

mmdetection 解析

Sisyphes,yehao

2019 年 8 月 1 日

目录

第一节 结构设计	4
1.1 总体逻辑	5
1.2 Configs	5
1.3 Backbone	6
1.3.1 backbone 及改进	6
1.4 Necks	10
1.4.1 多尺度结构及改进	10
1.4.2 小目标	12
1.5 Heads	12
1.5.1 AnchorHead	12
1.5.2 SSDHead	16
1.5.3 RetinaHead	17
1.5.4 GuidedHead	17
1.5.5 FCOSHead	17
1.6 Losses	19
1.6.1 基本认识	19
1.6.2 实现解析	20
1.7 Detectors	22
1.7.1 maskrcnn	22
1.7.2 RepPoints	24
1.7.3 RetinaFace	27

目录	2
第二节 数据处理	27
2.1 检测分割数据	27
第三节 FP16	33
第四节 训练 pipeline	33
4.1 训练逻辑	33
第五节 更改模型	35
5.1 增加模块	35
5.2 模型瘦身	35
第六节 抽离模型	35
6.1 retinanet_resnet18	35
第七节 新增模型	35
7.1 centernet	35
7.2 代码说明	36
第八节 numpy,torch 某些基础函数	37
8.1 损失函数部分	37
8.2 模型实现部分	37
8.3 数据类部分	37
第九节 版本变化说明	37
9.1 v2.0	37
第十节 计划	38
第十一节 检测模型的简略综述	38
11.1 通用物体检测	38
11.1.1 Yolo 系列	38
11.1.2 SSD 系列	39
11.1.3 Fast RCNN 系列	40
11.1.4 Anchor Free 系列	41
11.2 总结	41

第十二节 官方文档 2.0 伪译	41
12.1 配置系统	41
12.2 使用预训练模型	41
12.2.1 继承基础配置文件	42
12.2.2 更改头部	42
12.2.3 更改数据	43
12.2.4 改写训练 schedule	43
12.2.5 使用预训练模型	43
12.3 增加新数据类	43
12.3.1 转成公用格式	43
12.3.2 中间格式	44
12.4 自定义数据管道	47
12.4.1 扩展 pipelines	47
12.5 增加新模块	48
12.5.1 优化器	48
12.5.2 开发新组件	48
12.6 1.x 模型升级到 2.0	50
12.7 2.0 和 1.x 的不同之处	50
12.7.1 坐标系	50
12.7.2 Codebase Conventions	50
12.7.3 训练超参数	51

第一节 结构设计

- Backbone: 特征提取骨架网络, ResNet, ResNeXt 等.
- Neck: 连接骨架和头部. 多层次特征融合, FPN, BFP 等.
- DenseHead: 处理特征图上的密集框部分, 主要分 AnchorHead, AnchorFreeHead 两大类, 分别有 RPNHead, SSDHead, RetinaHead 和 FCOSHead 等.
- RoIExtractor: 汇集不同层级的特征框, 将其大小统一, 为二步定位, 类别优化服务.
- RoIHead (BBBoxHead/MaskHead): 类别分类或位置回归等.
- OneStage: Backbone + Neck + DenseHead
- TwoStage: Backbone + Neck + (DenseHead) + RoIExtractor + RoIHead

代码结构:

configs 网络组件结构等配置信息

tools: 训练和测试的最终包装

mmdet:

apis: 分布式环境设定, 推断和训练基类代码

core: anchor, bbox, mask 等在训练前和训练中的各种变换函数

datasets: coco 和 voc 格式的数据类以及一些增强代码

models: 模型组件, 采用注册和组合构建的形式完成模型搭建

ops: 优化加速代码, 包括 nms, roialign, dcn, gcb, mask, focal_loss 等

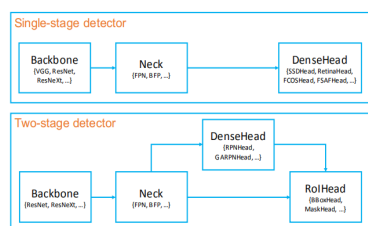


图 1: Framework

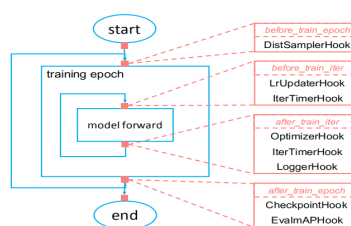


图 2: Training pipeline

1.1 总体逻辑

在最外层的 `train.py` 中能看到:

1. `mmcv.Config.fromfile` 从配置文件解析配置信息, 并做适当更新, 包括预加载模型文件, 分布式相关等
2. `mmdet.models.builder` 中的 `build_detector` 根据配置信息构造模型
 - 2.5 `build` 函数调用 `_build_module`(新版为 `build_from_cfg`) 函数, 按 `type` 关键字从注册表中获取相应的模型对象, 并根据配置参数实例化对象 (配置文件的模型参数只占了各模型构造参数的一小部分, 模型结构并非可以随意更改).
 - 2.6 `registr.py` 实现了模型的注册装饰器, 其主要功能就是将各模型组件类对象保存到 `registry.module_dict` 中, 从而可以实现 2.5 所示功能.
 - 2.7 目前包含 `BACKBONES,NECKS,ROI_EXTRACTORS,SHARED_HEADS,HEADS,LOSSES,DETECTORS` 七个 (容器). 注册器可按 `@NAME.register_module` 方式装饰, 新增. 所有被注册的对象都是一个完整的 `pytorch` 构图
 - 2.9 `@DETECTORS.register_module` 装饰了完整的检测算法 (`OneStage`, `TwoStage`), 各个部件在其 `init()` 函数中实例化, 实现 2.5 的依次调用.
3. 最后是数据迭代器和训练 pipeline四.

1.2 Configs

配置方式支持 `python/json/yaml`, 从 `mmcv` 的 `Config` 解析, 其功能同 `maskrcnn-benchmark` 的 `yacs` 类似, 将字典的取值方式属性化.

配置文件模型部分包含模型组件及其可改动模型结构的参数, 比如 `backbone` 的层数, 冻结的 `stage`; `bbox_head` 的 `in_channel`, 类别, 损失函数等; 训练部分主要包括 `anchor` 采样相关系数; 测试包括非极大抑制等相关参数; 剩下数据, 优化器, 模型管理, 日志等相关信息, 一看即明.

1.3 Backbone

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112x112	7x7, 64, stride 2				
		3x3 max pool, stride 2				
conv2.x	56x56	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$		$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$
conv3.x	28x28	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$		$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 4$	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 8$
conv4.x	14x14	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$		$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 23$	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 36$
conv5.x	7x7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$		$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$
	1x1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

图 3: resnet

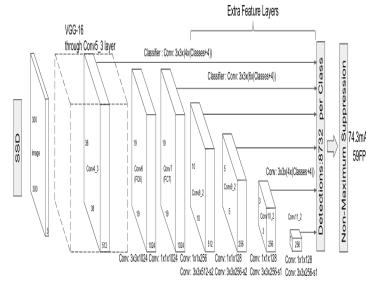


图 4: ssd

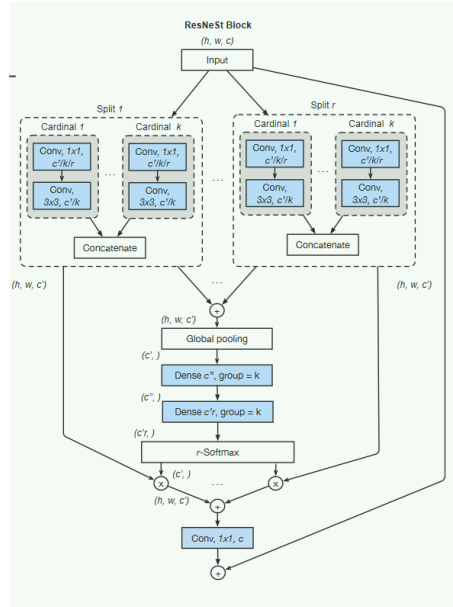


图 5: ResNeSt

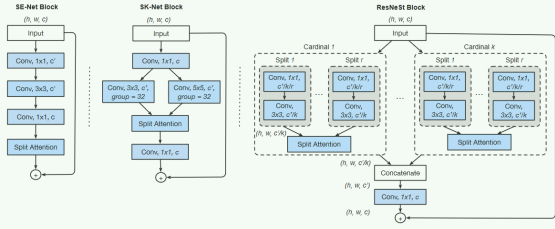


Fig. 1: Comparing our ResNeSt block with SE-Net [30] and SK-Net [38]. A detailed view of Split-Attention unit is shown in Figure 2. For simplicity, we show ResNeSt block in cardinality-major view (the featuremap groups with same cardinal group index reside next to each other). We use radix-major in the real implementation, which can be modularized and accelerated by group convolution and standard CNN layers (see supplementary material).

图 6: ResNeSt_SE_SK

1.3.1 backbone 及改进

出于对Universal approximation theorem的考察，人们人为加深网络，在复杂问题上会带来更好的效果，但实验情况并非如此，于是 2015 年ResNet出现，旨在解决模型网络加深带来的性能下降问题 (degradation problem)。其

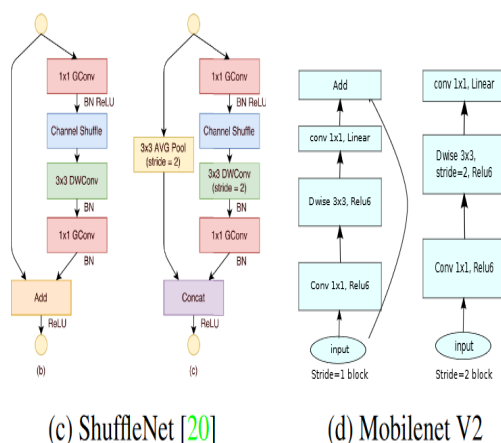


图 7: mobilenetv2

结构参数见3。其中 $\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix}$ 为 Basicblock 基础块, $\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix}$ 为

Bottleneck 基础块, 后接池化, 全连接和概率 softmax 映射, Bottleneck 计算量为 Basicblock 的 5.9%, 池化替代全连接, 更本质, 一来计算量极大减小, 二来过滤或者综合出了更有效的特征。一点需要注意一下, 基础块都 \times 了一个数, 重复计算可以增加感受野, 同时增加了非线性算子个数, 能提取更抽象的特征, 因为上下层尺度只减小了一半, 所以 stride=2 只会在重复块中的某一个上做操作, 这个是可以调节的。

那么 ResNet 是怎么解决梯度消失问题的呢? 由 Skip Connection 组成的残差块 (residual blocks)。

比如将第 l 层到 $l+1$ 层的激活函数改到 addition 之前 (参考 ResNet V2) 的残差模块的表示方式: $\mathbf{x}_{l+1} = \mathbf{x}_l + \mathcal{F}(\mathbf{x}_l, \mathcal{W}_l)$, 递归形式 $\mathbf{x}_L = \mathbf{x}_l + \sum_{i=l}^{L-1} \mathcal{F}(\mathbf{x}_i, \mathcal{W}_i)$, 则其梯度如下:

$$\frac{\partial loss}{\partial \mathbf{x}_l} = \frac{\partial loss}{\partial \mathbf{x}_L} \frac{\partial \mathbf{x}_L}{\partial \mathbf{x}_l} = \frac{\partial loss}{\partial \mathbf{x}_L} \left(1 + \frac{\partial}{\partial \mathbf{x}_l} \sum_{i=l}^{L-1} \mathcal{F}(\mathbf{x}_i, \mathcal{W}_i) \right)$$

从 BP 公式我们知道梯度消失来源于梯度向量叠乘后太小, 这里均有 +1, 极大抑制了梯度消失问题。

ResNet 的改进文章有很多, 主要可分为模型性能, 模型轻量化和功能扩展。如上 V2, Bag 包含了性能, 轻量上的改进, 而 resnest 集成了三方面的

