面向物体检测的特征金字塔网络

Tsung-Yi Lin，Piotr Dollar，Ross Girshick,
Kaiming He, Bharath Hariharan, and Serge Belongie

1 摘要

特征金字塔是在不同尺度检测物体的识别系统的基本组件。但是由于它们是计算和存储密集的，所以近期的深度学习物体检测器都避免了金字塔表示。在这篇文章中，我们发挥深度卷积网络中固有的多尺度金字塔层级的作用，使用很少的额外代价来构造特征金字塔。我们通过开发一个有旁路连接的自顶向下的结构来在所有尺度构建高层级语义特征图。这个被称为特征金字塔网络（FPN）的结构作为通用的特征提取器为多个应用带来了显著提高。通过在基础的 Faster R-CNN 中使用 FPN，在没有花里胡哨的情况下，我们的方法在 COCO 检测基准上达到了最佳的单模型结果，超过了包括 COCO 2016 挑战赢家在内的所有现存单一模型条目。除此之外，我们的方法可以在单个 GPU 上打到 6FPS 的速度，因此这是一个实际的精准多尺度物体检测解决方案。

2 简介

识别不同尺度的物体时计算机视觉中的一个基本挑战。构建在图片金字塔上的特征金字塔是标准解法的基础 [1]（图1(a)）。在物体的尺度变化会通过移动它在金字塔中层级来补偿的角度来说，这些金字塔是具有尺度不变性的。直觉上来说，通过让模型扫描各个位置和各个尺度，模型可以在很大的尺度区间中检测物体。

在手工设计特征的时代，特征化的图片金字塔被重度使用。它们如此重要，以至于例如 DPM 的物体检测器要求密集尺度采样来达到好的结果（例如，在每个八度采样 10 个尺度）。对于识别任务，手工设计特征已经被深度卷积网络计算的特征大量替代。随着表示高层级语义的能力而来的，还有对

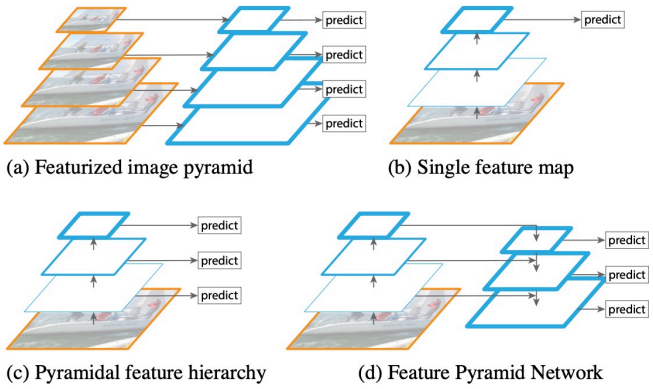

图 1: (a) 使用图片金字塔来构建特征金字塔。特征在每个图片尺度上被独立计算，这是很慢的。(b) 最近的检测系统选择使用单一尺度特征来进行更快的检测。(c) 一个可选方案是复用卷积网络计算得到的金字塔特征层级并将之视作特征化图片金字塔。(d) 我们提出的 FPN 和 (b)(c) 一样快，但是更加准确。在这张图中，特征图被标记为蓝色外边框，更粗的边框表示语义更强的特征。

于尺度变化的鲁棒性，因此似的根据在单一输入尺度计算的特征上识别更加简单（图1(b)）。然而即使有了如此的鲁棒性，为了得到最精准的结果我们仍需要金字塔。近期 ImageNet 和 COCO 检测挑战中的所有的顶部条目都在特征化的图片金字塔上进行多尺度测试（例如，[8, 10]。特征化图片金字塔中的每一个层级的主要优势在于它产生了一个包括高分辨率层级在内的所有层级语义都很强的多尺度特征表示。

