论文复现报告

谢敏儿

摘要

本文旨在总结复现论文《用残差对数似然估计回归人体姿态》过程中的收获。目前的基于回归的方法效率高，但性能较差。该工作提出的带有残差对数似然估计的回归范式使基于回归的方法首次优于基于热图的方法，尤其是在多人姿态估计方面。本文在第一部分介绍了选题背景；第二部分介绍了人体姿态估计的相关工作；第三部分详细介绍了该工作提出的方法；第四部分介绍了复现细节；第五部分介绍了实验结果；最后对本文内容进行了总结，并对未来工作进行展望。

关键词：人体姿态估计；最大似然函数；回归范式

1 引言

基于热图的方法通过似然热图对输出分布进行建模，从而在人体姿态估计领域占据主导地位。相比之下，基于回归的方法效率更高，但性能较差。在这项工作中[1]，作者探索最大似然估计（MLE）以开发一种高效且有效的基于回归的方法。从 MLE 的角度来看，采用不同的回归损失是对输出密度函数做出不同的假设。更接近真实分布的密度函数会导致更好的回归性能。鉴于此，作者提出了一种带有残差对数似然估计 (RLE) 的新型回归范式来捕获潜在的输出分布。具体来说，RLE 学习分布的变化而不是未参考的底层分布，以促进训练过程。通过提出的重新参数化设计，该工作的方法与现成的流模型兼容。所提出的方法有效、高效且灵活。

2 相关工作

2.1 基于热图的姿态估计

Tompson 等人提出了利用似然热图来表示人体关节位置的想法。从那时起，基于热图的方法在二维人体姿态估计领域占据主导地位。先前的工作设计了强大的 CNN 模型来估计单人姿势估计的热图。许多工作将这个想法扩展到遵循自上而下框架的多人姿态估计，即检测和单人姿态估计。在自下而上的框架中，从热图中检索多个身体关节并将其分组为不同的人体姿势。Pavlakos 等人首先将热图扩展到三维空间。孙等利用 soft-argmax 操作以可区分的方式从热图中检索关节位置，从而允许端到端训练。它可以防止量化错误，但模型仍然需要生成高分辨率特征和热图。

2.2 基于回归的姿态估计

在人体姿态估计的背景下，只有少数工作是基于回归的。Toshev 等人首先利用卷积网络进行人体姿态估计。Carreira 等人提出了一个迭代误差反馈（IEF）网络来提高回归模型的性能。Zhou 等和 Tian 等人在单阶段目标检测框架中提出直接姿态回归。Nie 等将长程位移分解为累积的较短位移。但是，它对有遮挡的情况很脆弱。Wei 等回归关于预定义的姿态锚点的位移。在三维姿态估计中，Sun 等人提出组合姿态回归来学习三维人体姿态的内部结构。Rogez 等人将人体姿态分类为一组 K 个锚点姿势，并提出了一个回归模块来将锚点细化为最终预测。两阶段方法通过回归将二维姿态提升到三维空间。但是二维姿态仍然由基于热图的二维姿态估计器预测。尽管以前的工作取得了很多进展，但纯基于回归的方法和基于热图的方法之间仍然存在巨大的性能差距。

2.3 人体姿态估计中的标准化流

最近的一些工作利用标准化流来构建三维人体姿势估计的先验。Xu 等提出了新的三维人体形状和关节姿态模型，该模型具有基于标准化流的运动学先验。Zanfir 等使用标准化流为其弱监督方法建立 SMPL 关节角度的先验。Biggs 等通过标准化流以从模糊图像中采样最佳输出来先学习姿态先验。与以前的方法不同，我们利用归一化流来估计潜在的输出分布。

2.4 自适应损失函数

在本文的方法中，输出分布是可学习的，从而产生可学习的损失函数。已经有几项针对自适应损失函数的工作。Imani 等人提出直方图损失，它使用直方图（即热图）来表示输出分布。一些工作定义了损失函数的超集，并通过调整函数的参数来改变损失。Wu 等使用教师模型动态改变学生模型的损失函数。Barron 提出了常见损失函数的泛化，它会在训练过程中自动调整自身。

3 本文方法概述

3.1 从传统的回归方法开始

在传统的方法中，L1 和 L2 损失被直接应用到模型输出，我们不知道如何去选择恰当的损失函数，只能基于经验去选择。但当我们从最大似然估计 (MLE) 的角度看待该问题，不同的回归损失其实就是不同的对输出分布的假设。网络预测表示回归值概率的密度函数。

$$\mathcal{L}_{\text{mle}} = -\log P_{\Theta}(\mathbf{x} \mid \mathcal{I})|_{\mathbf{x}=\boldsymbol{\mu}_g}.$$

