

语义分割综述：截止2022，语义分割总结与展望

极市平台 2023-02-19 22:00:36 发表于广东 手机阅读 跟

以下文章来源于阿柴的算法学习日记，作者阿柴的算法日记



阿柴的算法学习日记

欢迎光临！一只可爱阿柴的算法世界，分享CV、NLP、Python探索之路上的点滴~

↑ 点击蓝字 关注极市平台

作者 | 阿柴本柴

来源 | 阿柴的算法学习日记

编辑 | 极市平台

极市导读

从近两年语义分割领域的工作出发，对语义分割方向的存在的三大桎梏出发，对20-22年对解决领域问题的优质工作进行了总结和分析。 >>加入极市CV技术交流群，走在计算机视觉的最前沿

前一段时间看到有网友在我这篇文章下留言：


<https://zhuanlan.zhihu.com/p/133212654>

-- “两年了，更新更新”

写下之前这篇综述文章的时候是刚接触分割不久，遍读多篇论文、笔记之后，颇有理清语义分割任务内在发展线路之感时写下的文章。当时我就在想，身边同学也是这么想，语义分割性能应该是到达顶峰了，很难有大的突破。后续的这两年的发展也比较符合预期，虽然Transformer开始大行其道，但是实质在落地上带来的收益却是不如2020前的研究那样突飞猛进，毕竟事物的发展总是存在着它的边际效应。

但是，收益不迅猛，不代表所做的工作是没有意义的。这两年的论文，从之前基于CNN的语义分割深度学习模型框架所存在的缺陷出发，通过各种各样精妙绝伦的构思，尝试突破传统CNN框架的桎梏，为后来者夯实了坚定的基础。所以，这些思想应该是需要被总结、传播的，这是我想再次尝试在上一篇小综述的基础上总结这两年语义分割发展的初衷。

当然，写下这篇文章对我本身而言还有着额外的含义。语义分割是我入门深度学习接触到的第一个应用方向，承载了我学生生涯中篇顶会的希望（但是，事与愿违，最后只发了几篇中规中矩的期刊论文）。毕业之后，这个方向虽然也时时关注，但已经真正去复现代码的频率已经没有那么高了。因此，这一篇总结也算是给我学生生涯以及语义分割生涯画上一个暂时的逗号吧。之后会时时关注，常常总结推送，不过不会向以前那样时常去复现代码了。毕竟业务上还是有太多的badcase要去解决了。



极市平台
extreme

月发文数目： **

月平均阅读： **

文章工具

已发文

采集图文

合成多

采集样式

查看

好了，闲话就说到这儿，进入正题（有总结的不好的地方请海涵）。

一、从语义分割的几个桎梏说起

1.1、CNN的局限性

从上一篇总结文章中可以看出，基于CNN的结构大多都遵循编码器-解码器的框架(https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Long_Fully_Convolutional_Networks_2015_CVPR_paper.pdf)。编码器中CNN用于特征提取，它在逐渐降低特征图分辨率的同时使得特征图富含语义信息。随后解码器中的CNN利用编码器编码特征作为输入，解码出最后的分割预测结果。

当然，这种最为基础的框架，存在很多的问题。比如语义分割任务除了语义信息还需要细节信息，因此UNet等论文给出了解决方案。比如语义分割任务需要上下文信息，因此PSPNet、DeepLab系列、基于自注意力机制的一系列方法（Non-Local、DANet、CCNet等）等被提出来获取局部、多尺度乃至全局上下文。又比如语义分割框架对于物体边缘处的分割效果不理想，因此Gated-SCNN等一些方法也在着力解决这些问题。

这些问题在以上这些方法的支撑下得到了极大的缓解。当然不是说这些点不值得研究了，学界和业界总是能有些有才之人再往前突破一点的。只不过我私以为它们都已经开始出现边际效应了。投入其中，无论对于学界还是业界可能投入与产出不一定能够达到预期的效果。

然而，除了上述的点之外，CNN的方法本质上存在着一个巨大的桎梏，就是图像初始阶段输入到网络之时，由于CNN的卷积核不会太大，所以模型只能利用局部信息理解输入图像，这难免有些一叶障目，从而影响编码器最后提取的特征的可区分性。这是只要使用CNN就逃脱不了的缺陷。当然，有人会说基于自注意力机制的一些即插即用的模块插入到编码器和解码器之间，就能获取到全局上下文，使得模型从全局的角度理解图像进而改善特征。但是，模型如果一开始因为一叶障目获取了错误的特征，在后续利用全局上下文是否能够纠正的过来是存在一个很大的疑问的。

1.2、标注数据的局限性

我们知道，语义分割任务是像素级别的分类。因此，一张512*512的图像在进行分割任务标注时，所需的标注次数理论上是图像分类任务的512*512倍。正因为如此，分割的输入获取是需要很大的资源投入的，简言之就是要烧很多的钱的。

