基于深度学习的语义分割方法综述

A review of Semantic Segmentation based on Deep Learning

计梦予* 裘肖明** 于治楼
JI Meng-yu XI Xiao-ming YU Zhi-lou

摘 要 语义分割是视觉分析的基础。近年来，深度学习在视觉领域取得了较大的成功。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，最后对未来的研究方向进行了展望。

关键词 视觉分析；语义分割；深度学习

Abstract Semantic segmentation is the basic of computer vision. In the recent years, deep learning has achieved success in computer vision community. This paper gives a review of semantic segmentation based on deep learning. This paper is first to conclude the related deep learning models for semantic segmentation, after that the paper gives the problem analysis of existing models, finally, this paper discuss the future direction of semantic segmentation.

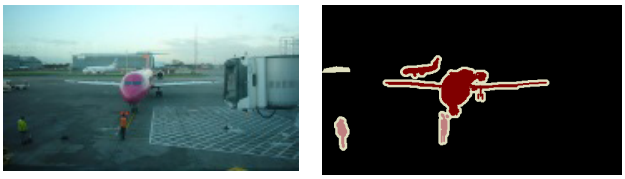
Key words Computer vision； Semantic segmentation； Deep learning

doi：10.3969/j.issn.1672-9528.2017.10.037

1 语义分割简介

在视觉分析领域，语义分割起着越来越重要的角色。语义分割是指将图像中的相同物体的像素分成同一类，并将不同物体分割出来。如图 1 所示，语义分割是将飞机、人等重要目标分割出来。语义分割是自动驾驶、医学图像处理、图像检索、目标分类等视觉分析的基础。例如，在自动驾驶领域，需要对道路、行人、车辆等复杂情况进行分析，从而才能对汽车发出操作指令。在对这些物体分析之前，首先需要进行语义分割，即将道路、行人以及车辆分割出来。在医学图像处理领域，首先要将病灶区分割出来，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。

近年来，深度学习由于其对复杂问题强大的拟合能力，使得其在计算机视觉领域取得了巨大的成功[1,2,3,4]。2012 年，Hinton 研究组使用 Alexnet 在 imagenet 图像分类的竞赛上取得了冠军，其准确率超过第二名 5 个百分点，使得深度学习获得了各大高校和科研机构的密切关注。此后，深度学习广泛用于图像分类、人脸识别、语音识别、目标检测等任务，并在这些任务上取得了突破性的进展。例如，2017 年深度学习在 imagenet 上的分类错误率小于百分之四，远远地高于人类的分类精度，深度学习在人脸库 LWF 上的识别精度达到了 99%，也超过了人类本身对于人脸的识别精度。因此，鉴于深度学习强大的学习能力，使用深度学习有望提高语义分割的精度。



原图 Groundtruth

图 1 语义分割说明图

然而，相比较其他视觉分析任务，语义分割具有更大的挑战性。虽然语义分割本质上是对单个像素点进行分类，但若直接使用图像分类的模型进行语义分割，则存在着一定的局限性：（1）对于一幅图像来说，由于像素点较多，会直接增大计算复杂度；（2）像素与像素之间是有相关性的，直

* 武汉大学计算机学院 湖北武汉 430072
** 浪潮集团有限公司 山东济南 250014

接使用图像分类模型，则不能获取像素点之间的相关性信息，从而降低语义分割的精度。因此，如何针对语义分割任务本身的特点，设计深度学习算法，是语义分割研究的热点。

为了对当今语义分割领域关键技术进行深入地分析、总结，加强相关领域研究者的交流，本文对基于深度学习的语义分割方法进行综述。本文首先总结了基于解码和上下文信息的经典深度学习方法，然后对现有模型的局限性进行分析，最后对未来的发展方向进行了展望。

