南 京 理 工 大 学

毕业设计(论文)开题报告

|  |  |  |  |
| --- | --- | --- | --- |
| **学 生 姓 名：** | 李明龙 | **学 号：** | 914106840521 |
| **专 业：** | 计算机科学与技术 | | |
| **设计(论文)题目**： | 基于深度学习的图像语义分割 | | |
| **指 导 教 师:** | 李泽超 | | |

2018年1月11日

1. **结合毕业设计（论文）课题情况，根据所查阅的文献资料，每人撰写 2000字左右的文献综述。**

**文献综述：**

**基于深度学习的语义分割方法文献综述**

**摘要：**

图像语义分割作为计算机视觉的关键任务，不仅能够判别图像中目标的类别信息，还能够准确地对其进行定位并描绘出其边缘，现已在学术界受到广泛关注，并在很多领域展现出巨大的应用前景。本文对基于深度学习的语义分割方法进行了综述：首先对语义分割领域的经典深度学习模型进行了总结，然后对现有模型存在的问题进行分析，并对未来的研究方向进行了展望。

**关键词：**

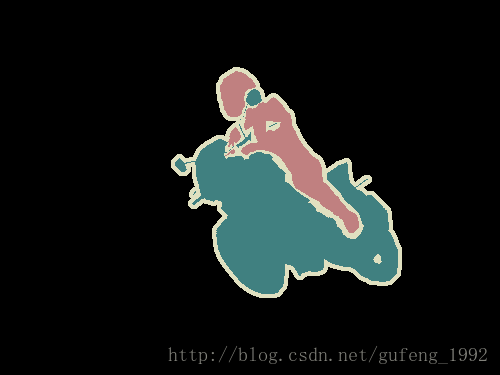
深度学习；语义分割；神经网络；监督学习

**1 简介**

在计算机视觉领域，语义分割扮演着越来越重要的角色。图像语义分割就是机器自动从图像中分割出对象区域，并识别其中的内容。语义分割方法在处理图像时，具体到像素级别，也就是说，该方法会将图像中每个像素分配到某个对象类别。例如对于图(1)，该模型不仅要识别出摩托车和驾驶者，还要标出每个对象的边界，得到图(2)的分割效果。因此，与分类目的不同，相关模型要具有像素级的密集预测能力。



图(1)



图(2)

越来越多的应用场景需要精确且高效的分割技术，如自动驾驶、室内导航、医学图像处理、图像搜索引擎、甚至虚拟现实与增强现实等。在自动驾驶领域，需要首先对输入图像中的道路、行人、车辆等进行分割识别，进而为车辆的行驶提供指导。在医学图像处理领域，则需要首先将病灶区分割出来，才能对病灶进行量化分析。鉴于语义分割的重要性，对语义分割算法的研究具有重要的意义。

语义分割问题还可以进一步发展细化，比如实例分割（即对同一类别的不同实例标以不同的标签，比如一号椅子，二号椅子），或者是基于部分的分割（即对已经分出不同类别的图像进行底层分解，找到每个类对应的组成成分，比如将人像分为头部，躯干，四肢）。

语义分割问题可以使用如下公式进行初略定义：对于随机变量集合X = {x1,x2,...,xN}中的每个随机变量，找到一种方法为其指派一个来自标签空间L = {l1,l2,...,lk} 中的标签。每个标签li表示唯一的一个类别，比如飞机、汽车、交通标志或背景等。如果这个标签空间有k个可能的状态，通常会将背景类加入其中将其扩展为 k+1个[1]。

语义分割问题可以采用传统的计算机视觉和机器学习方法进行解决。在深度学习应用到计算机视觉领域之前，研究人员一般使用纹理基元森林(TextonForest)或是随机森林(Random Forest)方法来构建用于语义分割的分类器。虽然这些方法很流行，但深度学习革命让相关领域发生了翻天覆地的变化，因此包括语义分割在内的许多计算机视觉问题都开始使用深度学习模型。深度学习应用于这一领域之后在准确率以及效率上也都远远超过了传统方法。

**2 方法概述**

现有的语义分割方法主要是将应用于图像分类的经典网络作为基网络，根据语义分割任务的特点对基网络进行改进。因此，本文首先简单介绍几种常用的基网络，这些基网络都是在卷积神经网络CNN[2],[3],[4],[5],[6]的基础上演化而来的。CNN的基本结构主要包括卷积层、池化层、全连接层。

