小或在网络顶部引入有效编码层来捕获更丰富的全局通信信息。此外，提出编码器 - 解码器结构来融合中级和高级语义特征。虽然上下文融合有助于捕获不同比例的对象，但它无法利用全局视图中对象或东西之间的关系，这对于场景分割也是必不可少的。另一种类型的方法使用递归神经网络来利用长程依赖性，从而提高分割精度。提出了基于 2DLSTM[1] 网络的方法来捕获标签上复杂的空间依赖性。

Figure 1: The goal of scene segmentation is to recognize each pixel including stuff, diverse objects. The various scales, occlusion and illumination changing of objects/stuff make it challenging to parsing each pixel.

图 1

1.3. 本文工作

为了解决上述问题，本文提出了一种双注意力网络模型，在传统attention方法的基础上，将宏观与微观特征结合在一起。相比于以前的工作[2][3]，本文的方法更具有灵活性。

具体来说，我们在传统的扩张FCN之上添加了两个并行的注意模块。一个是位置注意力模块(position attention module)，另一个是通道注意模块(channel attention module)。对于位置注意模块，本文引入自注意力制来捕获特征图的任意两个位置之间的空间依赖性。对于特定位置的特征，通过加权累积的所有位置的特征来聚合更新特征 权重由相应两个位置之间的特征相似性决定。也就是说，任何具有相似特征的两个位置都可以贡献出改进，无论它们在空间维度上的距离如何。对于通道注意力模块，本文使用相似的自注意力机制来捕获任意两个通道映射之间的通道依赖关系，并使用所有通道映射的加权和来更新每个通道映射。最后，这两个注意模块的输出被融合以进一步增强特征表示。

值得注意的是，在处理复杂多样的场景时，本文的方法比以前的方法[Rethinking atrous convolution for semantic image segmentation, PSPNet]更有效，更灵活。走图中的街景。以图 1 为例。首先，第一行中的一些“人”和“交通信号灯”因为光照和视角，是不显眼或不完整的物体。如果探索简单的上下文嵌入，来自主导的显着对象（例如汽车，建筑物）的上下文将损害那些不显眼的对象标记。相比之下，本文的注意模型选择性地聚合不显眼对象的相似特征，以突出其特征表示，并避免显着对象的影响。其次，“汽车”和“人”的尺度是多样的，并且识别这种不同的对象需要不同尺度的背景信息。也就是说，应该平等对待不同尺度的特征以表示相同的语义。本文的注意机制模型只是旨在从全局视角自适应地集成任何尺度的相似特征，这可以解决上述问题的程度。第三，本文明确地考虑空间关系和通道关系，以便场景理解可以受益于远程依赖。

本文的贡献点：

1.本文提出了一种具有自注意力机制的新型双重注意网络（DANet），以增强场景分割的特征表示的判别能力。

2.提出了一种位置注意力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力模块来模拟通道相互依赖性。 它通过在本地特征上建模丰富的上下文依赖性来显着改善分割结果。

3.本文在三个流行基准测试中获得了新的最新结果，包括Cityscapes数据集，PASCAL Context数据集和COCO Stuff数据集。

2. 模型

2.1. Attention

Attention模型提出于attention is all you need[4]，首先应用于自然语言处理的任务中，由于自然语言处理通常将单词转化为word embedding来输入，但是文本中有些单词对于文本的意义有重大影响，一些介词却并不会对文本意义有大的影响，为了让模型更加聚焦于起到关键作用的词汇，attention机制通过对不同输入赋予不同权重，以此来使模型有更好的聚焦，attention的模型如图 2 所示。之后被应用于多种其他任务[5][6][7]中，称为self-attention module[8][9][10]。

