计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(10):133−140 [doi: 10.15888/j.cnki.csa.007651]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

卷积神经网络的聚焦均方损失函数设计①

徐　锐[1,2], 冯　瑞[1,2]

[1](复旦大学 计算机科学与技术学院, 上海 201203)
[2](上海视频技术与系统工程研究中心, 上海 201203)
通讯作者: 冯　瑞, E-mail: imvl@fudan.edu.cn

摘　要: 为了提高卷积神经网络在人体姿势估计任务上的精度, 提出了一种基于均方损失函数 (Mean Squared Error, MSE) 的改进损失函数来处理网络学习中回归热点图的前景 (高斯核) 和背景之间像素点不均衡问题, 根据前景与背景不同像素点值对损失函数赋予不同权重, 并将其命名为聚焦均方损失函数 (Focus Mean Squared Error, FMSE). 与均方损失函数相比, 我们提出的聚焦均方损失函数可以有效地减少前景和背景之间像素点不均衡对网络性能的影响, 帮助网络定位关键点的空间位置, 提升了网络性能, 并使得训练阶段中损失函数收敛速度更快. 并在公开数据集上进行实验, 以验证我们所提出的聚焦均方损失函数的有效性.

关键词: 深度学习; 损失函数; 人体姿势估计; 关键点检测; 样本不均衡

引用格式: 徐锐,冯瑞.卷积神经网络的聚焦均方损失函数设计.计算机系统应用,2020,29(10):133–140. http://www.c-s-a.org.cn/1003-3254/7651.html

Focused Mean Square Loss Function Design in Convolutional Neural Network

XU Rui[1,2], FENG Rui[1,2]

[1](School of Computer Science, Fudan University, Shanghai 201203, China)
[2](Shanghai Engineering Research Center for Video Technology and System, Shanghai 201203, China)

Abstract: In order to improve the accuracy of the human pose estimation task of convolutional neural networks, we propose an improved loss function based on Mean Squared Error (MSE) to deal with the pixel imbalance between foreground (Gaussian kernel) and background in heatmaps, assign different weights to the loss function according to different pixel values in the foreground and background, and named it Focus Mean Squared Error (FMSE). Compared with the mean squared loss function, the proposed focused mean squared loss function can effectively reduce the impact of pixel imbalance between foreground and background on network performance, help the network locate the spatial location of key points, improve network performance, and make the loss function converge faster in the training phase. Experiments are performed on public data sets to verify the effectiveness of the proposed focused mean square loss function.

Key words: deep learning; loss function; human pose estimation; key point detection; sample imbalance

从图片中进行 2D 人体姿势估计是许多计算机视觉高阶任务的基础, 例如动作捕捉, 手势识别和活动识别等. 在神经网络出现之前就有很多基于图结构模型 (pictorial structure model) 的方法[1−7] 试图去解决这个问题. 但是随着卷积神经网络和大规模数据集的出现, 可以让网络模型即使在苛刻的场景中也表示出良好的

① 基金项目: 国家重点研发计划 (2017YFC0803702)
Foundation item: National Key Research and Development Program of China (2017YFC0803702)
收稿时间: 2020-03-25; 修改时间: 2020-04-21; 采用时间: 2020-04-28; csa 在线出版时间: 2020-09-30

预测效果, 而无视人体姿势的约束和大的外观的变换. DeepPose[8] 首先将卷积神经网络带入人体姿势估计领域, 就优于所有的传统方法. 之后 Tompson[9] 使用回归热点图方法来取代直接回归坐标值, 目前大部分网络都是直接使用均方损失函数回归热点图的方式来进行学习, 并未考虑到热点图中前景和背景之间像素点不均衡问题, 会导致网络倾向学习背景, 影响网络的性能.

所以本文提出了一个改进的损失函数去解决热点图中前景和背景之间样本不均衡问题, 并命名为聚焦均方损失函数, 通过对前景赋予高权重, 背景赋予低权重, 使得网络学习的重心放在前景部分, 减少背景对网络性能的影响.