假设预测的密度函数是高斯分布，网络预测两个值：μ 和 σ，去表示一个高斯分布。假设变化是常量，MLE 损失变为标准的 L2 损失：

$$\mathcal{L} = -\log P_{\Theta}(\mathbf{x} \mid \mathcal{I})|_{\mathbf{x}=\boldsymbol{\mu}_g} \propto \log \hat{\boldsymbol{\sigma}} + \frac{(\boldsymbol{\mu}_g - \hat{\boldsymbol{\mu}})^2}{2\hat{\boldsymbol{\sigma}}^2}.$$

$$\mathcal{L} = (\boldsymbol{\mu}_g - \hat{\boldsymbol{\mu}})^2.$$

同样地，如果假设该分布是一个常量变化的拉普拉斯分布，MLE 损失其实就是 L1 损失。

如果我们用一个模型去学习一个合适的分布，而不是人为地预定义一个分布的类型，回归的表现是否会更好？所以本文工作的核心理念是用标准化流去自主地学习输出分布，以获取更好的回归表现。

3.2 标准化流

标准化流是具有易处理分布的生成模型的一员，其中采样和密度评估都是高效和精确的。标准化流通过可逆转映射将简单分布转换为复杂分布。标准化流基础：变量替换。变量替换是概率论中的一个定理。假设有一个随机变量 x，可以对它进行一个变换 f(x)，变换称为 z，要用变量 z 去替换变量 x，它的分布 P(z) 和 P(x) 之间差的就是一个行列式。这个行列式是 f(x) 的 jacobian 矩阵。jacobian 矩阵其实就是告诉我们原先的空间到新的空间之后，它的方向上和模值上产生了一个多大的变化。行列式其实是它方向上和体积上的变化，取绝对值后就把方向给丢掉了。

流的组合：

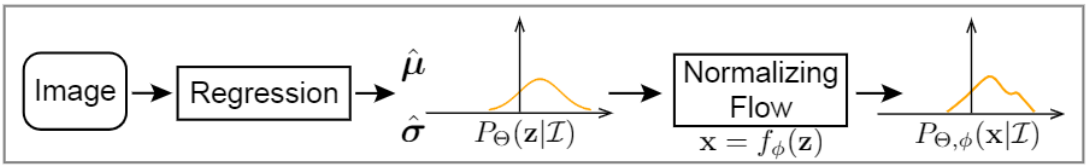
基本原理：可导的可逆的函数在进行组合后依然是一个可导可逆函数。实际上，我们可以连续组合几个简单的映射来构造任意复杂的函数，即：$\mathbf{x} = f_{\phi}(\mathbf{z}) = f_K \circ \cdots \circ f_2 \circ f_1(\mathbf{z})$.

在本文的工作中，采用 RealNVP 为流模型。RealNVP 的主要思想是通过精心设计一个神经网络，

让输出对应输入的导数行列式可以转化为类似于对角线的矩阵, 便于计算。

3.3 用标准化流进行回归

1）Basic Design



(a) Basic Design

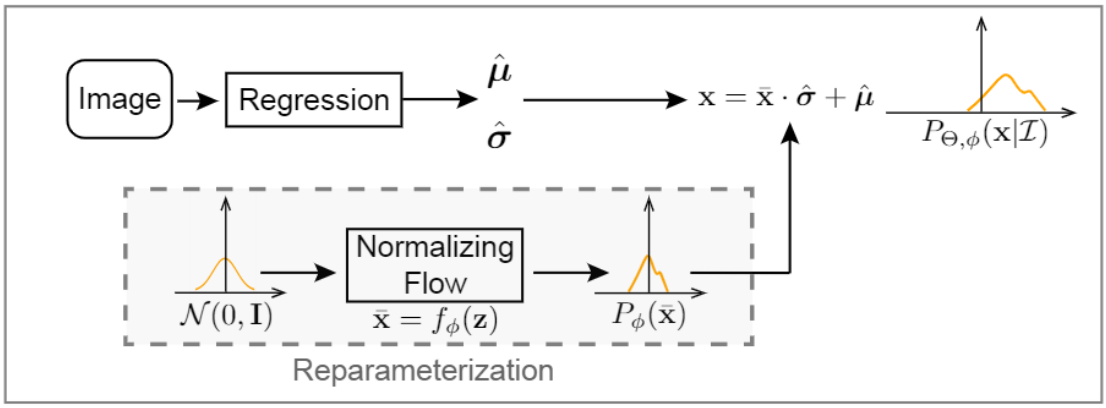
图 1: (a)Basic Design

Basic Design 是让回归模型去回归位置和标准差。这两个参数可以定义一个高斯分布。然后我们可以用标准化流去将这个高斯分布转换为一个可学习的、任意的复杂分布。损失函数为：