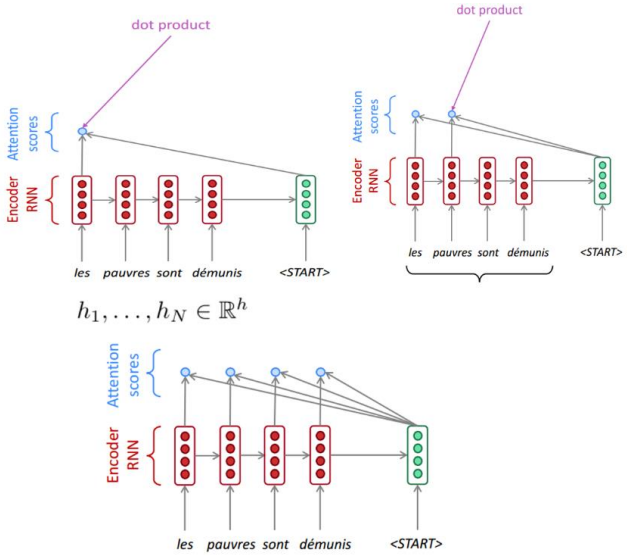
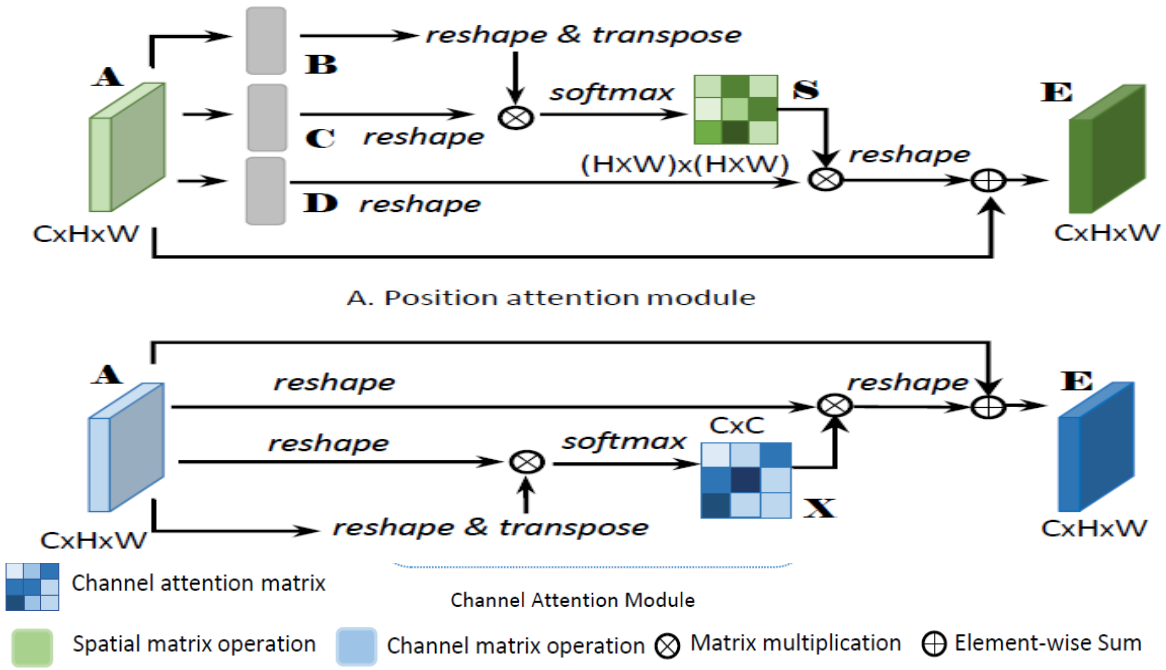


图 2

attention的通用定义：给定一组向量集合values，以及一个向量query，attention机制是一种根据该query计算values的加权求和的机制。attention的重点就是这个集合values中的每个value的“权值”的计算方法。有



时候也将这种attention的机制叫做query的输出关注。

图 3

2.2. 本文模型

DANet 在特征的空间维度和通道维度分别引入自注意力机制，即位置注意力模块和通道注意力模块，有效抓取特征的全局依赖关系。系统框架图如图 3 所示，两个模块的具体结构如图 4 所示

位置注意力模块旨在利用任意两点特征之间的关联，来相互增强各自特征的表达。具体来说，首先计算出任意两点特征之间关联强度矩阵，即原始特征 A 经过卷积降维获得特征 B 和特征 C，然后改变特征维度 B 为 ((HxW)xC') 和 C 为 (C'x(HxW)) 然后矩阵乘积获得任意两点特征之间的关联强度矩((HxW)x(HxW))。然后经过 softmax 操作归一化获得每个位置对其他位置的 attention 图 S，其中越相似的两点特征之间，其响应值越大。接着将 attention 图中响应值作为加权对特征 D 进行加权融合，这样对于各个位置的点，其通过 attention 图在全局空间中的融合相似特征。