改进。resnest 结构见6，从结构图基本可以看到该残差模块的含义。现‘按图索骥’，给出对应的公式 (对比论文)，即可完成清晰的理解。

首先对输入特征图按 channel 分解成 $G = KR$ 部分, 其中 K 是 cardinal group 数, R 是 radix group 数, 一句话, 组中组。每个原子特征图经过 F_i 映射成 U_i , $i \in \{1, 2, \dots, G\}$ 。从图中能看到 F_i 为一个 $1 \times 1, 3 \times 3$ 卷积组成, 这里好奇为何不是 131 结构, 然后一个 R 小组接一个 Split Attention 结构, 然后 K 个 Split Attention 结果 concatenate 在一起和输入图相加, 组成一个残差块。图中有个细节, c, c' , 这个也是 channel 降维减小计算量吧 (出于这个考虑, 也许可以圆一下为何不是 131 结构)。 K 个 R 基组经过 F 映射后做求和处理得到 K 基组单位元的表示, $\hat{U}^k = \sum_{j=R(k-1)+1}^{Rk} U_j$, 其中 $\hat{U}^k \in \mathbb{R}^{H \times W \times C/K}$, 每个 K 特征图在 channel 维度上做一个全局池化得到该特征图的 Global contextual information $s^k \in \mathbb{R}^{C/K}$, 其中第 c 个 channel $s_c^k = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \hat{U}_c^k(i, j)$, 这些 Global contextual information 做一个全连接映射 \mathcal{G} , 得到 K 组中每个 R 组中各特征的 channel 的重要系数 $a_i^k(c)$,

$$a_i^k(c) = \begin{cases} \frac{\exp(\mathcal{G}_i^c(s^k))}{\sum_{j=0}^R \exp(\mathcal{G}_j^c(s^k))} & \text{if } R > 1 \\ \frac{1}{1 + \exp(-\mathcal{G}_i^c(s^k))} & \text{if } R = 1 \end{cases}$$

最后用此系数加权求和 R 组各特征图 U , 其中第 c channel 为 $V_c^k = \sum_{i=1}^R a_i^k(c) U_{R(k-1)+i}$, 得到 Attention 后的 K 组结果, 然后 concat 一下得到映射特征, 即 $V = \text{Concat} \{V^1, V^2, \dots, V^K\}$, 最后执行 Skip Connection, $Y = V + X$ 得到残差块的最终输出。 X 上可做操作, 可调节。

以上基本囊括了 ResNet 的改进内容。接着说说 mmdet 的代码部分。

代码上, Resnet 由 BasicBlock, Bottleneck, make_res_layer 构成。其中 make_res_layer 重复构造残差模块。需要说明的是在 resnet.py 中引进了 dcn, gcb 和 gen_attention 等功能模块。从 Bottleneck 的 _inner_forward 可以看到, gen_attention_block 加到 kerner_size=1, kerner_size=3 组 (组: 卷积, 归一化, 激活) 之后, kerner_size=1, planes 扩展组之前。gcb 加到 planes 扩展组之后, 然后才是下采样。因 dcn 是替换卷积算子, 其参数和基本卷积相同, 因此你可以任意替换 conv2d, 只是试验表明, 替换此处的 kerner_size=3 组效果更好。

代码参数说明, 主要和 fpn 对齐部分, 在 config 中按照如下对齐方式更改即可。


```

2 @BACKBONES.register_module
3 class ResNet(nn.Module):
4     arch_settings = {
5         18: (BasicBlock, (2, 2, 2, 2)), # (2, 2, 2, 2)各层残差块的重复数目
6         34: (BasicBlock, (3, 4, 6, 3)),
7         50: (Bottleneck, (3, 4, 6, 3)),
8         101: (Bottleneck, (3, 4, 23, 3)),
9         152: (Bottleneck, (3, 8, 36, 3))
10    }
11    def __init__(self,
12        depth,
13        in_channels=3,
14        num_stages=4,
15        strides=(1, 2, 2, 2),
16        dilations=(1, 1, 1, 1),
17        out_indices=(0, 1, 2, 3),
18        style='pytorch',
19        frozen_stages=-1,
20        conv_cfg=None,
21        norm_cfg=dict(type='BN', requires_grad=True),
22        norm_eval=True,
23        dcn=None,
24        stage_with_dcn=(False, False, False, False),
25        gcb=None,
26        stage_with_gcb=(False, False, False, False),
27        gen_attention=None,
28        stage_with_gen_attention=(), (), (), (),
29        with_cp=False,
30        zero_init_residual=True):
31        super(ResNet, self).__init__()
32        # num_stages: 下采样特征层数目
33        # strides: 不同残差block的stride数
34        # out_indices: 需要的特征层索引, 对齐fpn
35        # frozen_stages: 冻结的残差层
36        # stage_with_dcn, stage_with_gcb, stage_with_gen_attention均需要和fpn层对齐
37
38    def forward(self, x):
39        # 首先是两个下采样: kernel_size=7和maxpool, 此时尺度为原始1/4
40        x = self.conv1(x)
41        x = self.norm1(x)
42        x = self.relu(x)
43        x = self.maxpool(x)
44        # 然后是几个尺度减半的特征层 make_res_layer, 对应fpn层
45        outs = []
46        for i, layer_name in enumerate(self.res_layers):

```

```

47     res_layer = getattr(self, layer_name)
48     x = res_layer(x)
49     if i in self.out_indices:
50         # out_indices 选取需要的层对应fpn层(通常index是连续的,除非数据特殊,
           刚好都只有一大一小物体)
51         outs.append(x)
52     return tuple(outs)

```

1.4 Necks

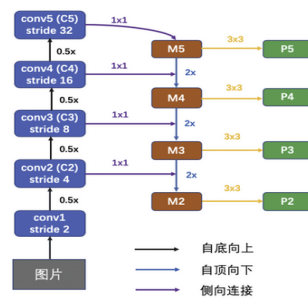


图 8: fpn

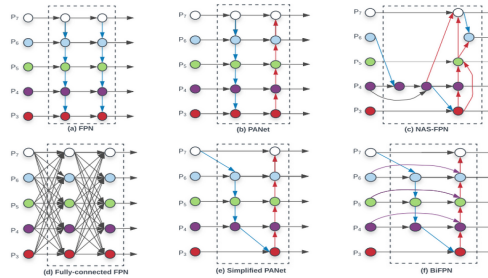


图 9: bifpn

1.4.1 多尺度结构及改进

提到数字图像的多尺度，首先想到的是 David Lowe 于 1999 发表的 SIFT 算法。而 DL 中的多尺度结构 FPN，算是接了 SIFT 的衣钵。2017 年，Tsung-Yi Lin 等人提出 FPN，其摘要中给出解释: A top-down architecture

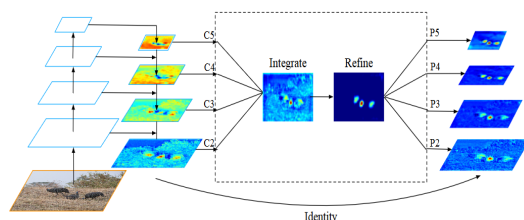


Figure 4: Pipeline and heatmap visualization of balanced feature pyramid.

图 10: bfpn

with lateral connections is developed for building high-level semantic feature maps at all scales。从中我们看到几个关键词，自顶向下，横向连接，高级语义特征，各种尺度。

想象一下，一张图被 $1/2$ 下采样了几次，那么同样的物体在不同从尺度上，表示的信息将至少对半折少 (像素信息)，那么合并各层信息将有助于有效信息的叠加，抑制某些物体的消失。因此容易看出这几个关键词的一些可变化的地方。自顶向下，自底向上，连接方式可从全连接开始缩减，获取高级语义特征的方式，不同尺度特征对齐的方式，更多尺度。若采用事后诸葛亮的逻辑，那么后续确实有一些文章是在围绕这个几个点做改进工作。比如看一眼9，不过 BiFPN增加了各层的权重因子，这和后来横飞的 Attention是一致的。bifpn 的实现参考YAEDet.Pytorch，如果想看更好看的实现，可参考bifpn.pytorch，实现细节上和论文有出入 (YAEDet.Pytorch)，所以自行采坑。类似于简化全连接可参考HRFPN，看图就行，大同小异。另外 mmdet 中实现了BFPN，从10可以看到，是对原始 fpn 后的特征做了一次残差操作。具体就是特征做了一次合并和解耦 (融合，分离)，然后和输入特征相加。

fpn 的改进远不止以上内容，因为其他很多改进可以合并到1.4.2小目标改进上，所以这里暂时省去。

最后，FPN 的代码相对比较简单，这里选择三点说明。

```

1 # 1. 一行代码:
2 # add extra conv layers (e.g., RetinaNet)
3 extra_levels = num_outs - self.backbone_end_level + self.start_level
4
5 # 参数说明, num_outs: 最终fpn输出的尺度个数, backbone_end_level: 输入尺度的
   # 第几层作为基础fpn输出的最后一层,
6 # start_level: 输入尺度作为基础fpn输出尺度的开始层
7 # 上面没解释清楚, 本来fpn没有额外加层的, 你可以理解基础fpn就是原始的fpn, 输

```

```

    入多少个层，选一个开始和结束，作为截断就是
8 #输出结果，现在有了额外层，所以，fpn从基础fpn变为最终fpn。num_outs就是你想输出
    出的最终尺度个数，额外差的尺度特征，从
9 # 基础fpn被截断的层数(x=backbone_end_level - start_level)末尾开始，往下再下
    采样y=num_outs-x)即得到最终的fpn特征。
10
11 # 2. 一个基础模块：ConvModule，有conv_cfg， norm_cfg， act_cfg三个配置， 作为
    最小卷积配置单元，注意一下。
12
13 # 3. 额外加层：add_extra_convs为False时，Maskrcnn，Fasterrcnn 会在基础fpn上
    外增加num_outs - used_backbone_levels(x,如上)个max_pool层，
14 # 为True时，Retina增加额外y层3*3， stride=2的基础ConvModule模块。

```

另外经典 fpn 的自顶向下连接方式从8图中容易理解，注意代码中使用的 F.interpolate 支持 3,4,5 维的张量。

1.4.2 小目标

这里只是随便一写，待修改。小目标建议 small aug small obj aug fpn on small obj SNIPER training

可以看到，以下都有论文，没代码。small obj paper

small obj detection code

以上内容粗略总结起来就是，根据标注 copy 出更多的小目标，分尺度采样训练 (SNIPER)，head 增加 SSH 等特征融合模块，改写 bbox loss，anchor 的细致设计 (faceboxes 的不同尺度上 anchor 均衡)，使用浅层特征，提高特征分辨率等有用，其他基本没啥用。

1.5 Heads

Heads 主要包含了是三部分,anchor_heads, bbox_heads, mask_heads, anchor 是 two-stage 的 rpn 部分或者 one-stage 的头部. 最终的损失函数均在这里实现.

1.5.1 AnchorHead

anchor head 主要包括 AnchorGenerator, anchor_target, 后者包含了 assign,sample. 因为不同尺度操作雷同, 于是作者设计了一个公共函数 multi_apply. 这一小节主要说明以上函数, 损失函数见1.6.

AnchorGenerator 类为不同特征层生成 anchor, 输入参数 base_size, scales, ratios 分别表示: anchor 在特征层上的基础大小 (特征层相对于原图的 stride), anchor 在特征层上的尺度大小 (可以多个, 增加感受野), anchor 在保持基础大小不变的情况下的长宽比. $\text{len}(\text{scales}) * \text{len}(\text{ratios})$ 即为一格的 anchor 个数.

比如输入图像大小 (640*640), 选择 (p2, p3) 作为其特征层, 则 p2 大小为 (160*160), base_size=4, 若设定 ratios=[0.5,1.0,2.0], scales=[8, 16], 则在 p2 上一格对应的 base_anchor 的 (w,h) 为 [(45.25,22.63), (90.51, 45.25), (32.00, 32.00), (64.00, 64.00), (22.63, 45.25), (45.25, 90.51)]. 其中 $64 = 4 * 16 * 1$, $90.51 = 4 * 16 * \sqrt{2}$, $22.63 = 4 * 8 / \sqrt{2}$. 那么每一格所对应的 6 个 base_anchor 相对于中心点的偏移量即为:

```
1
2         [[-21.,  -9.,  24.,  12.],
3          [-43., -21.,  46.,  24.],
4          [-14., -14.,  17.,  17.],
5          [-30., -30.,  33.,  33.],
6          [-9.,  -21.,  12.,  24.],
7          [-21., -43.,  24.,  46.]]
```

因为这里都是相对于自身一格的偏移量, 所有当要算在原图上的中心偏移量时, 直接根据 torch 或者 numpy 的广播机制, 首先得到特征图在原图的网格坐标, 然后和此相加, 即得此特征层的所有 all_anchors. 见 grid_anchors 函数. 这里面涉及到一些操作技巧, 比如:

```
1 ws = (w * w_ratios[:, None] * self.scales[None, :]).view(-1)
2 hs = (h * h_ratios[:, None] * self.scales[None, :]).view(-1)
```

包括 nms,assign 等部分都有很多细节操作, 这些实现检测算法的基础函数, 需要多加练, 才能复现新的算法. 可见八. 如上不难看到, 输入图像的所有 anchor 数量巨大.

