(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117172285 A
(43)申请公布日 2023.12.05

(21)申请号 202311029245.2

(22)申请日 2021.02.27

(62)分案原申请数据
202110221934.8 2021.02.27

(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼

(72)发明人 郭健元 韩凯 王云鹤 许春景

(51)Int.Cl.
G06N 3/045(2023.01)
G06N 3/0464(2023.01)
G06N 3/084(2023.01)
G06N 3/063(2023.01)
G06V 10/80(2022.01)
G06V 10/82(2022.01)

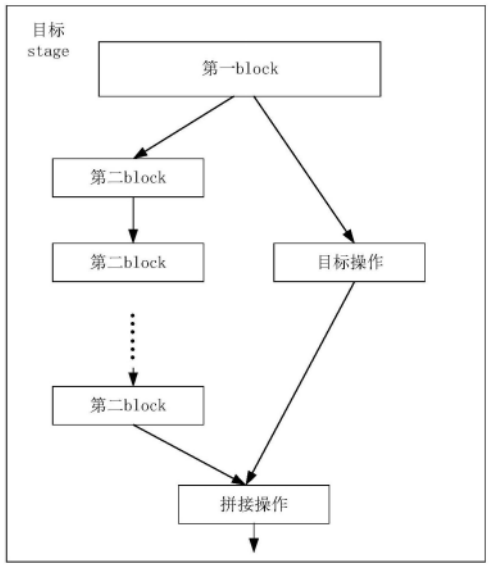
权利要求书3页 说明书27页 附图17页

(54)发明名称

一种感知网络及数据处理方法

(57)摘要

本申请公开了一种感知网络,可以应用于人工智能领域,包括:特征提取网络,其中特征提取网络中的第一block用于对输入数据进行卷积处理,以得到M个目标特征图,特征提取网络中的至少一个第二block用于对M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,特征提取网络中的目标操作用于对M个目标特征图中的M2个目标特征图进行处理,以得到M2个第二特征图,特征提取网络中的拼接操作用于将M1个第一特征图和M2个第二特征图进行拼接,以得到拼接后的特征图。本申请利用相同stage之间跨层的目标操作来让感知网络生成这些与关键特征相似性高的特征,降低了模型的参数量,以此提高在GPU设备、TPU设备以及NPU设备上的模型的运行速度。



CN 117172285 A

1.一种感知网络,其特征在于,所述感知网络包括:特征提取网络,所述特征提取网络包括:

第一block,所述第一block用于对输入数据进行卷积处理,以得到M个目标特征图;

至少一个第二block,所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,所述M1小于所述M;

用于执行目标操作的单元,所述目标操作用于对所述M个目标特征图中的M2个目标特征图进行处理以得到M2个第二特征图,所述M2小于所述M;

用于执行拼接操作的单元,所述拼接操作用于将所述M1个第一特征图和所述M2个第二特征图进行拼接以得到拼接后的特征图。

2.根据权利要求1所述的感知网络,其特征在于,所述第一block以及所述至少一个第二block为所述特征提取网络中同一个阶段stage内的block。

3.根据权利要求1或2所述的感知网络,其特征在于,所述目标操作的参数量小于所述至少一个第二block的参数量。

4.根据权利要求1-3任一项所述的感知网络,其特征在于,所述M1个目标特征图与所述M2个目标特征图的交集为空,且所述M1与所述M2的加和为所述M。

5.根据权利要求1-4任一项所述的感知网络,其特征在于,所述目标操作为参数量小于所述至少一个第二block的卷积操作;或,

所述目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作。

6.根据权利要求1所述的感知网络,其特征在于,所述特征提取网络还包括用于执行融合操作的单元,所述融合操作用于对所述每个第二block输出的特征图进行融合,以得到融合后的特征图,以及对所述融合后的特征图和所述M2个第二特征图进行加法操作,以得到处理后的M2个第二特征图;

所述拼接后的特征图基于所述M1个第一特征图和所述处理后的M2个第二特征图拼接得到。

7.根据权利要求6所述的感知网络,其特征在于,所述融合后的特征图的尺寸大小与所述M2个第二特征图的尺寸大小相同。

8.根据权利要求1-7任一项所述的感知网络,其特征在于,所述M1个第一特征图为所述至少一个第二block中距离所述第一block最远的第二block的输出。

9.根据权利要求6所述的感知网络,其特征在于,所述融合操作用于对所述每个第二block的输出进行拼接以及降维操作,以得到和所述M2个第二特征图的尺寸大小相同的所述融合后的特征图。

10.根据权利要求1-9任一项所述的感知网络,其特征在于,所述第一block以及所述至少一个第二block为所述特征提取网络中目标阶段stage内的block,所述拼接后的特征图用于作为所述特征提取网络中所述目标阶段stage的输出特征图;或,

所述目标阶段stage还包括至少一个第三block,所述至少一个第三block用于对所述拼接后的特征图进行卷积操作,以得到所述目标阶段stage的输出特征图。

11.根据权利要求1-10任一项所述的感知网络,其特征在于,所述特征提取网络用于获取输入的图像,并对所述输入的图像进行特征提取,输出所述输入的图像的特征图;

所述感知网络还包括：

任务网络，用于根据所述输入的图像的特征图，进行对应任务的处理，以得到处理结果。

12.根据权利要求11所述的感知网络，其特征在于，所述任务包括目标检测、图像分割或图像分类。

13.一种数据处理方法，其特征在于，应用于特征提取网络，所述方法包括：

对输入数据进行卷积处理，以得到M个目标特征图；

对所述M个目标特征图中的M1个目标特征图进行卷积处理，以得到M1个第一特征图，所述M1小于所述M；

对所述M个目标特征图中的M2个目标特征图进行处理，以得到M2个第二特征图，所述M2小于所述M；

将所述M1个第一特征图和所述M2个第二特征图进行拼接，以得到拼接后的特征图。

14.根据权利要求13所述的方法，其特征在于，所述特征提取网络包括：

第一block，所述第一block用于所述对输入数据进行卷积处理；

至少一个第二block，所述至少一个第二block用于所述对所述M个目标特征图中的M1个目标特征图进行卷积处理；

用于执行目标操作的单元，所述目标操作用于所述对所述M个目标特征图中的M2个目标特征图进行处理；

用于执行拼接操作的单元，所述拼接操作用于所述将所述M1个第一特征图和所述M2个第二特征图进行拼接。

15.根据权利要求14所述的方法，其特征在于，所述第一block以及所述至少一个第二block为所述特征提取网络中同一个阶段stage内的block。

16.根据权利要求14或15所述的方法，其特征在于，所述目标操作的参数量小于所述至少一个第二block的参数量。

17.根据权利要求14所述的方法，其特征在于，所述M1个目标特征图与所述M2个目标特征图的交集为空，且所述M1与所述M2的加和为所述M。

18.根据权利要求14-17任一项所述的方法，其特征在于，所述目标操作为参数量小于所述至少一个第二block的卷积操作；或，

所述目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作。

19.根据权利要求13-18任一项所述的方法，其特征在于，所述方法还包括：

对所述M1个第一特征图进行融合，以得到融合后的特征图；

对所述融合后的特征图和所述M2个第二特征图进行加法操作，以得到处理后的M2个第二特征图；

所述将所述M1个第一特征图和所述M2个第二特征图进行拼接，包括：

将所述M1个第一特征图和所述处理后的M2个第二特征图进行拼接，以得到拼接后的特征图。

20.根据权利要求14-19任一所述的方法，其特征在于，所述至少一个第二block中距离所述第一block最远的第二block的输出为所述M1个第一特征图。

21.根据权利要求19所述的方法,其特征在于,所述特征提取网络还包括用于执行融合操作的单元,所述融合操作用于所述对所述M1个第一特征图进行融合以及所述对所述融合后的特征图和所述M2个第二特征图进行加法操作。

22.根据权利要求19所述的方法,其特征在于,所述通过融合操作,对所述每个第二block输出的特征图进行融合,包括:

通过融合操作,对所述每个第二block的输出进行拼接以及降维操作,以得到和所述M2个第二特征图的尺寸大小相同的所述融合后的特征图。

23.根据权利要求14-18任一项所述的方法,其特征在于,所述第一block以及所述至少一个第二block为所述特征提取网络中目标阶段stage内的block,所述拼接后的特征图用于作为所述特征提取网络中所述目标阶段stage的输出特征图;或,

所述目标阶段stage还包括至少一个第三block,所述至少一个第三block用于对所述拼接后的特征图进行卷积操作,以得到所述目标阶段stage的输出特征图。

24.根据权利要求14-18任一项所述的方法,其特征在于,所述特征提取网络用于获取输入的图像,并对所述输入的图像进行特征提取,输出所述输入的图像的特征图;

所述方法还包括:

通过任务网络,用于根据所述输入的图像的特征图,进行对应任务的处理,以得到处理结果。

25.根据权利要求24所述的方法,其特征在于,所述任务包括目标检测、图像分割或图像分类。

26.一种数据处理装置,其特征在于,所述装置包括存储器和处理器;所述存储器存储有代码,所述处理器被配置为获取所述代码,以运行如权利要求13至25任一所述的数据处理方法。

27.一种计算机可读存储介质,用于存储计算机程序,其特征在于,所述计算机程序用于执行根据权利要求13至25中任一项所述方法的指令。

一种感知网络及数据处理方法

[0001] 本申请为2021年2月27日提交中国专利局、申请号为202110221934.8、发明名称为“一种感知网络及数据处理方法”的中国专利申请的分案申请,该中国专利申请的全部内容通过引用结合在本申请中。

技术领域

[0002] 本申请涉及人工智能领域,尤其涉及一种感知网络及数据处理方法。

背景技术

[0003] 人工智能(artificial intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个分支,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式作出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0004] 计算机视觉是各个应用领域,如制造业、检验、文档分析、医疗诊断,和军事等领域中各种智能/自主系统中不可分割的一部分,它是一门关于如何运用照相机/摄像机和计算机来获取我们所需的,被拍摄对象的数据与信息的学问。形象地说,就是给计算机安装上眼睛(照相机/摄像机)和大脑(算法)用来代替人眼对目标进行识别、跟踪和测量等,从而使计算机能够感知环境。因为感知可以看作是从感官信号中提取信息,所以计算机视觉也可以看作是研究如何使人工系统从图像或多维数据中“感知”的科学。总的来说,计算机视觉就是用各种成象系统代替视觉器官获取输入信息,再由计算机来代替大脑对这些输入信息完成处理和解释。计算机视觉的最终研究目标就是使计算机能像人那样通过视觉观察和理解世界,具有自主适应环境的能力。

[0005] 基于卷积神经网络的推理模型在以计算机视觉为基础的各个终端任务如图像识别、目标检测、实力分割等场景中都有广泛的应用。传统的基础神经网络往往由于规模较大的参数与计算量,导致各项终端任务无法实时工作。现有的轻量级推理网络(例如mobilenet,efficientnet,shufflenet)都是针对中央处理器(central processing unit,CPU)、ARM(advanced RISC machine)等移动设备设计而成,在图形处理器(graphics processing unit,GPU)设备、张量处理单元(tensor processing unit,TPU)设备以及神经网络处理器(neural network processing Unit,NPU)设备等基于大吞吐量设计的处理单元上的表现却不尽人意,推理速度甚至比传统的卷积神经网络更慢。

发明内容

[0006] 第一方面,本申请提供了一种感知网络,所述感知网络包括:特征提取网络,所述特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作,所述第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block,且所述目标操作的参数量小于所述M个第二block的参数量;

