CV Final Project: CenterNet

鲁东佑
2001213518
信息科学技术学院

苏可凡
2001111376
信息科学技术学院

摘要

我们的期末项目使关于目标检测的，复现了 CenterNet 模型并进行目标检测。CenterNet 在检测速度和准确率方面都有优势，是由主体网络和标头网络组成的 End-To-End 网络。我们基于官方的 ResNet 源代码复现了 CenterNet，使用 ImgAug 库进行了数据增强。在实验中，我们了使用 VOC2007 数据集，从中选取 5 个分类生成新的数据集，并且在这个新的数据集中进行了训练和测试，最终 ResNet18 和 ResNet50 为主体的两个网络的 mAP 在 25% 左右。

代码：`https://github.com/dwro0121/CVFinalProject`

1 介绍

我们的期末项目是关于目标检测的，我们选择复现 CenterNet 模型来进行目标检测。CenterNet 是一种 one-stage 的目标检测方法，其基本思想是用中心点来代表需要检测的物体，通过中心点来确定物体的位置和分类，之后通过回归的方式，利用中心点位置的特征得到物体的宽高等信息，最终实现目标物体的检测。

相较于之前的一些成功的目标检测方法，CenterNet 有着它的优势。之前大部分成功的方法，都需要提出大量的候选区域，再对这些候选区域进行特征提取和分类，最后再对被选中的区域利用非极大值抑制的方法进行后处理。这些方法计算量大，速度慢，并且因为采用了后处理的方法，因此很难实现端到端的学习，而 CenterNet 每个物体都只对应一个中心点，也即每个物体只会对应一个区域，因此不需要非极大值抑制来进行后处理，是端到端可微的，并且更加快速高效。

在复现的过程中，我们选用了 VOC-2007 数据集，并且从该数据集中选出 5 个类别的子集进行训练和测试。该数据集输入图片的大小为 416x416，我们的模型输出的热力图的大小为 104x104，为输入图片的 4 分之 1，与原论文的设定保持一致。在实验中，我们尝试了不同的主体网络对最终检测效果的影响。

2 相关研究

2.1 one-stage 和 two-stage

one-stage 和 two-stage 是目前目标检测的两大类主流方法。two-stage 方法的特点是由一个模型产生候选区域，这个部分可以是深度模型，也可以使 selective search 等传统算法，再由另一个模型对这些候选区域进行判别和分类，其代表主要有 Faster-RCNN, SFP-net, R-FCN 等。而 one-stage 方法的特点则是输入图片后由一个模型直接得到分类结果和目标的范围，代表方法有 YOLO，RetinaNet，CornerNet 等，我们复现的方法 CenterNet 也属于 one-stage 方法。

2.2 非极大值抑制

非极大值抑制（Non-Maximum Suppression，NMS）是一种用于筛选候选区域的后处理算法，其基本流程就是给定候选区域的集合，选出其中置信度最高的候选区域，加入到最终结果中，同时去除掉剩下的候选区域中与该区域的重叠率超过阈值的区域。NMS 是一种常用的去除冗余候选区域的方法，不过使用 NMS 后就无法实现端到端的训练，原 CenterNet 中每个目标物体只对应一个候选区域，因此不需要使用 NMS。

2.3 基于关键点的目标检测

CenterNet 属于利用关键点来表示目标物体的目标检测方法，而在 CenterNet 之前就存在类似的工作，例如 CornerNet 和 ExtremeNet。CornerNet 用物体边界框的左上右下 2 个点来代表物体，ExtremeNet 则是用上下左右以及中心点 5 个点来代表物体。CenterNet 与这两个方法不同之处就在于 CenterNet 不需要在最后阶段进行组合分组，因此获得了更快的检测速度。

3 主要方法

CenterNet 网络是一个 End-To-End 网络，网络包含主体网络和预测网络两个模块。在这一节中分别介绍主体网络，标头网络、损失函数和生成边框等具体的实现方法。

