用于语义分割的分支结构网络

雷璧闻，21821214，计算机科学与技术，1021989513@qq.com，13129903409

摘要：语义分割为了达到优秀的分割效果，往往需要有效的结合空间信息和空间信息，同时还要能够充分利用边缘细节信息。为了解决这一系列问题，我们提出了本文的网络结构。我们使用分支结构分别提取语义信息和空间信息，再进行融合。在网络中，我们尽可能的加入 attention 结构，用以提升网络学习能力，并且几乎不引入新的参数。与此同时，我们加入边缘检测作为辅助监督，从而提高网络对于边缘细节部分的分割效果。为了测试算法的有效性，我们在 ADE20k 数据集上进行训练与测试，最终达到 51.96%的逐像素准确率。

关键字

分支结构，边缘检测，attention 结构

1 引言

语义分割在是指为图像中每个像素进行分类，即判断像素属于哪种类别，目前广泛应用于无人驾驶、医疗图像等领域，是计算机视觉领域的重要分支。

在语义分割领域，常常会面临这样一个矛盾：高层次的卷积层能够带来大的感受野，即丰富的空间信息，但是此时的空间信息已经大量丢失，即该层的图像尺寸很小，无法有效提取细节信息；同理，在低层次的卷积层能够保留图像的尺寸，即有效学习到丰富的空间信息，但浅层的感受野较小，无法有效学习空间信息。

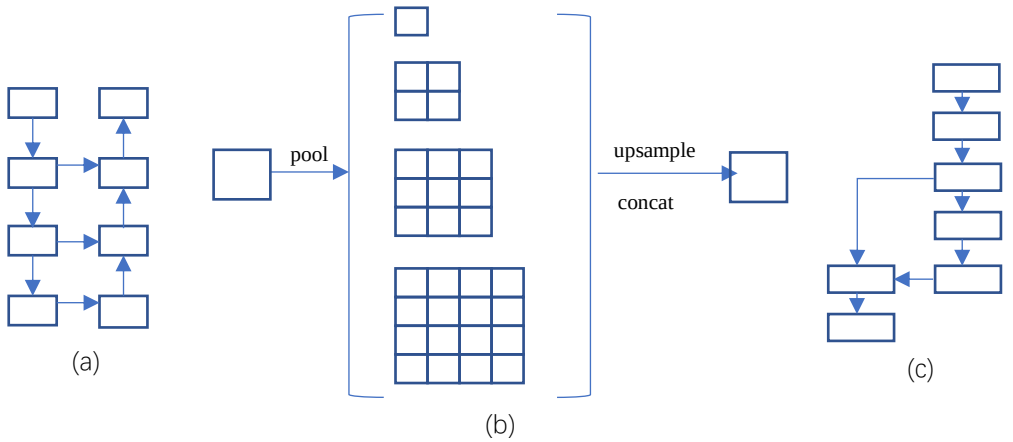

图 1 网络结构简图（a）U 型网络通用结构简图；（b）空间金字塔池化结构简图；（c）我们提出的网络结构简图

于是，若能同时利用高层次与低层次的信息能够有效提升语义分割效果。目前，能够有效同时提取深层空间信息与浅层空间信息的网络，主要以 U 型网络[1,2]为主，U 型网络通过直连结构，将浅层的空间信息逐步加入深层网络，使得网络在深层卷积中能够获得浅层细节信息，以提升分割效果，图 1（a）展示了 U 型网络结构；同时，空间金字塔池化结构[3,4]的引入，也能在一定程度上同时利用空间和空间信息，该结构对网络中某一层特征图进行多尺度池化，因而在该层能够获得不同大小感受野，在保留空间信息的同时，引入空间信息，也能明显提升分割效果，图 1（b）展示了空间金字塔池化结构。

但是，U 型网络中的堆叠（concat）操作造成网络参数量较大，为了将图像放入网络往往需要进行 resize，于是空间信息在一开始就已经损失，很难通过后期的网络学习进行修补，这也是造成 U 型网络对于边缘等细节部分的分割效果不佳的原因之一。空间金字塔池化尽管在分割效果有明显提升，但是这并不能说明，提高感受野大小等同于获取空间信息，也就是说，空间金字塔池化结构不能够明确的表示空间信息。