[0007] 其中,目标操作也可以称之为廉价操作(cheap operation),可以指一系列参数量较少多少的参数量会认为是参数量较少的操作的统称,用于区别传统的卷积操作;参数量(parameters)可以用于描述神经网络包含的参数量,用于评价模型的大小。

[0008] 拼接操作(concat)是指在不改变特征图的数据的前提下,进行特征图的合并,例如特征图1和特征图2进行拼接操作的结果为(特征图1,特征图2),其中特征图1和特征图2之间的顺序不限定。更具体的,具有三个语义通道的特征图与五个语义通道的特征图进行拼接操作的结果为具有八个语义通道的特征图。

[0009] 所述第一block用于对输入数据进行卷积处理,以得到M个目标特征图,每个目标特征图对应于一个通道;

[0010] 所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,所述M1小于所述M;

[0011] 所述目标操作用于对所述M个目标特征图中的M2个目标特征图进行处理,以得到M2个第二特征图,所述M2小于所述M;

[0012] 所述拼接操作用于将所述M1个第一特征图和所述M2个第二特征图进行拼接,以得到拼接后的特征图。

[0013] 在特征提取网络的相同阶段stage之间,不同block的输出特征之间具有较高的相似性,而不同阶段stage的block的输出特征之间,相似性较低。因此本申请实施例利用相同stage之间跨层的目标操作来让感知网络生成这些与关键特征相似性高的特征,降低了模型的参数量,以此提高在GPU设备、TPU设备以及NPU设备上的模型的运行速度。

[0014] 本申请实施例提供了一种感知网络,所述感知网络包括:特征提取网络,所述特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作,所述第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block,且所述目标操作的参数量小于所述M个第二block的参数量;所述第一block用于对输入数据进行卷积处理,以得到M个目标特征图,每个目标特征图对应于一个通道;所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,所述M1小于所述M;所述目标操作用于对所述M个目标特征图中的M2个目标特征图进行处理,以得到M2个第二特征图,所述M2小于所述M;所述拼接操作用于将所述M1个第一特征图和所述M2个第二特征图进行拼接,以得到拼接后的特征图。通过上述方式,利用相同stage之间跨层的目标操作来让感知网络生成这些与关键特征相似性高的特征,降低了模型的参数量,以此提高在GPU设备、TPU设备以及NPU设备上的模型的运行速度。

[0015] 在一种可能的实现中,所述M1个目标特征图与所述M2个目标特征图的交集为空,且所述M1与所述M2的加和为所述M,所述拼接后的特征图的通道数为所述M。

[0016] 相当于,至少一个第二block的输出特征图和输出特征图的通道数都是输出第一block的目标特征图的通道数的一部分,剩下的一部分通道的目标特征图由目标操作产来进行处理。且由于目标操作的参数量小于至少一个第二block的参数量,使得整体上感知网络的参数量降低了,进而可以提高感知网络在GPU设备、TPU设备以及NPU设备上的运行速度。

[0017] 在一种可能的实现中,所述目标操作为参数量小于所述至少一个第二block的卷积操作;或,

[0018] 所述目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作。

[0019] 例如,目标操作为1*1的卷积,或者目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作,相当于M2个第二特征图就是M1个第一特征图,即直接将M1个第一特征图作为M2个第二特征图。

[0020] 应理解,还可以将第一block的输出特征图拆分为多组特征图,并由多个目标操作来处理一组特征图,其中只要保证各个目标操作输出的特征图的通道数、至少一个第二block输出的特征图的加和与第一block输出的特征图的通道数相同即可,不同目标操作输出的特征图的通道数之间可以不相同。

[0021] 在一种可能的实现中,所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到每个第二block输出的特征图,其中,所述至少一个第二block中距离所述第一block最远的第二block的输出为所述M1个第一特征图;

[0022] 所述特征提取网络还包括:

[0023] 融合操作,用于对所述每个第二block输出的特征图进行融合,以得到融合后的特征图,所述融合后的特征图的尺寸大小与所述M2个第二特征图的尺寸大小相同;

[0024] 对所述融合后的特征图和所述M2个第二特征图进行加法操作,以得到处理后的M2个第二特征图;

[0025] 所述拼接操作用于将所述M1个第一特征图和所述处理后的M2个第二特征图进行拼接,以得到拼接后的特征图。

[0026] 其中,可以将每个第二block输出的特征图进行拼接操作,得到拼接后的特征图(通道数为每个第二block输出的特征图的加和),由于拼接后的特征图的通道数大于目标操作的输出的特征图的通道数(M2),为了能够对拼接后的特征图与目标操作的输出(M2个第二特征图)进行加法运算,则需要将拼接后的特征图进行降维操作,以使得拼接后的特征图的通道数等于M2,进而可以对拼接后的特征图与目标操作的输出进行矩阵的加法运算。

[0027] 在一种可能的实现中,所述融合操作,用于对所述每个第二block的输出进行拼接以及降维操作,以得到和所述M2个第二特征图的尺寸大小相同的所述融合后的特征图。

[0028] 在一种可能的实现中,所述第一block以及所述M个第二block为所述特征提取网络中目标阶段stage内的block,所述拼接后的特征图用于作为所述特征提取网络中所述目标阶段stage的输出特征图;或,

[0029] 所述目标阶段stage还包括至少一个第三block,所述至少一个第三block用于对所述拼接后的特征图进行卷积操作,以得到所述目标阶段stage的输出特征图。

[0030] 本申请实施例中,拼接后的特征图可以作为目标阶段stage的输出特征图,或者拼接后的特征图还可以被目标阶段stage包括的其他block(第三block)进行处理,拼接后的特征图还可以被至少一个第三block进行处理,以得到目标阶段stage的输出特征图。

[0031] 在一种可能的实现中,第一block可以为特征提取网络中的第一个block,或者是中间层的block,在第一block之前还可以连接有至少一个第三block,则第一block用于对至少一个第三block输出的数据进行卷积处理。

[0032] 在一种可能的实现中,所述特征提取网络用于获取输入的图像,并对所述输入的图像进行特征提取,输出所述输入的图像的特征图;

[0033] 所述感知网络还包括：

[0034] 任务网络，用于根据所述输入的图像的特征图，进行对应任务的处理，以得到处理结果。

[0035] 在一种可能的实现中，所述任务包括目标检测、图像分割或图像分类。

[0036] 第二方面，本申请提供了一种数据处理方法，所述方法包括：

[0037] 获取特征提取网络，所述特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作，所述第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block，且所述目标操作的参数量小于所述M个第二block的参数量；

[0038] 通过所述第一block，对输入数据进行卷积处理，以得到M个目标特征图，每个目标特征图对应于一个通道；

[0039] 通过所述至少一个第二block，对所述M个目标特征图中的M1个目标特征图进行卷积处理，以得到M1个第一特征图，所述M1小于所述M；

[0040] 通过所述目标操作，对所述M个目标特征图中的M2个目标特征图进行处理，以得到M2个第二特征图，所述M2小于所述M；

[0041] 通过所述拼接操作，将所述M1个第一特征图和所述M2个第二特征图进行拼接，以得到拼接后的特征图。

[0042] 在一种可能的实现中，所述M1个目标特征图与所述M2个目标特征图的交集为空，且所述M1与所述M2的加和为所述M，所述拼接后的特征图的通道数为所述M。

[0043] 在一种可能的实现中，所述目标操作为参数量小于所述至少一个第二block的卷积操作；或，

[0044] 所述目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作。

[0045] 在一种可能的实现中，所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理，以得到每个第二block输出的特征图，其中，所述至少一个第二block中距离所述第一block最远的第二block的输出为所述M1个第一特征图；

[0046] 所述方法还包括：

[0047] 通过融合操作，对所述每个第二block输出的特征图进行融合，以得到融合后的特征图，所述融合后的特征图的尺寸大小与所述M2个第二特征图的尺寸大小相同；对所述融合后的特征图和所述M2个第二特征图进行加法操作，以得到处理后的M2个第二特征图；

[0048] 所述通过所述拼接操作，将所述M1个第一特征图和所述M2个第二特征图进行拼接，包括：

[0049] 将所述M1个第一特征图和所述处理后的M2个第二特征图进行拼接，以得到拼接后的特征图。

[0050] 在一种可能的实现中，所述通过融合操作，对所述每个第二block输出的特征图进行融合，包括：

[0051] 通过融合操作，对所述每个第二block的输出进行拼接以及降维操作，以得到和所述M2个第二特征图的尺寸大小相同的所述融合后的特征图。

[0052] 在一种可能的实现中，所述第一block以及所述M个第二block为所述特征提取网

络中目标阶段stage内的block,所述拼接后的特征图用于作为所述特征提取网络中所述目标阶段stage的输出特征图;或,

[0053] 所述目标阶段stage还包括至少一个第三block,所述至少一个第三block用于对所述拼接后的特征图进行卷积操作,以得到所述目标阶段stage的输出特征图。

[0054] 在一种可能的实现中,所述特征提取网络用于获取输入的图像,并对所述输入的图像进行特征提取,输出所述输入的图像的特征图;

[0055] 所述方法还包括:

[0056] 通过任务网络,用于根据所述输入的图像的特征图,进行对应任务的处理,以得到处理结果。

[0057] 在一种可能的实现中,所述任务包括目标检测、图像分割或图像分类。

[0058] 第三方面,本申请提供了一种数据处理装置,所述装置包括:

[0059] 获取模块,用于获取特征提取网络,所述特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作,所述第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block,且所述目标操作的参数量小于所述M个第二block的参数量;

[0060] 卷积处理模块,通过所述第一block,对输入数据进行卷积处理,以得到M个目标特征图,每个目标特征图对应于一个通道;

[0061] 通过所述至少一个第二block,对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,所述M1小于所述M;

[0062] 通过所述目标操作,对所述M个目标特征图中的M2个目标特征图进行处理,以得到M2个第二特征图,所述M2小于所述M;

[0063] 拼接模块,用于通过所述拼接操作,将所述M1个第一特征图和所述M2个第二特征图进行拼接,以得到拼接后的特征图。

[0064] 在一种可能的实现中,所述M1个目标特征图与所述M2个目标特征图的交集为空,且所述M1与所述M2的加和为所述M,所述拼接后的特征图的通道数为所述M。

[0065] 在一种可能的实现中,所述目标操作为参数量小于所述至少一个第二block的卷积操作;或,

[0066] 所述目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作。

[0067] 在一种可能的实现中,所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到每个第二block输出的特征图,其中,所述至少一个第二block中距离所述第一block最远的第二block的输出为所述M1个第一特征图;

[0068] 所述装置还包括:

[0069] 融合模块,用于通过融合操作,对所述每个第二block输出的特征图进行融合,以得到融合后的特征图,所述融合后的特征图的尺寸大小与所述M2个第二特征图的尺寸大小相同;对所述融合后的特征图和所述M2个第二特征图进行加法操作,以得到处理后的M2个第二特征图;

[0070] 所述拼接模块,用于将所述M1个第一特征图和所述处理后的M2个第二特征图进行拼接,以得到拼接后的特征图。

[0071] 在一种可能的实现中,所述融合模块,用于通过融合操作,对所述每个第二block的输出进行拼接以及降维操作,以得到和所述M2个第二特征图的尺寸大小相同的所述融合后的特征图。