$$\begin{aligned}\mathcal{L}_{\text{mle}} &= -\log P_{\Theta,\phi}(\mathbf{x} \mid \mathcal{I})|_{\mathbf{x}=\boldsymbol{\mu}_g} \\ &= -\log P_\Theta\left(f_\phi^{-1}(\boldsymbol{\mu}_g) \mid \mathcal{I}\right) - \log\left|\det \frac{\partial f_\phi^{-1}}{\partial \boldsymbol{\mu}_g}\right|.\end{aligned}$$

但该方案是不可行的。因为分布取决于输入图像，如果我们直接训练该模型，标准化流会拟合数据集上所有图像的分布，这不是我们想要的。我们想学习的分布是关于输出是如何偏离输入图像的真实值，而不是所有输入图像的真实值的分布本身。

2）Direct likelihood estimation with reparameterization



(b) Direct likelihood estimation with reparameterization

图 2: (b)Direct likelihood estimation with reparameterization

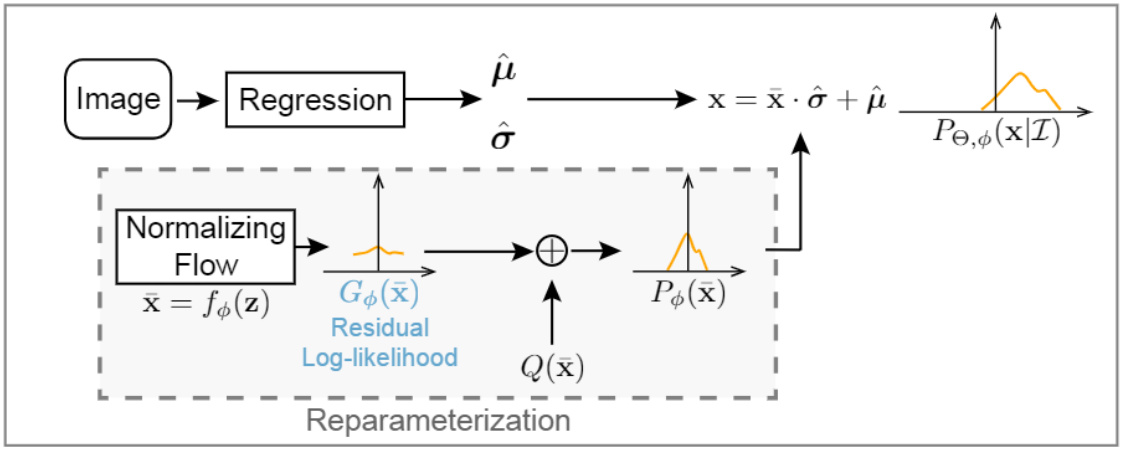
为了构建一个可行的方案，我们通过重参数化来解耦图像。我们假设所有的基础分布共享相同的密度函数族，但对于输入图像 I 具有不同的均值和标准差。首先，流模型学习零均值分布，然后回归模型预测两个值：μ 和 σ，去控制分布的位置和比例，流模型基于回归输出对零均值分布进行 reshape 和 rescale，最终分布是通过移动和缩放 x 得到的。这样，我们可以构建一个可学习的分布，并用最大似然估计损失同时训练回归模型和流模型。损失函数为：

$$\begin{aligned}\mathcal{L}_{\text{mle}} &= -\log P_{\Theta,\phi}(\mathbf{x} \mid \mathcal{I})|_{\mathbf{x}=\boldsymbol{\mu}_g} \\ &= -\log P_\phi\left(\overline{\boldsymbol{\mu}}_g\right) - \log\left|\det \frac{\partial \overline{\boldsymbol{\mu}}_g}{\partial \boldsymbol{\mu}_g}\right| \\ &= -\log P_\phi\left(\overline{\boldsymbol{\mu}}_g\right) + \log \hat{\boldsymbol{\sigma}},\end{aligned}$$

其中：$\overline{\boldsymbol{\mu}}_g = (\boldsymbol{\mu}_g - \hat{\boldsymbol{\mu}})/\hat{\boldsymbol{\sigma}}$, and $\partial \overline{\boldsymbol{\mu}}_g / \partial \boldsymbol{\mu}_g = 1/\hat{\boldsymbol{\sigma}}$.

