**北京君正专利技术交底书**

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | **发明人姓名** | **发明人身份证号** | **发明人电话/邮箱** |
| 1 | 徐时杰 | 410522200102103712 | 18937219579 |
| 2 |  |  |  |
| 3 |  |  |  |
| 4 |  |  |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **专利名称** | 一种分割模型的后处理流程优化 | | | |
| **发明人所属部门** | 算法部 | **发明人所属部门经理** | | 于晓静 |
| **该申请所属项目**(可具体到芯片) | |  | | |
| **已经检索文件名称**(检索2-3篇，若检索不出相关文献，请注明“已经检索，无相关”) | |  | | |
| 技术联系人 | 徐时杰 | 技术联系人邮箱 | mobach.sjxu@ingenic.com | |
| 申请国际专利 | 否 是 | 进入国家或地区 |  | |
|  |  |  |  | |

**注意事项：**

1. 专利技术交底书要求全面撰写专利所在技术领域的专有技术，所以技术交底书需要代理人能够看懂，尤其是背景技术和详细的技术方案，一定要写完整，清楚，可实现。
2. 尽量多写技术方案相关在内容，可以多写“废话”介绍背景或相关技术。力求清晰准确的说明技术方案的改进。

3.同一个事物全篇技术交底书保持用词一致，避免出现同一事物多种名称。

4.使用中文通用名称，不使用英文。除CPU这样大众熟知的，可以使用英文缩写。

1. **专有名词解释**

YOLOv8是Ultralytics公司推出的YOLO最新系列,目前支持图像分类、物体检测和实例分割任务。

后处理是指在模型推理或预测之后对输出结果进行进一步处理、修正或转换的过程，以提高结果的准确性、可解释性或适应特定应用场景。它可以包括各种技术和方法，如阈值化、滤波、归一化、边界框调整等，旨在优化模型输出并满足具体需求，在本专利中指的是网络输出结果后生成最终的图像掩码的过程。

预测边界框指的是目标检测或物体定位任务中，用矩形框标识出目标在图像中位置的一种表示方法。

NMS（Non-Maximum Suppression）是一种目标检测算法中常用的去重策略，其原理是在一组候选框中，选择具有最高置信度分数的框作为输出，并将与该框高度重叠（IoU 大于一定阈值）的其他框删除，以达到减少冗余检测结果的目的。

图像掩码（mask）是一种用于表示图像中特定区域的二进制图像，其中像素值为1表示该像素在目标区域内，像素值为0表示该像素在目标区域外。通过图像掩码可以实现目标的精确定位和分割。

双线性插值是一种计算图像上任意点灰度值的方法，它基于周围四个像素点的灰度值，通过对其进行加权平均来计算目标点的灰度值，以达到缩放和旋转等操作时的图像平滑处理效果。

模型量化是一种通过减少模型中存储的参数位宽来降低模型计算量和存储大小的技术，它通过对权重、激活函数等进行量化，从而减小浮点数精度，达到在保持模型准确率的同时降低模型计算量和存储大小的目的。

sigmoid是一个解决二分类问题的一种函数，其公式如下，同理可以反推其的逆函数反sigmoid函数为

1. **背景技术介绍**

**2.1 技术领域**

本发明属于深度学习网络模型量化部署终端阶段

**2.2技术背景**

语义分割作为计算机视觉领域的热门研究方向，近年来得到了越来越广泛的关注。在深度学习技术的推动下，语义分割的准确率和效率得到了极大的提升。其中，yolov8的分割模型是一种基于深度学习技术的较新语义分割模型，具有高效、高准确率等优点。

然而，当前语义分割的后处理步骤仍然存在一些问题，如处理速度慢、精度不高等。为了解决这些问题，需要对现有的后处理方法进行优化。优化的关键技术包括但不限于：图像超分辨率技术、目标检测技术、语义分割模型的结构优化等。这些技术可以帮助提高后处理的速度和准确率，进一步提高语义分割的应用价值。

因此，本专利旨在提供一种优化 yolov8分割模型后处理步骤的方法，通过结合上述关键技术，提高后处理的效率和准确率，进一步提高 yolov8 分割模型 的性能。

**2.3与本专利有关的现有技术介绍**

**无**

**2.4 现有技术的缺点**

现有的yolov8分割模型的后处理有默认和高精度两个选择，但默认情况下对于分割一切的实例分割任务来说难以满足精度的需求，而高精度由于处理流程对算力要求高处理时间长，难以满足实时需求（主要瓶颈在做nms，双线性插值和矩阵乘法时计算量过大）