然而，特征化图片金字塔中的每一个层级有着明显的局限。推理时间显著增加（例如在 [**4**] 中增加 4 倍），这使得这个方法对于实时应用来说不切实际。不仅如此，就内存来说，在一个图片金字塔上端到端地训练一个深度网络是不切实际的，如果要使用的话，只能在测试阶段使用图片金字塔，而这制造了训练和测试时推理的不一致性。处于这些原因，Fast R-CNN 和 Faster R-CNN 在默认设置下选择不适用特征化的图片金字塔。

然而，图片金字塔并不是计算多尺度特征表示的唯一方法。深度卷积网络会一层接着一层地计算一个特征层级，同时由于下采样层的存在，特征层级有一个固有的多尺度金字塔的形状。这个网络中特征层级产生了一系列有不同空间分辨率的特征图，然而也引入了由于不同深度产生的巨大语义

缺口。高分辨率特征图有低层级特征，这损害了它们对于物体检测的表征能力。

Single Shot Detector(SSD) 是最早进行将卷积网络的金字塔特征层级作为特征化图片金字塔尝试的检测器之一（图1(c)）。在理想情况下，SSD 风格的金字塔可以复用在前向传播过程中各层计算的多尺度特征图，因此这是没有代价的。但是为了避免使用低层级特征，SSD 放弃了复用已经计算好的各层，而是从网络的高层（例如 VGG 网络中的 conv4_3）开始并加入一些新层来构建金字塔。因此它错过了使用特征层级中的更高分辨率的特征图的机会。我们发现这对于检测小物体是十分重要的。

这篇文章的目标是自然地借助卷积网络的特征层级的金字塔形状同时构建一个在所有尺度都有强语义的特征金字塔。为了实现这一目标，我们借助通过自顶向下通路和旁路连接，将低分辨率强语义特征和高分辨率弱语义特征结合起来的结构（图1(d)）。结果是一个从单一输入图片尺度快速构建，然而在所有层级都有丰富语义的特征金字塔。换一种说法，我们展示了如何在不牺牲表征能力、速度和内存的情况下，构建一个可以替代特征化图片金字塔的网络内特征金字塔。

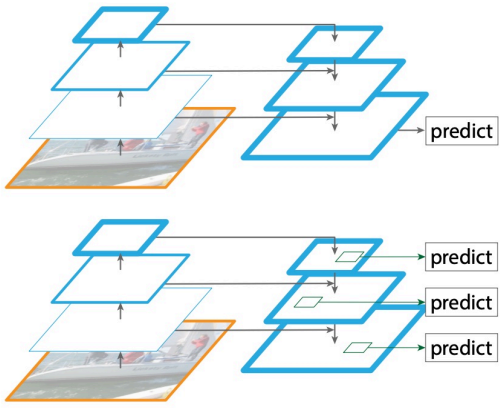

图 2: 顶部：有跳跃连接的自顶向下结构，仅在最好的层级预测（例如 [9]）。底部：我们的模型与之有类似的结构但是借用它为特征金字塔，在所有层级独立做出预测。

类似采用自顶向下以及跳跃连接的结构在最近的研究中十分普遍。它们的目标是产生分辨率良好的单一高层级特征图，并在其上进行预测（图2顶部）。与之相反的是，我们的方法利用如特征金字塔的结构，并独立在各个

层级进行预测（例如物体检测）（图2底部）。我们的模型与特征化图片金字塔类似，这并没有在这些工作中被探索。

我们我们在多个检测和分割网络 [**fasterrcnn**, 4, 6] 中评估了我们的方法，我们将之命名为 Feature Pyramid Network (FPN)。在没有任何花里胡哨的情况下，我们简单地基于 FPN 和 Faster R-CNN 检测器，在 COCO 检测基准上达到了最好的单模型结果，超越了所有已有的重度手工设计的单模型条目。在消融实验中，我们发现在强大的单尺度 ResNets Faster R-CNN 基线上，FPN 为候选边界框生成提高了 8% 地平均召回率（Average Recall，AR）；对于物体检测，它提高了 2.3% 的 COCO 风格 AP 和 3.8% 的 PASCAL 风格的 AP。我们的方法也可以轻易地扩展到掩码候选，并提高了严重基于图片金字塔的最优方法地速度和实例分割 AR。