模块 anchor 中还有一个 anchor_target 文件, 其主要是为了给设计的 anchor 标定 label 以及筛选 anchor(便于优化等).

这里利用的 anchor_target_single 函数来解释 multi_apply 函数:

```
1 from functools import partial
2 from six.moves import map, zip
3 def multi_apply(func, *args, **kwargs):
4     pfunc = partial(func, **kwargs) if kwargs else func
5     map_results = map(pfunc, *args)
6     return tuple(map(list, zip(*map_results)))
```

partial 函数的功能就是: 把一个函数的某些参数给固定住, 返回一个新的函数 (这里就是将 `**kwargs` 中的参数固定住). 当函数参数太多, 需要固定某些参数时, 可以使用 `functools.partial` 创建一个新的函数 `map(function, sequence)`, 对 `sequence` 中的 `item` 依次执行 `function(item)`, 并将结果组成一个迭代器返回, 最后 `zip` 并行循环.

在 `anchor_target` 中

```
1 result = multi_apply(
2     anchor_target_single,
3     anchor_list,
4     valid_flag_list,
5     gt_bboxes_list,
6     gt_bboxes_ignore_list,
7     gt_labels_list,
8     img_metas,
9     target_means=target_means,
10    target_stddevs=target_stddevs,
11    cfg=cfg,
12    label_channels=label_channels,
13    sampling=sampling,
14    unmap_outputs=unmap_outputs)
```

是先将配置文件, 采样与否的 `flag`, `anchor` 于 `gt_box` 在映射空间 (平移归一, 放缩 `log`) 中各偏差量的滑动平均处理 (见 `core/box/transforms/box2delta` 函数最后一行) 固定住. 然后 `anchor_target_single` 对 `*args` 参数: `anchor_list`, `gt_bboxes_list`, `img_metas` 等并行处理. 得到最终的 `label`, `bbox`, 正负样本 `index`. 这里 `label_weights` 用来记录筛选后的信息位置.

`*args` 参数的说明:

- `anchor_list`: 见 `anchor_head` 中的 `get_anchor` 函数, 是一个 `list[list[Tensors]]` 结构, 最外层是图片个数, 再内一层是尺度个数, 里面的 `Tensors` 的 `shape` 是 `[H*W*4, 4]`, 其中 `H` 和 `W` 代表对应尺度特征图的高和宽.
- `img_metas` 有五个字段: `ori_img_shape`, `scale_factor`, `flip`

在调用 `anchor_target_single` 函数之前:

```
1 num_imgs = len(img_metas)
2 assert len(anchor_list) == len(valid_flag_list) == num_imgs
3 # anchor number of multi levels
4 # 选出第一张图的各个特征层的anchor数量, 比如上面给的p2即为160*160*6, 一格点对应的anchor shape为(6,4)
```

```

5 # 因为有多层,比如p2,p3, 则[(160*160*4,4), (80*80*6,4)]
6 # 为images_to_levels函数,切片用:将所有以图片为第一维度的结果,转换成以特征图
   尺度个数为第一维度的结果(算loss)
7 num_level_anchors = [anchors.size(0) for anchors in anchor_list[0]]
8 # concat all level anchors and flags to a single tensor
9 #在调用multi\_apply之前需要将anchor\_list的不同层级的anchor cat在一起,最终变
   为list [Tensor] 结构.
10 #最外层为图像个数.这样就能利用multi\_apply中的map在每张图上做做同样的操作了.
11 #将所有图的anchor合在一起
12 for i in range(num_imgs):
13     assert len(anchor_list[i]) == len(valid_flag_list[i])
14     anchor_list[i] = torch.cat(anchor_list[i])
15     valid_flag_list[i] = torch.cat(valid_flag_list[i])
16 # valid_flag_list 结合meta信息,进一步筛选无效anchor(其实筛掉的很少)

```

anchor_target_single 函数: 主要涉及 assign_and_sample, bbox2delta, unmap 四个函数.

assign_and_sample 根据 cfg 从配置信息拿到对应的 assign, sample 类对象, one-stage 模型没有 sample(不代表损失函数利用所有的 anchor), sample 主要为 two-stage 服务.

采样主要包含 RandomSampler, OHEMSampler, InstanceBalancedPosSampler, IoUBalancedNegSampler, 均继承自 BaseSampler 为一抽象类, 继承自它的的类必须完成 _sample_pos, _sample_neg 两函数. 按照 python 的语法, 子类可以使用 raise NotImplementedError 来避免不能实例化的问题. PseudoSampler 即是如此, 它重写了 sample 函数, 该函数并没有做任何筛选. RandomSampler 的意思自明, 其他采样后补. 标定 MaxIoUAssigner 见 1.7.1.

bbox2delta 主要注意一点的是相对位置是相对于 anchor 的, 所以除的是 anchor 的 w, h

$$d_x = (g_x - p_x)/p_w, d_y = (g_y - p_y)/p_h, d_w = \log(g_w/p_w), d_h = \log(g_h/p_h).$$

其中 x, y 是中心坐标. 这里涉及到一些优化目标函数的变量的”变换”问题, 比如这里若 $g_w - p_w = 1 + \epsilon$, 则 $d_w = \epsilon$. 如此似乎可以将其替换为 $d_w = g_w/p_w - 1$, 实际上这样的效果可能还是不及 \log , 因为一开始它就将变化幅度压缩了. 另外, 相对位置的归一化差异是关键, 比如人脸识别中, landmarks 的优化, 同常损失函数为欧氏距离, 如何对变量进行处理? 若是先预测 bbox, 再预测 landmarks, 可以相对于 bbox 的中心坐标做变换, 若直接预测呢?

1.5.2 SSDHead

ssd 结构的检测网络, 目前已有 ssd300,ssd512, 结构细节参考???. 从配置文件中可有看到, 它没有 neck, 因层级结构在 backbone 实现.

ssdhead 继承自 anchorhead, 主要功能为处理多层级特征上的 anchor 构造和 target 标定与筛选, 基本的 featuremap 上的 anchor 生成由 mmdet.core.anchor 中的 AnchorGenerator 完成, 优化目标 anchor 由 anchor_target 完成. ssdhead 中 forward 前向返回各层级对应的类别分数和坐标信息, loss 函数则得到对应的损失函数, 以字典的形式返回, 最终求导时, 汇总成一个值, 同时也能计算各个部分损失函数的均值, 方差, 方便优化, debug.

此处的难点在于 anchor 的设定和 target 的标定, 筛选. 现就 anchor 这一块细说如下:

anchor 基本介绍: anchor 设计和 caffe ssd anchor 设计一致, 假设 min_size 为 a , max_size 为 b , 则先生成 ratio 为 1, 宽度和高度为 $(a, a), (\sqrt{ab}, \sqrt{ab})$ 的两个 anchor, ratio 为 2, 1/2, 3, 1/3 则分别生成宽度和高度为 $(a * \sqrt{ratio}, a / \sqrt{ratio})$ 的 anchor, mmdetection 中必须设定每一层的 min_size, max_size, 因此 ratios 为 [2] 则对应 4 个 anchor, ratios 为 [2, 3] 则对应 6 个 anchor.

在 init() 函数中, 先生成 min_size, max_size, 注意它这里是必须要指定 max_size(和 caffe SSD 不同, 无法生成奇数个 anchor), 确保 len(min_size)=len(max_size), 调用 AnchorGenerator() 类生成了 base_anchors, 数量是 6 或者 10, 使用 indices 操作从 6 个 anchor 里选择 (0, 3, 1, 2) 或者从 10 个 anchor 里选择 (0, 5, 1, 2, 3, 4) → 最终生成 4 个或者 6 个 anchor. 于在多个 feature map 上生成 anchor, 因此使用了一个 for 循环操作, 将 anchor_generator 放入到 anchor_generators[] 中.

AnchorGenerator 类, init() 函数需要如下参数:

- base_size: 即设置的 min_size
- scales: 是 $(1, \sqrt{max_size/min_size})$, 用来生成 ratio 为 1 的两个 anchor
- ratios: 是 (1, 2, 1/2) 或者 (1, 2, 1/2, 3, 1/3)
- ctr: ctr 由 stride 生成, 是 anchor 的中心坐标, $(\frac{stride-1}{2}, \frac{stride-1}{2})$ 在 gen_base_anchor() 函数里, 使用上面的参数来计算 base_anchor, 计算流程如下:

- 根据 ratios 来计算 h_ratios 和 w_ratios, 即上面所述的 $(1/\sqrt{ratios}, \sqrt{ratios})$.
- 根据 scales 来计算 base_size, 一共有 2 个分别是

$$(min_size, \sqrt{min_size * max_size}) = min_size * scales$$

- 计算 anchors 的宽度和高度, 只以宽度举例: $w = base_size * w_ratios$, 以 ratios 是 (1, 2, 1/2) 举例, base_size shape 为 (2, 1), w_ratios shape 为 (1, 3), 计算出的 w 是 (2, 3) 一共生成了 6 个 anchor, 如果 ratios 是 (1, 2, 1/2, 3, 1/3), 则生成 10 个 anchor (此处 anchor 数量和标准 ssd anchor 数量不一致 → 再筛选 (即 ssd_head.py 中使用 indices 操作进行筛选))

1.5.3 RetinaHead

后补。

1.5.4 GuidedHead

guided_anchor_head.py, ga_retina_head.py

1.5.5 FCOSHead

2.0 版 fcos 代码合并到了 singlestage, 这里就省略代码解析了。

将开始写的 fcos 论文阅读写的文字帖到这里, 只能和这一节的整体结构不违和一下了。

物体检测的算法基本可以用以下这段话来描述:

在某一度量 (..) 下, 找出与人为标定的物体的框的集合, 利用这个集合按不同策略生成更有效的正负样本, 完成物体的位置预测。

以 IOU 为度量的模型有 Faster R-CNN, SSD 和 YOLOv2, v3 等. 它从标定框与虚拟框的偏移集中学习物体的标定框与虚拟框的映射关系, 以此来完成物体定位. 这里的隐含点为映射变量为物体的表示向量, 这是根据类别标签优化而得. FCOS 可以称为以 center-ness 为度量的检测模型. 它将标定框内的聚中心的所有点作为物体的有效表示点, 学习其与标定框的中心偏移量, 完成物体的检测。

因为两者度量方式不同, 导致后续操作差别比较大, iou 度量的 anchor-based 模型, 需要给每个位置点设计不同形状的 anchor 作为候选框, 然后在所有尺度上进行筛选 (或不同尺度分别筛选), 这个过程较为细致, 且超参数较多, 对不同任务较为敏感. 后者 anchor-free 则省去了 anchor 构造的过程, 可以直接在选出的点上做位置偏移预测, 比较简单.

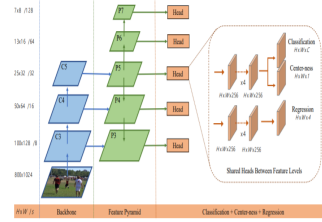


Figure 2 - The network architecture of FCOS, where C3, C4, and C5 denote the feature maps of the backbone network and P3 to P7 are the feature levels used for the final prediction. $H \times W$ is the height and width of feature maps. 2^a ($a = 0, 16, \dots, 128$) is the down-sampling ratio of the feature maps at the level to the input image. As an example, all the numbers are computed with an 800×1024 input.

图 11: fcoss

从以上结构图能清晰的看到 FCOS 算法的三大模块. 它的核心点在 Classification + Center-ness + Regression. (代码以 maskrcnn-benchmark 为基础, 核心在 rpn/fcos 的三个文件中 (后续).)

这里有点需要说明:

- 限制不同特征层的预测框范围 $\{P_3, P_4, P_5, P_6, P_7\}$ 对应范围区间 0, 64, 128, 256, 512 and ∞ .
- P_i 层的位置回归做如下动态调整 $\exp(s_i x)$ 以适应不同特征层回归不同范围.
- 位置预测的是中心点距离物体左上下右的距离, 且位置回归函数是 UnitBox 中的 IOU loss.

$$\square \text{ Ground that } \tilde{x} = (\tilde{x}_t, x_b, \tilde{x}_l, \tilde{x}_r)$$

$$\square \text{ predictors } x = (x_t, x_b, x_l, x_r)$$

- IOU loss:

$$\ell_2 \text{ loss} = \|\square - \square\|_2^2$$

$$\text{IoU loss} = -\ln \frac{\text{Intersection}(\square, \square)}{\text{Union}(\square, \square)}$$

- centerness $^* = \sqrt{\frac{\min(l^*, r^*)}{\max(l^*, r^*)} \times \frac{\min(t^*, b^*)}{\max(t^*, b^*)}}$ 以此来筛选更有代表的物体表示点.