位置注意力模型的权重计算方法如下：

$$s_{ji} = \frac{exp(B_i \cdot C_j)}{\sum_{i=1}^{N} exp(B_i \cdot C_j)} \tag{1}$$

最终的输出矩阵计算方法如下：

$$E_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j \tag{2}$$

通道注意力模块旨在通过建模通道之间的关联，增强通道下特定语义响应能力。具体过程与位置注意力模块相似，不同的是在获得特征注意力图 X 时，是将任意两个通道特征进行维度变换和矩阵乘

积，获得任意两个通道的关联强度，然后同样经过 softmax 操作获得的通道间的 attention 图。最后通过通道之间的 attention 图加权进行融合，使得各个通道之间能产生全局的关联，获得更强的语义响应的特征。

位置注意力模型的权重计算方法如下：

$$x_{ji} = \frac{exp(A_i \cdot A_j)}{\sum_{i=1}^{C} exp(A_i \cdot A_j)} \tag{3}$$

最终的输出矩阵计算方法如下：

$$E_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j \tag{4}$$

为了进一步获得全局依赖关系的特征，将两个模块的输出结果进行相加融合，获得最终的特征用于像素点的分类向量。

3. 实验结果

位置注意力模块的效果在图 5 中可视化，一些细节和对象边界在使用位置注意力模块时更加清晰，例如第一行中的“杆子”和第三行的“人行道”。对局部特征的选择性融合增强了对细节的区分

同时，图 5 证明，利用本文的信道注意模块，一些错误分类的类别现在被正确地分类，如第一行和第三行中的“公交车”。 通道映射之间的选择性集成有助于捕获上下文信息。 语义一致性得到了明显的改善。

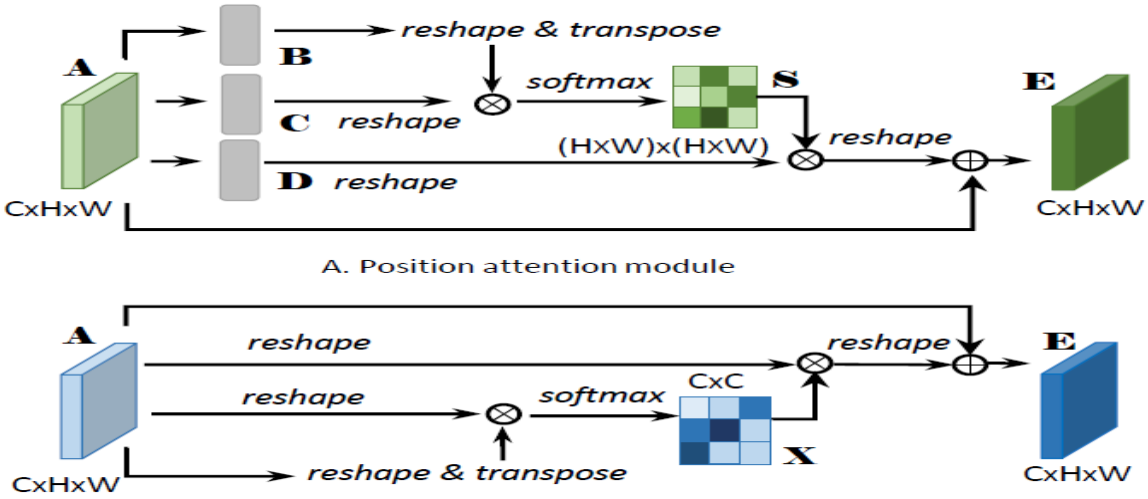

A
B
C
D
reshape & transpose
softmax
S
reshape
(HxW)x(HxW)
reshape
reshape
E
CxHxW
CxHxW
A. Position attention module
A
reshape
reshape
softmax
CxC
reshape
E
X
reshape & transpose
CxHxW
CxHxW