除此以外，在语义分割中，边缘部分的处理往往被忽视，许多方法[4,5,6]会在分割结果上进行 CRF 后处理，即概率学上对像素类别进行分配，将靠近的 RGB 值相近的像素分配为同一类别，从而提升边缘细节的分割效果。但是这种方法并不是对图像语义的精细分割，仅仅是从概率场上进行操作，具有很大的不稳定性。

于是，在这些问题的基础上，我们提出了多种策略用于解决这些问题。首先，为了能够同时有效利用深层空间信息和浅层空间信息，我们使用两条支路结构，一条专注于提取浅层空间信息，一条专注于提取深层空间信息，并进行融合，使得融合后的特征图同时包含空间信息和空间信息。除此之外，在提取特征的过程中加入 attention 结构，即对特征图进行全局池化从而对逐通道赋予权重，用以提升分割效果。最后，为了解决边缘细节的精细分割问题，我们加入辅助监督，专注于学习边缘检测，从而提升边缘分割效果。我们提出的网络结构简图形如图 1（c）。最终，我们的算法在 ADE20k 数据集上取得（），能够证明算法的有效性。

2 相关工作

空间和空间信息 在分割任务中，空间和空间信息往往处于难以两全其美的状态，有效提取这两种信息并融合往往能够提升算法效果。U 型网络中的代表有 U-Net[1]以及 Dense-UNet[2]，他们都是先进行下采样不断提取丰富的空间信息，同时对每次下采样得到的特征图进行保存，在后来的上采样过程中，为了恢复已经损失的空间信息，将之前保存的特征图与上采样的特征图进行堆叠（concat）操作，使得网络在学习到空间信息的基础上再次得到空间信息，从而达到空间信息空间信息的融合。PSPNet[3]中加入空间金字塔池化结构，即将特征图进行多个尺度的池化，相当于在该层中得到多个不同尺度的感受野，用于提取丰富的空间信息。DeeplabV2[4]中使用带孔金字塔池化结构，与上述空间金字塔结构的目的一样，都是为了获取不同大小的感受野用来提取空间信息，唯一的不同是引入了带孔卷积结构，这样能够进一步提升感受野大小。

Attention 模块 SE-Net[7]在网络中加入 Squeeze 和 Excitation 结构，在 Squeeze 结构中，首先对特征图进行全局池化以得到全局信息，每一个通道变成一个实数，这个实数某种程度上具有全局的感受野，并且输出的维度和输入的特征通道数相匹配，它表征着在特征通道上响应的全局分布。随后进入 Excitation 结构，该结构是一个类似于循环神经网络中门的机制。通过参数 w 来为每个特征通道生成权重，其中参数 w 被学习用来显式地建模特征通道间的相关性，最后是一个 Reweight 的操作，我们将 Excitation 的输出的权重看做是进过特征选择后的每个特征通道的重要性，然后通过乘法逐通道加权到先前的特征上，完成在通道维度上的对原始特征的重标定。

边缘检测 之前的边缘检测算法主要依赖于机器学习，包括使用 Sobel 算子、Laplacian 算子以及 Canny 算子。但是这些算法有一个明显缺点：仅依赖于像素值，或者说是图像梯度值进行边缘检测，而非根据图像的语义信息进行边缘检测，即简单地将梯度值变化大的部分作为边缘。HED[8]提出一种 end-to-end 的深度学习边缘检测网络，其效果远远好于基于机器学习的边缘检测方法，这里 HED 网络实质是一种语义分割模型，只是将网络的标签从填充 mask 变为边缘 mask，该文也变相证明分割网络在学习边缘检测时的有效性。部分算法受