本文组织如下: 在第 1 节简要介绍人体姿势估计领域经典网络. 第 2 节介绍所提出的聚焦均方损失函数, 并于均方损失函数对比, 以分析其优点. 第 3 节为实验部分, 通过实验验证我们所提出的聚焦损失函数在公开数据集 MPII[10] 和 MSCOCO[11] 上的性能. 最后在第 4 节我们对全文工作做了总结与展望.

1 相关工作

深度学习出现之前人体姿势估计的主流模型一直是基于树形结构的图模型, 通常是基于 Felzenszwalb 和 Huttenlocher 所提出的高效的图结构方法[12]. 但是随着卷积神经网络和大规模数据集的出现, 深度学习方法主要占据主流. DeepPose 首先将卷积神经网络应用于人体姿势估计领域, DeepPose 网络是基于 AlexNet[13] 结构, 直接回归坐标点. 之后 Tompson 等提出使用网络回归热点图的方法来替代直接回归坐标点值. 回归热点图的优势在于: 可以让网络全卷积, 减少参数量, 并捕捉关键点之间的相关关系以及前景与背景之间的对比关系. CPM[14] 网络使用级联网络去逐级精化网络的预测效果, 同时生成 center map 来约束网络, 把响应归拢到图像中心. HourglassNet[15] 网络由多个 Hourglass Block 串联而成, 每个 block 之间加入损失函数, 进行中间监督, 防止因为网络过深导致的梯度消失. 低分辨率的特征图拥有较大的感受野, 捕捉图像的全局特征, 高分辨率的特征图拥有较小的感受野, 捕捉图像的局部特征, 进行信息交融. 微软提出了基于 Hourglass 改进的 Pose_Resnet[16], 将反卷积层替换原网络的上采样层, 并剔除 shortcut 支路, 并只使用一个 block 就得到 SOTA (State Of The Art) 效果. 之后微软又提出了全新的网络 HRNet[17], 其主干网络保持高分辨特征图不变, 此时高分辨特征图不是从低分辨特征图上采样或者反卷积得到的, 这样可以保留更多细节信息, 以获得更丰富的局部特征信息, 分支网络进行降采样以获取全局特征, 之后再将分支网络的特征图上采样后与主干网络进行信息交互, 因此预测的热点图更加准确. OpenPose[18] 使用向量 (Part Affinity Field, PAF) 对关键点进行建模, 其网络分成两支, 同时预测热点图和 PAF, 根据回归得到的 PAF 对关键点进行聚类.

2 聚焦均方损失函数

目前人体姿势估计领域的网络都是通过回归热点图来完成训练的, 将各关键点的空间位置标识为前景, 其他像素点即为背景. 每张热点图对应一个关键点, 数据集图片中标识多少关键点, 则网络需回归相应的热点图数. 所回归的热点图中前景部分一般使用高斯函数来计算像素点值, 如式 (1) 所示:

$$f = e^{-\frac{(x-x_0)^2+(y-y_0)^2}{2\times\delta^2}} \tag{1}$$

其中, x_0, y_0 是高斯核中心坐标, x, y 是当前坐标, δ 是高斯核方差, 高斯核宽度为$2\delta+1$. 我们取 δ=6, x_0, y_0 设置为热点图的中心坐标, 生成的热点图如图 1 所示.

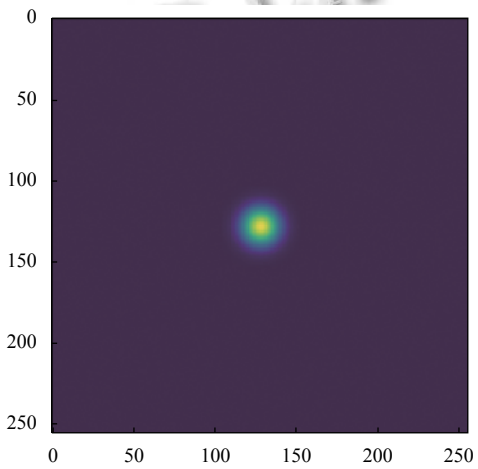


图 1 热点图

图 1 形象地展示了一个标准的热点图, 高斯核 (中心亮斑部分) 只占据热点图全部像素的很小一部分. 而当前大部分网络算法都直接使用均方损失函数直接计算预测的特征图与所标注的热点图像素值之间的欧式

距离, 并没有考虑到高斯核与背景像素点不均衡问题. 假设网络所回归的热点图大小为 64×64 像素大小, 高斯核的大小设置为 13×13 像素大小, 这样前景与背景的比例为 169:3927 (0.00430), 故背景占据了热点图的绝大部分的像素点, 而当前网络直接使用均方损失函数, 计算标注热点图与网络预测热点图的欧式距离, 使得网络平等地对待前景与背景, 可能导致网络更加倾向于回归背景而非前景, 降低了网络的识别率.

