Towards Efficient Instance Segmentation with Hierarchical Distillation

通过分层蒸馏实现的高效实例分割

问题

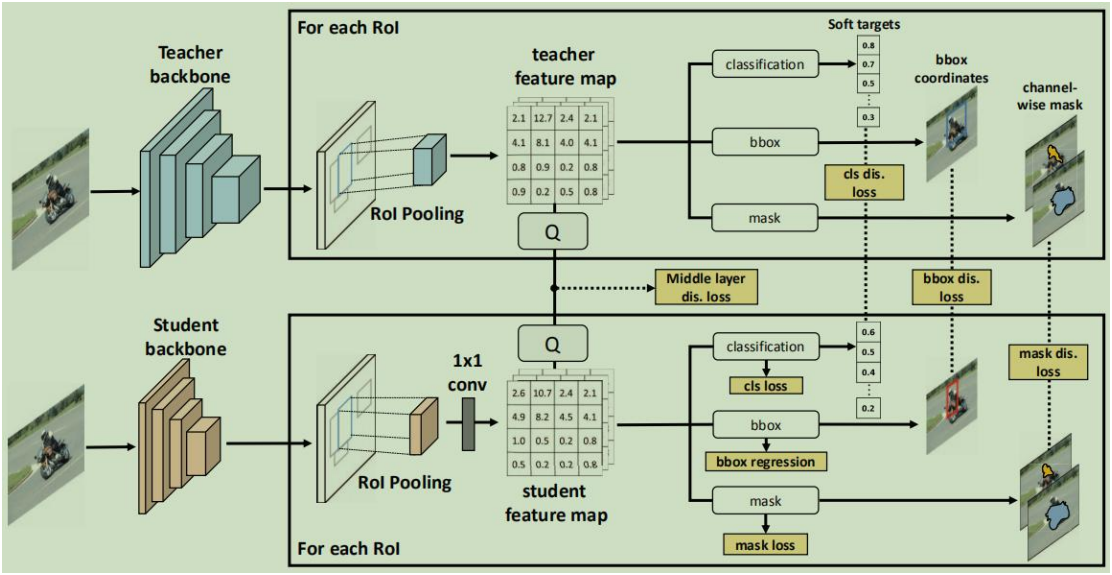
实例分割在公共基准上精度获得了巨大的提升。但推理速度较慢，难以用于实际场景。

本文提出了一种利用师生学习框架的更快的实例分割模型，该模型将训练好的老师模型中获得的知识转移到轻量的学生模型中。

在有限的计算资源下，此方法可以有效的加速实例分割模型，且精度损失较小。

网络框架

采用教师网络的中层和后层输出作为蒸馏目标，将结构信息转移到学生网络。



- 首先，作者使用所有的训练数据训练教师网络，在蒸馏阶段只前向传播输入图片。
- 其次，作者在教师网络和学生网络 RoIs 上的特征图用 L2 回归使用**特征蒸馏**作为中间层的蒸馏损失。

  由于大多数多任务网络更大更深，因此在特征图上进行量化操作使学生网络从老师网络学习收敛更快更简单。
- 最后，作者还为每个头部定制，从教师网络蒸馏后期层输出作为**语义信息**，包括每个 RoI 的软目标分类，边界框坐标，每个通道掩码。

  语义蒸馏损失由类别蒸馏的交叉熵损失和方框蒸馏与掩码蒸馏的 L2 损失组成。

特征蒸馏（Representation distillation）

对实例分割，从共享的卷积层提取特征图会影响定位，分类和分割的精度。
将包含许多的黑箱知识的中间层转移来促进学生网络是必要的。
实例分割是基于区域的工作，头部网络基于池化的候选区域，所以在此任务中区域候选有非常重要的作用。
从中间层蒸馏知识，在教师网络和学生网络上对特征图引入量化操作。
量化方法是教师网络和学生网络的输出离散化，使学生网络能更好的匹配教师网络。
特征蒸馏损失函数 L_{rp_dis} 定义如下：

$$L_{rp\_dis} = \frac{1}{2N} \sum_{i \in N} \|Q(f_t^i) - Q(r(f_s^i))\|_2^2$$

Q 是以元素为单位的量化函数。N 是提案数目，r 是将学生特征图转为教师特征图同样大小的函数。这里，作者使用 1x1 的卷积层作为 r。

语义蒸馏（Semantic distillation）

分类蒸馏

已经提出了常规的知识蒸馏方法来训练分类网络，将老师模型预测的类别作为“软目标”来指导学生网络的训练。
神经网络通常使用 softmax 输出层来产生类别概率，其中 softmax 输出层将预测第 i 个类别得分的输出 zi，计算转为其类别概率 pi，将 zi 与其它的对数变换 pi=softmax（zi/T）比较。
T 是温度参数，通常设为 1。
T 的值越高，类别间的分布越平滑。
在作者的工作中，将教师网络预测的第 i 个候选区域其类别的可能性记为$\{p_{i,l}^t, l \in L\}$，l 是第 l 个类别的概率，L 是类别的数目。
$p_{i,l}^t$ 作为软目标从教师网络转移到学生网络，通过优化下面作为类别蒸馏损失的交叉熵损失 L_{cls_dis}：

$$L_{cls\_dis} = -\frac{1}{N} \sum_{i \in N} \sum_{l \in L} p_{i,l}^t log(p_{i,l}^s)$$