本文的章节安排如下，第一章对语义分割及深度学习进行简单介绍，第二章主要对相关方法进行总结，第三章对存在问题进行分析，并对相关方向进行展望，第四章对全文进行总结。

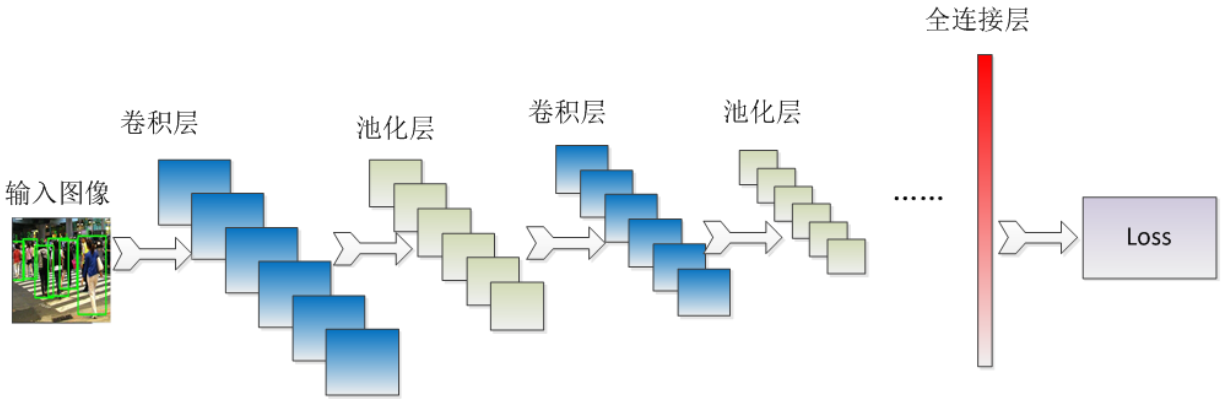


图 2.Alexnet 网络结构图

2 基于深度学习的语义分割方法

现有的语义分割方法主要是将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，使其能够更好地解决语义分割中存在的问题，因此，本文将首先简单介绍几种常用的基网络。

2.1 常用网络

在图像分类任务中，常用的网络主要有 Alexnet，VGG-net，Googlenet，和 Resent。在这些网络中，卷积层、池化层、全连接层等是共有的相关网络层。对于一幅输入图像，首先通过卷积层对图像中的局部区域进行卷积运算，学习相关的局部细节，获得一个新的 feature map. 然后将学习到的 map 再输入到池化层， 池化层对输入的 map 进行采样 . 常用的池化层有最大值采样 . 该过程不仅可以减少 feature map 的尺寸，还能选择出对旋转平移鲁棒的特征 . 经过几轮的卷积层 - 池化层的迭代，将最后的 map 映射成一个列向量 . 该过程是一个特征降维的过程，通过学习一个空间转换矩阵，将全连接层之前的多个 feature map 降维成一个简单的列向量 . 最后通过 Loss 层计算损失 . 计算 Loss 的过程是一个前向传播的过程，通过前向传播可以获得训练时的 Loss. 在求解每层的参数时，使用后向传播，利用随机梯度下降法，通过最小化前向传播的 Loss, 求解每层的参数 .

Alexnet 是 ILSVRC-2012 的冠军。Alexnet 的网络结构如图 2 所示。其主要由 3 个卷积层和 3 个池化层构成，然后后面接有两个全连接层，最后进行输出。首先通过卷积和池化的反复操作，将结果传到全连接层，然后全连接层的输出为向量式特征，最后通过输出层进行分类。在常用的基网络中，Alexnet 只有 8 个层，是最简单的网络。

VGGnet 是由牛津大学的研究者提出的。该模型取得了 ILSVRC-2013 的冠军。相比较 Alexnet，VGGnet 有更大的深度，该网络有 16 层。然而不同的是，VGGnet 并没有全连接层，其将全连接层替换为了全卷积层。另外，VGGnet 减少了卷积核的大小，从而减少了相关的参数。该网络的提出，证明了深度对于提升网络的性能是十分重要的。

Googlenet 取得了 ILSVRC-2014 的冠军。相比较 VGG-net，Googlenet 进一步加深了网络的深度，该网络有 22 层。此外，该网络还通过引入 inception 来增加网络的宽度。Inception 可以看作是一个小网络（如图 3 所示），其第一层是由 2 个卷积层和 1 个池化层组成，第二层是由 4 个卷积层组成。该结构可以利用不同尺寸的卷积核来学到不同尺度的相关信息，从而提高网络的性能。