3.1 主体网络

CenterNet 原文中使用 ResNet，Hourglass 和 DLA 三个网络作为主体网络，我们使用其中的 Resnet 为主体网络复现 CenterNet。我们根据 pytorch 官方的代码，在第四个模块后面加上采样网络，构建一个编码-解码网络。我们复现的上采样网络不包含原论文中使用的 deformable 卷积，主体网络的每个卷积层之后使用 Batch Norm 和 ReLU 激活函数。主体网络的输出为大小为原图的四分之一、通道数量为 64 的热力图，基于 ResNet 的主体网络的具体结构在表 1 中所示。

表 1: 基于 Res18 和 Res50 的主体网络

模块	ResNet18	ResNet50
卷积 1	$7\times7, 64, stride2, padding3$	
Res 模块 1	$3\times3, maxpool, stride2$	
	$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
Res 模块 2	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
Res 模块 3	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$
Res 模块 4	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
反卷积 1	$4\times4, 256, stride2, padding1$	
反卷积 2	$4\times4, 128, stride2, padding1$	
反卷积 3	$4\times4, 64, stride2, padding1$	

3.2 标头网络

主体网络的输出是一个通道数量为 64 的热力图，为了检测目标还需要更多的处理。在这里介绍使用主体网络输出的热力图预测目标区域的网络。因为 CenterNet 利用目标物体的中心点和边框的大小来检测物体，因此需要使用两个标头网络分别预测中心点位置和边框的大小。另外，网络输出的热力图大小为原图的四分之一，这样的热力图重新映射到原始图像上的时候会带来精度误差，因此我们需要增加一个标头网络预测偏移值减少这种误差。每个标头网络由两个卷积层组成，第一个卷积层之后使用 Batch Norm 和 ReLU 激活函数，最后的卷积层是一个 1×1 的卷积层。预测中心点的标头网络输出的通道数量为分类数，其他两个标头网络的输出通道数量为 2（分别预测高度、宽度和 x、y 坐标系的偏移值），标头网络的具体结构可以参考表 2，主体网络为 ResNet18 的整体 CenterNet 模型在图 1 所示。

3.3 损失函数

CenterNet 的三个标头网络分别预测三种边框的属性，需要使用三种损失函数进行模型的优化。预测中心点网络的学习使用基于 RetinaNet 的 Focal Loss 的损失函数，Focal Loss 是基于 CE 损失函数的略微修改的函数。该函数对于简单的样本，给出较小的权重，对于困难样本，给出较大的权重。这样，对于简单的样本进行较少的更新，对

表 2: 各标头网络的结构

模块	预测中心点的标头	预测大小的标头	预测偏移值的标头
卷积 1	3×3, 64, *padding*1		
Batch Norm	Batch Norm，64		
激活函数	ReLU		
卷积 2	1× 1, 分类数	1× 1, 2	