$P_{i,l}^s$ 是学生网络预测的类别概率，N 是候选区域的数目。
软目标包含教师网络发现的不同类别间的关系信息。
通过对软目标的学习，学生网络继承了这样的隐藏信息。

框回归蒸馏

对调整候选区域位置和大小的边界框回归，作者使用如下 L2 优化作为边界框蒸馏损失：

$$L_{bbox\_dis} = \frac{1}{2N} \sum_{i \in N} \| R_i^t - R_i^s \|_2^2$$

Rt 和 Rs 分别表示相应的教师网络和学生网络的回归层输出。L_{bbox_dis} 可以鼓励学生网络在框回归的表现上接近教师网络性能。

掩膜蒸馏

通过老师网络中的掩膜预测分支，学生网络被训练为通道级别的蒸馏。

掩膜头部的预测结果是多通道的掩码。

在每个通道内，掩膜表示特定类别或简单的背景或前景的分割结果。

与使用 softmax 的预测每个像素的类别的像素级别分类任务的语义分割相比，实例分割解耦了分类头部和分割头部来预测特定类别的掩膜。

类别间的竞争不利于一个好的掩膜预测，作者设计了如下的通道级别的掩膜蒸馏损失 L_{mask_dis}：

$$L_{mask\_dis} = -\frac{1}{2N \times C} \sum_{i \in N} \sum_{c \in C} \| M_{i,c}^t - M_{i,c}^s \|_2^2$$

$M_{i,c}^t M_{i,c}^s$ 是学生和老师网络在通道 c 处的候选区域 i 的尺寸 14x14 的掩膜输出。

C 是等于类别数的通道数目。

以这样的方式定义，无论类别预测如何，学生网络分割头部层特定类别的掩膜输出与教师网络的类似。

因此，语义蒸馏损失 L_{sm_dis} 是：

$$L_{sm\_dis} = L_{cls\_dis} + L_{bbox\_dis} + L_{mask\_dis}$$

这是每个头部蒸馏分割的和。

分层蒸馏 (H-Dis)

作者总体的蒸馏训练损失是分层蒸馏损失，其中包括特征蒸馏损失和语义蒸馏损失，记为 L_{h_dis}:

$$L_{h\_dis} = L_{gt} + \lambda L_{sm\_dis} + \sigma L_{rp\_dis}$$

超参数 λ 和 σ 表示不同损失之间的平衡参数，在作者的实验中固定为 1。

实验

作者对不同的主干和不同的数据集进行了评估，证明了方法的有效性和泛化能力。
作者使用带有 ResNet 的 FCIS（Fully convolutional instance-aware semantic segmentation）实例分割框架作为主干。
呈现实验在 Pascal VOC 2012 和 Cityscapes 上的结果。
准确度是通过在掩膜水平的 IoU 阈为 0.5 和 0.7 的平均精度评估的。
速度统计是在单张 1080Ti 显卡上对一张图片输入的所有处理过程，作者使用 mxnet 实现方法。

实现细节

遵循两阶段训练策略实现分层蒸馏。
第一阶段是训练 RPN 网络，不需要模仿 RoIs 中的特征图。
第二阶段是训练头部网络，不需要蒸馏。
由于实现更高的精度不是本文的目标，所以实现没有经过任何精度增强技巧的优化。
RPN 的锚点在所有的实验中共有三个尺寸和一个长宽比。

训练

主要遵循 FCIS 的训练配置。
在单 GPU 上训练所以 batch size 为 1。
在 Pascal VOC 上的实验，进行 240k 次迭代，学习率分别为 10^{-3} 和 10^{-4} 在前 160K 次和后 80K 次。在 Cityscapes 上的迭代次数是 144K。
在每个批次，有 128 个候选反向传播它们的梯度。
所有的教师网络和学生网络是在相同的训练集和测试/验证集上进行的。

推理

在测试时，在第一个迭代时一张图片提出了 300 个候选，之后另外 300 候选在框回归分支后生成。
作者的方法在测试时没有添加任何计算，所以学生网络的速度和基准模型一样。