1.3、模型的泛化能力

这不仅是分割任务上存在的问题，只要基于深度学习的任务就难免面临这么一个窘境。我辛辛苦苦训练好了一个模型，在换了一个场景的图像输入到模型之后，模型的性能往往出现一个很大的下降。这个问题在实际中太常见了。以遥感图像语义分割为例，我在上海采集的城市数据集上训练好模型，在对来自成都的影像进行分割时，往往效果与随机预测的无异。这太尴尬了，要按传统方式解决这一问题的话，各个地区的测绘局都得采集一遍数据，然后标注好，才能无缝对国内所有城市的进行语义分割。这工作量，想想就很难。

二、2020-2022对于这三点给出的答案

2.1、CNN的模型不能从一开始就从全局理解输入图像的问题

对于这一点，所给出的答案最好的当然还是基于Transformer的方案，它将输入的图像Token化，然后利用自注意力机制就能在模型的一开始使得模型能够以全局的角度去理解图片。这里

顺着发展脉络，给出语义分割任务上使用Transformer必须要了解的几篇文章：

1) 用Transformer就不能不看ViT，他是将Transformer用于视觉任务的开山之作，主要思想可以看如下解读：

ICLR2021-谷歌大脑团队Vision Transformer: AN IMAGE IS WORTH 16X16 WORDS

<https://zhuanlan.zhihu.com/p/427997345>

2) 在ViT的基础上尝试使用Transformer解决分割任务的几个方法值得一看，有针对Transformer解码器改进的、有针对Transformer编码器改进的，也有结合CNN与Transformer的。这些方法本质上仍然还是属于编码器解码器的基础框架，但是已经打破了CNN的桎梏：

语义分割中的Transformer（第一篇）：SETR与TransUNet — 使用Transformer时解码器的设计

语义分割中的Transformer（第二篇）：SegFormer — 简单有效的语义分割新思路

语义分割中的Transformer（第三篇）：PVT — 用于密集预测任务的金字塔 Vision Transformer

3) 经典之作Swin Transformer系列不可不看，就像基于编码器-解码器结构要为任务选择一个合适的编码器backbone一样，基于Transformer结构的编码器也是需要精益求精的，而Swin Transformer显然是一个比较优秀的选项：

Swin v1:<https://arxiv.org/abs/2103.14030v1%3Fref%3Dhackernoon.com>

Swin v1中利用滑动窗口和分层结构的设计使得Swin Transformer成为了CV领域新的SOTA Backbone，在图像分类、目标检测、语义分割等多种机器视觉任务中达到了SOTA水平。

Swin v2:<https://arxiv.org/abs/2111.09883>

Swin v2中提出了post-norm and cosine similarity、Continuous relative position bias 和 Log-spaced coordinates来分别解决模型不够大和不能适配不同分辨率的图片和不同尺寸的窗口的问题。

当然，最近也有新的backbone在论文中report的精度超过了SWIN，感兴趣也可多看看，比如：

CVPR2022 Oral - Shunted Transformer: 全新多尺度视觉 Transformer 主干网络

4) 最近新出的一些用于分割的Transformer也值得一看，他们的思想源于NLP预训练中会使用CLS这个Token去表征语义，在这些方法中也用到了随机初始化去构建的Token，在学习过程中逐渐富有了想表征的语义：

Segmenter: <https://arxiv.org/abs/2105.05633>

Segmenter 在解码阶段使用一系列与语义类别对应的可学习token，与图像自身解码的特征进行交互，从而实现最终分割预测。

还有一系列与Segmenter相同思想的文章，他们都用到了可学习token去表征他们想要表征的语义。

MaskFormer:<https://arxiv.org/abs/2107.06278>

语义分割新范式：上海 AI Lab 联合北邮、商汤提出 StructToken:<https://zhuanlan.zhihu.com/p/535029873>

然而，诚然打破CNN桎梏的Transformer是的模型可以在一开始就从全局的角度去理解图像。但是这样是否获取到了真正的全局信息呢？答案是多数情况下是不行的。为什么？我们知道真实的图片往往是分辨率比较大的，直接输入到模型中，显卡是扛不动的，因此一张图像我们需要裁剪成多个小图才能使得送到模型中。这一裁剪，先天就使得模型只能看到完整大图中的一部分内容。因此，此时的全局角度并不是完整的全局，而只是裁剪后对于小图的全局。显然这是会影响模型的性能的。那么怎么解决呢？可以看看以下文章的解决方案：

CVPR2021-MagNet与ICCV2021-FCtl：如何提高超分辨率图像的语义分割准确性

2.2、对于标注数据获取困难的解决方案

对于标注数据不好获取，显然结合弱监督、无监督的思想来做语义分割是比较好的解决方案。我对无监督语义分割了解的甚少，因为我认为暂时CV领域还做不到像NLP领域那样巧妙设计无监督任务的程度，因此无监督语义分割暂时应该是达不到一个能看的精度（如果这个判断有误的话，欢迎指正）。所以这里主要介绍弱监督语义分割。

首先可以先了解一下弱监督分割的简要概念与做法：

弱监督语义分割综述

CVPR2022-Class Re-Activation Maps：用于弱监督语义分割的类重新激活图 <https://zhuanlan.zhihu.com/p/479730141>

