Context Encoding for Semantic Segmentation

骆泽河

摘要

近期的工作在通过使用扩张卷积、利用多尺度特征和细化边界来提高全卷积网络（FCN）框架像素标记的空间分辨率方面取得了重大进展。在本文中，我们通过引入上下文编码模块来探讨全局上下文信息在语义分割中的影响，该模块捕获场景的语义上下文并选择性地突出类相关特征图。与 FCN 相比，所提出的上下文编码模块仅具有边际额外计算成本，显著提高了语义分割结果。我们的方法在 PASCAL VOC 2012 中获得了 85.9% mIoU。

关键词：上下文编码；语义分割

1 引言

语义分割为给定图像分配对象类别的每像素预测，这提供了包括对象类别、位置和形状信息的全面场景描述。最先进的语义分割方法通常基于完全卷积网络（FCN）框架[1]。深度卷积神经网络（CNNs）[2]受益于从不同的图像集合中学习到的对象类别和场景语义的丰富信息[3]。神经网络能够通过叠加具有非线性和下采样的卷积层来捕获具有全局感受野的信息表示。为了克服与下采样相关的空间分辨率损失问题，最近的工作使用扩张/萎缩卷积策略从预训练网络产生密集预测[4-5]。然而，该策略还将像素与全局场景上下文隔离，导致像素分类错误。例如，在图 1 的第三行中 FCN 方法将窗格中的一些像素分类为门。

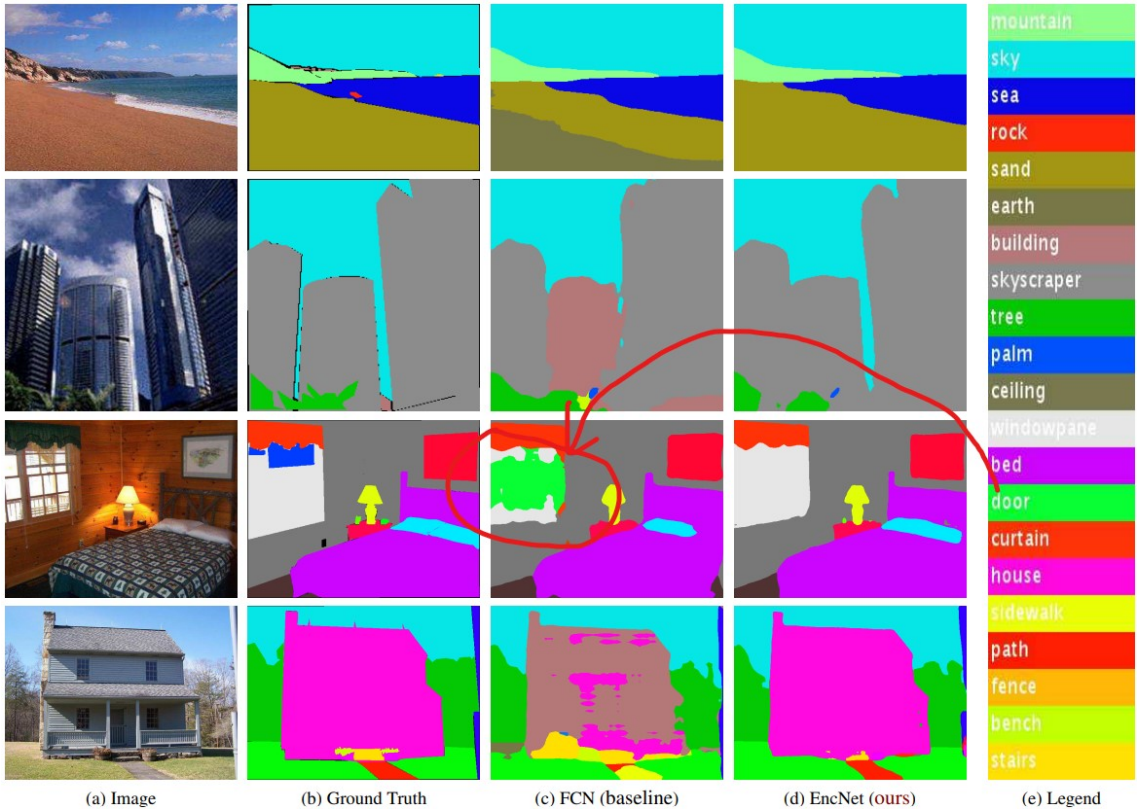

图 1: 各种方法的效果图

近期的工作使用基于金字塔多分辨率表示扩大接受野。例如，PSPNet 采用的 PSP 模块将特征图池化为不同尺寸，再做联接上采样[6]；DeepLab 采用 ASPP 模块并行的使用大扩张率卷积扩大接受野[7]。