- 在多个 box 内的点, 选择 box 面积最小的那个作为其 label box.

最终损失函数为:

$$L(\{\mathbf{p}_{x,y}\}, \{\mathbf{t}_{x,y}\}) = \frac{1}{N_{\text{pos}}} \sum_{x,y} L_{\text{cls}}(\mathbf{p}_{x,y}, c_{x,y}^*) + \frac{\lambda}{N_{\text{pos}}} \sum_{x,y} \mathbb{I}_{\{c_{x,y}^* > 0\}} L_{\text{reg}}(\mathbf{t}_{x,y}, \mathbf{t}_{x,y}^*)$$

1.6 Losses

1.6.1 基本认识

Fcoalloss. 想像一下特征层上的锚框, 远离 gt 的必然占绝大多数, 围绕在 gt 周围的 bbox 有模棱两可的 (正负), 故有正负难易之分. FcoalLoss 是为解决难易样本的不平衡问题. 公式如下:

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t), \text{ 其中 } p_t = \begin{cases} p & \text{if } y = 1 \\ 1-p & \text{else.} \end{cases} \quad (1)$$

α_t 缓解样本不平衡现象, $(1-p_t)^\gamma$ 为降低易分样本的损失值 (考虑到易分样本比例高). 公式可以这样理解: 对于正样本分对了, 则 $p_t \rightarrow 1$, $(1-p_t)^\gamma$ 有减缓效果, 当分错了, 即 $p_t \ll 0.5$, 则 $1-p_t \gg 0.5$, 效果和 CrossEntropy 没差, 对于负样本也是如此. 因此起到压缩图 6 中的左图右端, 右图左端的效果.

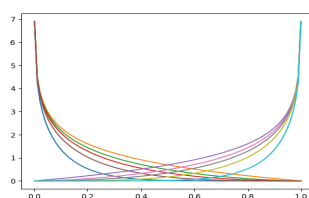


图 12: FcoalLoss

GHM. GHM 算是对 Fcoal 改进, 作者统计了样本的梯度信息, 提出梯度均衡机制, 让各种难度类型的样本有均衡的累计贡献. 具体细节看论文即可.

IoULoss. **BoundedIoULoss**.

1.6.2 实现解析

`mmdet.models.losses` 里面实现了所有损失函数。

`losses.utils` 有两个基本函数, `weight_reduce_loss` 和 `weighted_loss`, 前者将返回的损失向量点乘一权重向量, 再分别求和或者算平均值。后者为一装饰器, 装饰你自己定义的损失函数, 实现带权重效果。具体为先算自定义 `loss_func`, 然后再 `weight_reduce_loss` 一下。其中 `weighted_loss` 中的 `@functools.wraps` 装饰的好处是返回的函数保持被装饰函数名字。接下来逐一说明各个损失函数。

`CrossEntropyLoss`, 其理论来源有两点, 信息论或极大似然估计, 可参考花书。pytorch 官方文档给的公式, 明显看出是极大似然的写法。注意实际运算是张(向)量的形式。 $\text{loss}(x, \text{class}) = -\log\left(\frac{\exp(x[\text{class}])}{\sum_j \exp(x[j])}\right) = -x[\text{class}] + \log\left(\sum_j \exp(x[j])\right)$, 增加权重因子为

$$\text{loss}(x, \text{class}) = \text{weight}[\text{class}] \left(-x[\text{class}] + \log\left(\sum_j \exp(x[j])\right) \right).$$

在 `cross_entropy_loss.py` 中实现了三种交叉熵损失函数, `binary_cross_entropy`, `mask_cross_entropy` 以及 `cross_entropy`, 前两者相同点在于将标签扩展成 one hot 形式然后调用 `F.binary_cross_entropy_with_logits` 函数, 不同点为 `mask` 没有乘权重向量, `cross_entropy` 调用 `F.cross_entropy`, 并被 `weight_reduce_loss` 了。需要注意的点是, 最终的 `CrossEntropyLoss` 中的 `loss_weight` 参数是多任务中此 Loss 的权重。而 `weight_reduce_loss` 中的 `weight` 是各类别的权重。可以说前者 and 样本平衡有关, 后者和任务平衡有关。

`SmoothL1Loss` 来源于 FastRCNN, 用于解决方框回归的不稳定问题(也许是优化速度), 由 L2, L1 演变而来。

$$L_1 = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & x < -1 \text{ or } x > 1 \end{cases}$$

三者的关系, 画个图其意自明: L2 两边变化太快, 对不好优化的 anchor 不友好(离群点), L1 0 点不可导, Smooth 两边较为平缓。在 `smooth_l1_loss.py` 中, `smooth_l1_loss` 被上述 `weighted_loss` 装饰, 实现代码比较简单:

```
1 diff = torch.abs(pred - target)
2 loss = torch.where(diff < beta, 0.5 * diff * diff / beta,
3                       diff - 0.5 * beta)
```

注意这里的 where 实现了分段函数，所以其含义自明。

BalancedL1Loss 来自 Libra R-CNN，为 SmoothL1 的改进。改进函数为

$$L_b(x) = \begin{cases} \frac{\alpha}{b}(b|x| + 1) \ln(b|x| + 1) - \alpha|x| & \text{if } |x| < 1 \\ \gamma|x| + C & \text{otherwise} \end{cases}$$

其梯度为：

$$\frac{\partial L_b}{\partial x} = \begin{cases} \alpha \ln(b|x| + 1) & \text{if } |x| < 1 \\ \gamma & \text{otherwise} \end{cases}$$

其中 $b = e^{\frac{\gamma}{\alpha}}$ ，使得函数连续。容易得到在 $|x| < 1$ 附近的梯度比 SmoothL1 要大。具体分析可参考 [BalancedL1Loss](#)，代码和 SmoothL1 一致，故析略。

Focalloss 原理见上节，其扩展有 [CenterNet](#), GHM 系列。因 `_sigmoid_focal_loss` 由 cpp,cu 实现，故略（可看看 debug 版本）。

GHMLoss 有 GHMC, GHMR，分别作用于分类，回归。原理查看论文或者参考 [GHM](#)，其主要思想是简单和特难的样本均进行抑制，也算是 FocalLoss 的改进，析码略。

剩下的 IoULoss, BoundedIoULoss, GIoULoss 等以 iou 的方式进行 Box 回归，含义自现。三者的细致分析（各自优缺点及关系）看原始论文。代码上根据如下公式，即可容易写出。Giou:

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

$$GIoU = IoU - \frac{|C \setminus (A \cup B)|}{|C|}$$

其中 C 为 A, B 的闭包。在 `iou_loss.py` 中，最终的 `iou_loss` 为 $-\log(ious)$ ，`giou_loss` 为 $1 - giours$ ，`boundedloss` 和 `SmoothL1Loss` 差不多，只是此时被求的变量 (dx, dy, dw, dh) 的编码方式变了。具体关系如下：

$$\text{论文中的分段函数为 Huber loss } L_{\tau}(z) = \begin{cases} \frac{1}{2}z^2 & |z| < \tau \\ \tau|z| - \frac{1}{2}\tau^2 & \text{otherwise} \end{cases}$$

和 SmoothL1 类似，但代码用的是带了一个 β 因子的 SmoothL1。

分量损失为 $\text{cost}_i = 2L_1(1 - \text{IoU}_B(i, b_i)), i \in \{x, y, w, h\}$ 。原始的 $L1$ (见 FasterRCNN) 定义为

$$\text{cost}_x = L_1\left(\frac{\Delta x}{w_s}\right)$$

$$\text{cost}_w = L_1\left(\ln\left(\frac{w}{w_t}\right)\right)$$

这里使用的 IoU_B 定义为:

$$\begin{aligned}\text{IoU}_B(x, b_t) &= \max\left(0, \frac{w_t - 2|\Delta x|}{w_t + 2|\Delta x|}\right) \\ \text{IoU}_B(w, b_t) &= \min\left(\frac{w}{w_t}, \frac{w_t}{w}\right)\end{aligned}$$

最终分别算出 x, y, w, h 的各自 IoU_B 损失”, 作为最终损失的分量, 带入 SmoothL1 分段函数中即可。

以上的代码中出现了 `new_full, nonzero, numel, expand, where` 等常用函数。

1.7 Detectors

这节主要分析 maskrcnn 和 reppoints, retinaFace 三个算法。

1.7.1 maskrcnn

以配置文件 `mask_rcnn_r50_fpn_1x.py` 为例说说 `twao_stage` 的实现过程. 配合 `two_stage` 的 `forward_train()` 函数和配置文件, 即可.

首先 backbone 为 resnet50, (resnet 系列结构参见3), 其以 tuple 形式返回 4 个 stage 的特征图, 片段代码如下:

```
1
2 outs = []
3 for i, layer_name in enumerate(self.res_layers):
4     res_layer = getattr(self, layer_name)
5     x = res_layer(x)
6     if i in self.out_indices:
7         outs.append(x)
```

然后 neck 为 fpn, 结构参见8, fpn 根据 config 中的 `out_indices` 取出以 resnet50 输出的对应 stage, 分别构造输出 channel 维度统一的卷积算子, 然后按照8所示融合方式进行不同尺度的特征融合, 以元组形式输出结果. 在配置信息里有一条 `num_outs=5`, 是为 mask-rcnn 在最顶层特征增加的最大池化特征输出. 以上两块为提取特征, 被 `extract_feat` 整合在一块,

紧接着 `forward_train` 中包含了剩下的所有流程.

`rpn_head` \rightarrow `rpn_head.loss` \rightarrow `rpn_head.get_bboxes` \rightarrow `assign` \rightarrow `sample`
 \rightarrow `bbox_roi_extractor` \rightarrow `bbox_head` \rightarrow `bbox_head.get_target.` \rightarrow `bbox_head.loss` \rightarrow `mask_roi_extractor` \rightarrow `mask_head` \rightarrow `mask_head.get_target`

这里梳理一下部分函数.

候选框层 RPN,RPNHead 继承 AnchorHead, 它的几个核心操作都在 anchor_head.py 中实现, 主要包括 get_anchors, anchor_target 见1.5.1, 函数 get_bboxes 结合配置参数从 rpn 前向得到的 2 分类和位置预测结果中筛选出最终的 proposals.

get_bboxes 中先通过 self.anchor_generators[i].grid_anchors() 这个函数取到所有的 anchor_boxes, 再通过 self.get_bboxes_single() 根据 rpn 前向的结果选出候选框, 在 self.get_bboxes_single() 中, 先在每个尺度上取 2000(配置) 个 anchor 出来, concat 到一起作为该图像的 anchor, 对这些 anchor boxes 作 nms(thr=0.7) 就得到了所需的候选框. 需注意预测的 bbox 是对数化了的, 在做 iou 计算之前需用 delta2bbox() 函数进行逆变换.bbox_head 中的 bbox2roi 类似.

得到的候选框最终由配置中 train_cfg 的 rcnn.assigner, rcnn.sampler 进行标定和筛选, 保持正负样本平衡和框的质量, 方便优化.

MaxIoUAssigner:

1. 所有候选框置-1
2. 将与所有 gt bbox 的 iou 小于 neg_iou_thr 置 0
3. iou 大于 pos_iou_thr 的将其匹配
4. 为了避免标定框无训练目标, 将 gt bbox 匹配于与它 iou 最近的 bbox(会导致部分正样本的匹配 iou 值很小).

```

1  # 交并比矩阵(n,m), gt=n, bboxes=m
2  overlaps = bbox_overlaps(gt_bboxes, bboxes)
3  # 每个bbox和所有gt的最大交并比,(m,)
4  max_overlaps, argmax_overlaps = overlaps.max(dim=0)
5  # 每个gt和所有bbox的最大交并比
6  gt_max_overlaps, gt_argmax_overlaps = overlaps.max(dim=1)
7  # 1将所有bbox赋值为-1,注意new_full操作
8  assigned_gt_inds = overlaps.new_full(
9      (num_bboxes, ), -1, dtype=torch.long)
10 # 2交并比大于0同时小于负阈值的赋值为0
11 assigned_gt_inds[(max_overlaps >= 0)
12                  & (max_overlaps < self.neg_iou_thr)] = 0
13 # 将与gt交并比大于正阈值的赋值为1(可能没有)
14 pos_inds = max_overlaps >= self.pos_iou_thr
15 assigned_gt_inds[pos_inds] = argmax_overlaps[pos_inds] + 1
16 # 保证每个gt至少对应一个bbox
17 # 遍历gt,将与gt最近(max(iou))的bbox,将gt的label赋值给此bbox
18

```

```

19 for i in range(num_gts):
20     if gt_max_overlaps[i] >= self.min_pos_iou:
21         # 此判断较迷
22         max_iou_inds = overlaps[i, :] == gt_max_overlaps[i]
23         assigned_gt_inds[max_iou_inds] = i + 1
24         # 与gt最大iou的bbox 赋值为i+1

```

RandomSampler, 保持设定的平衡比例, 随机采样.