[0072] 在一种可能的实现中,所述第一block以及所述M个第二block为所述特征提取网络中目标阶段stage内的block,所述拼接后的特征图用于作为所述特征提取网络中所述目标阶段stage的输出特征图;或,

[0073] 所述目标阶段stage还包括至少一个第三block,所述至少一个第三block用于对所述拼接后的特征图进行卷积操作,以得到所述目标阶段stage的输出特征图。

[0074] 在一种可能的实现中,所述特征提取网络用于获取输入的图像,并对所述输入的图像进行特征提取,输出所述输入的图像的特征图;

[0075] 所述装置还包括:

[0076] 任务处理模块,用于通过任务网络,用于根据所述输入的图像的特征图,进行对应任务的处理,以得到处理结果。

[0077] 在一种可能的实现中,所述任务包括目标检测、图像分割或图像分类。

[0078] 第四方面,本申请实施例提供了一种数据处理装置,可以包括存储器、处理器以及总线系统,其中,存储器用于存储程序,处理器用于执行存储器中的程序,以运行如上述第一方面及其任一可选的感知网络。

[0079] 第五方面,本申请实施例提供了一种计算机可读存储介质,所述计算机可读存储介质中存储有计算机程序,当其在计算机上运行时,使得计算机运行如上述第一方面及其任一可选的感知网络。

[0080] 第六方面,本申请实施例提供了一种计算机程序,包括代码,当代码被执行时,用于运行如上述第一方面及其任一可选的感知网络。

[0081] 第七方面,本申请实施例提供了一种数据处理装置,可以包括存储器、处理器以及总线系统,其中,存储器用于存储程序,处理器用于执行存储器中的程序,以执行如上述第一方面及其任一可选的方法。

[0082] 第八方面,本申请实施例提供了一种计算机可读存储介质,所述计算机可读存储介质中存储有计算机程序,当其在计算机上运行时,使得计算机执行如上述第一方面及其任一可选的方法。

[0083] 第九方面,本申请实施例提供了一种计算机程序,包括代码,当代码被执行时,用于实现上述第一方面及其任一可选的方法。

[0084] 第十方面,本申请提供了一种芯片系统,该芯片系统包括处理器,用于支持执行设备或训练设备实现上述方面中所涉及的功能,例如,发送或处理上述方法中所涉及的数据;或,信息。在一种可能的设计中,所述芯片系统还包括存储器,所述存储器,用于保存执行设备或训练设备必要的程序指令和数据。该芯片系统,可以由芯片构成,也可以包括芯片和其他分立器件。

[0085] 本申请实施例提供了一种感知网络,所述感知网络包括:特征提取网络,所述特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作,所述第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block,且所述目标操作的参数量小于所述M个第二block的参数量;所述第一block用于对输入数据

进行卷积处理,以得到M个目标特征图,每个目标特征图对应于一个通道;所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,所述M1小于所述M;所述目标操作用于对所述M个目标特征图中的M2个目标特征图进行处理,以得到M2个第二特征图,所述M2小于所述M;所述拼接操作用于将所述M1个第一特征图和所述M2个第二特征图进行拼接,以得到拼接后的特征图。通过上述方式,利用相同stage之间跨层的目标操作来让感知网络生成这些与关键特征相似性高的特征,降低了模型的参数量,以此提高在GPU设备、TPU设备以及NPU设备上的模型的运行速度。

附图说明

[0086] 图1为人工智能主体框架的一种结构示意图;

[0087] 图2a为本申请实施例提供的一种应用场景的示意图;

[0088] 图2b为本申请实施例提供的一种应用场景的示意图;

[0089] 图2c为本申请实施例提供的卷积神经网络的示意图;

[0090] 图2d为本申请实施例提供的卷积神经网络的示意图;

[0091] 图3为本申请实施例提供的一种系统架构的示意图;

[0092] 图4为本申请实施例提供的一种数据处理方法的实施例示意;

[0093] 图5a为本申请实施例提供的一种感知网络示意;

[0094] 图5b为本申请实施例提供的一种感知网络示意;

[0095] 图5c为本申请实施例提供的一种感知网络示意;

[0096] 图5d为本申请实施例提供的一种感知网络示意;

[0097] 图6至图14为本申请实施例提供的一种感知网络示意;

[0098] 图15为本申请实施例提供的一种数据处理装置的示意;

[0099] 图16为本申请实施例提供的执行设备的一种结构示意图;

[0100] 图17为本申请实施例提供的训练设备一种结构示意图;

[0101] 图18为本申请实施例提供的芯片的一种结构示意图。

具体实施方式

[0102] 下面结合本发明实施例中的附图对本发明实施例进行描述。本发明的实施方式部分使用的术语仅用于对本发明的具体实施例进行解释,而非旨在限定本发明。

[0103] 下面结合附图,对本申请的实施例进行描述。本领域普通技术人员可知,随着技术的发展和新场景的出现,本申请实施例提供的技术方案对于类似的技术问题,同样适用。

[0104] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的术语在适当情况下可以互换,这仅仅是描述本申请的实施例中对相同属性的对象在描述时所采用的区分方式。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,以便包含一系列单元的过程、方法、系统、产品或设备不必限于那些单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它单元。

[0105] 首先对人工智能系统总体工作流程进行描述,请参见图1,图1示出的为人工智能主体框架的一种结构示意图,下面从“智能信息链”(水平轴)和“IT价值链”(垂直轴)两个维

度对上述人工智能主题框架进行阐述。其中,“智能信息链”反映从数据的获取到处理的一列过程。举例来说,可以是智能信息感知、智能信息表示与形成、智能推理、智能决策、智能执行与输出的一般过程。在这个过程中,数据经历了“数据—信息—知识—智慧”的凝练过程。“IT价值链”从人智能的底层基础设施、信息(提供和处理技术实现)到系统的产业生态过程,反映人工智能为信息技术产业带来的价值。

[0106] (1)基础设施

[0107] 基础设施为人工智能系统提供计算能力支持,实现与外部世界的沟通,并通过基础平台实现支撑。通过传感器与外部沟通;计算能力由智能芯片(CPU、NPU、GPU、ASIC、FPGA等硬件加速芯片)提供;基础平台包括分布式计算框架及网络等相关的平台保障和支持,可以包括云存储和计算、互联互通网络等。举例来说,传感器和外部沟通获取数据,这些数据提供给基础平台提供的分布式计算系统中的智能芯片进行计算。

[0108] (2)数据

[0109] 基础设施的上一层的数据用于表示人工智能领域的数据来源。数据涉及到图形、图像、语音、文本,还涉及到传统设备的物联网数据,包括已有系统的业务数据以及力、位移、液位、温度、湿度等感知数据。

[0110] (3)数据处理

[0111] 数据处理通常包括数据训练,机器学习,深度学习,搜索,推理,决策等方式。

[0112] 其中,机器学习和深度学习可以对数据进行符号化和形式化的智能信息建模、抽取、预处理、训练等。

[0113] 推理是指在计算机或智能系统中,模拟人类的智能推理方式,依据推理控制策略,利用形式化的信息进行机器思维和求解问题的过程,典型的功能是搜索与匹配。

[0114] 决策是指智能信息经过推理后进行决策的过程,通常提供分类、排序、预测等功能。

[0115] (4)通用能力

[0116] 对数据经过上面提到的数据处理后,进一步基于数据处理的结果可以形成一些通用的能力,比如可以是算法或者一个通用系统,例如,翻译,文本的分析,计算机视觉的处理,语音识别,图像的识别等等。

[0117] (5)智能产品及行业应用

[0118] 智能产品及行业应用指人工智能系统在各领域的产品和应用,是对人工智能整体解决方案的封装,将智能信息决策产品化、实现落地应用,其应用领域主要包括:智能终端、智能交通、智能医疗、自动驾驶、智慧城市等。

[0119] 本申请实施例主要应用在驾驶辅助、自动驾驶、手机终端等需要完成多种感知任务的领域。视频经过抽帧得到单张图片,该图片送入到本发明中的感知网络,得到该图片中感兴趣物体的2D、3D、Mask(掩膜)、关键点等信息。这些检测结果输出到后处理模块进行处理,比如在自动驾驶系统中送入规划控制单元进行决策、在手机终端中送入美颜算法进行处理得到美颜后的图片。下面分别对ADAS/ADS视觉感知系统和手机美颜两种应用场景做简单的介绍。

[0120] 应用场景1:ADAS/ADS视觉感知系统

[0121] 如图2a所示,在ADAS和ADS中,需要实时进行多类型的2D目标检测,包括:动态障碍

物(行人(Pedestrian)、骑行者(Cyclist)、三轮车(Tricycle)、轿车(Car)、卡车(Truck)、公交车(Bus)),静态障碍物(交通锥标(TrafficCone)、交通棍标(TrafficStick)、消防栓(FireHydrant)、摩托车(Motocycle)、自行车(Bicycle)),交通标志(TrafficSign、导向标志(GuideSign)、广告牌(Billboard)、红色交通灯(TrafficLight_Red)/黄色交通灯(TrafficLight_Yellow)/绿色交通灯(TrafficLight_Green)/黑色交通灯(TrafficLight_Black)、路标(RoadSign))。另外,为了准确获取动态障碍物的在3维空间所占的区域,还需要对动态障碍物进行3D估计,输出3D框。为了与激光雷达的数据进行融合,需要获取动态障碍物的Mask,从而把打到动态障碍物上的激光点云筛选出来;为了进行精确的泊车位,需要同时检测出泊车位的4个关键点;为了进行构图定位,需要检测出静态目标的关键点。使用本申请实施例提供的技术方案,可以在感知网络中完成上述的全部或一部分功能。

[0122] 应用场景2:手机美颜功能

[0123] 如图2b所示,在手机中,通过本申请实施例提供的感知网络检测出人体的Mask和关键点,可以对人体相应的部位进行放大缩小,比如进行收腰和美臀操作,从而输出美颜的图片。

[0124] 应用场景3:图像分类场景:

[0125] 物体识别装置在获取待分类图像后,采用本申请的物体识别方法获取待分类图像中的物体的类别,然后可根据待分类图像中物体的类别对待分类图像进行分类。对于摄影师来说,每天会拍很多照片,有动物的,有人物,有植物的。采用本申请的方法可以快速地将照片按照照片中的内容进行分类,可分成包含动物的照片、包含人物的照片和包含植物的照片。

[0126] 对于图像数量比较庞大的情况,人工分类的方式效率比较低下,并且人在长时间处理同一件事情时很容易产生疲劳感,此时分类的结果会有很大的误差;而采用本申请的方法可以快速地将图像进行分类,并且不会有误差。

[0127] 应用场景4商品分类:

[0128] 物体识别装置获取商品的图像后,然后采用本申请的物体识别方法获取商品的图像中商品的类别,然后根据商品的类别对商品进行分类。

[0129] 由于本申请实施例涉及大量神经网络的应用,为了便于理解,下面先对本申请实施例涉及的相关术语及神经网络等相关概念进行介绍。

[0130] (1)物体识别,利用图像处理和机器学习、计算机图形学等相关方法,确定图像物体的类别。

[0131] (2)神经网络

[0132] 神经网络可以是由神经单元组成的,神经单元可以是指以xs和截距1为输入的运算单元,该运算单元的输出可以为:

[0133] $$h_{W,b}(x)=f\left(W^{T}x\right)=f\left(\sum_{s=1}^{n}W_{s}x_{s}+b\right) \quad (1\text{-}1)$$