**2.1 通用的深度神经网络概述**

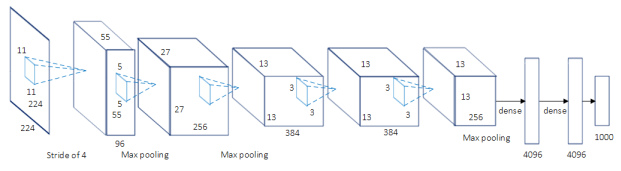
某些深度神经网络已成为众所周知的领域标准，这包括AlexNet，VGG，GoogLeNet，以及ResNet。卷积层、池化层、全连接层等是这些网络共有的网络层。

卷积层是CNN首创的结构，它用来对前一层输入图像进行卷积运算，通过卷积核在图像上的移动使得CNN能够“看到”局部图形而非一个点。卷积层中使用了局部感受野和权值共享两种技术有效降低了网络训练参数，通过对图像中的局部区域进行卷积运算学习到图像的局部细节。所谓的局部感受野指的是同卷积核做卷积运算的区域只是图像的一部分，权值共享指的是不同区域进行卷积运算时使用的卷积核是相同的。

池化层又称为亚采样层，本质上是对输入图像进行压缩，例如3\*3的池化操作会在输入图像每一个3\*3的小区域中通过某种标准选择一个像素保留，其他全部舍弃作为压缩结果。最常见的池化操作有平均池化和最大池化。通过亚采样可以在压缩图像大小的同时维持图像的旋转不变性。

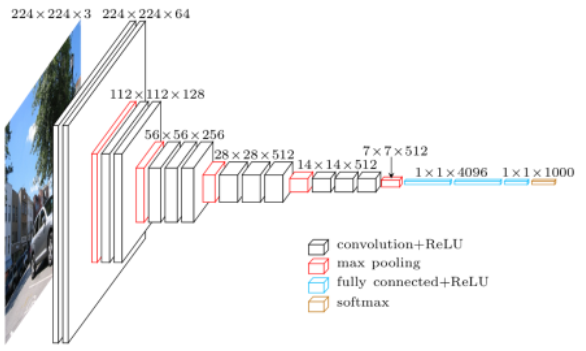
经过几轮卷积层-池化层的迭代，即一边理解局部信息一边压缩图像大小降低图像分辨率的过程，CNN得以一层一层逐步扩大一次性的理解范围，比如一开始只能识别出线条，随着层数的加深，逐渐能够识别出眼睛，鼻子，甚至可以从中总结出脸部的信息。卷积层+池化层处理过后，最后通常会接一个全连接层进行分类或者回归。全连接层中每一个结点都与上一层的所有结点相连，用于将原来二维矩阵压缩至一维，输出一个n\*1的向量，其中n指的是事先定义的标签的类别。

AlexNet[7]首创了深度卷积神经网络模型，在2012年ILSVRC竞赛中以超出第二名十个百分点(84.6%)的准确率优势取得冠军。AlexNet的非线性层包括卷积层、最大池化层、ReLU层各五层，此外还有全连接层三层以及dropout层。其中ReLU是一种比Sigmod更好的激活函数；dropout层的功能是按照一定的概率将神经网络单元从网络中舍弃，用于防止过拟合。图(3)给出了AlexNet的结构。



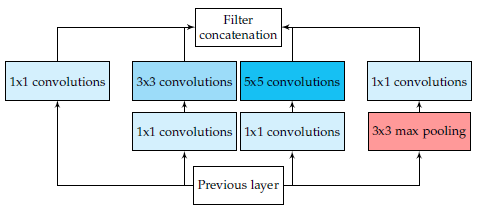
图(3)

VGG[8]由牛津大学的研究者提出。VGG网络共有16层，因此也被称为VGG-16。该模型通过减小卷积核大小、增加卷积核数目的方式减少了网络参数。这使得网络模型的参数更少，非线性性更强，也因此使得决策函数更具区分度，模型更容易被训练。VGG-16以92.7%的准确率获得了2013年ILSVRC的冠军。图(4)给出了VGG的网络结构。