这些方法都有提升，但是这对上下文表示都不够明确，这出现了一个问题：捕获上下文信息是否等同于增加接受野大小？

如图 2 所示，如果我们能够先捕获到图像上下文信息 (例如这是卧室)，然后，这可以提供许多相关小型目标的信息 (例如卧室里面有床、椅子等)。这可以动态的减少搜索区域可能。说白了，这就是加入一个场景的先验知识进去，这样对图片中像素分类更有目的性。依照这个思路，可以设计一种方法，充分利用场景上下文和存在类别概率的之间的强相关性，这样语义分割会就容易很多。

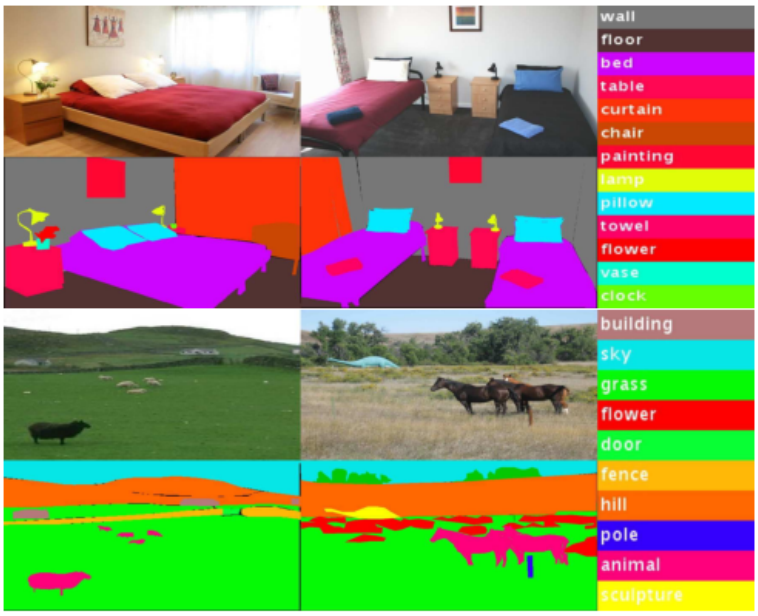

图 2: 各个场景效果图

2 相关工作

经典的计算机视觉方法具有捕获场景上下文语义的优点。例如 SIFT[8]提取密集特征或滤波器组响应密集提取图像特征。学习一个视觉字典，BoW[9-12]，VLAD[13]和 Fish Vector[14]通过类别编码描述特征统计信息。经典表示通过捕获特征统计信息编码全局信息，虽然手工提取特征通过 CNN 方法得到了很大的改进，但传统方法的总体编码过程更为方便和强大。在本文中，使用扩展编码层用于捕获全局特征的统计信息用于理解上下文语义。

2.1 上下文编码模块

引入了上下文编码模块，该单元用于捕获全局场景上下文信息和选择性的突出于类别相关的特征图。

集成了语义编码损失 (Semantic Encoding Loss,SE-loss)。例如，我们不考虑车辆出现在卧室的可能性，在现有标准的训练过程使用的是像素分割损失，这不强调场景的全局信息。我们引入语义编码损失 (SE-loss) 可进一步规范网络训练，让网络预测能够预测场景中对象类别的存在，强化网络学习上下文语义。与逐像素的损失不同，SE-Loss 对于大小不同的物体有相同的贡献，在实践中这能够改善识别小物体的表现，这里提出的上下文编码模块和语义编码损失在概念上是直接的并且和现存的 FCN 方法是兼容的。

2.2 EncNet 语义分割框架

设计了一个新的语义分割架构 Context Encoding Network (EncNet)。如图 3 所示，EncNet 通过上下文编码模块增强了预训练的 ResNet[15]。