[0134] 其中,s=1、2、……n,n为大于1的自然数,Ws为xs的权重,b为神经单元的偏置。f为神经单元的激活函数(activation functions),用于将非线性特性引入神经网络中,来将神经单元中的输入信号转换为输出信号。该激活函数的输出信号可以作为下一层卷积层的输入。激活函数可以是sigmoid函数。神经网络是将许多个上述单一的神经单元联结在一起

形成的网络,即一个神经单元的输出可以是另一个神经单元的输入。每个神经单元的输入可以与前一层的局部接受域相连,来提取局部接受域的特征,局部接受域可以是由若干个神经单元组成的区域。

[0135] (3)深度神经网络

[0136] 深度神经网络(Deep Neural Network,DNN),可以理解为具有很多层隐含层的神经网络,这里的"很多"并没有特别的度量标准,我们常说的多层神经网络和深度神经网络其本质上是同一个东西。从DNN按不同层的位置划分,DNN内部的神经网络可以分为三类:输入层,隐含层,输出层。一般来说第一层是输入层,最后一层是输出层,中间的层数都是隐含层。层与层之间是全连接的,也就是说,第i层的任意一个神经元一定与第i+1层的任意一个神经元相连。虽然DNN看起来很复杂,但是就每一层的工作来说,其实并不复杂,简单来说就是如下线性关系表达式:$\vec{y}=\alpha(W\vec{x}+\vec{b})$,其中,$\vec{x}$是输入向量,$\vec{y}$是输出向量,$\vec{b}$是偏移向量,W是权重矩阵(也称系数),α()是激活函数。每一层仅仅是对输入向量$\vec{x}$经过如此简单的操作得到输出向量$\vec{y}$。由于DNN层数多,则系数W和偏移向量$\vec{b}$的数量也就是很多了。那么,具体的参数在DNN是如何定义的呢?首先我们来看看系数W的定义。以一个三层的DNN为例,如:第二层的第4个神经元到第三层的第2个神经元的线性系数定义为w_{24}^{3}。上标3代表系数W所在的层数,而下标对应的是输出的第三层索引2和输入的第二层索引4。总结下,第L-1层的第k个神经元到第L层的第j个神经元的系数定义为W_{jk}^{L}。注意,输入层是没有W参数的。在深度神经网络中,更多的隐含层让网络更能够刻画现实世界中的复杂情形。理论上而言,参数越多的模型复杂度越高,"容量"也就越大,也就意味着它能完成更复杂的学习任务。

[0137] (4)卷积神经网络(Convosutionas Neuras Network,CNN)是一种带有卷积结构的深度神经网络。卷积神经网络包含了一个由卷积层和子采样层构成的特征抽取器。该特征抽取器可以看作是滤波器,卷积过程可以看作是使用一个可训练的滤波器与一个输入的图像或者卷积特征平面(feature map)做卷积。卷积层是指卷积神经网络中对输入信号进行卷积处理的神经元层。在卷积神经网络的卷积层中,一个神经元可以只与部分邻层神经元连接。一个卷积层中,通常包含若干个特征平面,每个特征平面可以由一些矩形排列的神经单元组成。同一特征平面的神经单元共享权重,这里共享的权重就是卷积核。共享权重可以理解为提取图像信息的方式与位置无关。这其中隐含的原理是:图像的某一部分的统计信息与其他部分是一样的。即意味着在某一部分学习的图像信息也能用在另一部分上。所以对于图像上的所有位置,我们都能使用同样的学习得到的图像信息。在同一卷积层中,可以使用多个卷积核来提取不同的图像信息,一般地,卷积核数量越多,卷积操作反映的图像信息越丰富。

[0138] 卷积核可以以随机大小的矩阵的形式初始化,在卷积神经网络的训练过程中卷积核可以通过学习得到合理的权重。另外,共享权重带来的直接好处是减少卷积神经网络各层之间的连接,同时又降低了过拟合的风险。

[0139] 由于CNN是一种非常常见的神经网络,下面结合图2c重点对CNN的结构进行详细的介绍。如上文的基础概念介绍所述,卷积神经网络是一种带有卷积结构的深度神经网络,是一种深度学习(deep learning)架构,深度学习架构是指通过机器学习的算法,在不同的抽象层级上进行多个层次的学习。作为一种深度学习架构,CNN是一种前馈(feed-forward)人

工神经网络,该前馈人工神经网络中的各个神经元可以对输入其中的图像作出响应。

[0140] 在图2c中,卷积神经网络(CNN)100可以包括输入层120,卷积层/池化层120(其中池化层为可选的),以及神经网络层130。其中,输入层120可以获取待处理图像,并将获取到的待处理图像交由卷积层/池化层120以及后面的神经网络层130进行处理,可以得到图像的处理结果。下面对图2c中的CNN 100中内部的层结构进行详细的介绍。

[0141] 卷积层/池化层120:

[0142] 卷积层:

[0143] 如图2c所示卷积层/池化层120可以包括如示例121-126层,举例来说:在一种实现中,121层为卷积层,122层为池化层,123层为卷积层,124层为池化层,125为卷积层,126为池化层;在另一种实现方式中,121、122为卷积层,123为池化层,124、125为卷积层,126为池化层。即卷积层的输出可以作为随后的池化层的输入,也可以作为另一个卷积层的输入以继续进行卷积操作。

[0144] 下面将以卷积层121为例,介绍一层卷积层的内部工作原理。

[0145] 卷积层121可以包括很多个卷积算子,卷积算子也称为核,其在图像处理中的作用相当于一个从输入图像矩阵中提取特定信息的过滤器,卷积算子本质上可以是一个权重矩阵,这个权重矩阵通常被预先定义,在对图像进行卷积操作的过程中,权重矩阵通常在输入图像上沿着水平方向一个像素接着一个像素(或两个像素接着两个像素……这取决于步长stride的取值)的进行处理,从而完成从图像中提取特定特征的工作。该权重矩阵的大小应该与图像的大小相关,需要注意的是,权重矩阵的纵深维度(depth dimension)和输入图像的纵深维度是相同的,在进行卷积运算的过程中,权重矩阵会延伸到输入图像的整个深度。因此,和一个单一的权重矩阵进行卷积会产生一个单一纵深维度的卷积化输出,但是大多数情况下不使用单一权重矩阵,而是应用多个尺寸(行×列)相同的权重矩阵,即多个同型矩阵。每个权重矩阵的输出被堆叠起来形成卷积图像的纵深维度,这里的维度可以理解为由上面所述的“多个”来决定。不同的权重矩阵可以用来提取图像中不同的特征,例如一个权重矩阵用来提取图像边缘信息,另一个权重矩阵用来提取图像的特定颜色,又一个权重矩阵用来对图像中不需要的噪点进行模糊化等。该多个权重矩阵尺寸(行×列)相同,经过该多个尺寸相同的权重矩阵提取后的卷积特征图的尺寸也相同,再将提取到的多个尺寸相同的卷积特征图合并形成卷积运算的输出。

[0146] 这些权重矩阵中的权重值在实际应用中需要经过大量的训练得到,通过训练得到的权重值形成的各个权重矩阵可以用来从输入图像中提取信息,从而使得卷积神经网络100进行正确的预测。

[0147] 当卷积神经网络100有多个卷积层的时候,初始的卷积层(例如121)往往提取较多的一般特征,该一般特征也可以称之为低级别的特征;随着卷积神经网络100深度的加深,越往后的卷积层(例如126)提取到的特征越来越复杂,比如高级别的语义之类的特征,语义越高的特征越适用于待解决的问题。

[0148] 池化层:

[0149] 由于常常需要减少训练参数的数量,因此卷积层之后常常需要周期性的引入池化层,在如图2c中120所示例的121-126各层,可以是一层卷积层后面跟一层池化层,也可以是多层卷积层后面接一层或多层池化层。在图像处理过程中,池化层的唯一目的就是减少图

像的空间大小。池化层可以包括平均池化算子和/或最大池化算子,以用于对输入图像进行采样得到较小尺寸的图像。平均池化算子可以在特定范围内对图像中的像素值进行计算产生平均值作为平均池化的结果。最大池化算子可以在特定范围内取该范围内值最大的像素作为最大池化的结果。另外,就像卷积层中用权重矩阵的大小应该与图像尺寸相关一样,池化层中的运算符也应该与图像的大小相关。通过池化层处理后输出的图像尺寸可以小于输入池化层的图像的尺寸,池化层输出的图像中每个像素点表示输入池化层的图像的对应子区域的平均值或最大值。

[0150] 神经网络层130:

[0151] 在经过卷积层/池化层120的处理后,卷积神经网络100还不足以输出所需要的输出信息。因为如前所述,卷积层/池化层120只会提取特征,并减少输入图像带来的参数。然而为了生成最终的输出信息(所需要的类信息或其他相关信息),卷积神经网络100需要利用神经网络层130来生成一个或者一组所需要的类的数量的输出。因此,在神经网络层130中可以包括多层隐含层(如图2c所示的131、132至13n)以及输出层140,该多层隐含层中所包含的参数可以根据具体的任务类型的相关训练数据进行预先训练得到,例如该任务类型可以包括图像识别,图像分类,图像超分辨率重建等等。

[0152] 在神经网络层130中的多层隐含层之后,也就是整个卷积神经网络100的最后层为输出层140,该输出层140具有类似分类交叉熵的损失函数,具体用于计算预测误差,一旦整个卷积神经网络100的前向传播(如图2c由120至140方向的传播为前向传播)完成,反向传播(如图2c由140至120方向的传播为反向传播)就会开始更新前面提到的各层的权重值以及偏差,以减少卷积神经网络100的损失,及卷积神经网络100通过输出层输出的结果和理想结果之间的误差。

[0153] 需要说明的是,如图2c所示的卷积神经网络120仅作为一种卷积神经网络的示例,在具体的应用中,卷积神经网络还可以以其他网络模型的形式存在。

[0154] 本申请实施例的图像处理方法具体采用的神经网络的结构可以如图2d所示。在图2d中,卷积神经网络(CNN)100可以包括输入层110,卷积层/池化层120(其中池化层为可选的),以及神经网络层130。与图2c相比,图2d中的卷积层/池化层120中的多个卷积层/池化层并行,将分别提取的特征均输入给全神经网络层130进行处理。

[0155] 需要说明的是,图2c和图2d所示的卷积神经网络仅作为一种示例,在具体的应用中,卷积神经网络还可以以其他网络模型的形式存在。

[0156] (5)反向传播算法

[0157] 卷积神经网络可以采用误差反向传播(back propagation,BP)算法在训练过程中修正初始的超分辨率模型中参数的大小,使得超分辨率模型的重建误差损失越来越小。具体地,前向传递输入信号直至输出会产生误差损失,通过反向传播误差损失信息来更新初始的超分辨率模型中参数,从而使误差损失收敛。反向传播算法是以误差损失为主导的反向传播运动,旨在得到最优的超分辨率模型的参数,例如权重矩阵。

[0158] (6)损失函数

[0159] 在训练深度神经网络的过程中,因为希望深度神经网络的输出尽可能的接近真正想要预测的值,所以可以通过比较当前网络的预测值和真正想要的目标值,再根据两者之间的差异情况来更新每一层神经网络的权重向量(当然,在第一次更新之前通常会有初始