额外地，我们的金字塔结构可以多尺度端到端训练，因此在训练和测试时是具有一致性的，若使用图片金子塔这将是不可能的。作为结果，FPN 可以达到比已有最佳方法都更高的准确率。不仅如此，这种进步是在不增加单尺度基线的测试时间的条件下达到的。我们相信这些进步将会帮助未来的研究和应用。我们的代码将会开源。

3 特征金字塔网络

我们的目标是借助卷积网络的金字塔特征层级，它有从低到高的语义层级，并构建一个在各处都有高层级语义的特征金字塔。最终的特征金字塔网络是通用目的的，在这片文章中我们专注于滑动窗口候选生成（候选区域网络，RPN）和基于区域的检测器（Faster R-CNN）。我们也在第 6 章中将 FPN 推广到分割候选。

我们的方法将任意大小的单一尺度图片作为输入，并在多个尺度通过全卷积的方式输出相应尺度的特征图。这个过程与骨干卷积网络架构（例如，[2, 3, 8]）是独立的，在这篇文章中我们展示了使用 ResNets[8] 的结果。正如后面将要介绍的，我们的金字塔构建过程涉及一个自底向上的通路、一个自顶向下的通路和旁路连接。

自底向上通路。 自底向上的通路是骨干网卷积络的前向传播计算，它会计算一个由以缩放步长为 2 变化的各个尺度的特征组成的特征层级。通常会存在多个产生相同大小的输出特征图的层，我们称这些层在同一个阶段（*stage*）。对于我们的特征金字塔，我们为每一个阶段定义一个金字塔层级。

我们选择每个阶段最后一层的输出作为我们对于这个阶段的参考。由于每一个阶段最深的层应该有最强的特征，所以这个选择是自然的。

特别的，对于 ResNets[8] 我们使用了每一个阶段的最后一个残差模块的特征激活输出。对于 conv2, conv3, conv4 和 conv5 的输出，我们将这些最后的残差模块的输出记为 $\{C_2, C_3, C_4, C_5\}$，并且它们对于输入图像的步长为 $\{4, 8, 16, 32\}$ 个像素。由于 conv1 巨大的内存占用，我们没有将它包括在金字塔内。

自顶向下通路和旁路连接。 自顶向下通路通过对空间更加粗糙但是语音更强的来自金子塔更高层级的特征进行上采样，仿佛得到了更高分辨率的特征。这些特征接着会被来自自底向上通路的特征经过旁路连接增强。每个旁路连接将会合并来自自底向上通路和自顶向下通路中有相同空间大小的特征图。自底向下的特征图的语义层级更低，但是由于它被下采样的次数更少，所以它的激活被更加准确地定位。