而在目标检测领域同样存在着类似的不均衡问题. 在检测网络的起始阶段, 需要通过 SS (Selective Search) 或卷积网络生成一系列的候选框 (Proposal), 在后续阶段根据一些规则, 对这些候选框执行保留, 合并或抛弃等操作, 最后得到网络的检测结果. 但是所生成的候选框大部分都是被合并或抛弃的, 对最终检测结果没有贡献, 即是负样本. 例如在 Faster-R-CNN[19] 网络的起始阶段会生成约 2000 个候选框, 但是正样本候选框可能只有几个, 正负样本比例严重失衡. 为解决这个问题, 有学者在交叉熵损失函数 (如式 (2) 所示) 的基础上, 提出了 focal loss[20], 根据标注标签对损失函数赋予权重, 使得降低了大量简单负样本在训练中的所占比重, 数学形式如式 (3) 所示.

$$Cross\_Entropy\_Loss=\begin{cases}\log_2 y' & y=1\\ \log_2(1-y') & y=0\end{cases} \tag{2}$$

$$Focal\_Loss=\begin{cases}-\alpha(1-y')^{\gamma}\log_2 y' & y=1\\ -(1-\alpha)y'^{\gamma}\log_2(1-y') & y=0\end{cases} \tag{3}$$

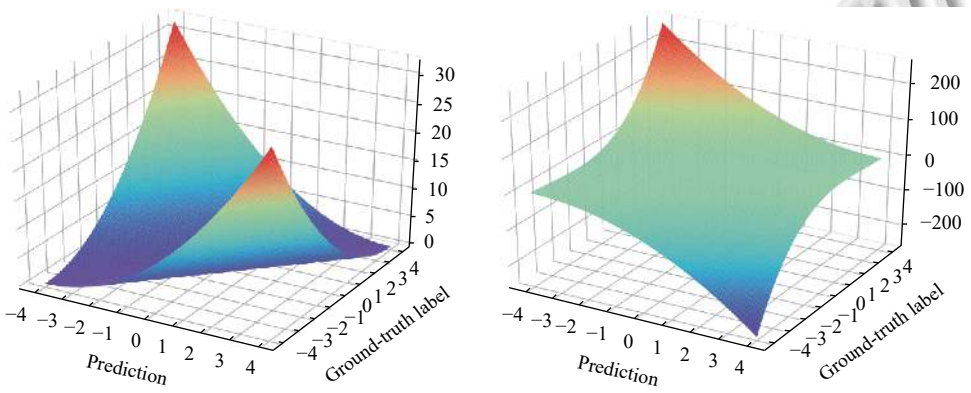
式 (3) 是 focal loss 的数学形式, 其中 α 是平衡因子, 以平衡正负样本不均衡问题, y 是标注标签数据, y'是网络所预测的数据, γ 是调节简单样本权重降低的速率, 当 γ 为 0 时 focal loss 退化为交叉熵损失函数, 当 γ 增加时, 调整因子的影响也在增加.

本文亦受目标检测领域中 focal loss 所启发, 在均方损失函数的基础上进行修改, 增加前置权重, 根据热点图中各像素点值对损失函数赋予不同的权重, 使得网络更倾向于学习高斯核的位置, 对于回归背景位置则施予更大的惩罚.

$$MSE\_Loss=\frac{1}{2}\sum_{i=1}^{n}(y'_i-y_i) \tag{4}$$

$$FMSE\_Loss=\frac{1}{2}\sum_{i=1}^{n}(y_i+\delta)^{\gamma}(y'_i-y_i) \tag{5}$$

均方损失函数和聚焦均方损失函数的数学形式分别如式 (4) 和式 (5) 所示, 而各自函数图像如图 2 所示.