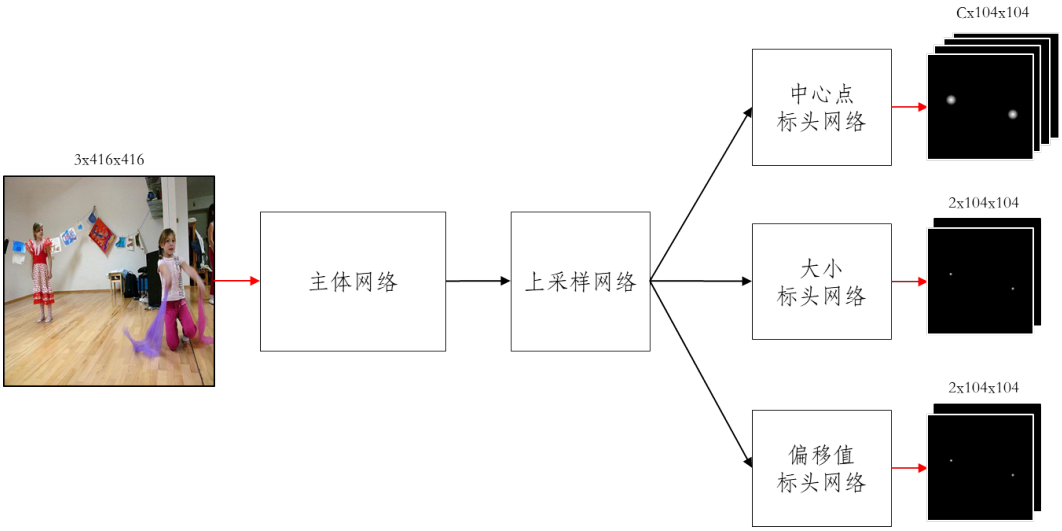


图 1: CenterNet 模型结构

于困难的样本进行较多的更新，并且可以通过关注困难的样本来进行学习。Focal 损失函数计算损失值的公式在 Eq1：

$$L_{hm} = \frac{-1}{N} \sum_{xyc} \begin{cases} (1-\hat{Y}_{xyc})^{\alpha} log(\hat{Y}_{xyc}), & \text{if } Y_{xyc} = 1 \\ (1-Y_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} log(1-\hat{Y}_{xyc}), & \text{otherwise} \end{cases} \tag{1}$$

损失函数中的 α 和 β 为参数，默认为 2 和 4。$\hat{Y}_{xyc} = 1$ 表示检测到中心点，$\hat{Y}_{xyc} = 0$ 表示背景。其中 $\hat{Y}_{xyc}$ 是对热力图进行高斯分布处理后的热力图。但因为我们复现过程中，使用高斯分布处理的结果不太好，最后决定不使用高斯分布，直接带着中心点的热力图作为标签。

预测偏移值和边框大小网络的学习均使用 L1 损失函数，两个损失函数的公式分别在 Eq2 和 Eq3：

$$L_{wh} = \frac{1}{N} \sum_{k=1}^{N} \left| \hat{S}_{p_k} - s_k \right| \tag{2}$$

$$\text{物体 k 的边框坐标} = (x_1^{(k)}, y_1^{(k)}, x_2^{(k)}, y_2^{(k)}), \qquad s_k = (x_2^{(k)} - x_1^{(k)}, y_2^{(k)} - y_1^{(k)})$$

$$L_{off} = \frac{1}{N} \sum_{p} \left| \hat{O}_{\tilde{p}} - (\frac{p}{R} - \tilde{p}) \right| \tag{3}$$

$$O_k = (\frac{x_k}{n} - \lfloor \frac{x_k}{n} \rfloor, \frac{y_k}{n} - \lfloor \frac{y_k}{n} \rfloor)$$

网络整体的损失函数（Eq4）：

$$L = L_{hm} + \lambda_{wh} L_{wh} + \lambda_{off} L_{off} \tag{4}$$

其中 $\lambda_{wh} = 0.1, \lambda_{off} = 1$
这样，CenterNet 可以只使用一个 End-To-End 网络预测中心点、大小和偏移值。

3.4 从特征获得边框

模型的训练可以使用以上的内容进行，需要后处理获得目标边框。首先有通过模型标头的 $C + 2 + 2$ 个热力图，其中 C 个热力图包含中心点的信息。每个分类在这 C 个热力图中选取 peak，peak 就是每一个大小为 3×3 的窗口里置信度最高的坐标。可以把所有的 peak 按置信度高到低排序，每个分类选取了 n 个 peak。然后可以使用 peak 的位置信息，在预测大小和偏移值的热力图获得对应的值，可以参考（Eq5，Eq6）：