然后通过 SingleRoIExtractor(roi_extractors/single_level.py) 统一 RoI Align 四个尺度且大小不同的 proposals, 使其大小为 $7*7(\text{bbox})$ 或 $14*14(\text{mask})$. 配置信息 rpn_head 中的 anchor_strides 为 5 个尺度, 包含了 fpn 额外加入的最大池化层, 而 bbox_roi_extractor 的 featmap_strides 却只包含四个尺度, 表明只需对前四层进行 align. 最终送入 bbox head 和 mask head 做第二次优化 (two stage).

RoIAlign 在 ops 中, 经 cuda 加速, 详解待后. 其中 roi_extractors 中的特征层级映射函数如下:

```

1 def map_roi_levels(self, rois, num_levels):
2     """Map rois to corresponding feature levels by scales.
3     self.finest_scale = 56, 映射到0级的阈值
4     $(0, 56, 56*2, 56*4, \infty) \rightarrow (0, 1, 2, 3)$
5     bbox2roi变换后的 rois
6
7     Returns:
8     Tensor: Level index (0-based) of each RoI, shape (k, )
9     因不同层级对应不同的ROIAlign
10    """
11    scale = torch.sqrt(
12        (rois[:, 3] - rois[:, 1] + 1) * (rois[:, 4] - rois[:, 2] + 1))
13    target_lvls = torch.floor(torch.log2(scale / self.finest_scale + 1e-6))
14    target_lvls = target_lvls.clamp(min=0, max=num_levels - 1).long()
15    # 这个变换在原始论文中有.
16    return target_lvls

```

1.7.2 RepPoints

2019 年, Ze Yang 等人利用 DCNv2 的特性, 实现了结构点表示物体的检测算法 RepPoints. 其中 DCNv2 从 DCNv1 进化而来, 出发点如 13 所示, 改变了卷积的采样点. 其卷积数学表达式如下: $y(p) = \sum_{k=1}^K w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k$ 对应的池化层为: $y(k) = \sum_{j=1}^{n_k} x(p_{kj} + \Delta p_k) \cdot \Delta m_k / n_k$ 其中 $w_k, p, p_k, \Delta p_k, \Delta m_k$ 分别表示卷积核参数, 卷积中心点 (图上), 当前

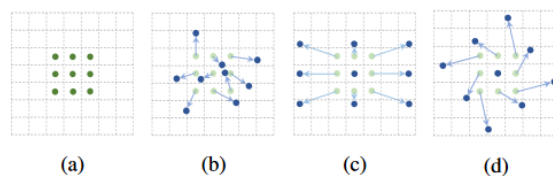


Figure 1: Illustration of the sampling locations in 3×3 standard and deformable convolutions. (a) regular sampling grid (green points) of standard convolution. (b) deformed sampling locations (dark blue points) with augmented offsets (light blue arrows) in deformable convolution. (c)(d) are special cases of (b), showing that the deformable convolution generalizes various transformations for scale, (anisotropic) aspect ratio and rotation.

图 13: dcn

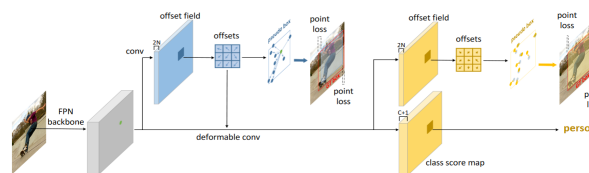


Figure 2: Overview of the proposed RPDet (RepPoints detector). While feature pyramidal networks (FPN) [24] are adopted as the backbone, we only draw the afterwards pipeline of one scale of FPN feature maps for clear illustration. Note all scales of FPN feature maps share the same afterwards network architecture and the same model weights.

图 14: reppoints

卷积点，当前点的偏移量，当前点的权重因子，就这么简单。不规则采样点和图像中物体的不规则形状是吻合的，所以直观上讲，这个卷积是很有意义的。实验上你可以将以前的 conv 策略性的替换为 dcn，也许会带来性能的提升 (ps 后续很多新检测算法实验，证明了这一点)。

问题是如何利用 DCNv2 不规则卷积，来做检测算法呢？

1. 通过定位和分类的直接监督来学习可形变卷积的偏移量，使得偏移量具有可解释性 (@ 陀飞轮)。
2. 可以通过采样点来直接生成伪框 (pseudo box)，不需要另外学习边界框，并且分类和定位是有联系的 (@ 陀飞轮)。

从14可以看出，DCN 卷积得到 $2N$ 的偏移量，将其限制到标定的 bbox，类似于 rpn，然后重复 DCN 得到 $2N, C + 1$ 分别作送入 bbox 回归和 class 分类 (refine)。也即上面所说的两点。其中的关键点是学习的 points，需要限制到标定的 bbox 中，也即学习的引导。因为是点表示，所以一个自然的问题是比如人脸检测，那么所得到的点的几何结构是一致的吗？更精细的问题是，可以得到物体的边缘点吗？

这需要你对算法更细致的研究了，可以参考一下[官方说明](#)：

“这里面“可学习的采样点集合”其实是更本质的概念，这些采样点同时用来提取语义对齐的特征，又用来表示物体的几何形态。deformable convolution 是其中一种基于采样点的特征提取方法，但同时还可以有其它的实现。几何形态监督的方式也可以是更多样的，除了物体检测中的 bounding box 标注外，还可以利用更精细的几何标注来进行监督。我们最近有一个工作就提出了不同的特征提取方法，以及不同的几何监督方式。

通常来说，用 anchor 去覆盖 4d 空间是困难的，所以一般需要不同尺度和不同长宽比的多种 anchor，而且赋予这个 anchor box 类别标签也相对麻烦，需要计算和 ground-truth box 的 IoU。与之对应的，要覆盖 2d 空间和赋予类别都很容易，典型的 2d 表示包括 center point, corner point 等等。

回归 vs. 验证是一个很古老的问题，通常来讲求解验证问题比求解回归问题更容易，但是效率更低。在物体检测里面，通常的做法是先做粗验证，后用回归方法来做精确定位。基于 anchor 的设计就是典型的验证思路，后续回归只在一个很小的范围内进行。而 anchor-free 方法更多依赖回归来求解问题，通常讲更难学习，但是因为有了 deep 的特征，以及多阶段的定位方法，效果能比肩依赖验证更多的 anchor-based 方法，同时框架更简洁。

RepPoints 不仅能用来解决物体检测问题，它在表示物体的精细结构上也有很大潜力，例如预测物体的 contour(后续会公布)。”

1.7.3 RetinaFace

再说，和 ssd 重复较大。可以选择放自己改的代码。

第二节 数据处理

2.1 检测分割数据

看看配置文件，数据相关的有 data dict，里面包含了 train,val,test 的路径信息，用于数据类初始化。然后就是 pipeline，将各个函数及对应参数以字典形式放到列表里面，是对 pytorch 原装的 transforms+compose，在检测，分割相关数据上的一次封装，使得形式更加统一。

从 builder.py 中 build_dataset 函数能清晰的看到，构建数据有三种方式，ConcatDataset，RepeatDataset 和从注册器中提取。

其中 dataset_wrappers.py 中 ConcatDataset 和 RepeatDataset 意义自明。前者继承自 pytorch 原始的 ConcatDataset，将多个数据集整合到一起，将不同序列 (可参考[容器的抽象基类](#)) 的长度相加，__getitem__ 函数对应 index 替换一下。后者就是单个数据类 (序列) 的多次重复。就功能来说，前者提高数据丰富度，后者可解决数据太少使得 loading 时间长的问题。而与注册相关的，被注册的数据类在 datasets 下大家熟知的数据类型的 py 文件中。基类为 custom.py 中的 CustomDataset，coco 继承自它，cityscapes 继承自 coco，xml_style 的 XMLDataset 继承 CustomDataset，然后 wider_face，voc 均继承自 XMLDataset。因此这里先分析一下 CustomDataset。

CustomDataset 记录数据路径等信息，解析标注文件，将每一张图的所有信息以字典作为数据结构存在 results 中。然后进入 pipeline，也即数据增强相关操作。从下面代码可以清晰的看到。这里数据结构的选取需要注意一下，字典结构，在数据增强库 albu 中也是如此处理，因此可以快速替换 albu 中的算法。另外每个数据类增加了各自的 evaluate 函数。evaluate 基础函数在 mmdet.core.evaluation 中，后做补充。

```
1 self.pipeline = Compose(pipeline)
2 # Compose是实现了__call__方法的类，其作用是使实例能够像函数一样被调用，同时不影响实例本身的生命周期
```

```

3 def pre_pipeline(self, results):
4     # 扩展字典信息
5     results['img_prefix'] = self.img_prefix
6     results['seg_prefix'] = self.seg_prefix
7     results['proposal_file'] = self.proposal_file
8     results['bbox_fields'] = []
9     results['mask_fields'] = []
10    results['seg_fields'] = []
11
12 def prepare_train_img(self, idx):
13     img_info = self.img_infos[idx]
14     ann_info = self.get_ann_info(idx)
15     # 基本信息, 初始化字典
16     results = dict(img_info=img_info, ann_info=ann_info)
17     if self.proposals is not None:
18         results['proposals'] = self.proposals[idx]
19     self.pre_pipeline(results)
20     return self.pipeline(results)    # 数据增强
21
22 def __getitem__(self, idx):
23     if self.test_mode:
24         return self.prepare_test_img(idx)
25     while True:
26         data = self.prepare_train_img(idx)
27         if data is None:
28             idx = self._rand_another(idx)
29         continue
30     return data

```

mmdet 的数据处理, 字典结构, pipeline, evaluate 是三个关键部分。其他所有类的文件解析部分, 数据筛选等, 看看即可。因为我们知道, pytorch 读取数据, 是将序列转化为迭代器后进行 io 操作的。所以在 dataset 下除了 pipelines 外还有 loader 文件夹, 里面实现了分组, 分布式分组采样方法, 以及调用了 mmcv 中的 collate 函数, 且 build_dataloader 封装的 DataLoader 最后在 train_detector 中被调用, 这部分将在后面补充, 这里说说 pipelines。所有 pipelines 的注册函数在 __init__.py 中可见。

返回 maskrcnn 的配置文件, 可以看到, 训练和测试的不同, LoadAnnotations, MultiScaleFlipAug, DefaultFormatBundle 和 Collect。虽然测试没有 LoadAnnotations, 实际上根据前面的 CustomDataset 可知道, 它仍然需要标注文件。这和 inference 的 pipeline 是不同的。

```

1 # 序列中的dict可以随意删减, 增加, 所谓数据增强调参
2 train_pipeline = [

```

```

3     dict(type='LoadImageFromFile'),
4     dict(type='LoadAnnotations', with_bbox=True, with_mask=True),
5     dict(type='Resize', img_scale=(1333, 800), keep_ratio=True),
6     dict(type='RandomFlip', flip_ratio=0.5),
7     dict(type='Normalize', **img_norm_cfg),
8     dict(type='Pad', size_divisor=32),
9     dict(type='DefaultFormatBundle'),
10    dict(type='Collect', keys=['img', 'gt_bboxes', 'gt_labels', 'gt_masks'])
11 ]
12
13 test_pipeline = [
14     dict(type='LoadImageFromFile'),
15     dict(
16         type='MultiScaleFlipAug',
17         img_scale=(1333, 800),
18         flip=False,
19         transforms=[
20             dict(type='Resize', keep_ratio=True),
21             dict(type='RandomFlip'),
22             dict(type='Normalize', **img_norm_cfg),
23             dict(type='Pad', size_divisor=32),
24             dict(type='ImageToTensor', keys=['img']),
25             dict(type='Collect', keys=['img']),
26         ])
27 ]