在 Pascal VOC 上进行的消融实验

消融实验在 Pascal VOC 上进行。遵循前人的规程，对训练集执行模型训练，在验证集上进行评估，包括 20 个类的目标。训练图片被缩放为短边 600 像素。

表 1 中显示了作者的工作和其它常规实例分割工作的消融实验的结果。

在流行的实例分割工作中，作者选择 FCIS 作为教师和学生实例分割架构，因为它在速度和精度上有竞争力并取得好的平衡。

可以观察到，以 ResNet-18 为主干的 FCIS 框架作为学生网络能达到 16fps 的高速度，以及与以 ResNet-101 作为教师网络相比 10.2%的精度损失。

本文提出的方法可以提高小的学生网络的精度到 58.1%，与 PFN 的结果接近，但相比于 1fps 的 PFN，仍然可以保持 16fps 速度。

Model	$AP_{0.5}(\%)$	speed(fps)
SDS[10]	49.7	<1
CFM[5]	60.7	<1
PFN[19]	58.7	1
MNC[6]	63.5	1
Mask R-CNN[11]	69.0	3
Teacher(ResNet-101)	65.7	6
Student(ResNet-18)	55.5	16
Student w/ H-Dis	**58.1**	**16**

表 1：在 Pascal VOC 2012 数据集上，我们提出的方法和其他流行的实例分割方法比较结果。Teacher（ResNet-101）是被训练有素的教师网络。Students（ResNet-18）是没有经过蒸馏就被作为我们的基准训练的学生网络。Student w/ H-Dis 是经过我们分层蒸馏损失训练的学生网络。

表 2 展示了不同蒸馏策略，显示了不同损失的有效性。
掩码头部的损失显示了最高的提升（为 AP0.5 增加了 1%）。
原因是作者利用多类别的掩码预测，而不是和类别无关的掩码预测来蒸馏，这其中可能包括教师网络获得的完整的特定类别信息。
分类和边界框回归头部的蒸馏提升较弱，因为实例分割任务复杂且受多因素影响，仅蒸馏这些不能从教师网络转移出丰富的知识。
和离散的分类蒸馏不同，边界框回归输出可能会为学生网络提供非常错误的指导。
然而，通过将损失综合起来，教师网络的知识变得有足够的影响力贡献到学生网络训练中，准确度的提升变得更有希望（为 AP0.5 总共增加了 2.6%）。

GT	Cls	Box	Mask	Mid	$AP_{0.5}$	$AP_{0.7}$
✓					55.5	36.1
✓	✓				55.8	36.1
✓	✓	✓			56.2	36.3
✓	✓	✓	✓		57.2	39.2
✓	✓	✓	✓	✓	58.1	42.3

表 2：在 Pascal VOC 2012 数据集上，进行我们的分层蒸馏时，其组件不同设置的有效性。Mid 表示中间层的特征蒸馏。

不同的主干
除了 ResNet-18 外，我们还在名为 ResNet-18-4 的压缩 ResNet-18 进行了实验，与原始层相比每一层的通道数减少到了 1/4。
表 3 显示了 ResNet-18-4 的精度提高和速度结果，相比与基准提高了 3.5%。
这个结果表明了较轻和较弱的学生网络能从训练有素的教师网络中获得更多的精度增强。
还可以得出结论，从更好的教师网络上蒸馏能获得更大的提升，它们之间的差距越大，教师网络所传递的知识就更有效和更丰富。

Model	student	w/ H-Dis	speed(fps)
ResNet-18	55.5	58.1(+2.6)	16
ResNet-18-4	39.0	42.5(+3.5)	19

表 3：在 Pascal VOC 2012 数据集上，具有不同的主干的学生网络的比较结果。教师网络是 ResNet-101，使用 H-Dis。

为了显示提出的 H-Dis 方法的有效性，将 PSACAL VOC 数据集的实例分割结果例子显示在图 2。
顶部 3 行显示，蒸馏的模型可以分割到基准结果漏掉的实例。
最后三行显示本文的方法能减少一些假阳性（误检）结果。