该方法是可行的，但不够好。因为回归模型的训练完全取决于流模型学习的分布。当我们从零开始训练模型时，分布是远远不正确的，这将对回归模型的训练有害。

3）Residual log-likelihood estimation 所以进一步提出了 RLE(Residual log-likelihood estimation)。



(c) Residual log-likelihood estimation with reparameterization

图 3: (c)Residual log-likelihood estimation

假设 Popt 是我们想学习的分布，我们可以将它分成三项：

$$\begin{aligned}\log P_{\text{opt}}(\bar{\mathbf{x}}) &= \log\left(Q(\bar{\mathbf{x}}) \cdot \frac{P_{\text{opt}}(\bar{\mathbf{x}})}{s \cdot Q(\bar{\mathbf{x}})} \cdot s\right) \\ &= \log Q(\bar{\mathbf{x}}) + \log \frac{P_{\text{opt}}(\bar{\mathbf{x}})}{s \cdot Q(\bar{\mathbf{x}})} + \log s,\end{aligned}$$

第一项是一个简单且可行的分布，如高斯分布；第二项是残差对数项，它描述了简单分布 Q(x) 是怎样变化到最优分布的；第三项是常数项，它被用来 rescale 第二项以确保它是一个密度函数。在 RLE 中，用分布 Q(x) 是基本正确但不够完美的，所以与其拟合原始分布，我们让流模型去学习分布的变化，就是残差项。在这种方法下，在训练的开始，尽管流模型什么都没学习，分布也是基本正确的，并且回归模型可以学习得很快。除此之外，就像 Resnet 的思想，学习残差将会比学习原来的 mappings 更简单。

4 复现细节

4.1 与已有开源代码对比

1）主要用了作者开源的代码完成该复现工作，实现了在 COCO 数据集上用该方法进行姿态估计。由于还没有很好的想法去改进本文工作，所以看了一些其他工作的改进方法，希望之后能想到一些较好的方法来改进。

2）其他工作的改进方法：Poseur: Direct Human Pose Regression with Transformers[2] 该工作指出，RLE 采用的全局平均池化会破坏卷积特征图的空间结构，对性能产生一定影响。Poseur 中使用了 Transformer 模型，其注意力机制可以更好地利用关节点之间的结构化依赖关系。Poseur 遵循 RLE，预测一个反映每个位置出现 gt 概率的概率分布，并以最大概率监督网络在 gt 上。

3）总结：目前的一些新工作一般都是把 RLE 中提出的损失函数应用在其他框架中，没有直接对 RLE 结构进行修改。之后可以考虑更换主干网络/流模型的实现方式，观察是否可以使 RLE 方法获得更好的效果；也可以考虑将 RLE 应用在其他关键点检测的任务中，观察效果。

5 实验结果分析

5.1 原论文中实验结果

如图所示，在实验中发现 RLE 较大地提升了传统的回归方法。在二维人体姿态估计中，mAP 从 58 提升生到 71。

Method	# Params	GFLOPs	AP	AP_{50}	AP_{75}
Direct Regression (with ℓ_1)	23.6M	4.0	58.1	82.7	65.0
Regression with DLE	23.6M	4.0	62.7	86.1	70.4
Regression with RLE	23.6M	4.0	70.5	88.5	77.4
***Regression with RLE**	23.6M	4.0	**71.3**	**88.9**	**78.3**

Table 1: **Comparison with the conventional regression paradigm.** RLE provides significant improvements with the same test-time computational complexity.

图 4: RLE 与传统回归范式对比

作者更深入地比较了本文的回归方法与基于热力图的方法，如图所示，RLE 提升了所有基于热图的方法。

	Method	AP	AP_{50}	AP_{75}
(a)	SimplePose [67]	71.0	**89.3**	**79.0**
	Integral Pose [57]	63.0	85.6	70.0
	***Regression with RLE**	**71.3**	88.9	78.3
(b)	HRNet-W32 [55]	74.1	**90.0**	**81.5**
	HRNet-W32 + RLE (Regression)	**74.3**	89.7	80.8
(c)	Mask R-CNN [15]	66.0	**86.9**	71.5
	Mask R-CNN + RLE	**66.7**	86.7	**72.6**
(d)	PointSet Anchor [64]	67.0	87.3	73.5
	PointSet Anchor + RLE	**67.4**	**87.5**	**73.9**

Table 2: **Comparison with heatmap-based methods** on COCO validation set. The proposed paradigm achieves competitive performance to the heatmap-based methods.