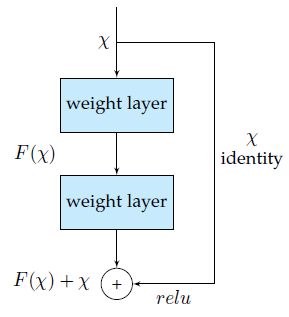
图（4）

GoogLeNet[9]以其复杂程度著称，该网络有22层，并引入了inception模块。Inception可以看作是一个小网络，其第一层是由两个卷积层和一个池化层组成，第二层由四个卷积层组成。该结构可以利用不同尺寸的卷积核学习不同尺度的图像信息，在增加网络宽度的同时减少了参数及操作数量，从而提高了网络的性能。GoogLeNet以93.3%的准确率获得了2014年ILSVRC的冠军。图(5)给出了其inception模块的结构图。



图(5)

微软提出的ResNet[10]将网络深度拓展到152层，并引入了残差模块。残差模块解决了训练真正深层网络时存在的问题，通过引入跳跃式连接使得低层与高层之间直接建立连接，保证了下一层可以从输入中学到与已有知识不同的新知识，同时协助解决了梯度消失的问题。ResNet以96.4%的准确率获得了2016年ILSVRC的冠军。图(6)给出了其残差模块的结构图。



图(6)

**2.2 基于深度学习的语义分割方法**

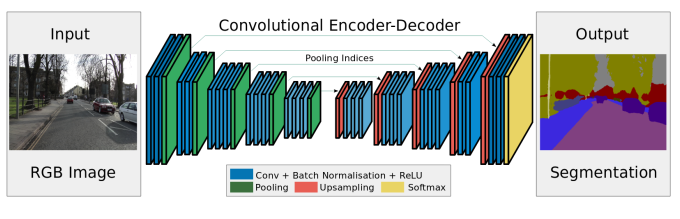
DL在语义分割领域最初的应用是基于CNN的图像块分类。该方法中为了对一个像素进行归类，需要使用该像素周围的一个图像块作为CNN的输入。可想而知这样做会有许多重复计算，效率很低，而且由于全连接层的存在对输入图像块的大小有着严格的限制。2014年问世的FCN[11]将CNN的全连接层替换为卷积层，输出空间映射而非分类分数，与CNN结构相比不仅提高了效率而且可以支持任意大小的图像输入，是一个开创性的工作，奠定了后续研究的基础。尽管如此，FCN中依然保留了池化层，而下采样操作之后很难进行精确上采样。为了解决这个问题，研究者们提出了两种思路。

**2.2.1 基于解码的方法**

该思想将神经网络视为编码器和解码器的连接体。编码器位于网络前端，用于产生低分辨率的图像表示或者特征映射；解码器位于编码器之后，用于将这些低分辨率的图像做一个像素级别的分类预测。一个典型的例子是Segnet[12]。Segnet的编码器部分由许多卷积层-池化层组成。不同于一般网络的池化层，Segnet中的池化层能够记录池化后的结果在原图像中的空间位置。

Segnet的解码器部分则由许多上采样层及卷积层组成，用于对编码器产生的特征图进行还原和理解。解码器中每个上采样层对应于编码器中的一个最大池化层，于是在上采样恢复图像尺寸时能够根据对应池化层的记录值精准地还原到原始位置，提高了上采样精确性。这些上采样得来的映射接下来将由不同的卷积核进行卷积，从而产生密集的特征映射。当特征映射被修复为与原输入相同分辨率的时候，使用一个softmax分类器用来预测像素级别的标签，得到最终的分割结果。

图(7)展示了Segnet架构示意图。

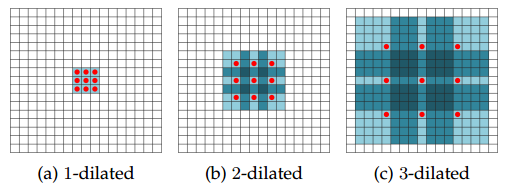


图(7)

**2.2.2 基于空洞卷积的方法[13]**

众所周知，引入池化层的一个目的是扩大感受野，得益于此不同卷积层的每个神经元可以逐层看到更大范围的图像。空洞卷积通过在卷积运算中引入空洞，没有改变卷积核大小，而是通过增大每个卷积核单元感受野大小的方式增大与卷积核做卷积运算的图像片段的大小，实现整体扩大感受野的目的。空洞卷积之后不再需要紧跟一个池化层，可以在不降低分辨率同时增大感受野，使得卷积输出包含较大范围的信息。

