# 语义分割学习笔记

*huang_rui4@dahuatech.com*

## 导航

# 1 前言

以下是在学习基于深度学习方法实现语义分割的过程中，总结的论文、学习报告、博客等等，以及针对部分问题提出的个人看法，特此记录！

# 2 从语义分割的几个桎梏说起

## 2.1 CNN 的局限性

从

### 2.1.1 UNet

到

### 2.1.2 PSPNet

再到

### 2.1.3 Deeplab 系列

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，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，难免有些一叶障目

### 2.1.4 HRNet

理论参考：Link

HRNet(图2) 通过**并行多个分辨率分支**以及不同分支之间的信息交互，获取强语义信息 & 精准位置信息。而在此之前几乎所有的网络 (图1) 都是**通过下采样得到强语义信息，再上采样恢复高分辨率信息**，导致**大量的有效信息会在不断的上下采样过程中丢失**。

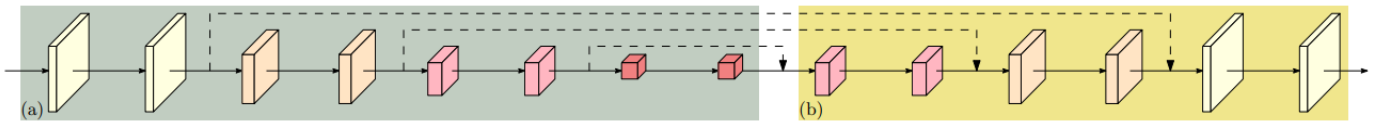


图 1: 由低分辨率恢复到高分辨率的网络结构图

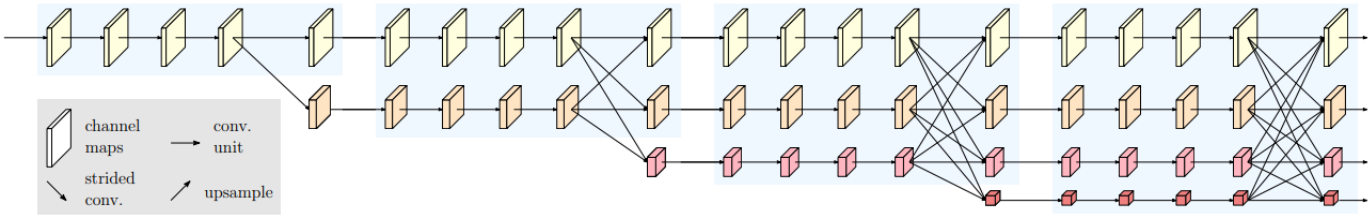


图 2: HRNet 网络结构图

## 2.2 标注数据的局限性

- 现实的数据往往缺乏标签
- 数据标注需要很大的资源投入————烧很多的钱
- 很多任务难以获取真实标签

## 2.3 模型的实时性

### 2.3.1 DDRNet[4]

- **应用场景：**高分辨率图像实时语义分割
- **动机：**
- **创新点：**1. 使用两个分辨率不同的分支，一个用于生成高分辨率的特征图，另一个通过多次降采样提取丰富的语义信息，并在两个分支之间建立多个双边连接，实现高效信息融合；2. 提出一个新的多尺度模块 DAPPM **标签：**一个效果还不错的实时语义分割网络

1. 深度双分辨率网络

2. DAPPM 作者认为，仅使用一个 $3\times3$ 卷积或 $1\times1$ 卷积将所有的多尺度上下文信息融合在一起是不够的。所以采用类似 **Res2Net** 的结构，首先对特征图进行上采样（灰色块），然后使用更多的 3×3 卷积，以层次残差 hierarchial-residual 的方式融合不同尺度的上下文信息（紫色块）。最后将所有得到的特征 concatenate 并使用一个 1×1 卷积压缩。此外，为了便于优化，还添加了一个 1×1 卷积映射作为 shortcut。

## 2.4 模型的泛化性

辛辛苦苦训练好了一个模型，换了一个场景后，模型的性能出现很大下滑。按传统方式解决这一问题，是将新场景的数据采集一遍，然后标注好，再重新训练。这工作量一言难尽。。。

# 3 Transformer，它来了

针对 CNN 模型不能在一开始就从全局理解输入图像的问题，基于 Transformer 的方案，是将输入的图像 Token 化，然后利用自注意力机制使得模型一开始就以全局的角度去理解图片。

## 3.1 一问一答

### 3.1.1 多头注意力机制为什么可以提高模型性能？

等价于卷积操作中使用多个卷积核，用于提取不同的特征，信息量更加丰富！

#### 3.1.2 Q、K、V 的意义是什么？

#### 3.1.3 如何解释注意力机制的有效性？

#### 3.1.4 为什么通过 Q、K 的内积索引 V？

#### 3.1.5 如何实现并行训练的？

#### 3.1.6 为什么使用 LN 而不是 BN，dropout 与 LN 同时使用时，有什么特殊操作？

#### 3.1.7 用于训练的数据格式？

#### 3.1.8 为什么要基于学习的词向量空间？

#### 3.1.9 可学习的位置编码？

#### 3.1.10 位置编码如何生效的？

#### 3.1.11 teacher forcing 原理

### 3.2 开山之作 VIT

# 4 附录

## 4.1 半监督学习

提醒：

- 定义：半监督学习（Semi-Supervised Learning，SSL）属于弱监督学习的一个分支，具体地：弱监督学习可以分为三类：
  - 不完全监督【只有很小的子集有标签】，包含两个方向：
    - 主动学习
    - 半监督学习，按学习方式可分为纯半监督学习和直推学习
  - 不确切监督【只有粗粒度标签】
  - 不准确监督【标签不总是真实】
- 基本假设：未标记数据必须是有意义的，具有潜在价值的样本，而非无用的噪声样本
- 存在的问题:
  - 无效标签样本的有效利用问题
  - 大量无效标签样本的高效使用问题
  - 很多半监督算法都有无监督的影子，因此对于特征的选择非常的敏感

半监督语义分割方法大致可以分为 5 类：

- 类似于 GAN 结构生成对抗

- 一致性正则化【最小化同一图像的不同预测之间的差异】
- 伪标记方法
  - 自训练方法【依赖于现有模型的泛化性，缺乏检测自身错误的机制，同时需要解决伪标签中的类不均衡问题】
  - 互训练方法【基于分歧的策略】
- 基于对比学习
- 混合模式

### 4.1.1 基于 GAN 的半监督语义分割框架. 2017

### 4.1.2 Mean Teacher

核心思想是强制学生网络和教师网络的预测在存在扰动的情况下一致。其中教师网络的权重是通过学生网络权重的指数移动平均值（EMA）计算得出的。

扰动的方式有四种：

- 基于输入的扰动：CutOut、CutMix、ClassMix、Complex、其他方法
- 基于特征的扰动：交叉一致性训练（CCT）
- 基于网络的扰动，比如使用不同的起始权重
- 混合模式

# 参考文献

[1] 语义分割综述: 截止 2022, 语义分割总结与展望, https://zhuanlan.zhihu.com/p/538050231.

[2] J. Wang, K. Sun, T. Cheng, et al. Deep High-Resolution Representation Learning for Visual Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, 43(10), 3349-3364.

[3] M. Xu, Z. zhang, H., Hu, et al. End-to-End semi-supervised object detection with soft teacher, *Computer Vision and Pattern Recognition*, 2021.

[4] Y. Hong, H. Pan, W., Sun, et al. Deep Dual-resolution Networks for Real-time and Accurate Semantic Segmentation of Road Scenes, *Computer Vision and Pattern Recognition*, 2021.