弱监督语义分割截止2020年论文汇总、简要解读

总的来说，弱监督语义分割就是使用比像素级标签更容易获取的标签，比如图像分类标签来训练分割模型。目前而言，使用图像分类标签训练的分割模型已经开始可以逐渐全监督语义分割的精度，比如：

CVPR2022-Class Re-Activation Maps：用于弱监督语义分割的类重新激活图

CVPR2022-Pixel-to-Prototype Contrast：将对比学习应用于弱监督语义分割

它们能够在PASCAL VOC2012的数据集上达到72以上的miou，可以说十分惊艳。当然出了图像分类的标签，还可以使用其他的容易获取的标签，他们的精度能够达到更高，不过相应的能够标签获取难度会上升一点，比如：

CVPR2022-Tree Energy Loss：能够扩展弱监督语义分割中稀疏真值标签的新方法

2.3、如何增强模型的泛化能力

增强模型的泛化能力其实有很多基础的方法，比如数据增强、正则化等等。但是它们起到的效果是有限的。这里我们分为两种情况来讨论如何增加模型的泛化能力：

1) 测试集数据不可获取：

那此事就只能从模型本身出发，迫使模型能够学习到更为鲁棒的特征，具体这一篇文章值得一看：

CVPR2022 Oral-即插即用！感知语义的域泛化语义分割模型 (SAN & SAW)

2) 测试集数据可获取（没有标签）：

这显然就是无监督域适应的范围了。无监督域适应语义分割主要分为三个研究方向：

1) 基于对抗学习：这一类的方法出发点在于目标域与源域在同一Encoder后编码的特征能够尽量相似。主要在FCAN与ADVENT的基础上寻求突破与创新。以下链接深入的讲了基于对抗学习的无监督域自适应语义分割的原理：

语义分割中的无监督域自适应系列-AdaptSegNet

2) 风格迁移：这一类的方法出发点在于转换源域图片的风格使得其与目标域相似。代表方法有CycleGAN。

3) 自监督学习：在目标域上形成伪标签来训练模型。

这些方向上方向上值得一看的论文有：

ICCV2021-语义分割无监督域适应：Dual Path Learning（DPL）

CVPR2021语义分割无监督域适应：Self-supervised Augmentation Consistency（SAC）

我自己也试过一些方法效果还不错，也总结过：

实战篇：使用GAN的思想进行遥感图像语义分割的域适应

三、总结

总的来说，本文总结了2020年后，从语义分割三个缺陷出发来破局的一系列方法。本文中只包含了我阅读过的一些论文，难免有些偏颇之处，往读者取其精华去其糟粕。



极市
EXTREME MART

视觉AI算法实训加速营

从模型开发到部署落地提升CV工程能力

TIME 2023.02.16-03.12

免费报名

CV算法全流程技术能力提升


全程实战技术指导

上千元奖学金



长按识别二维码入群报名

公众号后台回复“数据集”获取200+数据集资源汇总



极市平台

为计算机视觉开发者提供全流程算法开发训练平台，以及大咖技术分享、社区交流、竞...
848篇原创内容

公众号

极市干货

技术干货：损失函数技术总结及Pytorch使用示例 | 深度学习有哪些trick？ | 目标检测正负样本区分策略和平衡策略总结

实操教程：GPU多卡并行训练总结（以pytorch为例） | CUDA WarpReduce 学习笔记 | 卷积神经网络压缩方法总结



极市原创作者激励计划

.....

极市平台深耕CV开发者领域近5年，拥有一大批优质CV开发者受众，覆盖微信、知乎、B站、微博等多个渠道。通过极市平台，您的文章的观点和看法能分享至更多CV开发者，既能体现文章的价值，又能让文章在视觉圈内得到更大程度上的推广，并且极市还将给予优质的作者可观的稿酬！

我们欢迎领域内的各位来进行投稿或者是宣传自己/团队的工作，让知识成为最为流通的干货！

对于优质内容开发者，极市可推荐至国内优秀出版社合作出书，同时为开发者引荐行业大牛，组织个人分享交流会，推荐名企就业机会等。

投稿须知：

- 1.作者保证投稿作品为自己的原创作品。
- 2.极市平台尊重原作者署名权，并支付相应稿费。文章发布后，版权仍属于原作者。
- 3.原作者可以将文章发在其他平台的个人账号，但需要在文章顶部标明首发于极市平台

投稿方式：

添加小编微信Fengcall（微信号：fengcall19），备注：姓名-投稿



点击阅读原文进入CV社区
收获更多技术干货

阅读原文

喜欢此内容的人还喜欢

ICCV 2023 | 南开程明明团队提出适用于SR任务的新颖注意力机制（已开源）
极市平台



YOLOv5帮助母猪产仔？南京农业大学研发母猪产仔检测模型并部署到Jetson Nano开发板
极市平台



ICCV23 | 将隐式神经表征用于低光增强，北大张健团队提出NeRC
极市平台