1. **专利内容**

**3.1 本专利解决的技术问题**

首先分析梳理下原始的yolov8的分割后处理流程以640\*640的输入分辨率为例

原始网络输出结果：

输出一为（1\*160\*160\*32）的候选掩码特征

输出二为（1\*8400\*37）的掩码置信度复合特征，（37个通道为前4个为预测边界框坐标，后面1个是类别分数，最后是32个掩码置信度）是通过检测头的生成框步骤生成。

输出二进行一步nms操作，

进去的（1\*8400\*37）的特征图根据轴2的第4个数据做筛选（类别分数>0.4），将筛选完的N\*37的数据进行NMS，由于做分割的类别被认为1因此直接做nms即可

输出的结果形状为(预测边界框数量,32个掩码置信度)即(N，32)，N为NMS和阈值处理后得到的n个特征实例数。

然后根据原图大小将当前的预测边界框缩放回原图尺寸，将输出一候选掩码特征与nms出来的结果做矩阵乘法即（N\*32@32\*160\*160输出为N\*160\*160），

然后输入sigmoid函数，之后把（N\*160\*160）的特征图将图像预处理时填充的特征部分去掉然后再通过双线性插值还原回原始图像尺寸，之后通过原始图像尺度的预测边界框扣出对应的图像掩码，最后通过一个阈值化比较让值大于0.5的赋予1，否则赋予0，得到最终原始图像尺寸的图像掩码。

通过分析以上步骤，我们不难发现由于后处理步骤中nms，矩阵乘法，双线性插值还原到原始尺寸这些步骤的计算量过大，导致后处理的时间过长。因此为了能够让基于yolov8 分割的系列模型可以在业务终端做到实时又精准的分割，本专利通过分析分割模型的后处理流程，使用了数据量化的计算，将原本的流程进行优化。

**3.2 详细描述本专利**

将优化步骤分为以下几块：

1.生成框部分：

首先为了提升检测速度，在部署yolov8分割模型时模型的输出节点二不选择原始的最终的输出二，而选择了模型检测头里面生成框之前的输出，并且输出的候选掩码特征和掩码置信度（上文提到的8400\*32的掩码置信度）是做了8比特量化的值

此时的类别分数还没有计算，对要筛选的阈值进行反sigmoid操作即->这样我们可以节省对原先的8400个值做sigmoid的时间，然后直接让原始的类别分数与反sigmoid的0.4做阈值比较，大于该值的才保留，这样提前将数据进行筛选，然后在外面手动做生成预测边界框的操作再次减少了部分时间。

2.矩阵乘法部分：

在得到预测框后还是和之前一样做nms，在得到要做矩阵乘法的(N，32)的掩码置信度矩阵后就要做矩阵乘法了，在这一步相当于把原始的浮点数据的矩阵乘法变成了量化后的8比特整形的矩阵乘法有效的减少了该步骤的计算量（经测试量化后精度基本一致），下面简单介绍下量化的基本原理

 





如上所示R代表实际的数据，scale为量化因子，offset为量化的偏移量，q为量化后的数值。

对于上述的8比特的矩阵乘法来说就是

因此想要让8比特量化的两个值在做完矩阵乘法后依然是8比特的值应该是让原始的两个输入的量化值减去其各自的offset做矩阵乘法后乘以两个输入的除以输出的scale后取整后加上输出的offset即可。

而这些scale和offset可以通过让模型推理一批数据统计各个数据分别的全局最大最小值计算即可。

3.生成最终掩码部分：

在做完8比特的矩阵乘法后依旧是将填充的块去掉，然后将根据候选框抠图放在做双线性插值前面（节省时间的同时最终效果基本一致），并且省去了将生成框给缩放回原始大小的步骤，在抠图的同时还将最后一步的阈值化比较的阈值0.5给反sigmoid后用Q3的量化因子scale和偏移量给量化为8比特，让抠图得到的值与其比较，大于其的赋予1，否则为0，最后让得到的最后的二值化的掩码图像做双线性插值还原到原始图像尺寸得到最终的图像掩码。

1. **本专利的优点**

本方法使用了数据量化以及sigmoid反函数，和流程优化的方法，有效的减少了后处理的时间，保证了实时性，为终端部署提供了可靠的优化流程

1. **替代方案**

暂时未见替代方案。

1. **本发明的关键点和保护点**

本发明的关键点主要在将比较的阈值做反sigmoid和矩阵乘法的量化替代，双线性插值流程优化操作，通过上述的优化，可以做到大幅度减少后处理时间的同时基本保证了与原始流程一致的图像结果。

1. **附图**