$$\hat{\rho} = (\hat{x_i}, \hat{y_i})_{i=1}^{n} \; of \; class \; c \tag{5}$$

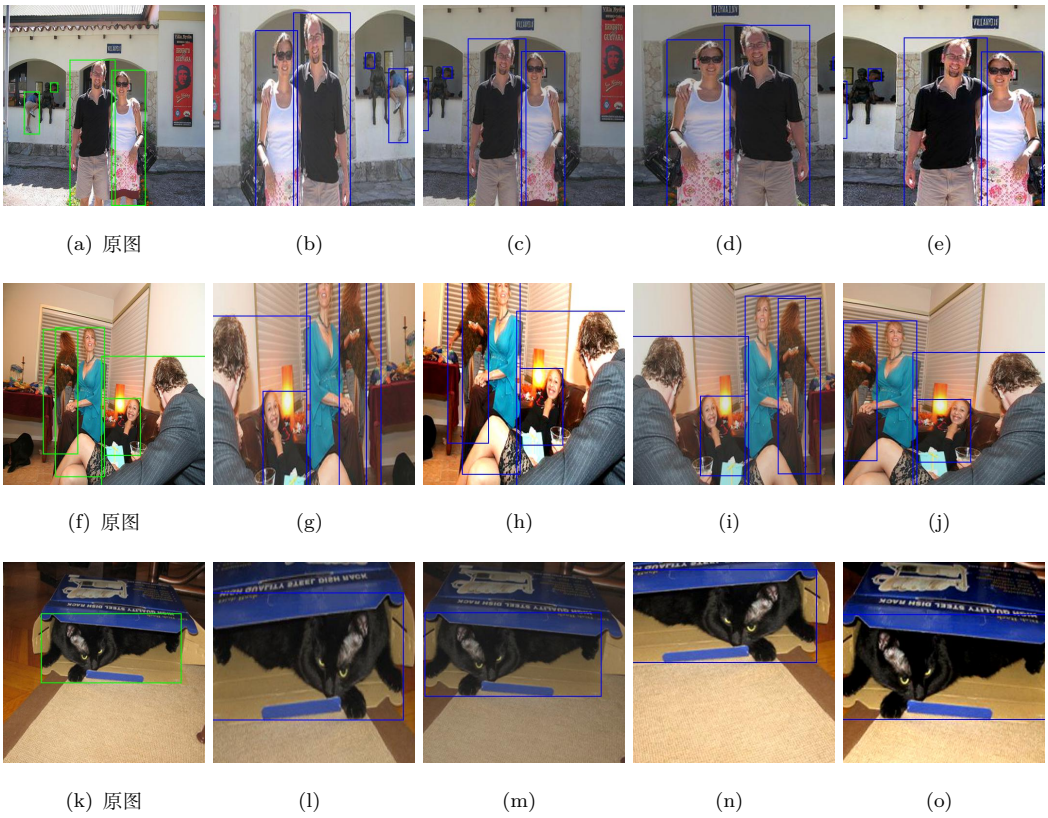
$$BoundingBox = (\frac{\hat{x}_i + \delta_{x_i} - \hat{w}_i}{2}, \frac{\hat{y}_i + \delta_{y_i} - \hat{h}_i}{2}, \frac{\hat{x}_i + \delta_{x_i} + \hat{w}_i}{2}, \frac{\hat{y}_i + \delta_{y_i} + \hat{h}_i}{2}) \tag{6}$$

$$\hat{O}_{\hat{x}_i, \hat{y}_i} = (\delta_{x_i}, \delta_{y_i}) \qquad \hat{S}_{\hat{x}_i, \hat{y}_i} = (\hat{w}_i, \hat{h}_i)$$

可以使用 stirde 为 2，大小为 3×3 的 MaxPool 获取所有的 peak，根据置信度排序之后选取 20 个。然后原论文是在所有的分类，再次选取 n 个 peak。但是，由于我们复现过程中置信度比较低，设置了较低的阈值。结果有了很多边框，还是使用非极大值抑制作为后处理。