```

这些所有操作被 Compose 串联起来 (列表是序列结构)。

```

1 @PIPELINES.register_module
2 class Compose(object):
3
4     def __init__(self, transforms):
5         assert isinstance(transforms, collections.abc.Sequence)
6         self.transforms = []
7         for transform in transforms:
8             if isinstance(transform, dict):
9                 transform = build_from_cfg(transform, PIPELINES)
10                self.transforms.append(transform)
11            elif callable(transform):
12                self.transforms.append(transform)
13            else:
14                raise TypeError('transform must be callable or a dict')
15
16     def __call__(self, data):
17         for t in self.transforms:
18             data = t(data)

```

```

19     if data is None:
20         return None
21     return data

```

上面代码能清晰的看到,配置文件中 pipeline 中的字典传入 build_from_cfg 函数,逐一实现了各个增强类。扩展的增强类均需实现 `__call__` 方法,这和 pytorch 原始方法是一致的。

```

1 def build_from_cfg(cfg, registry, default_args=None):
2     assert isinstance(cfg, dict) and 'type' in cfg
3     assert isinstance(default_args, dict) or default_args is None
4     args = cfg.copy()
5     obj_type = args.pop('type')
6     if mmcv.is_str(obj_type):
7         # 从注册类中拿出obj_type类
8         obj_cls = registry.get(obj_type)
9         if obj_cls is None:
10             raise KeyError('{} is not in the {} registry'.format(
11                 obj_type, registry.name))
12     elif inspect.isclass(obj_type):
13         obj_cls = obj_type
14     else:
15         raise TypeError('type must be a str or valid type, but got {}'.format(
16             type(obj_type)))
17     if default_args is not None:
18         # 增加一些新的参数
19         for name, value in default_args.items():
20             args.setdefault(name, value)
21     return obj_cls(**args) # **args是将字典解析成位置参数(k=v)。

```

有了以上认识,重新梳理一下 pipelines 的逻辑,由三部分组成,load, transforms, 和 format。load 相关的 LoadImageFromFile, LoadAnnotations 都是字典 results 进去,字典 results 出来。具体代码看下便知,LoadImageFromFile 增加了 'filename', 'img', 'img_shape', 'ori_shape', 'pad_shape', 'scale_factor', 'img_norm_cfg' 字段。其中 img 是 numpy 格式。LoadAnnotations 从 results['ann_info'] 中解析出 bboxes, masks, labels 等信息。注意 coco 格式的原始解析来自 pycocotools, 包括其评估方法, 这里关键是字典结构。transforms 中的类作用于字典的 values, 也即数据增强。format 中的 DefaultFormatBundle 是将数据转成 mmcv 扩展的容器类格式 DataContainer。另外 Collect 会根据不同任务的不同配置, 从 results 中选取只含 keys 的信息生成新的字典, 具体看下该类帮助文档。这里看一下从 numpy 转成 tensor 的代码:

```

1 def to_tensor(data):
2     """Convert objects of various python types to :obj:`torch.Tensor`.
3
4     Supported types are: :class:`numpy.ndarray`, :class:`torch.Tensor`,
5     :class:`Sequence`, :class:`int` and :class:`float`.
6     """
7     if isinstance(data, torch.Tensor):
8         return data
9     elif isinstance(data, np.ndarray):
10        return torch.from_numpy(data)
11    elif isinstance(data, Sequence) and not mmcv.is_str(data):
12        return torch.tensor(data)
13    elif isinstance(data, int):
14        return torch.LongTensor([data])
15    elif isinstance(data, float):
16        return torch.FloatTensor([data])
17    else:
18        raise TypeError('type {} cannot be converted to tensor.'.format(
19            type(data)))
20    以上代码告诉我们，基本数据类型，需掌握。

```

那么 DataContainer 是什么呢？它是对 tensor 的封装，将 results 中的 tensor 转成 DataContainer 格式，实际上只是增加了几个 property 函数，cpu_only, stack, padding_value, pad_dims, 其含义自明，以及 size, dim 用来获取数据的维度，形状信息。考虑到序列数据在进入 DataLoader 时，需要以 batch 方式进入模型，那么通常的 collate_fn 会要求 tensor 数据的形状一致。但是这样不是很方便，于是有了 DataContainer。它可以做到载入 GPU 的数据可以保持统一 shape，并被 stack，也可以不 stack，也可以保持原样，或者在非 batch 维度上做 pad。当然这个也要对 default_collate 进行改造，mmcv 在 parallel.collate 中实现了这个。

也许可以看看通常的 collate_fn，它是 DataLoader 中将序列 dataset 组织成 batch 大小的函数。

```

1 def collate_fn_1(batch):
2     # 这是默认的，明显batch中包含相同形状的img\_\_tensor和label
3     return tuple(zip(*batch))
4
5 def coco_collate_2(batch):
6     # 传入的batch数据是被albu增强后的(字典结构)
7     imgs = [s['image'] for s in batch]    # tensor, h, w, c->c, h, w ,
8     handle at transform in __getitem__
9     annots = [s['bboxes'] for s in batch]
10    labels = [s['category_id'] for s in batch]

```

```

10
11 # 以当前batch中图片annot数量的最大值作为标记数据的第二维度值，空出的就补-1
    。
12 max_num_annot = max(len(annot) for annot in annots)
13 annot_padded = np.ones((len(annots), max_num_annot, 5))*-1
14
15 if max_num_annot > 0:
16     for idx, (annot, lab) in enumerate(zip(annots, labels)):
17         if len(annot) > 0:
18             annot_padded[idx, :len(annot), :4] = annot
19             # 不同模型，损失值计算可能不同，这里ssd结构需要改为xyxy格式并且要做
            尺度归一化
20             # 这一步完全可以放到\_\_getitem\_\_中去，只是albu的格式需求问题。
21             annot_padded[idx, :len(annot), 2] += annot_padded[idx, :len(
annot), 0] # xywh→x1,y1,x2,y2 for general box,ssd target assigner
22             annot_padded[idx, :len(annot), 3] += annot_padded[idx, :len(
annot), 1] # contains padded -1 label
23             annot_padded[idx, :len(annot), :] /= 640 # priorbox for
ssd primary target assinger
24             annot_padded[idx, :len(annot), 4] = lab
25 return torch.stack(imgs, 0), torch.FloatTensor(annot_padded)
26
27 def detection_collate_3(batch):
28     targets = []
29     imgs = []
30     for _, sample in enumerate(batch):
31         for _, img_anno in enumerate(sample):
32             if torch.is_tensor(img_anno):
33                 imgs.append(img_anno)
34             elif isinstance(img_anno, np.ndarray):
35                 annos = torch.from_numpy(img_anno).float()
36                 targets.append(annos)
37 return torch.stack(imgs, 0), targets # 做了stack, DataContainer可以
不做stack

```

以上就是数据处理的相关内容。最后再用 DataLoader 封装拆成迭代器，其相关细节，sampler 等就略去了。

```

1 data_loader = DataLoader(
2     dataset,
3     batch_size=batch_size,
4     sampler=sampler,
5     num_workers=num_workers,
6     collate_fn=partial(collate, samples_per_gpu=imgs_per_gpu),
7     pin_memory=False,
8     worker_init_fn=init_fn,
9     **kwargs)

```


第三节 FP16

模型预测加速等.

第四节 训练 pipeline

4.1 训练逻辑

图见2注意它的四个层级. 主要查看 api/train.py, mmcv 中的 runner 相关文件. 主要两个类: Runner 和 Hook Runner 将模型, 批处理函数 batch_processor, 优化器作为基本属性, 是为训练过程中记录相关节点信息, 这些信息均被记录在 mode, _hooks, _epoch, _iter, _inner_iter, _max_epochs, _max_iters 中. 从而实现训练过程中插入不同的操作, 也即各种 hook. 理清训练流程只需看 Runner 的成员函数 run. 在 run 里会根据 mode 按配置 (workflow) epoch 循环调用 train 和 val 函数, 跑完所有的 epoch. 其中 train 代码如下:

```

1 def train(self, data_loader, **kwargs):
2     self.model.train()
3     self.mode = 'train'    # 改变模式
4     self.data_loader = data_loader
5     self._max_iters = self._max_epochs * len(data_loader)    # 最大batch循环次数
6     self.call_hook('before_train_epoch')    # 根据名字获取hook对象函数
7     for i, data_batch in enumerate(data_loader):
8         self._inner_iter = i    # 记录训练迭代轮数
9         self.call_hook('before_train_iter')    # 一个batch前向开始
10        outputs = self.batch_processor(
11            self.model, data_batch, train_mode=True, **kwargs)
12        self.outputs = outputs
13        self.call_hook('after_train_iter')    # 一个batch前向结束
14        self._iter += 1    # 方便resume时, 知道从哪一轮开始优化
15
16    self.call_hook('after_train_epoch')    # 一个epoch结束
17    self._epoch += 1    # 记录训练epoch状态, 方便resume

```

上面让人困惑的是 hook 函数, hook 函数继承自 mmcv 的 Hook 类, 其默认了 6+8+4 个函数, 也即2所示的 6 个层级节点, 外加 2*4 个区分 train 和 val 的节点记录函数, 以及 4 个边界检查函数. 从 train.py 中容易看出, 在训练之前, 已经将需要的 hook 函数注册到 Runner 的 self._hook 中了, 包括从配置文件解析的优化器, 学习率调整函数, 模型保存, 一个 batch 的时间

记录等 (注册 hook 算子在 self.__hook 中按优先级升序排列). 于是只需理解 call_hook 函数即可.

```
1 def call_hook(self, fn_name):
2     for hook in self.__hooks:
3         getattr(hook, fn_name)(self)
```

如上看出, 在训练的不同节点, 将从注册列表中调用实现了该节点函数的类成员函数. 比如

```
1 class OptimizerHook(Hook):
2
3     def __init__(self, grad_clip=None):
4         self.grad_clip = grad_clip
5
6     def clip_grads(self, params):
7         clip_grad.clip_grad_norm_(
8             filter(lambda p: p.requires_grad, params), **self.grad_clip)
9
10    def after_train_iter(self, runner):
11        runner.optimizer.zero_grad()
12        runner.outputs['loss'].backward()
13        if self.grad_clip is not None:
14            self.clip_grads(runner.model.parameters())
15        runner.optimizer.step()
```

将在每个 train_iter 后实现反向传播和参数更新.

学习率优化相对复杂一点, 其基类 LrUpdaterHook, 实现了 before_run, before_train_epoch, before_train_iter 三个 hook 函数, 意义自明. 这里选一个余弦式变化, 稍作说明:

```
1 class CosineLrUpdaterHook(LrUpdaterHook):
2
3     def __init__(self, target_lr=0, **kwargs):
4         self.target_lr = target_lr
5         super(CosineLrUpdaterHook, self).__init__(**kwargs)
6
7     def get_lr(self, runner, base_lr):
8         if self.by_epoch:
9             progress = runner.epoch
10            max_progress = runner.max_epochs
11        else:
12            progress = runner.iter
13            max_progress = runner.max_iters
14        return self.target_lr + 0.5 * (base_lr - self.target_lr) * \
15            (1 + cos(pi * (progress / max_progress)))
```

从 `get_lr` 可以看到, 学习率变换周期有两种, `epoch->max_epoch`, 或者更大的 `iter->max_iter`, 后者表明一个 epoch 内不同 batch 的学习率可以不同, 因为没有什么理论, 所有这两种方式都行. 其中 `base_lr` 为初始学习率, `target_lr` 为学习率衰减的上界, 而当前学习率正如函数的返回表达式.