此启发将边缘检测引入分割网络，作为辅助监督[9]，有效提升了算法分割效果，尤其是在边缘部分，提升了分割精度。

3 方法

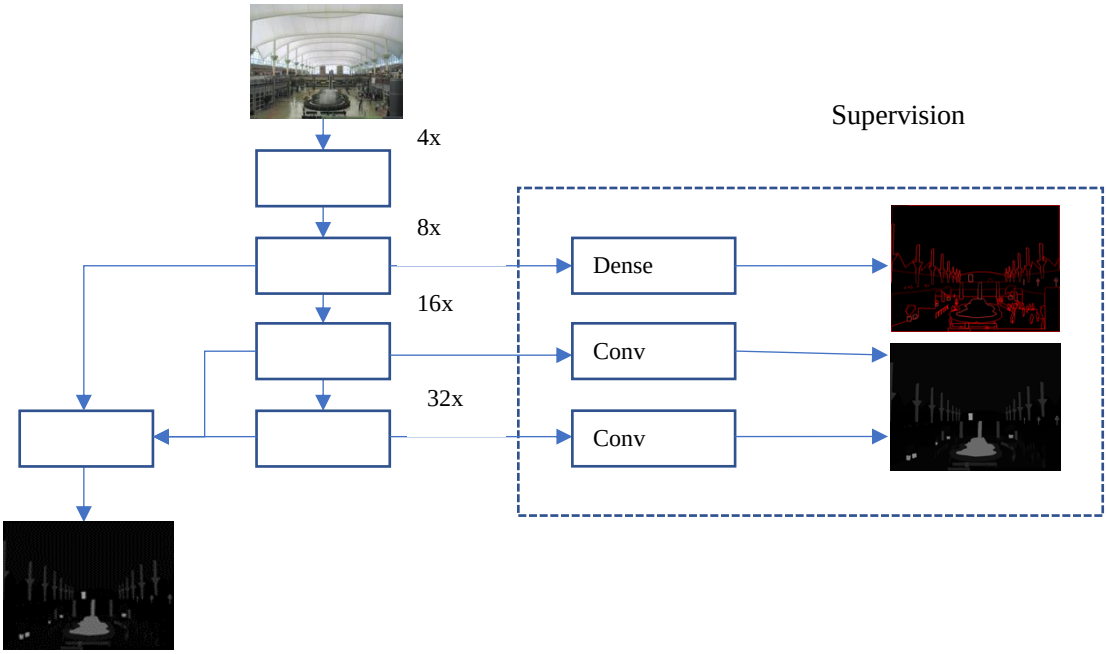

图 2 网络结构示意图。下采样主网络为预训练网络，将下采样 8 倍的特征图保存，经过 Dense Block，得到新的特征图用于边缘 mask 辅助监督，将下采样 16 倍、32 倍的特征图保存，经过一层卷积层输出新的特征图，用于填充 mask 辅助监督。

分支结构

为了取得较好的分割效果，多种网络尝试结合空间信息与空间信息[1,2,3,4]，其中 U 型网络以“补充”的形式，在深层特征图加入浅层空间信息；而空间金字塔结构则是通过取得不同大小的感受野，使得网络在保留浅层的信息的同时获得空间信息。这些结构的设计也从侧面说明了结合空间信息与空间信息的重要性。浅层空间信息往往注重细节的学习，而随着层次的加深，并且伴随着下采样过程，细节信息逐渐丢失，网络则会逐渐学到空间信息，即目标的位置等信息，两者的结合互补能够显著提升网络的分割效果。

基于以上分析，为了能够同时利用空间信息和空间信息，我们考虑将下采样过程中浅层网络提取的特征图进行保存，即保存空间信息。随着下采样的进行，网络不断加深，空间信息渐渐损失，随之而来的则是空间信息的逐渐丰富，因此我们考虑将这部分空间信息与之前保存的空间进行融合处理。在这里，我们选择一些预训练网络作为下采样主网络，在下采样 8 倍的时候对特征图进行保存，此时特征图包含了丰富的空间信息，提取了大量的细节特征信息，能够有效提升后续的精细分割效果。对于下采样 16 倍以及 32 倍的特征图，往往包含大量空间信息，此时的空间信息已经逐渐损失，于是我们对这一系列的深层次特征图进行相互融合，最大化的提取空间信息。之后，我们对融合后的特征图进行上采样操作，上采样后的空间信息特征图与空间信息特征图保持同样大小的尺寸，方便后续的融合操作。