3.5 数据增强

数据增强是一种策略，用于提高模型性能的方法，从磁盘读取训练图像，并通过各种方法对原图进行变形处理后将其用作模型的输入图像。如果某一个图像中将所有像素向右移动一个空格，则它们在人眼上看起来相同。但是，由于计算机以像素质量的形式表示和识别图像，因此将移动一个像素的图像识别为与原图不同，为了解决这个问题，使用数据增强。我们在项目中使用 imgaug 库进行数据增强，对每个输入原图和边框随机地进行亮度、对比度、水平反转和裁剪等操作，获得新的图像和对应的边框。原图和数据增强的结果在图 2 所示。



(a) 原图 (b) (c) (d) (e)

(f) 原图 (g) (h) (i) (j)

(k) 原图 (l) (m) (n) (o)

图 2: Image Augmentation with ImgAug.)

4 实验与结果

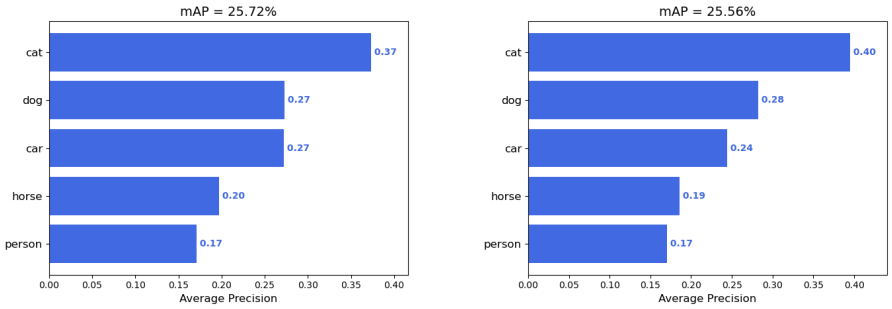
数据集 我们选用了 VOC-2007 数据集，并且从该数据集中选出 5 个类别的子集进行训练和测试。5 各类别分别为猫、狗、汽车、马、人，只保留这五个类别的图像。使用 ImgAug 库，把图像大小缩小为 416×416，同时把对应的边框信息也修改。改完之后训练集和验证集的比率为 8：2，分别有 2536 张和 635 张图像数据。测试集也处理同样的过程，获得了共有 3163 张图像的数据集。

训练策略 我们主体网络使用预训练的权重，优化器使用 Adam 优化器。训练的前 50 个 epoch 冻结主体网络的编码部分，在后 50 个 epoch 训练全部网络。使用 ReduceLROnPlateau 调整学习率，基于训练过程中的测试集误差值对学习率进行状态的下降。

评价指标 AP（Average Precision）是一种用于测量物体检测器精度的流行指标。平均精度将计算 0 到 1 范围内召回值的平均精度值，它会计算 Precision x 召回曲线的曲线下面积（AUC）。因此，评价指标使用 mAP（mean Average Precision）。

图 3 给出了主体网络使用 Res18 和 Res50 网络的结果，两个都在检测猫的时候获得了比较合理的边框。但是检测人的时候 mAP 比较低，说明检测人的时候预测不太

准。图 4 中展示，我们模型预测结果的一部分，左边两个图为检测比较好的结果，右边两个图为比较差的结果。我们复现的模型也带着典型的 One Stage Dectector 具有的缺点，就是检测比较大的物体没有问题，检测小物体有点困难。我们觉得这个问题，可以把输入图像大小更大或者加预处理、后处理过程提高性能。然后我们这个模型预测的热力图中置信度有点低，我们觉得有几个原因。首先，我们模型中没有使用 deformable 卷积，这可能导致了预测的准确率下降。然后，选取 5 个类别生成数据集之后，每个类别的数量不平衡，这样可能导致了训练效率低。图 5 中给出了数据集中每个类别数量的百分比值。最后，可能是我们的数据增强方法生成的图像和边框的问题。因为我们对原图和边框进行增强处理后，会使某个目标物体变得特别小或者不在图像中。这时边框会指定这个小物体或者只有背景的部分，我们觉得这样也会导致准确率低。详细的例子在图 6 中给出，可以参考一下。