(a) 均方损失函数图像 (b) 聚焦均方损失函数图像

图 2 均方损失函数与聚焦均方损失函数图像

在式 (4) 中, y_i 为标注标签, y'_i是网络所预测的标签, n 是每个 mini batch 所含的样本数.

在式 (5) 中, 我们在均方损失函数的基础上加入了 $(y_i+\delta)^{\gamma}$前置权重, 使的它对前景与背景赋予不同的权重. y_i 为标注标签, y'_i是网络所预测的标签, n 是每个 mini batch 所含的样本数, δ 是极小值, 因为背景中像素点y_i为 0, 会导致损失函数为 0, 使得网络无法学习, 本文中σ取 1e−3. γ 是平衡因子, 调整标注标签对损失函数的贡献大小. 当 y_i 处于热点图的高斯核中, y_i 值较大, 此时对 loss 赋予较高的权重. 当 y_i 处于背景位置时,

y_i 值较小, 此时对 loss 赋予较低的权重. 当 γ 为 0 时, 则聚焦均方损失函数退化为普通均方损失函数. 当 γ 增加时, 像素点 y_i 的值对 loss 权重的影响也在增加.

为探究本文所提出的聚焦均方损失函数的 γ 值对网络性能的影响影响, 我们在 MPII 公开数据集上使用 HouglassNet 作为实验网络, 设置不同的聚焦均方损失函数 γ 值, 并固定其他实验条件, 得到最终实验结果如图 3 所示.

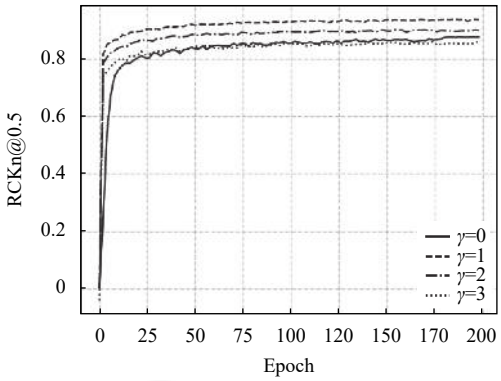


图 3 聚焦均方损失函数的 γ 值影响

观察实验结果图 3 可以得到, 当 γ=0 时, 此时聚焦均方损失函数退化为普通的均方损失函数, 在训练 200 轮后, 网络在验证集上的成绩为 88.6. 而 γ=1 时, 所提出的损失函数加速了网络的收敛速度, 并比使用均方损失函数的成绩提高了 4.6 的成绩, 达到了 93.2. 当 γ 值继续增加, 网络的性能反而下降, 这说明背景信息同样可以帮助网络定位关键点位置, 而此时聚焦损失函数把过多的权重分配给前景, 背景在损失函数中所占过低, 网络丢失了背景信息, 导致关键点定位精度, 因此一个合适的 γ 值是非常重要的.

所以在本文的后续实验部分, 我们设置聚集均方损失函数的 γ 为 1.

3 实验及分析

本节详细介绍实验结果以及分析, 首先简单介绍实验所使用的网络, 之后在介绍实验配置, 展示实验结果并做出相应的分析.

3.1 实验所选用的网络

本文实验使用沙漏网络 (HourglassNet) 和高分辨率网络 (HRNet) 作为基准网络, 来测试我们所提出的聚焦均方损失函数的有效性.

沙漏沙漏网络 (HourglassNet) 是一种新颖的卷积网络架构, 利用多尺度特征来捕捉人体各个关键点的空间位置信息, 网络结构形似沙漏状, 重复使用 top-down 到 bottom-up 来推断人体的关节点位置. 每一个 top-down 到 bottom-up 的结构都是一个 stacked hourglass 模块 (Hourglass Block), 并在每个 Block 之间都加入 loss 进行中间监督, 以防止网络过深导致梯度消失. 考虑到参数问题, 本文中我们使用含有 8 个沙漏模块的沙漏网络, 结构如图 4 所示.