化的过程,即为深度神经网络中的各层预先配置参数),比如,如果网络的预测值高了,就调整权重向量让它预测低一些,不断地调整,直到深度神经网络能够预测出真正想要的目标值或与真正想要的目标值非常接近的值。因此,就需要预先定义“如何比较预测值和目标值之间的差异”,这便是损失函数(loss function)或目标函数(objective function),它们是用于衡量预测值和目标值的差异的重要方程。其中,以损失函数举例,损失函数的输出值(loss)越高表示差异越大,那么深度神经网络的训练就变成了尽可能缩小这个loss的过程。

[0160] (7)反向传播算法

[0161] 神经网络可以采用误差反向传播(back propagation,BP)算法在训练过程中修正初始的神经网络模型中参数的大小,使得神经网络模型的重建误差损失越来越小。具体地,前向传递输入信号直至输出会产生误差损失,通过反向传播误差损失信息来更新初始的神经网络模型中参数,从而使误差损失收敛。反向传播算法是以误差损失为主导的反向传播运动,旨在得到最优的神经网络模型的参数,例如权重矩阵。

[0162] 下面介绍本申请实施例提供系统架构。

[0163] 参见图3,本申请实施例提供了一种系统架构100。如所述系统架构100所示,数据采集设备160用于采集训练数据,本申请实施例中训练数据包括:物体的图像或者图像块及物体的类别;并将训练数据存入数据库130,训练设备120基于数据库130中维护的训练数据训练得到CNN特征提取网络(解释说明:这里的特征提取网络就是前面介绍的经训练阶段训练得到的模型,可以是用于特征提取的神经网络等)。下面将以实施例一更详细地描述训练设备120如何基于训练数据得到CNN特征提取网络,该CNN特征提取网络能够用于实现本申请实施例提供的感知网络,即,将待识别图像或图像块通过相关预处理后输入该CNN特征提取网络。本申请实施例中的CNN特征提取网络具体可以为CNN卷积神经网络。需要说明的是,在实际的应用中,所述数据库130中维护的训练数据不一定都来自于数据采集设备160的采集,也有可能是从其他设备接收得到的。另外需要说明的是,训练设备120也不一定完全基于数据库130维护的训练数据进行CNN特征提取网络的训练,也有可能从云端或其他地方获取训练数据进行模型训练,上述描述不应该作为对本申请实施例的限定。

[0164] 根据训练设备120训练得到的目标模型/规则可以应用于不同的系统或设备中,如应用于图3所示的执行设备110,所述执行设备110可以是终端,如手机终端,平板电脑,笔记本电脑,增强现实(augmented reality,AR)AR/虚拟现实(virtual reality,VR),车载终端等,还可以是服务器或者云端等。在图3中,执行设备110配置输入/输出(input/output,I/O)接口112,用于与外部设备进行数据交互,用户可以通过客户设备140向I/O接口112输入数据,所述输入数据在本申请实施例中可以包括:图像或者图像块。

[0165] 在执行设备120对输入数据进行预处理,或者在执行设备120的计算模块111执行计算等相关的处理(比如进行本申请中神经网络的功能实现)过程中,执行设备120可以调用数据存储系统150中的数据、代码等以用于相应的处理,也可以将相应处理得到的数据、指令等存入数据存储系统150中。

[0166] 最后,I/O接口112将处理结果,如上述得到的图像或图像块或者图片中感兴趣物体的2D、3D、Mask、关键点等信息返回给客户设备140,从而提供给用户。

[0167] 可选地,客户设备140,可以是自动驾驶系统中的规划控制单元、手机终端中的美

颜算法模块。

[0168] 值得说明的是,训练设备120可以针对不同的目标或称不同的任务,基于不同的训练数据生成相应的目标模型/规则,该相应的目标模型/规则即可以用于实现上述目标或完成上述任务,从而为用户提供所需的结果。

[0169] 在图3中所示情况下,用户可以手动给定输入数据,该手动给定可以通过I/O接口112提供的界面进行操作。另一种情况下,客户设备140可以自动地向I/O接口112发送输入数据,如果要求客户设备140自动发送输入数据需要获得用户的授权,则用户可以在客户设备140中设置相应权限。用户可以在客户设备140查看执行设备110输出的结果,具体的呈现形式可以是显示、声音、动作等具体方式。客户设备140也可以作为数据采集端,采集如图所示输入I/O接口112的输入数据及输出I/O接口112的输出结果作为新的样本数据,并存入数据库130。当然,也可以不经过客户设备140进行采集,而是由I/O接口112直接将如图所示输入I/O接口112的输入数据及输出I/O接口112的输出结果,作为新的样本数据存入数据库130。

[0170] 值得注意的是,图3仅是本申请实施例提供的一种系统架构的示意图,图中所示设备、器件、模块等之间的位置关系不构成任何限制,例如,在图3中,数据存储系统150相对执行设备110是外部存储器,在其它情况下,也可以将数据存储系统150置于执行设备110中。

[0171] 首先以模型推理阶段为例对介绍本申请实施例提供的感知网络。

[0172] 参照图6,图6为本申请实施例提供的一种感知网络的结构示意,如图6所示,本申请实施例提供的感知网络包括:特征提取网络,所述特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作,所述第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block(图6中示出的为目标stage)。

[0173] 在特征提取网络中,通常包括多个阶段stage,其中每个stage可以包括多个block,每个block可以由至少一个卷积操作构成,参照图5a,其中,图5a示出了现有技术中的一种感知网络的结构示意,其中,特征提取网络包括四个阶段stage(stage1、stage2、stage3、stage4),其中,每个stage中包括的block的输入特征图和输出特征图的尺寸大小一致,所谓尺寸大小一致,是指特征图的通道数以及每个通道的特征图的大小一致,例如stage1中各个block的输入特征图和输出特征图的大小为56*56,通道数为24,stage2中各个block的输入特征图和输出特征图的大小为28*28,通道数为40,stage1中各个block的输入特征图和输出特征图的大小为14*14,通道数为80,stage1中各个block的输入特征图和输出特征图的大小为7*7,通道数为160,每一个block可以包括至少一个卷积操作,例如图5a中示出的三个卷积操作(1*1卷积、3*3卷积以及1*1卷积),且还存在由输入连接到输出的残差连接操作。

[0174] 具体的可以参照图5b,其中,图5b中的卷积模块可以表示为一个阶段stage,则卷积模块1可以对应于图5a中的stage1,卷积模块2可以对应于图5a中的stage2,卷积模块3可以对应于图5a中的stage3,卷积模块4可以对应于图5a中的stage4,特征图C1为stage1输出的特征图,特征图C2为stage2输出的特征图,特征图C3为stage3输出的特征图,特征图C4为stage4输出的特征图,以stage1为例,参照图5c,其中卷积模块1可以包括多个卷积层(或者称之为block)。

[0175] 特征提取网络可以对输入图像进行特征提取,以得到输出特征图,输出特征图可

以输入到任务网络,由任务网络进行相应任务的处理,以得到处理结果,例如任务可以是目标检测,则处理结果可以是图像中目标所在的检测框,任务可以是图像分割,则处理结果可以是图像中目标所在的图像分割区域。

[0176] 参照图5d,在特征提取网络的相同阶段stage之间,不同block的输出特征之间具有较高的相似性,而不同阶段stage的block的输出特征之间,相似性较低。因此本申请实施例利用相同stage之间跨层的目标操作来让感知网络生成这些与关键特征相似性高的特征,降低了模型的参数量,以此提高在GPU设备、TPU设备以及NPU设备上的模型的运行速度。

[0177] 接下来详细描述本申请实施例中的感知网络。

[0178] 本申请实施例中,特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作,所述第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block,且所述目标操作的参数量小于所述M个第二block的参数量。

[0179] 参照图6,所述第一block用于对输入数据进行卷积处理,以得到M个目标特征图,每个目标特征图对应于一个通道。

[0180] 其中,所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,所述M1小于所述M,所述目标操作用于对所述M个目标特征图中的M2个目标特征图进行处理,以得到M2个第二特征图,所述M2小于所述M,所述拼接操作用于将所述M1个第一特征图和所述M2个第二特征图进行拼接,以得到拼接后的特征图。

[0181] 在一种可能的实现中,所述M1个目标特征图与所述M2个目标特征图的交集为空,且所述M1与所述M2的加和为所述M,所述拼接后的特征图的通道数为所述M。

[0182] 相当于,至少一个第二block的输出特征图和输出特征图的通道数都是输出第一block的目标特征图的通道数的一部分,剩下的一部分通道的目标特征图由目标操作产来进行处理。且由于目标操作的参数量小于至少一个第二block的参数量,使得整体上感知网络的参数量降低了,进而可以提高感知网络在GPU设备、TPU设备以及NPU设备上的运行速度。

[0183] 在一种可能的实现中,所述目标操作为参数量小于所述至少一个第二block的卷积操作;或,所述目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作。

[0184] 例如,目标操作为1*1的卷积,或者目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作,相当于M2个第二特征图就是M1个第一特征图,即直接将M1个第一特征图作为M2个第二特征图。

[0185] 应理解,参照图7,还可以将第一block的输出特征图拆分为多组特征图,并由多个目标操作来处理一组特征图,其中只要保证各个目标操作输出的特征图的通道数、至少一个第二block输出的特征图的加和与第一block输出的特征图的通道数相同即可,不同目标操作输出的特征图的通道数之间可以不相同。

[0186] 在一种可能的实现中,所述第一block以及所述M个第二block为所述特征提取网络中目标阶段stage内的block,所述拼接后的特征图用于作为所述特征提取网络中所述目标阶段stage的输出特征图;或,所述目标阶段stage还包括至少一个第三block,所述至少

一个第三block用于对所述拼接后的特征图进行卷积操作,以得到所述目标阶段stage的输出特征图。

[0187] 本申请实施例中,拼接后的特征图可以作为目标阶段stage的输出特征图,或者拼接后的特征图还可以被目标阶段stage包括的其他block(第三block)进行处理,例如参照图8,拼接后的特征图还可以被至少一个第三block进行处理,以得到目标阶段stage的输出特征图。

[0188] 在一种可能的实现中,第一block可以为特征提取网络中的第一个block,或者是中间层的block,参照图9,在第一block之前还可以连接有至少一个第三block,则第一block用于对至少一个第三block输出的数据进行卷积处理。

[0189] 在一种可能的实现中,所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到每个第二block输出的特征图,其中,所述至少一个第二block中距离所述第一block最远的第二block的输出为所述M1个第一特征图;融合操作,用于对所述每个第二block输出的特征图进行融合,以得到融合后的特征图,所述融合后的特征图的尺寸大小与所述M2个第二特征图的尺寸大小相同,对所述融合后的特征图和所述M2个第二特征图进行加法操作,以得到处理后的M2个第二特征图,进而,拼接操作用于将所述M1个第一特征图和所述处理后的M2个第二特征图进行拼接,以得到拼接后的特征图。

[0190] 参照图10,每个第二block都可以输出一个特征图,融合操作的对象为每个第二block输出的特征图以及目标操作的输出(M2个第二特征图),其中所述融合操作,用于对所述每个第二block的输出进行拼接以及降维操作,以得到和所述M2个第二特征图的尺寸大小相同的所述融合后的特征图。

[0191] 具体的,可以参照图11,其中,可以将每个第二block输出的特征图进行拼接操作,得到拼接后的特征图(通道数为每个第二block输出的特征图的加和),由于拼接后的特征图的通道数大于目标操作的输出的特征图的通道数(M2),为了能够对拼接后的特征图与目标操作的输出(M2个第二特征图)进行加法运算,则需要将拼接后的特征图进行降维操作,以使得拼接后的特征图的通道数等于M2,进而可以对拼接后的特征图与目标操作的输出进行矩阵的加法运算。