(a) Res18 的 mAP　　(b) Res50 的 mAP

图 3: 主体网络为 Res18 和 Res50 模型的 mAP 结果)

(a)　(b)　(c)　(d)

图 4: 网络对测试集检测的结果的一部分。(a) 和 (b) 是比较好的结果，(c) 和 (d) 是比较差的结果

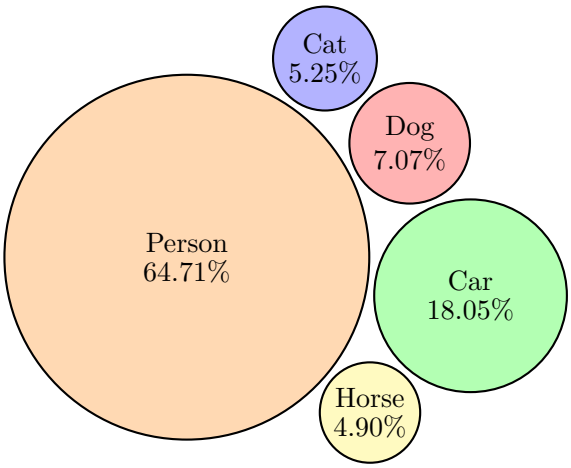


图 5: 数据集中每个类别的分布

(a) 原图　(b) 处理后的图

(c) 原图　(d) 处理后的图

图 6: 对原图进行数据增强处理后问题的例子。图 (b) 是处理图 (a) 获得的图像，图 (b) 中共有 7 个边框。但是最左边的边框人眼也无法识别出他是不是汽车，这样会使模型的准确率下降。图 (d) 是处理图 (c) 获得的图像，图 (d) 包含三个边框。其中两个边框就没看到物体的形状，边框的大部分都是背景，这样也会使模型的准确率下降

5 总结

CenterNet 是一个利用关键点表示目标物体的 one-stage 目标检测方法，利用中心点来表示物体，根据中心点用回归的方式得到检测物体所需的其他特征，其优势就在于快速、高效并且端到端可微。我们复现 CenterNet 的过程中，整体思路与原论文基本一致，但在实验中遇到了一些困难，因此我们也尝试了一些别的技巧，例如处理数据集，使用数据增强，使用 NMS 等，最终的模型在目标检测中也能取得一定的效果。

参考文献

[1] Zhou, X., Wang, D., Krähenbühl, P. (2019). Objects as points. arXiv preprint arXiv:1904.07850.

[2] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[3] Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[4] Law, H., Deng, J. (2018). Cornernet: Detecting objects as paired keypoints. In Proceedings of the European conference on computer vision (ECCV) (pp. 734-750).

[5] Hosang, J., Benenson, R., Schiele, B. (2017). Learning non-maximum suppression. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4507-4515).

[6] Zhou, X., Zhuo, J., Krahenbuhl, P. (2019). Bottom-up object detection by grouping extreme and center points. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 850-859).

[7] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[8] Newell, A., Yang, K., Deng, J. (2016, October). Stacked hourglass networks for human pose estimation. In European conference on computer vision (pp. 483-499). Springer, Cham.

附录：小组成员分工

课堂报告的部分，PPT 制作由鲁东佑完成，课堂展示由鲁东佑和苏可凡共同完成。

代码实现的部分，模型的复现，后续对模型的修改和调试以及最后对实验结果的分析主要由鲁东佑完成，数据处理以及模型的训练主要由苏可凡完成。

最终报告的部分，摘要，主要方法，实验结果，参考文献等部分由鲁东佑完成，介绍，相关工作和总结等部分由苏可凡完成。