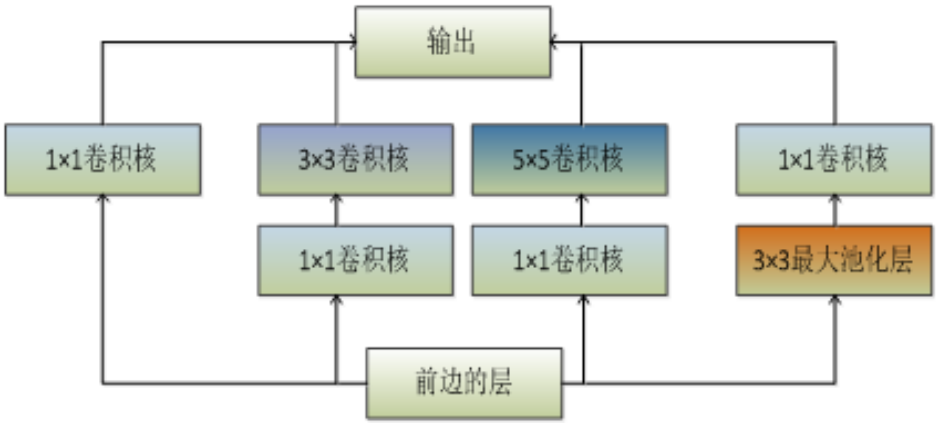


图 3. Inception 结构

Resnet 取得了 ILSVRC-2016 的冠军。相比较 Googlenet，Resnet 将网络的深度扩展的 50 层。另外，该网络引入了残差网络结构，并在输入与输出之间加入了 shotcut，从而可以防止梯度弥散，使得网络结构可以扩展的更深。如图.4 所示。

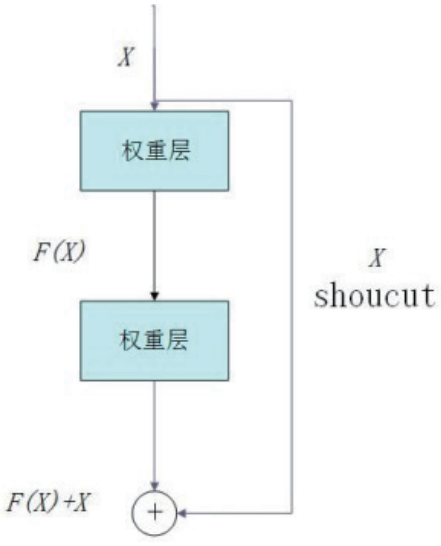


图 4. 残差网络结构

2.2 基于深度学习的语义分割方法

迄今为止，深度学习方法在语义分割大致可以分为 2 类，即基于解码的方法和基于领域知识的方法。

2.2.1 基于解码的方法

Long 等人[5] 在 2015 年提出了基于全卷积神经网络的语义分割方法。该方法在语义分割时，输入为一副图像，而不是 patch，大大降低了计算复杂度。该方法也成为了经典的语义分割方法。该方法的思路是将卷积神经网络中的全连接层替换成了卷积层，从而对每个像素进行分类，得到一个分类 MAP。分类完成后再通过上采样将分类获得的 MAP 映射到原图像大小，根据分类结果可以获得语义分割的结果。不同于传统的卷积神经网络必须输入固定大小的图像，在该方法中，去掉了全连接层，使得输入的图像可以是任意大小。另外，该方法还融合了多分辨率的信息，将不同大小的 MAP 进行上采样，并进行融合，从而获得精确的分割图像。该方法将得到的 MAP 进行上采样的过程可以看成是一个解码的过程，是基于解码方法的开山之作。然而，该方法也存在着一定的局限性，虽然该方法融合了多分辨率信息，但是通过在进行上采样时，容易造成像素位置信息的丢失，从而影响了分割精度。