图 4

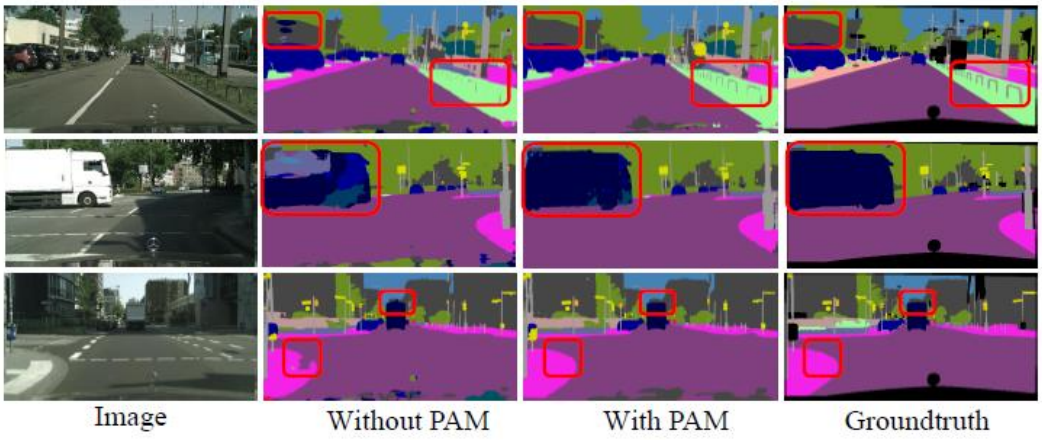


Image
Without PAM
With PAM
Groundtruth

图 5

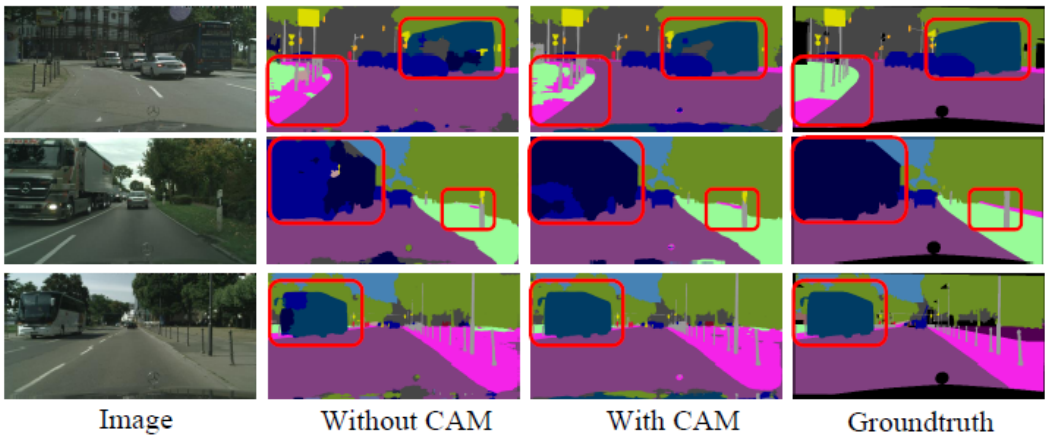


Image
Without CAM
With CAM
Groundtruth

图 6

References

[1] Byeon W, Breuel T M, Raue F, et al. Scene labeling with lstm recurrent neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3547-3555.

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. CoRR, abs/1706.05587, 2017 .

[3] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In IEEE Conference on Computer Vision and Pattern Recognition, pages 6230–6239, 2017.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[5] Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Ian D. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, pages 3194–3203, 2016.

[6] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130, 2017.

[7] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. Disan: Directional self-attention network for rnn/cnn-free language understanding. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[8] Jinhui Tang, Richang Hong, Shuicheng Yan, Tat-Seng Chua, Guo-Jun Qi, and Ramesh Jain. Image annotation by k nnsparse graph-based label propagation over noisily tagged web images. ACM Transactions on Intelligent Systems and Technology (TIST), 2(2):14, 2011.

[9] Jinhui Tang, Lu Jin, Zechao Li, and Shenghua Gao. Rgbd object recognition via incorporating latent data structure and prior knowledge. IEEE Transactions on Multimedia, 17(11):1899–1908, 2015.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems 30, pages 6000–6010, 2017.