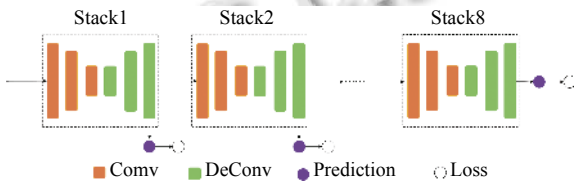


图 4 沙漏网络结构

沙漏网络主要由卷积层, Batch Normaliztion, 上采样层和 skip connection, 激活函数选取 ReLU.

而在 HRNet 网络的预处理阶段对图像进行提取特征, 之后逐步将高到低分辨率子网逐个添加以形成更多的阶段, 并将多分辨率子网并行连接. 网络进行了反复的多尺度融合, 以便每个高到低分辨率表示不断地从其他并行表示中接收信息, 从而获得丰富的高分辨率表示. 网络的高分辨率特征图不再是由低分辨率特征图上采样或反卷积得到, 而是在主干网上保持, 这样可以保留更多的细节信息, 网络所预测的关键点热点图在空间上的位置更加精确. 其网络结构如图 5 所示.

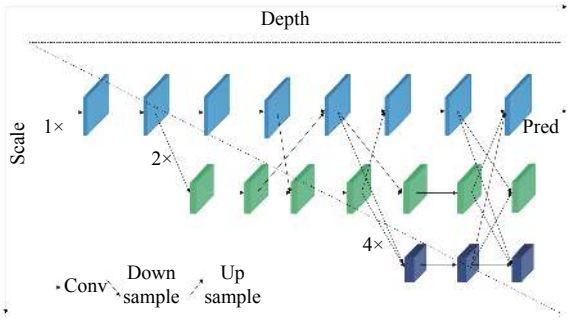


图 5 高分辨率网络结构

HRNet 结构分为纵向 Depth 和横向 Scale 两个维度, 横向上不同分辨率子网络并行, 纵向上进行多分辨率信息融合, 从上到下, 每个 stages 分辨率减半, 通道

数加倍.

3.2 MPII 和 MSCOCO 数据集

本文实验中我们选用 MPII 和 MSCOCO 为实验数据集, 并分别在各自评价指标下 (PCKh@0.5 和 OKS) 报告我们所提出的聚焦损失函数在应用在 HourglassNet 和 HRNet 网络的性能.

3.3 训练与测试信息

训练阶段我们使用 Faster-R-CNN 作为人体检测网, 并将人体检测框的高度或者宽度拓展到固定宽高比为 4:3, 然后从图片中裁剪人体检测框, 并调整图片到固定大小 256×256 像素大小. 对图片使用数据增强技术, 包括随机旋转 ([−30°, 30°]), 随机比例大小 ([0.65, 1.35]), 和随机左右翻转 (p=0.5). 我们选用 Adam 优化算法, 在前 10 轮训练中使用较小的学习率 (1e−6) 对网络进行预热, 并在 10 轮之后学习率设定为 1e−3, 之后分别在 100 轮训练后下降到 1e−5 和在 150 轮训练后下降到 1e−6. 在 200 轮后训练过程结束.

测试阶段我们与训练阶段一致, 使用 Faster-R-CNN 作为人体检测网, 从图片中裁剪人体检测框, 并调整图片到固定大小和左右翻转, 输入到关键点检测网络中, 回归热点图. 我们通过平均原始图片及其翻转图片的网络回归的热点图来作为最终预测的热点图.

3.4 实验环境

硬件信息: 4 块 Titan Xp 12 GB 显卡, CPU: Intel Xeon E5-2620, RAM: 128 GB, DISK: 4 TB.

软件信息: 操作系统是 Linux Ubuntu16.04, cuda 环境是 cuda9.0+cudnn7, 深度学习框架是 Pytorch 1.0.0.

3.5 实验结果及分析

根据前文设置的实验环境, 训练与测试信息等, 我们在表 1 报告在 MPII 数据集上的聚焦均方损失函数应用于 HourglassNet 和 HRNet 网络的实验成绩, 并在图 5 展示详细的训练与测试信息.