第五节 更改模型

5.1 增加模块

本身就包含了很多更改配置文件。

5.2 模型瘦身

已有模型参数思路: 各自参数往小的方向改, 比如结构的重复次数, `channel` 的数目, 或者采用更多的 1×1 替换 3×3 , 甚至可使用 $m \times n$ 联合 $n \times m$ 替换 $m \times m$ 卷积等. 检测模型基本组件 (mmdet 中的理解) 替换: `backbone` 更改为熟知的轻量 `backbone`, 比如 `mobilev2`. 自行设计思路: 这就考虑任务特性, 自身对模型的认识水平和经验问题了, 后面补充一个样例, 作为参考。

第六节 抽离模型

6.1 retinanet_resnet18

只需要将需要的部分代码抽离出来, 作为单独的 `inference` 模型, 可参考代码 `retinanet_resnet18`。

第七节 新增模型

7.1 centernet

先看原始论文 `Objects as Points` 以及官方代码。然后 `mmdet/centernet/core` 下的 `gaussian_radius` 函数的解读可参考 [centernet heatmap 解读](#), 对推广的 `FocalLoss` 的更细致的分析可参考 [centernet 解读](#), 想做一些其他任务尝试的可参考 [centernet 一些尝试](#)。

7.2 代码说明

这里只是将 mmdet 的注册方式和配置文件类拿来用了一下，从原始代码 `centernet office code` 将需要的部分抽离出来，将 pose 模型改成了 backbone+heathead 结合成 HeatMap 的形式，其他文件做了和自己对 mmdet 的理解保持一致方面的些调整。很多部分没有按照 mmdet 的方式去改写，想了下改动有点大，数据，pipeline，训练方式等，就算了。

datasets 将原始的数据类 `__getitem__` 分开实现的方式合并了，原始这样实现主要是不同的方向不同的 `getitem` 所致吧。ops 里的 dcn 等就没有加进来了。这个有参考抽离模型一节。

一些说明点：

- 0 没有 fpn，resnet 的 stem_layer 出来后，做了三次下采样，然后又三次转置卷积回去了，也就是得到了 heatmap 特征图。
- 1 dcn 替换在转置卷积部分
- 2 detector 中的 base_detector，cdetdetector 只是用来测试的
- 3 loss 的实现上，张量的选取 `tensor[index]` 没有 `tensor * mask` 速度快
- 4 核心地方在 core 中，这和 mmdet 中 anchor 相关的意思一致，core/image.py 中的射影点是原始图 scale 后的中心点，中心点关于 scale/2 左移，以及这两点关于图像原点组成的一个三角形（见 `get_3rd_point`）。
- 5 `__getitem__` 中将输入图像射影到 (512, 512) 大小 ($2 \times n$)，然后在构造的 $512/4=128$ 大小的 heatmap mask 上赋值对应的 hm，wh，reg，其中 hm 是将原始框射影到 128 大小的 heatmap mask 上，根据对应的框中心，进行赋值；wh 就是射影后的 w，h 值，reg 为中心向上调整后的误差，也即中心偏移量。

整体上 centernet 的思想是非常简洁的，效果也很好，以及后续的 centertrack 的 end2end 式追踪方案，不过后者的前后帧处理方式优化点似乎较多。

第八节 numpy,torch 某些基础函数

8.1 损失函数部分

torch 相关:new_full,nonzero,numel,expand,where 见损失函数一节。

8.2 模型实现部分

待补。

8.3 数据类部分

待补。

第九节 版本变化说明

9.1 v2.0

2.0 相对于 1.x 版本，从代码上主要有这些改变：

1. core 内部做了一些改写，比如 anchor 中的 tagret 移走了。
2. 数据类做了更彻底的分解，将 1.x 版本的 loader 中的 sample 分离成和 pipeline 同级别，data_loader 放到数据类中的 builder，结构更清晰。
3. models 中 anchor_heads 改为 dense_heads;bbox_heads, mask_heads 等 two stage 相关的合并到 roi_heads 中 (和 2 思想一致)
4. 所有 registry.py 均合并到 builder.py 中，更为简洁。
5. necks 增加了 pafpn, detectors 增加了 fsaf,core/mask 增加了分割数据结构,bbox 增加了 coder, 抽象了下 bbox 的编码解码, iou_calculators。
6. 配置文件采用了继承的方式。
7. 优化了训练，测试流程，速度提升明显，版本支持 pytorch1.5, 1.2 以下均不在支持，定制化模块做了优化。

第十节 计划

0. 模型评估, 分布式数据处理
1. 小样本数据增强代码, coco 格式数据的统计代码
2. setup.py, docker, onnx 等部分

第十一节 检测模型的简略综述

检测算法是由一些基本的组件组合而成. 这也是 `mmdection` 出现的原因. 这章节会在不同算法系列中总结一些基础组件, 最后再做个提取.

11.1 通用物体检测

按照物体编码方式可分为三类. 矩形编码 (anchor 机制), 结构点编码 (RepPoints) 和所有点编码 (mask).

11.1.1 Yolo 系列

yolo 相对孤立,v1 就不说了,v2,v3 基本思想还是来源于 resnet,fpn,ssd 等. 代码可参考 [yolov3](#).

与 SSD 的不同之处:

- 根据数据聚类 9 个先验框, 分成 3 个尺度, 分别作为三个检测层的 base anchor.
- 因为网络结构只有卷积和池化, 所以可以做多尺度输入 320-608,steps=32.
- 基础网络 DarkNet19,53 较 resnet 更为轻量, 主要是 1×1 卷积的大量使用.
- 多尺度的处理方式是在每个尺度上均计算一次检测 (yolo layer), 这和 ssd 合并起来做统一处理不同.
- 框回归编码不同 (相对于网格中心的偏移量, 物体由一个中心网格预测, 当物体重合度高时, 此假设不成立).
- 网络结构实现方式大多采用配置文件解析 (DarkNet).

11.1.2 SSD 系列

ssd 是第一个包含了几乎所有的检测组件的算法 (各种检测算法的所有组件集), 故能在后续的发展中多次被更改, 用于其他任务中, 比如 ctpn, textboxes++, faceboxes 等. 代码可参考: [SSD-Tutorial](#) 手把手教你实现 ssd, 以及相关原理讲解. [ssd.pytorch](#) 最先看的 ssd 源码, 简洁完完整. [maskrcnn-benchmark](#) 风格版 ssd.

原始 SSD:

- 多尺度
 - 不同尺度的 feature map 上生成 anchor (比如: $300 \times 2 \rightarrow 38 \times 2 \times 4 + 19 \times 2 \times 6 + 10 \times 2 \times 6 + 5 \times 2 \times 6 + 3 \times 2 \times 4 + 1 \times 2 \times 4$), 进行位置回归和类别判断.
 - 一个 $m \times n$ 大小的 feature map, 若每个 cell (可以理解成物体的离散表示点) 分配 k 个 anchor, 则每个 cell 输出 $(c + 4) \times k \times m \times n$ 个预测值 (class, box relative offset).
 - 注意每个 cell 其实是一个向量, 长度为 channel of feature map, 可以理解成一个物体的某一部分 (或全部) 的向量表示或者整体表示的一部分.
 - 事实上这里完成了两个任务, 分类和位置回归, 所以 cell 向量可能具有分段表示功效 (这和权重共享是不矛盾的, 共享的权重可能就具有可分离性).
 - anchor 计算损失函数前的有效编码: 首先中心坐标在特征层上归一化 (等同于相对于原图的归一化), 尺度根据原图尺寸以及当前特征图相对原图的尺寸进行设计, 比如第一特征层相对于原图的 0.1, 然后计算相对偏移量以及坐标和尺度的”等效”处理, 尺度求对数.
- 数据增强
 - DistortImage: 修改图像本身的 brightness, contrast, saturation, hue, reordering channels.
 - ExpandImage: 将 DistortImage 的图片用像素 0 进行扩展, 同时以黑边的左上角为原点计算 $[0, 1]$ 的 bbox 的左上角和右下角两个点坐标.

- BatchSampler: sampled_bboxes 的值是随机在 $[0, 1]$ 上生成的 bbox, 并且和某个 gt_bboxes 的 IOU 在 $[min, max]$ 之间
- resize 到固定大小 $300 * 300$, label 也同时线性缩放.
- 以 0.5 概率随机水平翻转, 或者 crop 等
- 样本平衡
 - 难例挖掘, 正负样本 1:3 等. 我的理解是制造更有效的优化空间 (anchor 是让优化空间变得合理).
- 损失函数
 - 利用 GT box 给个生成的 8732(可变)anchor 打标签, 筛选出有效优化对象, 计算分类和回归值.
- 后处理
 - NMS, Soft-NMS, OHEM

以上五部分均是以后论文的改进点. 比如特征提取基础结构, 采用其他有效的分类模型 resnet, 或者轻量级的 mobilenet 等, 或者替换在新的结构上替换一些结构组成算子, 卷积, 激活, BN 等操作. 比如多尺度的 FPN 类似思想, 融合不同特征层, 这在一定程度上解决了重复框, 小物体问题 (将同一物体的不同尺度表达进行融合, 当然能减缓底层表达能力不足的现象). 这些有 FSSD, RSSD (没有必要都要去看, 检测类的文章, 理清基本组件, 花时间分析组件功能, 做实验验证想法, 就 ok 了) 等. 数据增强的方式各不相同, 主要是提高数据的丰富性, 增加模型的泛化能力, 这个属于工程问题, 基本方法都来源于传统的图像处理. 样本平衡的扩展可参考 mmdetection, 用制造更有效的优化空间来理解, 就可以随意发挥了.

具体案例:

FaceBoxes:

textboxes++:

11.1.1.3 Fast RCNN 系列

Mask R-CNN:

[maskrcnn-benchmark](#)能学习的东西都在这里了 (仔细研读 3 遍). Mask R-CNN = Faster R-CNN with FCN on ROIs. 其主要流程参见[1.7.1.ROI](#)

Align 原理: 去掉了图像下采样到特征图的坐标量化, 保持分数坐标, 同时 ROI 分格池化时, 对格子的坐标也取消量化, 从而减少了坐标的二度漂移. 若两次量化, 最坏的情况, 下采样 5 次, 会有近 64 的位置偏移, 这样会漏掉小目标. 最后的池化采用双线性插值, 原理就是每个点的像素值是其临近 4 个像素点的距离权重平均.

Cascade RCNN 发现只有 proposal 自身的阈值和训练器的训练阈值较为接近时, 训练器的性能才最好.

11.1.4 Anchor Free 系列

FCOS, CenterNet, 见上。

11.2 总结

提供一套有用的代码。

第十二节 官方文档 2.0 伪译

2.0 相比 1.x, 就代码组织上, 在模块化这方面有了更好的贯彻。能拆分的就拆分, 比如配置文件, 数据的信息整合、变换、采样迭代, 以前版本命名不严格的一律改掉, 比如 `anchor_heads`。

12.1 配置系统

1.x 版本是将所有配置信息放到一个 `x.py` 配置文件中, 2.0 增加了配置文件的模块化和继承能力, 这样在实验中能提高组合不同部分的效率。执行 `python tools/print_config.py /PATH/TO/CONFIG` 能看到配置信息。`-options xxx.yyy=zzz` 可看到更新信息。

基础配置文件在 `config/_base_` 中, 有 `dataset`, `model`, `schedule`, `default_runtime` 四个部分, 对应 1.x 版本单个配置文件的不同部分。

12.2 使用预训练模型

将 coco 数据训练的模型作为 CitySpace 等数据预训练模型, 需要做以下五处改动

12.2.1 继承基础配置文件

基础模型继承自 `mask_rcnn_r50_fpn`，数据继承自 `cityscapes` 风格，训练 `schedules` 继承自默认的 `default_runtime`，在配置文件顶部增加如下代码：

```

1  _base_ = [
2      '../_base_/models/mask_rcnn_r50_fpn.py',
3      '../_base_/datasets/cityscapes_instance.py', '../_base_/default_runtime.
4      py'
5  ]
6

```

12.2.2 更改头部

如果新旧模型的类别不同，则需要改一下类别数目。

```

1
2  model = dict(
3      pretrained=None,
4      roi_head=dict(
5          bbox_head=dict(
6              type='Shared2FCBBoxHead',
7              in_channels=256,
8              fc_out_channels=1024,
9              roi_feat_size=7,
10             num_classes=8,          # new num_classes
11             target_means=[0., 0., 0., 0.],
12             target_stds=[0.1, 0.1, 0.2, 0.2],
13             reg_class_agnostic=False,
14             loss_cls=dict(
15                 type='CrossEntropyLoss', use_sigmoid=False, loss_weight=1.0)
16         ),
17         loss_bbox=dict(type='SmoothL1Loss', beta=1.0, loss_weight=1.0)),
18     mask_head=dict(
19         type='FCNMaskHead',
20         num_convs=4,
21         in_channels=256,
22         conv_out_channels=256,
23         num_classes=8,
24         loss_mask=dict(
25             type='CrossEntropyLoss', use_mask=True, loss_weight=1.0))
26     )
27

```

预训练的模型权重除最后的预测层不会加载预用模型权值，其他均会被加载。

12.2.3 更改数据

仿照 VOC, WIDER FACE, COCO and Cityscapes 数据类重写自己的数据整合方式。改一下数据类的名字即可。具体改写细节见下小节。

12.2.4 改写训练 schedule

优化器，训练超参数等的修改。

```

1
2 # optimizer
3 # lr is set for a batch size of 8
4 optimizer = dict(type='SGD', lr=0.01, momentum=0.9, weight_decay=0.0001)
5 optimizer_config = dict(grad_clip=None)
6 # learning policy
7 lr_config = dict(
8     policy='step',
9     warmup='linear',
10    warmup_iters=500,
11    warmup_ratio=0.001,
12    # [7] yields higher performance than [6]
13    step=[7])
14 total_epochs = 8 # actual epoch = 8 * 8 = 64
15 log_config = dict(interval=100)