基于全卷积神经网络的框架，Badrinarayanan 等人[6] 提出了 Segnet 用于道路、车辆的分割。Segnet 是典型的编码－解码网络。该网络首先通过卷积层和池化层进行运算。不同于一般网络的池化层，Segnet 中的池化层能记录池化层的值在原 MAP 中的空间位置，使得在用上采样方法恢复图像尺寸时，能够将相关的值精准的映射到对应的位置，提高恢复图像的精度。在解码时，由于未被选择到池化层的像素没有被记录相关位置，使用反卷积对相关位置的值进行填充。Segnet 记录了池化层相关值的空间位置，使得其在上采样对图像进行恢复时，能够进行准确的恢复，然而，Segnet 对于物体边界的分割精度仍然有待提高。

2.2.2 基于上下文信息的方法

在语义分割任务中，目标的空间信息对于提高分割精度具有重要的作用。因此，利用有效的上下文信息是提高语义分割任务的一种思路。

在语义分割中，为了对最后的分割结果进行细节增强，

条件随机场（Conditional Random Fields,CRF）是一种常用的方法。相比较卷积神经网络，条件随机场能够较为有效地学习到像素之间的相关性。

考虑到像素之间的局部特性，即相邻像素属于同一类别的概率应该更高，Zheng 等人[7]提出了 CRFasRNN。首先提出了平均场的概念，将平均场近似为条件随机场，转化为 RNN，并把 CRF-RNN 嵌入到卷积神经网络中。最后利用随机梯度下降法来求解参数。

在利用条件随机场一类方法中，最为经典的是 Deeplab 系列方法。在 Deeplab v1 中，引入了 hole 思想，并在卷积核里增加 hole。根据分割物体的尺度来修改 hole 的大小，从而自动的调整感受野的大小，避免了上采样带来的信息丢失，从而可以提高分割精度。多分辨率信息对于提高语义分割的精度具有重要作用。Deeplab v2 在 Deeplab v1 的基础之上利用了多分辨率信息。其主要思路是增加多分辨率的感受野。多分辨率的感受野可以更为有效地学习不同尺度目标的有效信息，从而进一步提高目标的分割精度。

3 展望

虽然现有的方法在很大程度上提高了语义分割的精度，然而仍存在一定的局限性，因此，如何解决这些局限性是未来研究的热点。（1）参数的有效学习。多分辨率信息对于提高分割精度具有重要的作用，然而现有的方法只是根据人工经验设置多分辨率的相关参数，如何针对现有数据的特点，自动学习出有效的参数，是未来的一个研究热点。（2）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复杂度。因此，如何在保证精度的同时，降低模型的复杂度，实现实时分割，是未来的另一个研究热点。（3）3D 数据的有效分割。现有的算法大部分是基于 2D 图像，在真实世界中存在大量的 3D 数据，深度学习在 3D 数据上的算法研究较少，因此，如何研究有效的深度学习算法使其能够较好的分割 3D 数据是未来的一个发展方向。

4 结束语

语义分割是视觉分析的基础。近年来，深度学习方法在视觉分析领域取得了较大的成功。本文对近年来语义分割领域中的经典基于解码和基于上下文信息的深度学习方法进行了深入的分析及总结。并根据现有方法的局限性，对有效参数学习、实时分割及 3D 数据的有效分割三个方向进行了展望。

参考文献：

[1]P. Dollar, R. Appel, S. Belongie, and P. Perona, Fast Feature Pyramids for Object Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014, 36(8):1532－1545,

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016,38(1): 142－158

[3] Dahl, G. E., Yu, D., Deng, L. and Acero, A. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(1):30－42

[4] Abdel-rahman Mohamed,George E. Dahl,Geoffrey Hinton, Acoustic Modeling Using Deep Belief Networks[J].IEEE Transactions on Audio, Speech, and Language Processing, 2012,20(1):14－22

[5] J. Long, E. Shelhamer, and T. Darrell, Fully convolutional networks for semantic segmentation[c]// in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3431－3440.

[6] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for scene segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017.

[7] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su,D. Du, C. Huang, and P. H. Torr, Conditional random fields asrecurrent neural networks[C]// in Proceedings of the IEEE International Conference on Computer Vision. 2015:1529－1537.

（收稿日期：2017-10-18）