从表 1 可看出, 在 MPII 数据集上的 PCKh@0.5 评价标准下, 使用所提出的聚焦均方损失函数的 HourglassNet 比使用均方损失函数的在难以预测的关键点 (Wrist 和 Ankle) 分别提升了 1.3 和 1.9 的成绩, 而在易于预测的关键点上, 我们同样也提升了网络的性能, 平均成绩达到了 88.0, 高于原有成绩 87.5. 对于 HRNet, 聚焦均方损失函数表现出相似的结果. 这充分说明了我们所提出的聚焦均方损失函数可以有效地帮助网络去学习那些困难的关键点, 提升了网络的性能.

表 1 MPII 数据集上实验结果

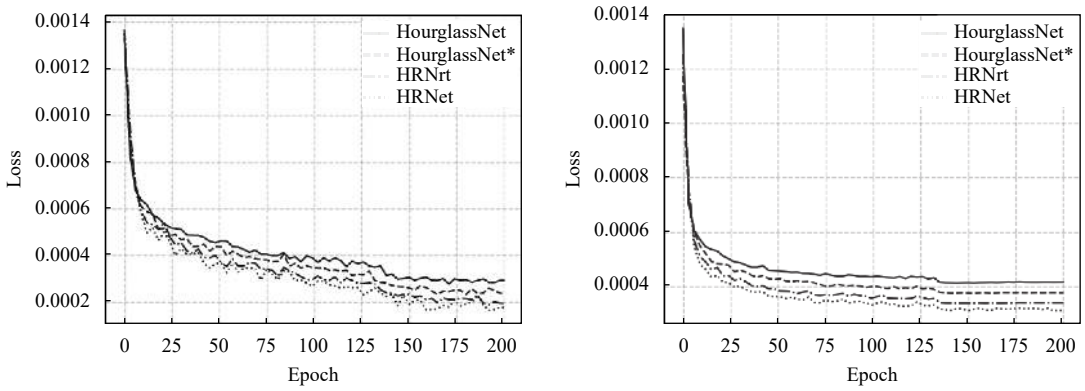
Arch	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
HourglassNet	96.4	95.3	89.1	83.2	88.4	84.0	79.6	87.5
HourglassNet*	95.9	96.3	89.1	84.5	88.0	86.2	81.5	88.0
HRNet	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3
HRNet*	97.0	96.0	89.4	87.0	89.0	86.5	85.0	91.8

注: 加*说明使用聚焦均方损失函数, 未加*使用均方损失函数.

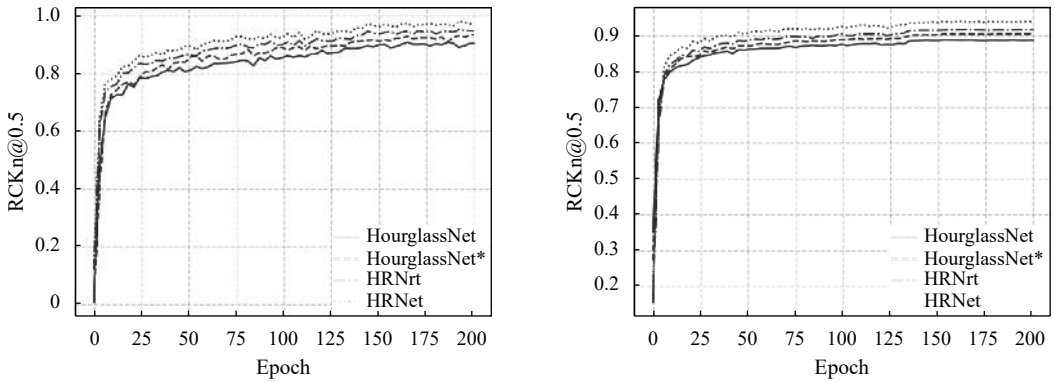
通过图 6 可以看到, 在 MPII 数据集上使用聚焦均方损失函数的网络 (HourglassNet 和 HRNet) 在训练集和验证集上的准确率均比使用均方损失函数的网络的更高, 其在训练集上的准确率提升了 1.2%, 验证集上提升了 0.9%, 而且其 loss 的收敛的速度更快, 说明了聚焦均方损失函数可以有效地提升网络的性能. 我们又分别在表 2 和图 4 分别报告在 MSCOCO 数据集上的实验成绩和详细信息.