图 5: RLE 与基于热图的方法对比

如图所示，coco 数据集上的量化结果，即使面对遮挡、截断、运动模糊等图像，本文的回归方法依旧有较好的效果。

Figure 6: **Qualitative** results on COCO dataset: containing crowded scenes, occlusions, appearance change and motion blur.

图 6: COCO 数据集上的量化结果

如图所示，在 COCO 测试集上，RLE 达到了 SOTA。这是基于回归的方法首次比基于热图的方法表现更好。

Method	Backbone	AP	AP_{50}	AP_{75}	AP_M	AP_L
Heatmap-based						
CMU-Pose [4]	3CM-3PAF	61.8	84.9	67.5	57.1	68.2
Mask R-CNN [15]	ResNet-50	63.1	87.3	68.7	57.8	71.4
G-RMI [48]	ResNet-101	64.9	85.5	71.3	62.3	70.0
RMPE [13]	PyraNet	72.3	89.2	79.1	68.0	78.6
AE [41]	Hourglass-4	65.5	86.8	72.3	60.6	72.6
PersonLab [47]	ResNet-152	68.7	89.0	75.4	64.1	75.5
CPN [6]	ResNet-Inception	72.1	91.4	80.0	68.7	77.2
SimplePose [67]	ResNet-152	73.7	91.9	81.1	70.3	80.0
Integral [57]	ResNet-101	67.8	88.2	74.8	63.9	74.0
HRNet [55]	HRNet-W48	75.5	**92.5**	**83.3**	71.9	**81.5**
EvoPose [37]	EvoPose2D-L	**75.7**	91.9	83.1	72.2	81.5
Regression-based						
CenterNet [73]	Hourglass-2	63.0	86.8	69.6	58.9	70.4
SPM [45]	Hourglass-8	66.9	88.5	72.9	62.6	73.1
PointSet Anchor [64]	HRNet-W48	68.7	89.9	76.3	64.8	75.3
ResNet + RLE (Ours)	ResNet-152	74.2	91.5	81.9	71.2	79.3
***ResNet + RLE (Ours)**	ResNet-152	75.1	91.8	82.8	72.0	80.2
HRNet + RLE (Ours)	HRNet-W48	**75.7**	92.3	82.9	**72.3**	81.3

Table 3: **Comparison with the SOTA** on COCO test-dev.

图 7: 在 COCO 测试集上与 SOTA 方法的比较

5.2 复现实验结果

论文作者提供的模型的验证结果：

```
index created!
100%|                                                  | 100/100 [01:15<00:00,  1.32it/s]
100%|                                                  | 1627/1627 [07:47<00:00,  3.48it/s]
##### gt box: 0.7218652944822992 mAP | det box: 0.7127219006071578 mAP #####
```

图 8: 论文作者提供的模型验证结果

复现中训练的模型的验证结果：

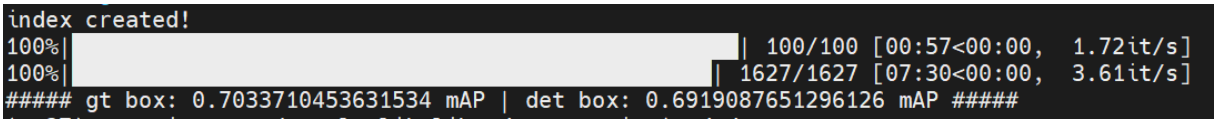


图 9: 复现中训练的模型的验证结果

两个验证结果的准确率约有 0.03 的区别，但仍比传统的回归方法有较大的提升。

6 总结与展望

本文首先介绍了一些人体姿态估计的相关工作，并对要复现的 RLE 方法进行概述：从传统的回归方法开始，介绍了 l1 和 l2 损失，并提出从最大似然估计的角度看待如何选择损失函数；然后提出用标准化流去自主地学习输出分布，以获取更好的回归表现；并介绍了用标准化流进行回归的三种设计。然后介绍了本次复现的细节，并对实验结果进行了分析。

本次复现中由于数据集获取不全的原因，还未在 H36M 数据集上成功复现，希望之后能成功在 H36M 数据集上复现该方法。在以后的工作中，可以考虑是否能应用到 RLE 中的提出的损失函数并找到好的想法去改进 RLE。在其他的工作中也可以学习 RLE 中应用到最大似然估计的思想。

参考文献

[1] LI J, BIAN S, ZENG A, et al. Human Pose Regression with Residual Log-likelihood Estimation[C]//ICCV. 2021.

[2] MAO W, GE Y, SHEN C, et al. Poseur: Direct human pose regression with transformers[C]//. 2022.