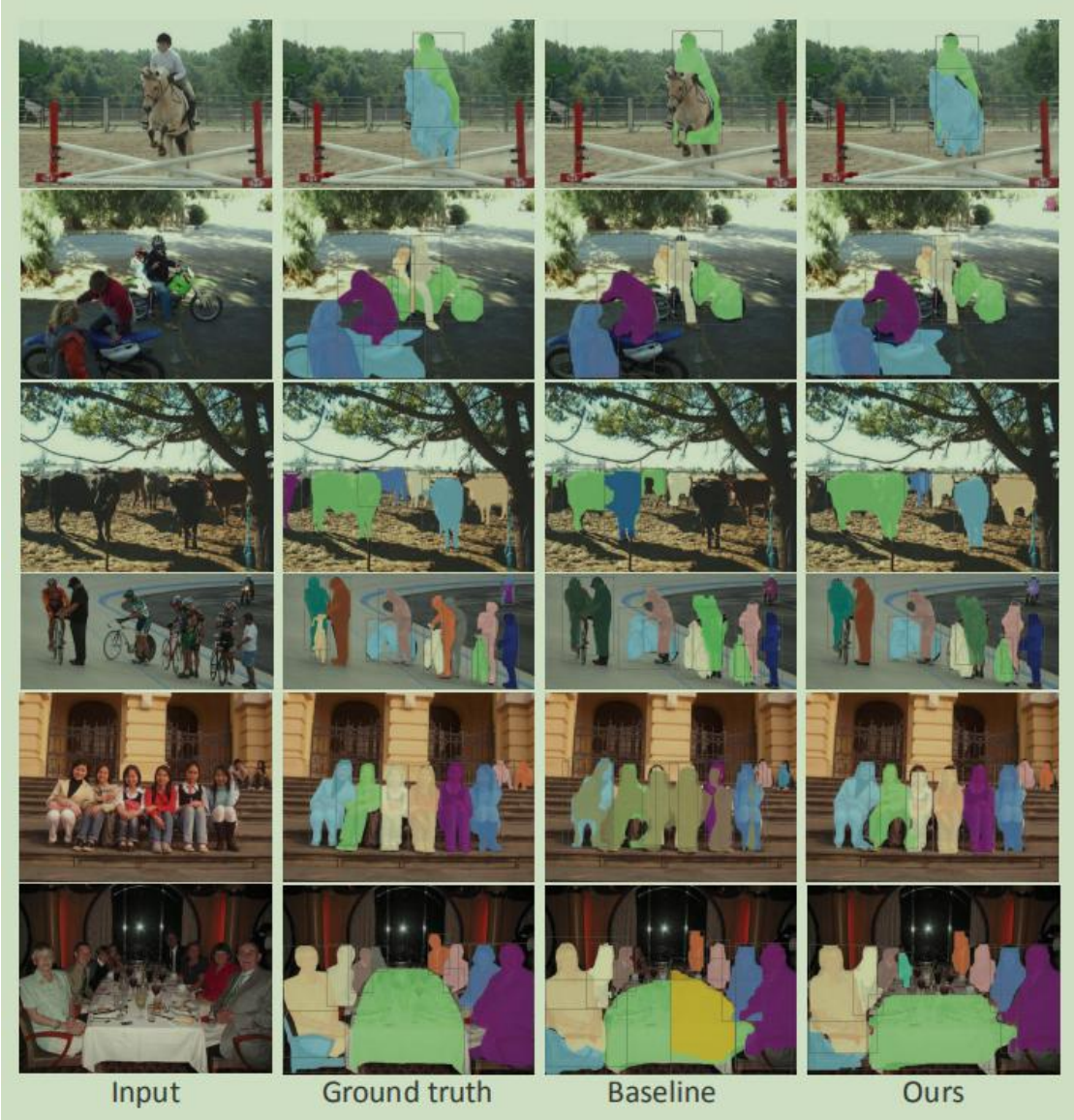


图 2：在 PSACAL VOC 2012 数据集上示例结果来自 Baseline: Student(ResNet-18) 和 Ours: Student w/ H-Dis。真实值和预测结果相对应的使用相同颜色的 mask 颜色打印。用其他颜色打印的 mask 表示假阳性（误检）结果。

Cityscapes 上的结果（不同的数据集）

作者进一步报告在 Cityscapes 数据集上的结果，使用 2975 张精细标注的图片训练，500 验证图片和 1525 张测试图片进行评估。

图片具有 1024x2048 的高分辨率，在训练和测试中重新调整图片的短边为 512 像素。

数据集包含 8 类用于实力分割任务，占主导地位的样本是人和车类别。

表 4 是 Cityscapes 数据集的评估结果。

正如预期，学生模型在测试和验证集上分别调高了 1.6%和 1.9%AP。

在 Pascal VOC 数据集上结果提升不明显，主要是因为教师模型不够强大无法为学生模型提供有效的知识。

结合表 3 中的结果，可以证明，教师模型没有训练足够好就来教授学生网络，提升程度会较低或几乎没有。

Model	*AP*[test]	*AP*[val]
Teacher(ResNet-101)	26.5	31.5
Student(ResNet-18)	16.5	18.6
Student w/ H-Dis	**18.1**	**20.5**

表 4：Cityscapes 数据集的比较结果。