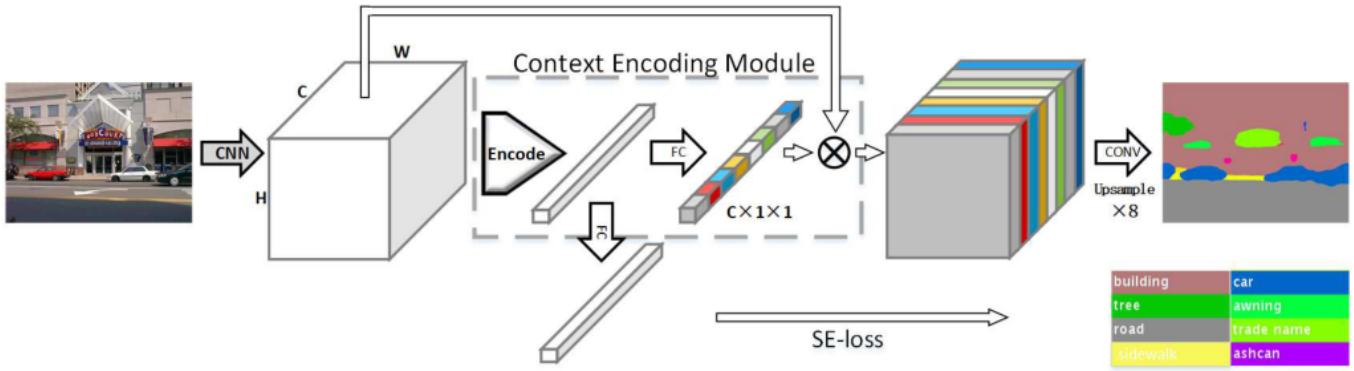

图 3: 上下文编码网络

3 本文方法

此部分对本文将要复现的工作进行概述，使用的网络如图 3 所示。

3.1 Context Encoding

编码层认为一个 shape 是 C×H×W 的输入特征图为一组 C 维度的输入特征 X=x1,..,xN，其中 N 是特征的总个数即 H×W，其学习固有的 codebook D=d1,...,dK 包含 K 个 codewords，和一组视觉中心平滑因子 S=s1,...,sK。编码层输出残差编码，这是通过聚合具有 soft-assignment 权重$ek = \sum_{i=1}^{N} eik$

其中，$eik = \frac{exp(-sk||rik||^2)}{\sum_{j=1}^{K} exp(-sj||rij||^2)} rik$

通过 rik=xi−dk 给定残差，我们在编码器上使用聚合而不是联接。

3.2 Featuremap Attention

为了使用编码层捕获的编码语义，我们预测一组特征图的放缩因子作为循环用于突出需要强调的类别。在编码层端上使用 FC 层，使用 sigmoid 作为激活函数，预测特征图的放缩因子γ=δ(We)，其中 W 表示层的权重，δ 表示 sigmoid 激活函数。模块通过 Y=X×γ 得到输出，每个通道在特征图 X 和放缩因子γ 之间做逐像素相乘。这样的方法受 SE-Net[16]等工作的启发，即考虑强调天空出现飞机，不强调出现车辆的可能性。

3.3 Semantic Encoding Loss

标准的语义分割训练过程，使用的是逐像素的交叉熵，这将像素独立开学习。这样网络在没有全局上下文情况下可能会难以理解上下文，为了规范上下文编码模块的训练过程，使用 Semantic Encoding Loss (SE-loss) 在添加少量额外计算消耗的情况下强制网络理解全局语义信息。

在编码层之上添加了一个带 Sigmoid 激活的 FC 层用于单独预测场景中出现的目标类别，并学习二进制交叉熵损失。不同于逐像素损失，SE loss 对于大小不同的目标有相同的贡献，这能够提升小目标的检测性能。

3.4 Context Encoding Network (EncNet)

为了进一步的提升和规范上下文编码模块的训练，使用了单独的分离分支用于最小化 SE-loss，该 Loss 采用已编码的语义作为输入并预测对象类别的存在。因为上下文模块和 SE-loss 是轻量级的，论文在 stage3 上端添加另一个上下文编码模块用于最小化 SE-loss 作为额外的正则化，这类比于 PSPNet[6]的辅助分支但比那个轻量了许多。

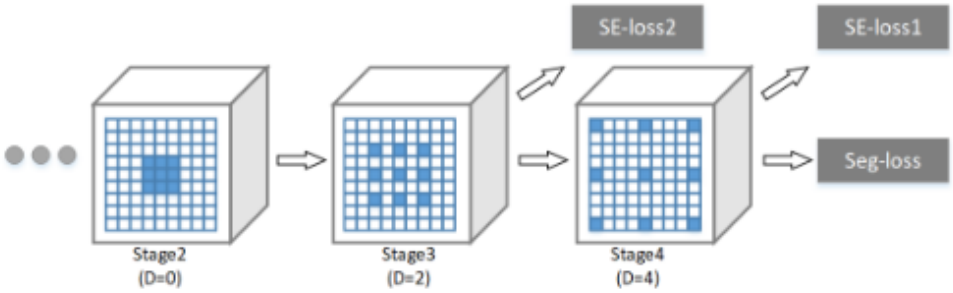

图 4: 扩张策略和损失

4 复现细节

4.1 与已有开源代码对比

复现过程中参考了作者上传到 GitHub 的源代码，根据作者提供的源代码实现了网络，编码模块以及测试，所使用的测试集是 Camvid 测试集。

4.2 创新点

本次复现只是把原文中提到的功能实现了，以及在测试集得到跟原文一样的结果，并没有进行创新。

5 实验结果分析

本次复现使用的数据集是 Camvid 数据集，在该数据集上进行测试，得到的训练集和测试集的训练结果如图 5 所示。本次复现在测试集达到的准确率为 85.4%，与原文 85.9% 相接近。