Attention 结构

SE-Net[7]在网络中加入 Squeeze 和 Excitation 结构，用于对特征图的各通道进行重标定。Squeeze 操作就是在得到多个 feature map 之后采用全局平均池化操作对其每个 feature map 进行压缩，使其 C 个 feature map 最后变成 1*1*C 的实数数列。一般 CNN 中的每个通道学习到的滤波器都对局部感受野进行操作，因此每个 feature map 都无法利用其它 feature map 的上下文（空间）信息，而且网络较低的层次上其感受野尺寸都很小，这样情况就会更严重。多个 feature map 可以被解释为局部描述子的集合，这些描述子的统计信息对于整个图像来说是有表现力的。而 Squeeze 操作中选择最简单的全局平均池化操作，从而使其具有全局的感受野，使得网络低层也能利用全局信息。 之后的 Excitation 操作则是用来全面捕获通道依赖性，因此该结构需要满足两个条件：

(1) 它必须是灵活的（特别是它必须能够学习通道之间的非线性交互）；

(2) 它必须学习一个非互斥的关系，与 one-hot（独热）结构不同，这里允许对多个通道同时进行强调。

为了满足这些要求，Excitation 结构选择采用一个简单的门结构，即引入两层全连接层，用于更好的学习非线性关系，并且进行降维操作减少参数量，最后使用了 sigmoid 激活函数，将权重值固定在 0 到 1 之间。

在本文提出的网络结构中，我们为了提升网络的学习能力，引入 Attention 结构，也就是 SE-Net 中的 SE-block，这一结构除了有以上所述的特性之外，还拥有参数量小的优点，在提升分割效果的同时几乎可以不考虑引入的参数量。在我们的网络结构中，我们选择对下采样 8 倍、16 倍、32 倍的特征图额外引入新的分支，分别进行 attention 操作，再对 attention 的结果进行融合。具体地说，首先对保存的 8 倍下采样的特征图进行 attention 操作，随后对 32 倍下采样的特征图进行 attention 操作后，上采样至 16 倍，与进行 attention 操作后的 16 倍下采样特征图进行堆叠（concat）操作，用于融合深层空间信息，之后同时进行上采样至 8 倍，与 8 倍下采样图堆叠后再次进行 attention 操作，完成空间信息与空间信息的融合。特别地，我们对 16 倍、32 倍下采样特征图分别加入额外的辅助监督，以提升分割效果。

边缘检测

HED[8]提出的 end-to-end 边缘检测系统表明，分割网络可以用于图像边缘的学习，在该网络结构中使用图像到图像的学习过程，意味着网络是基于语义信息而得出的边缘分割，并非单纯的从像素梯度的角度出发来进行边缘检测，因而大幅度提升了边缘检测的效果。而部分算法[9]将边缘检测加入分割算法中作为额外的监督，用来提升分割效果，具体的，在该网络结构中，使用了两个 mask 作为真实 mask 标签，一个为填充 mask，一个为边缘 mask，最终有效提升了边缘细节的分割效果。受此启发，本文提出的算法使用边缘检测作为辅助监督，我们选择对 8 倍下采样的特征图进行边缘辅助监督，首先进行一系列卷积操作，将得到的边缘预测图与边缘 mask 进行辅助监督计算 loss。

Loss 函数

本网络结构中对所有的 loss 计算前首先对特征图进行 LogSoftmax 操作，如公式 1 所示。计算 loss 时，统一采用 NLLloss（负对数概率 loss）。总 loss 则是由一个主 loss 函数加上另外三个辅助监督 loss，两个辅助填充 mask 监督 loss，一个辅助边缘监督 loss，我们在计算总 loss 时，为三个辅助监督 loss 增加权重α_i，用以适应网络训练，在本文中α_1、α_2取 1，α_3取 0.5。

$$\mathrm{LogSoftmax}(x_i) = \log\left(\frac{\exp(x_i)}{\sum_j \exp(x_j)}\right) \quad (1)$$