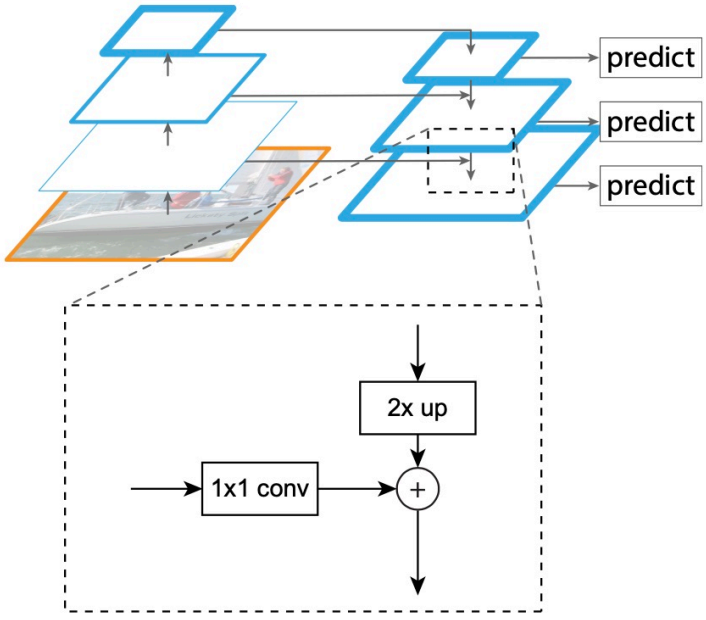

图 3: 一个展示了旁路连接和自顶向下通路的构造模块，通过相加合并。

图3展示了构建我们的特征图的构造模块。有了更粗糙分辨率的特征图，我们将它的空间分辨率以系数 2 进行上采样（出于简便使用最近邻上采样）。

接着上采样得到的特征图与对应的自底向上通路中的特征图（它会经历一个 1×1 卷积层来减少通道维度）通过对应元素相加的方式合并。这个过程将会被迭代只到最好分辨率的特征图被生成。为了开始迭代，我们简单地在 C_5 上依附一个 1×1 卷积层来生成最粗糙的特征图。最终，我们在每一个合并得到的特征图上添加一个 3×3 卷积来生成最终的特征图，卷积被用来减弱上采样的对齐影响。这个最终特征图的集合被称为 $\{P_2, P_3, P_4, P_5\}$，对应于分别有相同空间大小的 $\{C_2, C_3, C_4, C_5\}$。

由于在传统的特征化图片金字塔中，金字塔的所有层级使用共享的分类器/回归器，所以我们在所有特征图中固定了特征维度（通道数，记为 d）。我们在这篇文章中设置 $d = 256$，因此所有额外的卷积层输出通道数都为 256。这些额外层没有非线性，我们经验性地发现这影响并不大。

简便性是我们设计的中心，同时我们发现我们的模型对于多个不同的设计选择具有鲁棒性。我们使用更加精细的模块（例如，使用多层残差模块 [8] 作为连接）进行了实验，并观察到了些微更好的结果。设计更好的连接模型不是本文的关注点，所以我们选择了如上所述的简单设计。

4 应用

我们的方法对于在深度卷积网络中构建特征金字塔是普遍适用的。接下来我们在生成候选框的 RPN[7] 和物体检测的 Fast R-CNN[4] 中使用我们的方法。为了阐述我们的方法的简单性和有效性，当我们将原有系统适配到我们的特征金字塔时，我们对它们做了最小化的修改。

4.1 为 RPN 使用 FPN

RPN[7] 是滑动窗口的类别不可知物体检测器。在原本的 RPN 设计中，小的子网络在密集的 3×3 滑动窗口上评估，在单一尺度的卷积特征图上，进行物体/非物体而分类和边界框回归。这是通过一个 3×3 卷积层跟随着两个分别为了分类和回归的兄弟 1×1 卷积实现的，我们将这称为*头*。物体/非物体标准和边界框回归目标是在一系列被称为*锚* [**fpn**] 的参考框上定义的。为了涵盖不同形状的物体，这些锚有多个实现定义好的尺寸和高宽比。

我们通过将单一尺度特征图替换为我们的 FPN 来适配 RPN。我们为特征金字塔中的每一个层级依附一个有相同设计的头（3×3 卷积和两个兄

弟 1×1 卷积）。由于头将会在所有金字塔层级上的所有位置密集滑动，所以在一个特定层级上设置多尺度锚是不必要的。取而代之，我们为每一个层级分配单一尺度的锚。正式地来说，我们分别在 $\{P_2, P_3, P_4, P_5, P_6\}$ 上定义了面积为 $\{32^2, 64^2, 128^2, 256^2, 512^2\}$ 的锚。正如 [7] 中所讲的，我们也在每个层级使用不同高宽比的锚 $\{1:2, 1:1, 2:1\}$。所以金字塔共有 15 个锚。