```

12.2.5 使用预训练模型

```

1 load_from = 'https://s3.ap-northeast-2.amazonaws.com/open-mmlab/mmdetection/
    models/mask_rcnn_r50_fpn_2x_20181010-41d35c05.pth'

```

12.3 增加新数据类

12.3.1 转成公用格式

最简单的方式就是将自己的数据脚本转换成 coco 或者 voc 格式。然后更改配置文件中的数据信息。比如 coco 格式，在 configs/my_custom_config.py 中有：

```

1  ...
2  # dataset settings
3  dataset_type = 'CocoDataset'
4  classes = ('a', 'b', 'c', 'd', 'e')    # 自己的五类名字
5  ...
6  data = dict(
7      samples_per_gpu=2,
8      workers_per_gpu=2,
9      train=dict(
10         type=dataset_type,
11         classes=classes,
12         ann_file='path/to/your/train/data',
13         ...),
14     val=dict(
15         type=dataset_type,
16         classes=classes,
17         ann_file='path/to/your/val/data',
18         ...),
19     test=dict(
20         type=dataset_type,
21         classes=classes,
22         ann_file='path/to/your/test/data',
23         ...))
24  ...
25

```

12.3.2 中间格式

mmdet 提供了和 coco, voc 等兼容的中间格式:

```

1  [
2      {
3          'filename': 'a.jpg',
4          'width': 1280,
5          'height': 720,
6          'ann': {
7              'bboxes': <np.ndarray, float32> (n, 4),
8              'labels': <np.ndarray, int64> (n, ),
9              'bboxes_ignore': <np.ndarray, float32> (k, 4),
10             'labels_ignore': <np.ndarray, int64> (k, ) (optional field)
11         }
12     },
13     ...
14 ]
15

```

使用方式:

1. 在线转换

写一个继承自 CustomDataset 的类, 并重写 load_annotations(self, ann_file) 和 get_ann_info(self, idx) 两个方法。参考 CocoDataset, VOCDataset。

2. 离线转换

将标注文件转成中间格式, 保存成 pickle 或 json 文件, 参看 pascal_voc.py, 然后调用 CustomDataset 即可。

一个例子, 假设标注文件格式如下

```
1 #
2 000001.jpg
3 1280 720
4 2
5 10 20 40 60 1
6 20 40 50 60 2
7 #
8 000002.jpg
9 1280 720
10 3
11 50 20 40 60 2
12 20 40 30 45 2
13 30 40 50 60 3
```

我们可以写一个继承自 CustomDataset 的新类如下:

```
1 import mmcv
2 import numpy as np
3
4 from .builder import DATASETS
5 from .custom import CustomDataset
6
7 @DATASETS.register_module()
8 class MyDataset(CustomDataset):
9
10     CLASSES = ('person', 'bicycle', 'car', 'motorcycle')
11
12     def load_annotations(self, ann_file):
13         ann_list = mmcv.list_from_file(ann_file)
14
15         data_infos = []
16         for i, ann_line in enumerate(ann_list):
17             if ann_line != '#':
```

```

18         continue
19
20         img_shape = ann_list[i + 2].split(' ')
21         width = int(img_shape[0])
22         height = int(img_shape[1])
23         bbox_number = int(ann_list[i + 3])
24
25         anns = ann_line.split(' ')
26         bboxes = []
27         labels = []
28         for anns in ann_list[i + 4:i + 4 + bbox_number]:
29             bboxes.append([float(ann) for ann in anns[:4]])
30             labels.append(int(anns[4]))
31
32         data_infos.append(
33             dict(
34                 filename=ann_list[i + 1],
35                 width=width,
36                 height=height,
37                 ann=dict(
38                     bboxes=np.array(bboxes).astype(np.float32),
39                     labels=np.array(labels).astype(np.int64))
40             ))
41
42         return data_infos
43
44     def get_ann_info(self, idx):
45         return self.data_infos[idx]['ann']
46
47 # 配置文件做如下更改:
48
49 dataset_A_train = dict(
50     type='MyDataset',
51     ann_file = 'image_list.txt',
52     pipeline=train_pipeline
53 )

```

数据合并, Repeat 或 Concatenate, 顾名思义, 将同一种数据重复多次, 或不同数据 concat 成一个更大的数据。

如果你只想训练某数据的指定类别, 只需要做如下改动:

```

1 classes = ('person', 'bicycle', 'car')
2 # classes = 'path/to/classes.txt' # 或者类别从文件中读取
3 data = dict(
4     train=dict(classes=classes),
5     val=dict(classes=classes),
6     test=dict(classes=classes))

```

12.4 自定义数据管道

经典的数据管道如下：

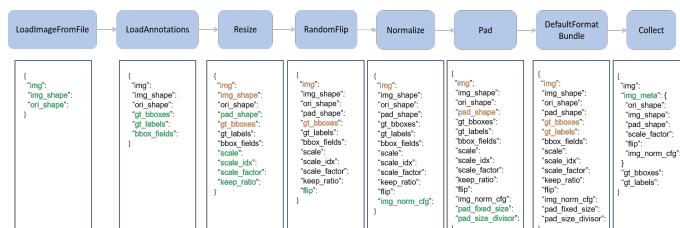


图 15: data pipeline

蓝色块为管道算子，一个算子为一个数据增强算法，从左到右，依次字典进，字典出。关于数据结构，可参考第二节数据处理。这里绿色为算子作用后新增的 keys，橙色为算子作用与已有 keys 的 values 后的更新标记。

12.4.1 扩展 pipelines

```

1
2 # 1. 实现新增强函数 my\_pipeline.py
3 # 和pytorch原始transforms中的增强方式一样，实现__call__方法的类即可
4 from mmdet.datasets import PIPELINES
5
6 @PIPELINES.register_module()
7 class MyTransform:
8
9     def __call__(self, results):          # 输入的是mmdet设定的字典格式
10         results['dummy'] = True
11         return results
12
13 # 2. 导入新类.
14 from .my_pipeline import MyTransform
15
16 # 3. 配置文件调用
17
18 img_norm_cfg = dict(
19     mean=[123.675, 116.28, 103.53], std=[58.395, 57.12, 57.375], to_rgb=True
20 )
21 train_pipeline = [
22     dict(type='LoadImageFromFile'),
23     dict(type='LoadAnnotations', with_bbox=True),
24     dict(type='Resize', img_scale=(1333, 800), keep_ratio=True),

```

```
24 dict(type='RandomFlip', flip_ratio=0.5),
25 dict(type='Normalize', **img_norm_cfg),
26 dict(type='Pad', size_divisor=32),
27 dict(type='MyTransform'),
28 dict(type='DefaultFormatBundle'),
29 dict(type='Collect', keys=['img', 'gt_bboxes', 'gt_labels']),
30 ]
```

12.5 增加新模块

每一个组件都如上小节所给的三部曲，实现，导入，修改配置文件。每一步，mmdet 中都有一些范例，分别看一下，就能实现改写。

12.5.1 优化器

一些模型可能需要对某部分参数做特殊优化处理，比如批归一化层的权重衰减。我们可以通过自定义优化器构造函数来进行细粒度参数调优。

```
1 from mmcv.utils import build_from_cfg
2
3 from mmdet.core.optimizer import OPTIMIZER_BUILDERS, OPTIMIZERS
4 from mmdet.utils import get_root_logger
5 from .cocktail_optimizer import CocktailOptimizer
6
7 @OPTIMIZER_BUILDERS.register_module
8 class CocktailOptimizerConstructor(object):
9
10     def __init__(self, optimizer_cfg, paramwise_cfg=None):
11
12     def __call__(self, model):
13
14         return my_optimizer
15
```

12.5.2 开发新组件

- backbone, 参考 MobilenetV2
- neck, 参考 PAFPN
- head 参考 Retinaface
- roi extractor, 参考 DCN ROIALign

关于 head 相关组件，核心的点我认为在于数据的流向。基础变换层由 forward 函数得到变换的结果，怎么放到 Loss 中去，其中所涉及到的数据细节操作，是比较关键的。mmdet 中会在 head 模块中实现对应的 loss 函数，最终被汇集到检测模型的 loss 中去。分别调用各自的 loss.step, 进行权重更新。

这里贴一个新增 loss 样例:

```

1
2 # 1. 在 mmdet/models/losses/my_loss.py 实现新的 box 回归函数
3 import torch
4 import torch.nn as nn
5
6 from ..builder import LOSSES
7 from .utils import weighted_loss
8
9 @weighted_loss      # 加权损失函数，可参考损失函数章节
10 def my_loss(pred, target):
11     assert pred.size() == target.size() and target.numel() > 0
12     loss = torch.abs(pred - target)
13     return loss
14
15 @LOSSES.register_module
16 class MyLoss(nn.Module):
17
18     def __init__(self, reduction='mean', loss_weight=1.0):
19         super(MyLoss, self).__init__()
20         self.reduction = reduction
21         self.loss_weight = loss_weight
22
23     def forward(self,
24                 pred,
25                 target,
26                 weight=None,
27                 avg_factor=None,
28                 reduction_override=None):
29         assert reduction_override in (None, 'none', 'mean', 'sum')
30         reduction = (
31             reduction_override if reduction_override else self.reduction)
32         loss_bbox = self.loss_weight * my_loss(
33             pred, target, weight, reduction=reduction, avg_factor=avg_factor)
34
35 # 2. 然后在 mmdet/models/losses/__init__.py 中注册
36 from .my_loss import MyLoss, my_loss
37
38 # 3. 配置文件使用

```

```
39 loss_bbox=dict(type='MyLoss', loss_weight=1.0))
```

12.6 1.x 模型升级到 2.0

执行脚本 `tools/upgrade_model_version.py`。可能有小于 1% 的绝对 AP 减小，具体可参见 `configs/legacy`。

12.7 2.0 和 1.x 的不同之处

主要有四点不同：坐标系，基础代码约定，训练超参数，模块设计。

12.7.1 坐标系

新坐标系与 `detectron2` 一致 treats the center of the most left-top pixel as (0, 0) rather than the left-top corner of that pixel. 这句话的意思是将 `bbox = [x1, y1, x1 + w - 1, y1 + h - 1]` 改为 `bbox = [x1, y1, x1 + w, y1 + h]`，这样更加自然和精确（假设长或宽为 1，则 box 退缩为点或线，这是有问题的），同理 `xyxy2xywh` 的长宽就不在 +1 了，生成的 anchor 的中心偏移也不在是 0.5 而是 0 了。与此相关的改动有 Box 的编解码，与 iou 计算相关的 nms, assinger。另外，现在的坐标为 float，1.xx 为 int，与此相关的有 anchor 与特征网格的中心对齐问题，这对 anchor-based 的方法在性能上有一定影响（变好），ROIAlign 也能更好的对齐，mask cropping and pasting 更精准，利用新的 RoIAlign 去 crop mask targets，会得到更好的结果，因为没有取整等误差了，而且在训练上也有 0.1s/iter 的速度提升（少了取整操作）。

12.7.2 Codebase Conventions

- 类别设定，1.x 中 0 为背景， $[1, k]$ 为 k 类对象，2.0 中 k 为背景， $[0, k-1]$ 为 k 类对象。
- bbox 分配方案就低质量分配上得到了改进。支持了更多长宽尺度输入，以上均有微弱性能提升。
- 配置名称约定改动为 `[model]_(modelsetting)_[backbone]_[neck]_(normsetting)_(misc)_(gpuxbatch)_[schedule]_[dataset].py`,

12.7.3 训练超参数

一些训练参数的优化

- nms 后的 rpn 的 proposals 从 2000 改为 1000, (nms_post=1000, max_num=1000), mask, bbox AP 有 0.2% 的提升。
- Mask, Faster R-CNN 框回归损失函数 Smooth L1 改为 L1 带来 0.6% 的提升, Cascade R-CNN and HTC 保持原样。
- RoIAlign layer 采样数设置为 0, 0.2% 提升
- 默认设置不在使用梯度截断, 这样训练更快, 但 RepPoints 保持是为了训练稳定以及更好的结果。
- 默认的 warmup ratio 从 1/3 改为 0.001, 这样更平滑, 同时也是因为去掉了梯度截断。