其中x_i为像素值，j为通道数（分割种类数）。

$$\mathrm{Loss} = L_{main} + \alpha \sum_1^3 L_{aux} \quad (2)$$

α_i为三个辅助 loss 的权重，其中α_1、α_2为填充 mask 监督，α_3为边缘 mask 监督。

4 实验结果与分析

为了测试算法效果，我们在 ADE20k 数据上进行验证。首先我们对 ADE20k 数据集进行介绍，并介绍我们算法的实现细节。其次，我们会对主网络的选择进行测试。最后，我们会将本文提出的算法与经典分割网络进行对比，以证明算法的有效性。

数据集介绍

ADE 20k 该数据集包括 150 个语义分割类别，标签为像素级别的语义分割 mask，训练集共有 20210 张图片，验证集共有 2000 张图片。

实现细节

网络：我们选择 resnet18 作为下采样主网络，并作为 baseline，输出图片为 1/8 下采样的图片，我们人为的将结果 resize 到输入图片尺寸用于测试结果。

训练细节：我们选择 RMSprop 优化器，batch size 选择 1。学习率初始设置为 1e-3，使用“poly”学习率训练策略，随着训练逐渐降低，以减少训练后期的损失函数值不稳定性。

数据增强：在训练过程中，我们对图片进行归一化操作，并进行随机的左右翻转,为方便处理，将图片 resize 至固定尺寸，这里我们取 512×512。

主网络测试

在该部分，我们测试不同的预训练下采样主网络对于分割效果的影响，下采样特征提取过程是该网络的主要学习部分，因而对主网络进行多网络测试极为重要，能够利于我们选出学习能力最强，并且最适合于我们的网络结构的主网络。测试主网络包括：ResNet18，ResNet50，ResNet101，SE-Net。

表 1. 主网络效果测试。评测指标为逐像素正确率

主网络	Acc(%)
ResNet-18	0.4893
ResNet-101	0.5196

算法比较

为了验证本算法的效果，与主流的分割网络 U-Net 进行比较。

表 2. 分割效果比较。评测指标为逐像素正确率

网络	Acc
UNet	0.4973
Ours(without edge detection)	0.5035

Ours	0.5196

5 结论

本文提出的网络结构同时结合浅层空间信息与深层空间信息，显著提升网络的语义分割效果，同时加入 attention 机制，学习到通道间的全局语义信息，除此之外，辅助边缘检测 loss 的引入，提升了网络的边缘细节分割效果。我们在 ADE20k 数据集上取得了 51.96%的逐像素正确率，也证明了我们提出的网络结构的优越性。

参考文献

[1] Ronneberger O , Fischer P , Brox T . U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. 2015.

[2] Simon J , Michal D , David V , Adriana R , Yoshua Bengio . The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation[J]. 2016.

[3] Simon J , Michal D , David V , Adriana R , Yoshua Bengio . The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation[J]. 2016.

[4] Chen L C , Papandreou G , Kokkinos I , et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 40(4):834-848.

[5] Long J , Shelhamer E , Darrell T . Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014.

[6] Chen L C , Papandreou G , Kokkinos I , et al. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs[J]. Computer Science, 2014(4):357-361.

[7] Hu J, Shen L, Albanie S, et al. Squeeze-and-Excitation Networks[J]. 2017.

[8] Xie S, Tu Z. Holistically-Nested Edge Detection[J]. International Journal of Computer Vision, 2015, 125(1-3):3-18.

[9] Roth H R , Lu L , Farag A , et al. Spatial Aggregation of Holistically-Nested Networks for Automated Pancreas Segmentation[C]// International Conference on Medical Image Computing & Computer-assisted Intervention. Springer, Cham, 2016.