空洞卷积示意图如下：

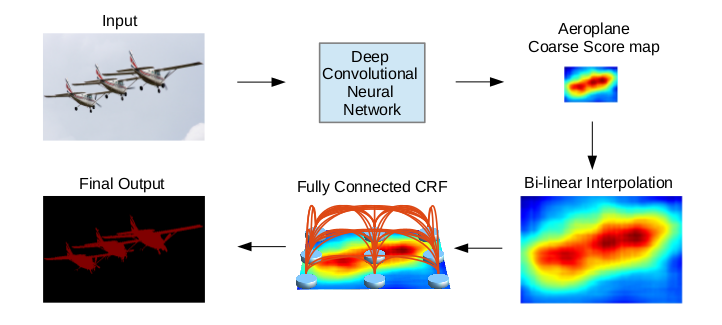


图(8)

**2.3深度学习与传统方法的结合**

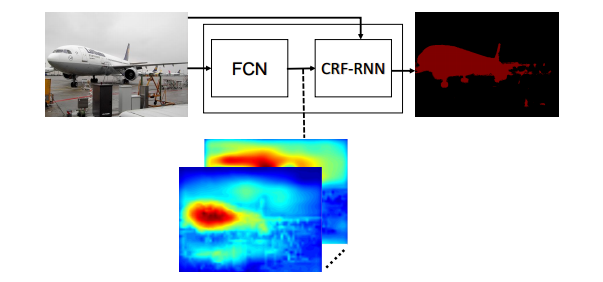
FCN开辟了卷积网络用于像素级语义分割的新思路，但其输出的分割结果的边界比较粗糙。在语义分割中，我们可以利用上下文信息对最后的分割结果进行细节增强，例如对物体边界进行更加清晰的刻画。条件随机场CRF[14][15]是一种常用的方法。CNN擅长关注局部细节，CRF则能够更为有效地学习到像素之间的联系，即使这两个像素相距较远。

FCN的问世为CNN与CRF结合提供了必要的技术基础。在利用CRF方法的一类网络中，最为经典的是DeepLab[17][18]系列方法。DeepLab使用了两两之间全连接的CRF模型作为其流程中的一个独立的后处理步骤，以此对分割结果进行调优。该模型将每个像素建模为某区域内的一个节点，无论两个像素距离多远，其两两之间的关系都会被衡量，因此，本模型也被称为密集或全连接因子图。DeepLab模型中考虑到了无论短期还是长期的像素联系，因此对细节信息更加敏感。此外，尽管全连接模型通常是低效的，DeepLab模型由于可以用概率推理来近似，所以可以达到相对的高效。



图(9)DeepLab结构示意图

DeepLab使用CRF来改进FCN的结果，但是CRF只作为一个后处理工具，不参与训练，所以未能发挥出CRF的全部威力。考虑到相邻像素属于同一类别的概率应该更高，Zheng等人提出了CRFasRNN[16]。CRFasRNN将FCN和CRF融合为一个网络，同时训练。作者巧妙地将CRF的求解步骤设计成了一个RNN，从而在一个迭代中能够求出CRF的解。即通过展开均值场推理的各个步骤，并将其视为RNN结构，该工作成功地将CRF与RNN整合成为一个完整高效的端对端网络。



图(10) CRF as RNN结构示意图

**3 总结与展望**

根据前人实验数据，二维图像分割领域中，DeepLab在几乎每个RGB图像数据集上都表现出最好的性能，被认为是目前最可靠的方法。深度学习很大程度上提高了语义分割的精度，然而仍存在一些开放的问题：

（1）三维数据的有效分割：由于大规模三维语义分割数据集的缺失，现有的算法大部分仅针对二维图像。在真实世界中存在大量的三维数据，2018年ILSVRC也将会创建三维数据竞赛，因此，如何研究有效的深度学习算法使其能够较好的分割三维数据是未来的一个发展方向。

（2）时间复杂度的降低，即实时分割：现有的方法虽然在分割精度上取得了较大的进展，然而模型具有较高的时间复杂度。因此，如何降低模型的复杂度，在准确率与运行时间之间寻求一个平衡，将是未来的另一个研究热点。

（3）空间复杂度的降低：分割网络一般需要较大的存储空间，从而可以同时进行推理与训练。不少设备存储空间有限，因此如何使得网络在保留多数信息的同时变得轻量化，是另一个值得探讨的问题。

**参考文献：**

[1] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, Jose Garcia-Rodriguez. A Review on Deep Learning Techniques Applied to Semantic Segmentation. arXiv:1704.06857, 2017.

[2] F. Ning, D. Delhomme, Y. LeCun, F. Piano, L. Bottou, and P. E. Barbano, “Toward automatic phenotyping of developing embryos from videos,” IEEETransactionsonImageProcessing, vol. 14, no. 9, pp. 1360–1371, 2005.