和 [7] 一样，我们基于锚与 ground-truth 边界框的 IoU 来为锚分配训练标签。正式地来说，如果一个锚在所有锚中与某个 ground-truth 框有最大的 IoU 或者与某个 ground-truth 框 IoU 大于 0.7，那么它将被分配正标签。如果一个锚与所有的 ground-truth 框 IoU 都小于 0.3，那么它将被分配负样本。注意 ground-truth 框的尺度在分配它们至各个金字塔层级时并没有被显式使用；取而代之，ground-truth 框与锚相关联，而锚则已经被分配至金字塔的层级。这样，我们没有为 [7] 的标签分配引入新的额外规则。

我们注意到头的参数是在所有金字塔层级共享的；我们也评估了非共享的替代方案并观测到了类似的精确度。共享参数的良好性能表明了金字塔的所有层级共享类似的语义层级。这种优势与使用特征化图片金字塔类似，其中公共的分类器头可以用在使用任意尺度图片计算得到的特征上。

有了上述的适配，RPN 可以使用 RPN 自然地训练和测试。我们在实验中详细阐述了实现细节。

4.2 为 Fast R-CNN 使用 FPN

Fast R-CNN[4] 是使用 RoI pooling 来提取特征的基于区域的物体检测器。Fast R-CNN 普遍在单一尺度特征图上使用。为了配合 FPN 使用，我们需要为金字塔层级分配不同尺度的 RoI。

我们将特征金字塔视作由图片金字塔得到。因此我们可以调整基于区域的检测器 [5, 4] 在图片金字塔上运行时的分配策略。正式地来说，我们按照如下公式将宽度为 w 高度为 h（在网络的输入图像中）的 RoI 分配给特征金字塔的 P_k 层：

$$k = \lfloor k_0 + \log_2(\sqrt{wh}/224) \rfloor \tag{1}$$

这里 224 是 ImageNet 预训练的典型尺寸，k_0 是 $w \times h = 224^2$ 的 RoI 应该被映射到的目标层级。类比给予 ResNet 的 Faster R-CNN[7] 系统，它使用 C_4 作为单一尺度特征图，所以我们将 k_0 设置为 4。直觉上来说，公式1意味着如果 RoI 的尺寸变小（例如，224 的 1/2），那么它应该被映射到更加精细的分辨率层级（例如，$k = 3$）。

我们为所有层级的所有 RoI 依附了预测器头（在 Fast R-CNN 中头是类别特定的分类器以及边界框回归器）。再一次，所有的头共享参数，无论它们的层级。在 [8] 中，conv5 层（一个 9 层深度子网络）被当作头部用在 conv4 特征之上，但是我们的方法已经利用 conv5 来构造特征金字塔。所以与 [8] 不同的是，在最终的分类和边界框回归层之前，我们简单地采用 RoI pooling 来提取 7×7 的特征，并附加两个隐藏的 1024 维全连接层（每一个都跟随 ReLU）。由于 ResNets 中不存在可用的预训练全连接层，所以这些层被随机初始化。注意相比于标准的 conv5 头，我们的两层全连接多层感知机头部更加轻量化，更加快速。

基于这些改变，我们可以在特征金字塔的基础上训练和测试 Fast R-CNN。实现细节将会在实验章节给出。

References

[1] Edward H Adelson et al. “Pyramid methods in image processing”. In: *RCA engineer* 29.6 (1984), pp. 33–41.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[3] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[4] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.

[5] Kaiming He et al. “Spatial pyramid pooling in deep convolutional networks for visual recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 37.9 (2015), pp. 1904–1916.

[6] Pedro O Pinheiro, Ronan Collobert, and Piotr Dollár. “Learning to segment object candidates”. In: *arXiv preprint arXiv:1506.06204* (2015).

[7] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015), pp. 91–99.

[8] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[9] Pedro O Pinheiro et al. “Learning to refine object segments”. In: *European conference on computer vision*. Springer. 2016, pp. 75–91.

[10] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. “Training region-based object detectors with online hard example mining”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 761–769.