在 MSCOCO 数据集上的 OKS 评价标准下, 我们使用所提出的聚焦均方损失函数的 HourglassNet 相比比于直接使用均方损失函数的精准率 (AP) 和召回率 (AR) 分别提升了 0.021 和 0.020 的成绩, 并可以看出聚焦均方损失函数提升了 AP(M) 和 AR(M) 成绩, 说明网络有助于学习那些小的人体关键点. 对于 HRNet, 聚焦均方损失函数分别提升了 0.09 的 AP 成绩和 0.07 的 AR 成绩.

通过图 7 可以得到, 在 MSCOCO 数据集上使用聚焦均方损失函数的 HourglassNet 和 HRNet 网络模型在训练集和验证集上的准确率均比均方损失函数的更高, 准确率分别提升了 2.1% 和 0.8%, 召回率分别提升了 3.4% 和 1.2%, 而且其 loss 的收敛的速度更快且值更小. 使用聚焦均方损失函数的 loss 震荡也比使用均方损失函数要小. 本文所提出的聚焦均方损失函数可以有效地提升网络在 MSCOCO 数据集上的预测精度. 图 8 是关键点检测结果示例.



(a) MPII 训练集上损失函数 (b) MPII 验证集上损失函数



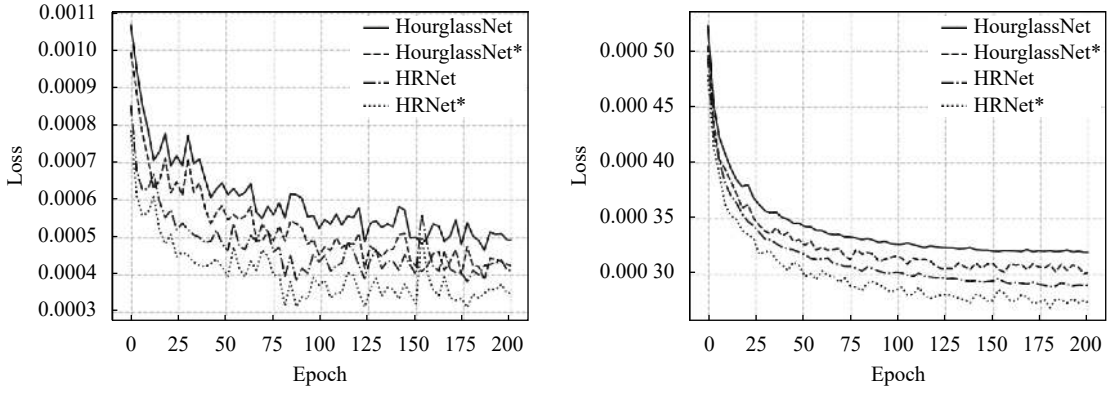
(c) MPII 训练集上准确率 (d) MPII 验证集上准确率

图 6 MSCOCO 数据集上训练与验证信息

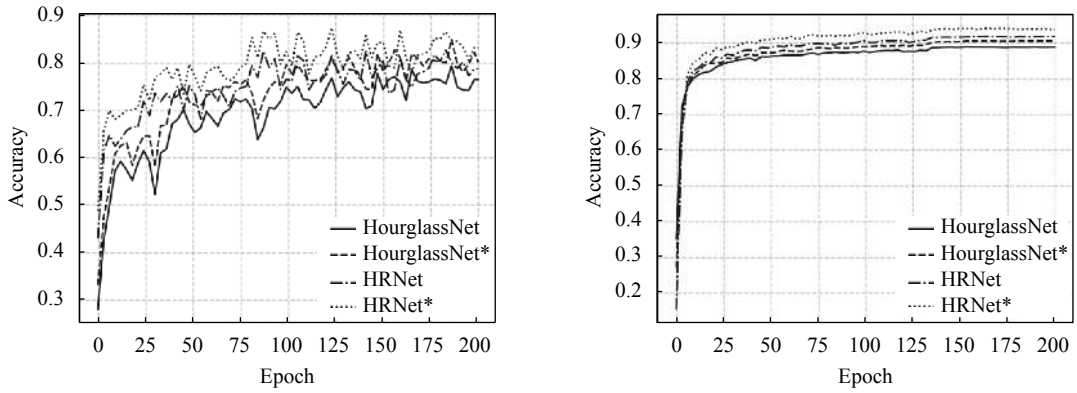
表 2 MSCOCO 数据集上实验结果

Arch	AP	AP(M)	AP(L)	AR	AR(M)	AR(L)
HourglassNet	0.714	0.651	0.772	0.763	0.711	0.824
HourglassNet*	0.735	0.681	0.780	0.783	0.752	0.853
HRNet	0.734	0.708	0.810	0.798	0.757	0.858
HRNet*	0.742	0.719	0.811	0.805	0.766	0.870

注: 加*说明使用聚焦均方损失函数, 未加*使用均方损失函数.