[3] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, “Deepneuralnetworkssegmentneuronalmembranesinelectron microscopy images,” in Advances in neural information processing systems, 2012, pp. 2843–2851.

[4] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1915– 1929, 2013.

[5] B. Hariharan, P. Arbel´aez, R. Girshick, and J. Malik, “Simultaneous detection and segmentation,” in European Conference on Computer Vision. Springer, 2014, pp. 297–312.

[6] S. Gupta, R. Girshick, P. Arbel´aez, and J. Malik, “Learning rich features from rgb-d images for object detection and segmentation,” in European Conference on Computer Vision. Springer, 2014, pp. 345–360.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classiﬁcation with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.

[8] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[11] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.

[12] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” arXiv preprint arXiv:1511.00561, 2015.

[13] S. Zhou, J.-N. Wu, Y. Wu, and X. Zhou, “Exploiting local structures with the kronecker layer in convolutional networks,” arXiv preprint arXiv:1512.09194, 2015.

[14] C. Rother, V. Kolmogorov, and A. Blake, “Grabcut: Interactive foreground extraction using iterated graph cuts,” in ACM transactionsongraphics(TOG), vol. 23, no. 3. ACM, 2004, pp. 309–314.

[15] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context,” International Journal of Computer Vision, vol. 81, no. 1, pp. 2–23, 2009.

[16] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, “Conditional random ﬁelds as recurrentneuralnetworks,”inProceedingsoftheIEEEInternational Conference on Computer Vision, 2015, pp. 1529–1537.

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” arXiv preprint arXiv:1412.7062, 2014.

[18] ——, “Deeplab: Semantic image segmentation with deep convolutionalnets, atrousconvolution,andfully connectedcrfs,”arXiv preprint arXiv:1606.00915, 2016.

**二．本课题要研究或解决的问题和拟采用的研究手段（途径）**

**2.1 要研究或解决的问题**

1. 掌握利用深度学习算法进行图像语义分割的原理，具体模型包括CNN,FCN,Deeplab,CRFasRNN,AlexNet,VGG,GoogLeNet,以及ResNet等；
2. 利用现有深度学习开源架构实现DeepLab，并提交至知名数据集观察分割效果；
3. 尝试对DeepLab模型进行时间复杂度优化。

**2.2 研究手段**

1. 细读参考文献中的论文，加深对每一个神经网络模型的认识；
2. 通过书籍或者官方文档学习现有深度学习架构caffe/tensorflow/theano/matconvnet/keras等的原理和使用方法，根据特性选择一款用于本次毕业设计；
3. 通过开源代码学习神经网络中的每一层（卷积层，池化层，全连接层）具体是如何实现的，然后配置开发环境，先实现一个结构简单的神经网络；
4. 通过博客、书籍和请教老师的方式深入理解条件随机场CRF的细节；
5. 逐步实现DeepLab，提交至VOC2012和MSCOCO数据集观察结果；
6. 学习应用于其他问题的深度神经网络架构，从中汲取灵感，并在语义分割问题中进行尝试调优，以VOC2012和MSCOCO数据集为反馈手段逐步优化现有框架；
7. 及时记录研究进展，早日开始最终论文的写作。

**2.3 课题整体（或软硬件系统）解决方案及其研究思路**

1、运用所学理论和技术手段对基于深度学习的语义分割系统进行分析、设计和开发，通过知名数据集进行准确率测试，达到较高准确率后尝试对南京市地铁监控图像、自然风光图像等现实数据进行语义分割。

2、评价课题提出的基于深度学习的语义分割系统解决方案和对社会、使用人员健康、安全、法律以及文化的影响，理解和评价课题设计的实践程序对环境、社会可持续发展的影响，并分析应用系统的各种角色应承担的责任。

3、与业界同行针对基于深度学习的语义分割系统设计过程中遇到的复杂工程问题进行有效沟通和交流，撰写阶段性研发报告和设计文稿，清晰地表达需求分析结论并正确回应用户指令。

4、定期沟通，不断完善模型和优化算法，提高系统的实用性和模型的高效性。

将对本课题所涉及的问题进行系统分析，按照任务书时间节点按时完成、定期与指导老师沟通，遵守学术规范，并考虑该课题在实际中的应用需求，提高算法鲁棒性，通过最优化理论等技术提高算法的高效性，为该问题设计一个实用的解决方案。