[0192] 以第二block的数量为5为例,此时共有5个第二block的输出会融合进最终输出。先拼接这5个特征为Z=[Y1,Y2,Y3,Y4,Y5],再对特征Z进行降维操作,使其的维度和M2一致并进行相加。具体公式可以为如下:

[0193] $\tau(Z)=W_2\sigma(W_1\mathrm{Pooling}(Z)+b_1)+b_2$,

[0194] $$Y_n^2 \leftarrow Y_n^2 + \tau(Z),$$

[0195] 本申请实施例中,感知网络中的特征提取网络可以由基于上述实施例中提供的阶段stage堆叠而成,示例性的,网络结构的一个示意可以如表1所示。其中,output代表输出特征图的大小,#out代表输出特征图的通道数(Block和Cheap的通道数实际上为#out的1/2)。

[0196] 表1

[0197]

stage	output	operator	#out	stride
stem	112 × 112	Conv3×3	16	2
1	56 × 56	Block	24	2
	56 × 56	Block×1 Cheap	24	1
	56 × 56	Concat	24	1
2	28 × 28	Block	40	2
	28 × 28	Block×1 Cheap	40	1
	28 × 28	Concat	40	1
3	14 × 14	Block	80	2
	14 × 14	Block×5 Cheap	80	1
	14 × 14	Concat	80	1
4	7 × 7	Block	160	1
	7 × 7	Block×5 Cheap	160	1
	7 × 7	Concat	160	1
head	7 × 7	Conv1×1	960	1
	1 × 1	Pool & Conv1×1	1280	1
	1 × 1	FC	1000	-

[0198] 在一种可能的实现中,所述特征提取网络用于获取输入的图像,并对所述输入的图像进行特征提取,输出所述输入的图像的特征图;所述感知网络还包括:任务网络,用于根据所述输入的图像的特征图,进行对应任务的处理,以得到处理结果。其中,所述任务包括目标检测、图像分割或图像分类等等图像处理任务。

[0199] 接下来描述一种感知网络的结构示意:

[0200] 参照图11,本申请实施例中的特征提取网络可以为图11中所示的主干网络,特征金字塔(feature pyramid network,FPN)与主干网络backbone连接,FPN可以对主干网络backbone生成的多个不同分辨率的特征图进行卷积处理,来构造特征金字塔。

[0201] 参照图12,图12为一种FPN的结构示意,其中,使用卷积模块1对最顶层特征图C4进行处理,卷积模块1可以包括至少一个卷积层,示例性的,卷积模块1可以使用空洞卷积和1×1卷积将最顶层特征图C4的通道数下降为256,作为特征金字塔的最顶层特征图P4;横向链接最顶层下一层特征图C3的输出结果并使用1×1卷积(卷积模块2)降低通道数至256后,与特征图p4逐像素相加得到特征图p3;以此类推,从上到下,构建出特征金字塔Φp={特征图p4,特征图p3,特征图p2,特征图p1}。

[0202] 本申请实施例中,FPN包括多个卷积模块,每个卷积模块包括多个卷积层,每个卷积模块可以对输入的特征图进行卷积处理,本申请实施例中FPN包括的第二卷积层为FPN包括的多个卷积层中的一个。

[0203] 需要说明的是,图12中示出的FPN仅为一种实现方式,并不构成对本申请的限定。

[0204] 本申请实施例中,以任务网络要实现的任务为目标检测为例,header与FPN连接,header可以根据FPN提供的特征图,完成一个任务的2D框的检测,输出这个任务的物体的2D框以及对应的置信度等等,接下来描述一种header的结构示意,图13为一种header的示意,如图13中示出的那样,Header包括候选区域生成网络(Region Proposal Network,RPN)、ROI-ALIGN和RCNN三个模块。

[0205] 其中,RPN模块可以用于在FPN提供的一个或者多个特征图上预测所述任务物体所在的区域,并输出匹配所述区域的候选2D框;或者可以这样理解,RPN在FPN输出的一个或者多个横图上预测出可能存在该任务物体的区域,并且给出这些区域的框,这些区域称为候选区域(Proposal)。比如,当Header负责检测车时,其RPN层就预测出可能存在车的候选框;

当Header负责检测人时,其RPN层就预测出可能存在人的候选框。当然,这些Proposal是不准确的,一方面其不一定含有该任务的物体,另一方面这些框也是不紧致的。

[0206] 2D候选区域预测流程可以由Header的RPN模块实施,其根据FPN提供的特征图,预测出可能存在该任务物体的区域,并且给出这些区域的候选框(也可以叫候选区域,Proposal)。在本实施例中,若Header负责检测车,其RPN层就预测出可能存在车的候选框。

[0207] 参照图14,在FPN提供的特征图上通过卷积模块1(例如一个3*3的卷积),生成特征图RPNHidden。后面Header的RPN层将会从RPN Hidden中预测Proposal。具体来说,Header的RPN层分别通过卷积模块2和卷积模块3(例如分别是一个1*1的卷积),预测出RPN Hidden每个位置处的Proposal的坐标以及置信度。这个置信度越高,表示这个Proposal存在该任务的物体的概率越大。比如,在Header中某个Proposal的score越大,就表示其存在车的概率越大。每个RPN层预测出来的Proposal需要经过Proposal合并模块,根据Proposal之间的重合程度去掉多余的Proposal(这个过程可以采用但不限制于NMS算法),在剩余的K个Proposal中挑选出score最大的N(N<k)个proposal作为候选的可能存在物体的区域。从图14可以看出,这些Proposal是不准确的,一方面其不一定含有该任务的物体,另一方面这些框也是不紧致的。因此,RPN模块只是一个粗检测的过程,需要后续的RCNN模块进行细分。在RPN模块回归Proposal的坐标时,并不是直接回归坐标的绝对值,而是回归出相对于Anchor的坐标。当这些Anchor与实际的物体匹配越高,RPN能检测出物体的概率越大。</k)个proposal作为候选的可能存在物体的区域。

[0208] ROI-ALIGN模块用于根据所述RPN模块预测得到的区域,从所述FPN提供的一个特征图中扣取出所述候选2D框所在区域的特征;也就是说,ROI-ALIGN模块主要根据RPN模块提供的Proposal,在某个特征图上把每个Proposal所在的区域的特征扣取出来,并且resize到固定的大小,得到每个Proposal的特征。可以理解的是,ROI-ALIGN模块可以使用但不局限于ROI-POOLING(感兴趣区域池化)/ROI-ALIGN(感兴趣区域提取)/PS-ROIPOOLING(位置敏感的感兴趣区域池化)/PS-ROIALIGN(位置敏感的感兴趣区域提取)等特征抽取方法。

[0209] RCNN模块用于通过神经网络对所述候选2D框所在区域的特征进行卷积处理,得到所述候选2D框属于各个物体类别的置信度;通过神经网络对所述候选区域2D框的坐标进行调整,使得调整后的2D候选框比所述候选2D框与实际物体的形状更加匹配,并选择置信度大于预设阈值的调整后的2D候选框作为所述区域的2D框。也就是说,RCNN模块主要是对ROI-ALIGN模块提出的每个Proposal的特征进行细化处理,得到每个Proposal的属于各个类别置信度(比如对于车这个任务,会给出Backgroud/Car/Truck/Bus 4个分数),同时对Proposal的2D框的坐标进行调整,输出更加紧致的2D框。这些2D框经过非极大值抑制(non maximum suppression,NMS)合并后,作为最后的2D框输出。

[0210] 2D候选区域细分类主要由图14中的Header的RCNN模块实施,其根据ROI-ALIGN模块提取出来的每个Proposal的特征,进一步回归出更加紧致的2D框坐标,同时对这个Proposal进行分类,输出其属于各个类别的置信度。RCNN的可实现形式很多,其中一种实现形式如图13所示。ROI-ALIGN模块输出的特征大小可以为N*14*14*256(Feature of proposals),其在RCNN模块中首先经过Resnet18的卷积模块4(Res18-Conv5)处理,输出的特征大小为N*7*7*512,然后通过一个Global Avg Pool(平均池化层)进行处理,把输入特

征中每个通道内的7*7的特征进行平均,得到N*512的特征,其中每个1*512维的特征向量代表每个Proposal的特征。接下来通过2个全连接层FC分别回归框的精确坐标(输出N*4的向量,这4个数值分表表示框的中心点x/y坐标,框的宽高),框的类别的置信度(在Header0中,需要给出这个框是Backgroud/Car/Truck/Bus的分数)。最后通过框合并操作,选择分数最大的若干个框,并且通过NMS操作去除重复的框,从而得到紧致的框输出。

[0211] 在一些实际应用场景中,该感知网络还可以包括其他Header,可以在检测出2D框的基础上,进一步进行3D/Mask/Keypoint检测。示例性的,以3D为例,ROI-ALIGN模块根据Header提供的准确的2D框,在FPN输出的特征图上提取出每个2D框所在区域的特征,假设2D框的个数为M,那么ROI-ALIGN模块输出的特征大小为M*14*14*256,其首先经过Resnet18的卷积模块5(例如为Res18-Conv5)处理,输出的特征大小为N*7*7*512,然后通过一个Global Avg Pool(平均池化层)进行处理,把输入特征中每个通道的7*7的特征进行平均,得到M*512的特征,其中每个1*512维的特征向量代表每个2D框的特征。接下来通过3个全连接层FC分别回归框中物体的朝向角(orientation,M*1向量)、质心点坐标(centroid,M*2向量,这2个数值表示质心的x/y坐标)和长宽高(dimention)。

[0212] 本申请实施例中,header包括至少是一个卷积模块,每个卷积模块包括至少一个卷积层,每个卷积模块可以对输入的特征图进行卷积处理,本申请实施例中header包括的第三卷积层为header包括的多个卷积层中的一个。

[0213] 需要说明的是,图13和图14示出的header仅为一种实现方式,并不构成对本申请的限定。

[0214] 接下来进行本申请实施例中提供的感知网络的效果描述:

[0215] 本申请实施例在图像分类数据集CIFAR10的结果可以如表2和表3所示,相比于ResNet等其他方法,本申请实施例中提供的感知网络结构在计算量和参数量最小的情况下达到了很高的精度,其中,表2为图像分类数据集CIFAR10上的结果,表3为图像分类数据集CIFAR10上与现有轻量级网络的对比。

[0216] 表2

[0217]

Model	FLOPs (M)	Params (M)	Acc (%)
ResNet56	125.0	0.85	94.14±0.48
ResNet56-0.7×	63.3	0.41	93.15±0.33
GhostV2 w/o mix	53.7	0.36	93.67±0.27
GhostV2 w/ mix	53.8	0.37	93.93±0.21

[0218] 表3

[0219]

Model	FLOPs (M)	Params (M)	Acc (%)
ResNet56	125.0	0.85	94.24±0.48
GAL [25]	78.3	0.75	92.98
FPGM [11]	59.4	-	93.49±0.13
LEGR [2]	58.9	-	93.7±0.2
HRank [24]	62.7	0.49	93.17
GhostV2 w/ mix	53.8	0.37	**93.93±0.21**

[0220] 本申请实施例中提供的感知网络在图像分类数据集ImageNet的结果可以如表4、

表5和表6所示,对比具有同样推理速度的基线网络ResNet,本申请实施例中提供的感知网络可以在该大型分类数据集上提升1.4%的推理精度。相比于其他一系列轻量级推理网络,本申请实施例中提供的感知网络能同时达到最快的推理速度和较高的推理精度,表4为图像分类数据集ImageNet上与基线网络ResNet的对比,表5为图像分类数据集ImageNet上与另一基线网络RegNet的对比,表6为图像分类数据集ImageNet上与其他轻量级网络的对比。