(a) MSCOCO 训练集上损失函数 (b) MSCOCO 验证集上损失函数



(c) MSCOCO 训练集上准确率 (d) MSCOCO 验证集上准确率

图 7 MSCOCO 数据集上训练与验证信息

图 8 关键点检测结果示例

4 总结与展望

文中提出的方法并不像大多数方法那样对网络结构进行修改来提升准确率, 而是重新设计损失函数来解决热点图中前景和背景之间不均衡问题, 让网络学习的重点放在前景中高斯核部分, 减少背景噪声对网络性能的干扰. 之后将所提出的聚焦损失函数应用在经典网络 (HourglassNet 和 HRNet), 展示了在公开数据集 (MPII 和 MSCOCO) 下的实验结果, 从各个角度对实验结果进行了分析, 证明了本文提出的聚焦损失函数具有较高的精度和鲁棒性.

在未来的研究工作中, 我们将会在其他的网络模型上进行实验, 用于验证聚焦均方损失函数的实用性和鲁棒性. 同时, 针对如何对网络结构本身进行设计和改进来提升性能, 也需要在未来的科研中更加深入的研究.

参考文献

1 Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA. 2008. 1–8.

2 Andriluka M, Roth S, Schiele B. Pictorial structures revisited: People detection and articulated pose estimation. Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. 2009. 1014–1021.

3 Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. Proceedings of the 2nd European Conference on Computational Learning Theory. Barcelona, Spain. 1995. 23–37.

4 Eichner M, Ferrari V, Zurich S. Better appearance models for pictorial structures. Proceedings of British Machine Vision Conference. London, UK. 2009.5..

5 Sapp B, Jordan C, Taskar B. Adaptive pose priors for pictorial structures. Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 422–429.

6 Yang Y, Ramanan D. Articulated pose estimation with flexible mixtures-of-parts. Proceedings of CVPR 2011. Providence, RI, USA. 2011. 1385–1392.

7 Pishchulin L, Andriluka M, Gehler P, *et al*. Poselet conditioned pictorial structures. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition.

Portland, OR, USA. 2013. 588–595.

8 Toshev A, Szegedy C. Deeppose: Human pose estimation via deep neural networks. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 1653–1660.

9 Tompson J, Jain A, LeCun Y, *et al*. Joint training of a convolutional network and a graphical model for human pose estimation. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2014. 1799–1807.

10 Andriluka M, Pishchulin L, Gehler P, *et al*. 2D human pose estimation: New benchmark and state of the art analysis. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 3686–3693.

11 Lin TY, Maire M, Belongie S, *et al*. Microsoft coco: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland. 2014. 740–755.

12 Felzenszwalb PF , Huttenlocher DP. Efficient matching of pictorial structures. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Hilton Head Island, SC, USA. 2000, 2. [doi: 10.1109/CVPR.2000.854739]

13 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097–1105.

14 Wei SE, Ramakrishna V, Kanade T, *et al*. Convolutional pose machines. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 4724–4732.

15 Newell A, Yang KY, Deng J. Stacked hourglass networks for human pose estimation. Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands. 2016. 483–499.

16 Xiao B, Wu HP, Wei YC. Simple baselines for human pose estimation and tracking. Proceedings of the 15th European Conference on Computer Vision. Munich, Germany. 2018. 472–487.

17 Sun K, Xiao B, Liu D, *et al*. Deep high-resolution representation learning for human pose estimation. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 5686–5696.

18 Cao Z, Simon T, Wei SE, *et al*. Realtime multi-person 2D pose estimation using part affinity fields. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 1302–1310.

19 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

20 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy. 2017. 2999–3007.