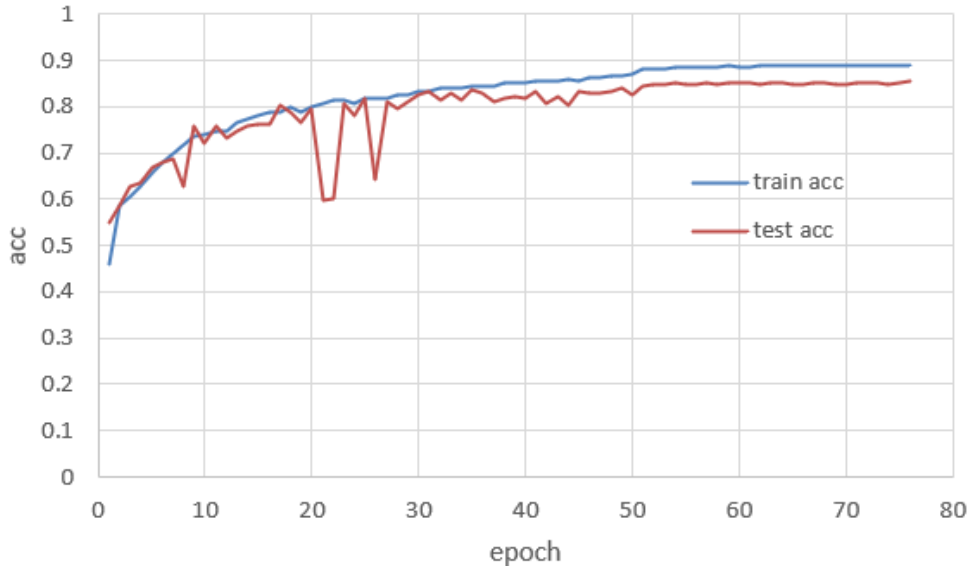

图 5: 实验结果

对图片进行处理后的结果如图 6 所示。

图 6: 处理后的结果

6 总结与展望

引入了上下文编码模块，该模块选择性地突出了类相关特征图，简化了网络的问题。与现有的基于 FCN 的方法兼容。并且实验结果证明了所提出的 EncNet 的优越性能。实现过程只把文章所提出的方法实现了，没有在此基础上进行改进和创新。

参考文献

[1] J. LONG E S, DARRELL T. Fully convolutional networks for semantic segmentation[J]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 1, 4, 6, 7: 3431-3440.

[2] Y.LECUN Y, L.Bottou, P.Haffner. Gradientbased learning applied to document recognition[J]. Proceedings of the IEEE, 1998. 1, 86(11): 2278-2324.

[3] J.DENG R S, W.Dong, FEI-FEI L. ImageNet: A Large-Scale Hierarchical Image Database[J]. In CVPR09, 2009. 1, 2.

[4] L.-C. CHEN G P I K K M, UILLE A L Y. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. In ICLR, 2015. 1, 2, 4, 5, 7.

[5] U F Y, KOLTUN V . Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint, 2015. 1, 2, 4, 5, 7: arXiv:1511.07122.

[6] H. ZHAO X Q, J. Shi, WANG X. Pyramid scene parsing network[J]. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2,4, 5, 7.

[7] L.-C. CHEN I K, G. Papandreou, MURPHY K. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]., 2016. 2, 4, 5, 6, 7: 1606.00915.

[8] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004. 2, 60(2): 91-110.

[9] J. SIVIC A A E A Z, B. C. Russell, FREEMAN W T. Discovering objects and their location in images [J]. In Tenth IEEE International Conference on Computer Vision (ICCV’05), IEEE, 2005. 2, 1: 370-377.

[10] JOACHIMS T. Text categorization with support vector machines: Learning with many relevant features [J]. In European conference on machine learning, Springer, 1998. 2: 137-142.

[11] G. CSURKA L F J W, C. Dance, BRAY C. Visual categorization with bags of keypoints[J]. In Workshop on statistical learning in computer vision, ECCV, Prague,2004. 2, 1: 1-2.

[12] FEI-FEI L, PERONA P . A bayesian hierarchical model for learning natural scene categories[J]. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), IEEE, 2005. 2, 2: 524-531.

[13] H. JÉGOU C S, M. Douze, PÉREZ P . Aggregating local descriptors into a compact image representation [J]. In Computer Vision and Pattern Recognition (CVPR), 2010, 2010. 2: 3304-3311.

[14] F. PERRONNIN J S, MENSINK T. Improving the fisher kernel for large-scale image classification[J]. In European conference on computer vision, 2010. 2: 143-156.

[15] K. HE S R, X. Zhang, SUN J. Deep residual learning for image recognition[J]. arXiv preprint, 2015. 2, 4, 8: arXiv:1512.03385.

[16] J. HU L S, SUN G. Squeeze-and-excitation networks[J]. arXiv preprint, 2017. 3, 4, 8: arXiv:1709.01507.