[0221] 表4

[0222]

Model	FLOPs (B)	Params (M)	Latency (ms)	Top1 (%)	Top5 (%)
ResNet34	3.7	21.8	29.1	73.3	91.4
ResNet34-0.8×	2.3	14.0	24.2	71.7	90.3
GhostV2 w/o mix	2.3	14.3	24.1	72.8	90.9
GhostV2 w/ mix	2.3	14.6	24.5	73.1	91.2

[0223] 表5

[0224]

Model	FLOPs (B)	Params (M)	Latency (ms)	Top1 (%)	Top5 (%)
RegNetX-3.2GF	3.2	15.3	59.09	78.2	94.0
RegNetX-3.2GF-0.75×	1.8	8.8	46.06	76.3	93.0
GhostV2 w/o mix	1.8	9.7	43.88	76.8	93.2
GhostV2 w/ mix	1.8	10.5	44.37	77.4	93.7

[0225] 表6

[0226]

Model	FLOPs	Params	CPU Latency	GPU Latency	Top-1	Top-5
MobileNetV2 1.0× [34]	300M	3.5M	50.5 ms	18.8 ms	71.8%	91.0%
MobileNetV3-Large 0.75× [13]	155M	4.0M	32.2 ms	16.4 ms	73.3%	-
GhostNetV1-1.0× [7]	141M	5.2M	31.1 ms	17.2 ms	73.9%	91.4%
GhostNetV2-1.25×	206M	5.1M	42.2 ms	**14.8** ms	**74.1%**	**91.6%**
MobileNetV3-Large 1.0× [13]	219M	5.4M	46.4 ms	19.1 ms	75.2%	-
GhostNetV1-1.3× [7]	226M	7.3M	46.9 ms	21.3 ms	**75.7%**	**92.7%**
GhostNetV2-1.5×	295M	6.3M	61.7 ms	**17.9** ms	**75.7%**	**92.7%**
MobileNetV2 1.4× [34]	582M	6.1M	81.6 ms	28.0 ms	75.0%	92.5%
MobileNetV3-Large 1.25× [7]	356M	7.5M	66.3 ms	25.5 ms	76.6%	-
GhostNetV1-1.65× [7]	351M	10.5M	67.4 ms	27.9 ms	76.9%	93.2%
EfficientNet-B0 [37]†	388M	5.3M	78.6 ms	26.9 ms	77.1%	93.3%
GhostNetV2-2.0×	466M	9.2M	87.7 ms	22.8 ms	76.9%	93.3%
GhostNetV2-2.0× + RA†	466M	9.2M	87.7 ms	**22.8** ms	**77.4%**	**93.6%**

[0227] 以任务网络实现的任务为目标检测为例,实验结果可以如表7所示,本申请实施例中提供的感知网络在达到最高mAP的同时还具有最快的推理速度,达到每秒25.9帧。

[0228] 表7

[0229]

Backbone	mAP	AP_S	AP_M	AP_L	FLOPs	FPS
RegNetX-3.2GF	39.1	22.6	43.5	50.8	3.2B	21.9
ResNet50	36.8	20.5	40.6	48.5	4.1B	22.4
ResNeXt50	37.6	21.1	41.4	49.4	4.2B	19.6
RegNetX-3.2GF-0.75×	37.7	21.2	41.6	49.5	1.8B	25.2
GhostV2-RegNetX-3.2GF	**38.4**	22.3	42.3	50.1	1.8B	**25.9**

[0230] 本申请实施例提供了一种感知网络,所述感知网络包括:特征提取网络,所述特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作,所述

第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block,且所述目标操作的参数量小于所述M个第二block的参数量;所述第一block用于对输入数据进行卷积处理,以得到M个目标特征图,每个目标特征图对应于一个通道;所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,所述M1小于所述M;所述目标操作用于对所述M个目标特征图中的M2个目标特征图进行处理,以得到M2个第二特征图,所述M2小于所述M;所述拼接操作用于将所述M1个第一特征图和所述M2个第二特征图进行拼接,以得到拼接后的特征图。通过上述方式,利用相同stage之间跨层的目标操作来让感知网络生成这些与关键特征相似性高的特征,降低了模型的参数量,以此提高在GPU设备、TPU设备以及NPU设备上的模型的运行速度。

[0231] 接下来从产品应用的角度,介绍几种本申请实施例的应用场景。

[0232] 本申请实施例主要应用在终端手机、云服务、检测、分割、底层视觉等以计算机视觉为基础的领域。面对海量数据的并行运算,计算机视觉及其相关任务对于算力的要求不断提升,对硬件的运算速度及功耗提出了更高的要求。现在越来越多的终端设备如手机芯片上都已经开始大量部署GPU和NPG处理单元。本申请实施例提出的轻量级网络模型在GPU上能极大提升推理速度,并可以作为基础网络直接替换现有的如目标检测器的基础特征提取网络,部署在自动驾驶等场景。在实际应用中,可以适配广泛应用场景和设备(如手机终端、云服务器等),利用该快速推理网络进行数据处理、图像检索等任务。

[0233] 在一种场景中,本申请实施例可以部署在手机终端上,为用户带来高效准确的推理。如手机拍照后的图像处理、图片识别、手机端目标检测等。

[0234] 在一种场景中,本发明可以在云服务上进行轻量级部署,为用户提供高效数据处理的服务,助力深度学习提速增效。用户上传待处理的数据,即可通过云服务上的推理模型进行快速的数据处理。

[0235] 在一种场景中,自动驾驶任务中,对视野范围内的行人、车辆等目标进行实时的检测对车辆做出正确的驾驶决策至关重要。本发明可以直接替换现有目标检测器的特征提取模块,压缩并加速检测器的推理过程。

[0236] 参照图4,本申请实施例还提供了一种数据处理方法,所述方法包括:

[0237] 1401、获取特征提取网络,所述特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作,所述第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block,且所述目标操作的参数量小于所述M个第二block的参数量;

[0238] 步骤1401的具体描述可以参照上述实施例中关于特征提取网络的描述,这里不再赘述。

[0239] 1402、通过所述第一block,对输入数据进行卷积处理,以得到M个目标特征图,每个目标特征图对应于一个通道;

[0240] 步骤1402的具体描述可以参照上述实施例中关于第一block的描述,这里不再赘述。

[0241] 1403、通过所述至少一个第二block,对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,所述M1小于所述M;

[0242] 步骤1403的具体描述可以参照上述实施例中关于至少一个第二block的描述,这

里不再赘述。

[0243] 1404、通过所述目标操作,对所述M个目标特征图中的M2个目标特征图进行处理,以得到M2个第二特征图,所述M2小于所述M;

[0244] 步骤1404的具体描述可以参照上述实施例中关于目标操作的描述,这里不再赘述。

[0245] 1405、通过所述拼接操作,将所述M1个第一特征图和所述M2个第二特征图进行拼接,以得到拼接后的特征图。

[0246] 步骤1405的具体描述可以参照上述实施例中关于拼接操作的描述,这里不再赘述。

[0247] 在一种可能的实现中,所述M1个目标特征图与所述M2个目标特征图的交集为空,且所述M1与所述M2的加和为所述M,所述拼接后的特征图的通道数为所述M。

[0248] 在一种可能的实现中,所述目标操作为参数量小于所述至少一个第二block的卷积操作;或,

[0249] 所述目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作。

[0250] 在一种可能的实现中,所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到每个第二block输出的特征图,其中,所述至少一个第二block中距离所述第一block最远的第二block的输出为所述M1个第一特征图;

[0251] 所述方法还包括:

[0252] 通过融合操作,对所述每个第二block输出的特征图进行融合,以得到融合后的特征图,所述融合后的特征图的尺寸大小与所述M2个第二特征图的尺寸大小相同;对所述融合后的特征图和所述M2个第二特征图进行加法操作,以得到处理后的M2个第二特征图;

[0253] 所述通过所述拼接操作,将所述M1个第一特征图和所述M2个第二特征图进行拼接,包括:

[0254] 将所述M1个第一特征图和所述处理后的M2个第二特征图进行拼接,以得到拼接后的特征图。

[0255] 在一种可能的实现中,所述通过融合操作,对所述每个第二block输出的特征图进行融合,包括:

[0256] 通过融合操作,对所述每个第二block的输出进行拼接以及降维操作,以得到和所述M2个第二特征图的尺寸大小相同的所述融合后的特征图。

[0257] 在一种可能的实现中,所述第一block以及所述M个第二block为所述特征提取网络中目标阶段stage内的block,所述拼接后的特征图用于作为所述特征提取网络中所述目标阶段stage的输出特征图;或,

[0258] 所述目标阶段stage还包括至少一个第三block,所述至少一个第三block用于对所述拼接后的特征图进行卷积操作,以得到所述目标阶段stage的输出特征图。

[0259] 在一种可能的实现中,所述特征提取网络用于获取输入的图像,并对所述输入的图像进行特征提取,输出所述输入的图像的特征图;

[0260] 所述方法还包括:

[0261] 通过任务网络,用于根据所述输入的图像的特征图,进行对应任务的处理,以得到

处理结果。

[0262] 在一种可能的实现中,所述任务包括目标检测、图像分割或图像分类。

[0263] 参照图15,图15为本申请实施例提供的一种数据处理装置1500的示意,如图15中示出的那样,本申请提供的数据处理装置1500包括:

[0264] 获取模块1501,用于获取特征提取网络,所述特征提取网络包括第一块block、串行连接的至少一个第二block、目标操作以及拼接操作,所述第一block以及所述M个第二block为所述特征提取网络中同一个阶段stage内的block,且所述目标操作的参数量小于所述M个第二block的参数量;

[0265] 卷积处理模块1502,通过所述第一block,对输入数据进行卷积处理,以得到M个目标特征图,每个目标特征图对应于一个通道;

[0266] 通过所述至少一个第二block,对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到M1个第一特征图,所述M1小于所述M;

[0267] 通过所述目标操作,对所述M个目标特征图中的M2个目标特征图进行处理,以得到M2个第二特征图,所述M2小于所述M;

[0268] 拼接模块1503,用于通过所述拼接操作,将所述M1个第一特征图和所述M2个第二特征图进行拼接,以得到拼接后的特征图。

[0269] 在一种可能的实现中,所述M1个目标特征图与所述M2个目标特征图的交集为空,且所述M1与所述M2的加和为所述M,所述拼接后的特征图的通道数为所述M。

[0270] 在一种可能的实现中,所述目标操作为参数量小于所述至少一个第二block的卷积操作;或,

[0271] 所述目标操作为由所述第一block的输出到所述拼接操作的输出之间的残差连接操作。

[0272] 在一种可能的实现中,所述至少一个第二block用于对所述M个目标特征图中的M1个目标特征图进行卷积处理,以得到每个第二block输出的特征图,其中,所述至少一个第二block中距离所述第一block最远的第二block的输出为所述M1个第一特征图;

[0273] 所述装置还包括:

[0274] 融合模块,用于通过融合操作,对所述每个第二block输出的特征图进行融合,以得到融合后的特征图,所述融合后的特征图的尺寸大小与所述M2个第二特征图的尺寸大小相同;对所述融合后的特征图和所述M2个第二特征图进行加法操作,以得到处理后的M2个第二特征图;

[0275] 所述拼接模块,用于将所述M1个第一特征图和所述处理后的M2个第二特征图进行拼接,以得到拼接后的特征图。

[0276] 在一种可能的实现中,所述融合模块,用于通过融合操作,对所述每个第二block的输出进行拼接以及降维操作,以得到和所述M2个第二特征图的尺寸大小相同的所述融合后的特征图。

[0277] 在一种可能的实现中,所述第一block以及所述M个第二block为所述特征提取网络中目标阶段stage内的block,所述拼接后的特征图用于作为所述特征提取网络中所述目标阶段stage的输出特征图;或,

[0278] 所述目标阶段stage还包括至少一个第三block,所述至少一个第三block用于对

所述拼接后的特征图进行卷积操作,以得到所述目标阶段stage的输出特征图。

[0279] 在一种可能的实现中,所述特征提取网络用于获取输入的图像,并对所述输入的图像进行特征提取,输出所述输入的图像的特征图;

[0280] 所述装置还包括:

[0281] 任务处理模块,用于通过任务网络,用于根据所述输入的图像的特征图,进行对应任务的处理,以得到处理结果。

[0282] 在一种可能的实现中,所述任务包括目标检测、图像分割或图像分类。

[0283] 接下来介绍本申请实施例提供的一种执行设备,请参阅图16,图16为本申请实施例提供的执行设备的一种结构示意图,执行设备1600具体可以表现为手机、平板、笔记本电脑、智能穿戴设备、服务器等,此处不做限定。其中,执行设备1600上可以部署有图15对应实施例中所描述的数据处理装置,用于实现图15对应实施例中数据处理的功能。具体的,执行设备1600包括:接收器1601、发射器1602、处理器1603和存储器1604(其中执行设备1600中的处理器1603的数量可以一个或多个,图11中以一个处理器为例),其中,处理器1603可以包括应用处理器16031和通信处理器16032。在本申请的一些实施例中,接收器1601、发射器1602、处理器1603和存储器1604可通过总线或其它方式连接。

[0284] 存储器1604可以包括只读存储器和随机存取存储器,并向处理器1603提供指令和数据。存储器1604的一部分还可以包括非易失性随机存取存储器(non-volatile random access memory,NVRAM)。存储器1604存储有处理器和操作指令、可执行模块或者数据结构,或者它们的子集,或者它们的扩展集,其中,操作指令可包括各种操作指令,用于实现各种操作。

[0285] 处理器1603控制执行设备的操作。具体的应用中,执行设备的各个组件通过总线系统耦合在一起,其中总线系统除包括数据总线之外,还可以包括电源总线、控制总线和状态信号总线等。但是为了清楚说明起见,在图中将各种总线都称为总线系统。

[0286] 上述本申请实施例揭示的方法可以应用于处理器1603中,或者由处理器1603实现。处理器1603可以是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述方法的各步骤可以通过处理器1603中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器1603可以是通用处理器、数字信号处理器(digital signal processing,DSP)、微处理器或微控制器、以及视觉处理器(vision processing unit,VPU)、张量处理器(tensor processing unit,TPU)等适用于AI运算的处理器,还可进一步包括专用集成电路(application specific integrated circuit,ASIC)、现场可编程门阵列(field-programmable gate array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。该处理器1603可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成,或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器,闪存、只读存储器,可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器1604,处理器1603读取存储器1604中的信息,结合其硬件完成上述方法的步骤。

[0287] 接收器1601可用于接收输入的数字或字符信息,以及产生与执行设备的相关设置

以及功能控制有关的信号输入。发射器1602可用于通过第一接口输出数字或字符信息;发射器1602还可用于通过第一接口向磁盘组发送指令,以修改磁盘组中的数据;发射器1602还可以包括显示屏等显示设备。

[0288] 执行设备可以运行图6所描述的感知网络,或者执行图14对应实施例中的数据处理方法。

[0289] 本申请实施例还提供了一种训练设备,请参阅图17,图17是本申请实施例提供的训练设备一种结构示意图,具体的,训练设备1700由一个或多个服务器实现,训练设备1700可因配置或性能不同而产生比较大的差异,可以包括一个或一个以上中央处理器(central processing units,CPU)1717(例如,一个或一个以上处理器)和存储器1732,一个或一个以上存储应用程序1742或数据1744的存储介质1730(例如一个或一个以上海量存储设备)。其中,存储器1732和存储介质1730可以是短暂存储或持久存储。存储在存储介质1730的程序可以包括一个或一个以上模块(图示没标出),每个模块可以包括对训练设备中的一系列指令操作。更进一步地,中央处理器1717可以设置为与存储介质1730通信,在训练设备1700上执行存储介质1730中的一系列指令操作。

[0290] 训练设备1700还可以包括一个或一个以上电源1726,一个或一个以上有线或无线网络接口1750,一个或一个以上输入输出接口1758;或,一个或一个以上操作系统1741,例如Windows ServerTM,Mac OS XTM,UnixTM,LinuxTM,FreeBSDTM等等。

[0291] 图17中描述的模型训练装置1700可以为训练设备中的模块,训练设备中的处理器可以执行通过模型训练得到图6所描述的感知网络。

[0292] 本申请实施例中还提供一种包括计算机程序产品,当其在计算机上运行时,使得计算机执行如前述执行设备所执行的步骤,或者,使得计算机执行如前述训练设备所执行的步骤。

[0293] 本申请实施例中还提供一种计算机可读存储介质,该计算机可读存储介质中存储有用于进行信号处理的程序,当其在计算机上运行时,使得计算机执行如前述执行设备所执行的步骤,或者,使得计算机执行如前述训练设备所执行的步骤。

[0294] 本申请实施例提供的执行设备、训练设备或终端设备具体可以为芯片,芯片包括:处理单元和通信单元,所述处理单元例如可以是处理器,所述通信单元例如可以是输入/输出接口、管脚或电路等。该处理单元可执行存储单元存储的计算机执行指令,以使执行设备内的芯片执行上述实施例描述的数据处理方法,或者,以使训练设备内的芯片执行上述实施例描述的数据处理方法。可选地,所述存储单元为所述芯片内的存储单元,如寄存器、缓存等,所述存储单元还可以是所述无线接入设备端内的位于所述芯片外部的存储单元,如只读存储器(read-only memory,ROM)或可存储静态信息和指令的其他类型的静态存储设备,随机存取存储器(random access memory,RAM)等。

[0295] 具体的,请参阅图18,图18为本申请实施例提供的芯片的一种结构示意图,所述芯片可以表现为神经网络处理器NPU 1800,NPU 1800作为协处理器挂载到主CPU(Host CPU)上,由Host CPU分配任务。NPU的核心部分为运算电路1803,通过控制器1804控制运算电路1803提取存储器中的矩阵数据并进行乘法运算。

[0296] NPU 1800可以通过内部的各个器件之间的相互配合,来实现图6所描述的实施例中提供的模型训练方法,或者对训练得到的模型进行推理。

[0297] 其中,NPU 1800中的运算电路1803可以执行获取第一神经网络模型以及对所述第一神经网络模型进行模型训练的步骤。

[0298] 更具体的,在一些实现中,NPU 1800中的运算电路1803内部包括多个处理单元(Process Engine,PE)。在一些实现中,运算电路1803是二维脉动阵列。运算电路1803还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中,运算电路1803是通用的矩阵处理器。

[0299] 举例来说,假设有输入矩阵A,权重矩阵B,输出矩阵C。运算电路从权重存储器1802中取矩阵B相应的数据,并缓存在运算电路中每一个PE上。运算电路从输入存储器1801中取矩阵A数据与矩阵B进行矩阵运算,得到的矩阵的部分结果或最终结果,保存在累加器(accumulator)1808中。

[0300] 统一存储器1806用于存放输入数据以及输出数据。权重数据直接通过存储单元访问控制器(Direct Memory Access Controller,DMAC)1805,DMAC被搬运到权重存储器1802中。输入数据也通过DMAC被搬运到统一存储器1806中。

[0301] BIU为Bus Interface Unit即,总线接口单元1810,用于AXI总线与DMAC和取指存储器(Instruction Fetch Buffer,IFB)1809的交互。

[0302] 总线接口单元1810(Bus Interface Unit,简称BIU),用于取指存储器1809从外部存储器获取指令,还用于存储单元访问控制器1805从外部存储器获取输入矩阵A或者权重矩阵B的原数据。

[0303] DMAC主要用于将外部存储器DDR中的输入数据搬运到统一存储器1806或将权重数据搬运到权重存储器1802中或将输入数据数据搬运到输入存储器1801中。

[0304] 向量计算单元1807包括多个运算处理单元,在需要的情况下,对运算电路1803的输出做进一步处理,如向量乘,向量加,指数运算,对数运算,大小比较等等。主要用于神经网络中非卷积/全连接层网络计算,如Batch Normalization(批归一化),像素级求和,对特征平面进行上采样等。

[0305] 在一些实现中,向量计算单元1807能将经处理的输出的向量存储到统一存储器1806。例如,向量计算单元1807可以将线性函数;或,非线性函数应用到运算电路1803的输出,例如对卷积层提取的特征平面进行线性插值,再例如累加值的向量,用以生成激活值。在一些实现中,向量计算单元1807生成归一化的值、像素级求和的值,或二者均有。在一些实现中,处理过的输出的向量能够用作到运算电路1803的激活输入,例如用于在神经网络中的后续层中的使用。

[0306] 控制器1804连接的取指存储器(instruction fetch buffer)1809,用于存储控制器1804使用的指令;

[0307] 统一存储器1806,输入存储器1801,权重存储器1802以及取指存储器1809均为On-Chip存储器。外部存储器私有于该NPU硬件架构。

[0308] 其中,上述任一处提到的处理器,可以是一个通用中央处理器,微处理器,ASIC,或一个或多个用于控制上述程序执行的集成电路。

[0309] 另外需说明的是,以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实

际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。另外,本申请提供的装置实施例附图中,模块之间的连接关系表示它们之间具有通信连接,具体可以实现为一条或多条通信总线或信号线。

[0310] 通过以上的实施方式的描述,所属领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件的方式来实现,当然也可以通过专用硬件包括专用集成电路、专用CPU、专用存储器、专用元器件等来实现。一般情况下,凡由计算机程序完成的功能都可以很容易地用相应的硬件来实现,而且,用来实现同一功能的具体硬件结构也可以是多种多样的,例如模拟电路、数字电路或专用电路等。但是,对本申请而言更多情况下软件程序实现是更佳的实施方式。基于这样的理解,本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在可读取的存储介质中,如计算机的软盘、U盘、移动硬盘、ROM、RAM、磁碟或者光盘等,包括若干指令用以使得一台计算机设备(可以是个人计算机,训练设备,或者网络设备等)执行本申请各个实施例所述的方法。

[0311] 在上述实施例中,可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时,可以全部或部分地以计算机程序产品的形式实现。

[0312] 所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时,全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中,或者从一个计算机可读存储介质向另一计算机可读存储介质传输,例如,所述计算机指令可以从一个网站站点、计算机、训练设备或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、训练设备或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存储的任何可用介质或者是包含一个或多个可用介质集成的训练设备、数据中心等数据存储设备。所述可用介质可以是磁性介质,(例如,软盘、硬盘、磁带)、光介质(例如,DVD)、或者半导体介质(例如固态硬盘(Solid State Disk,SSD))等。

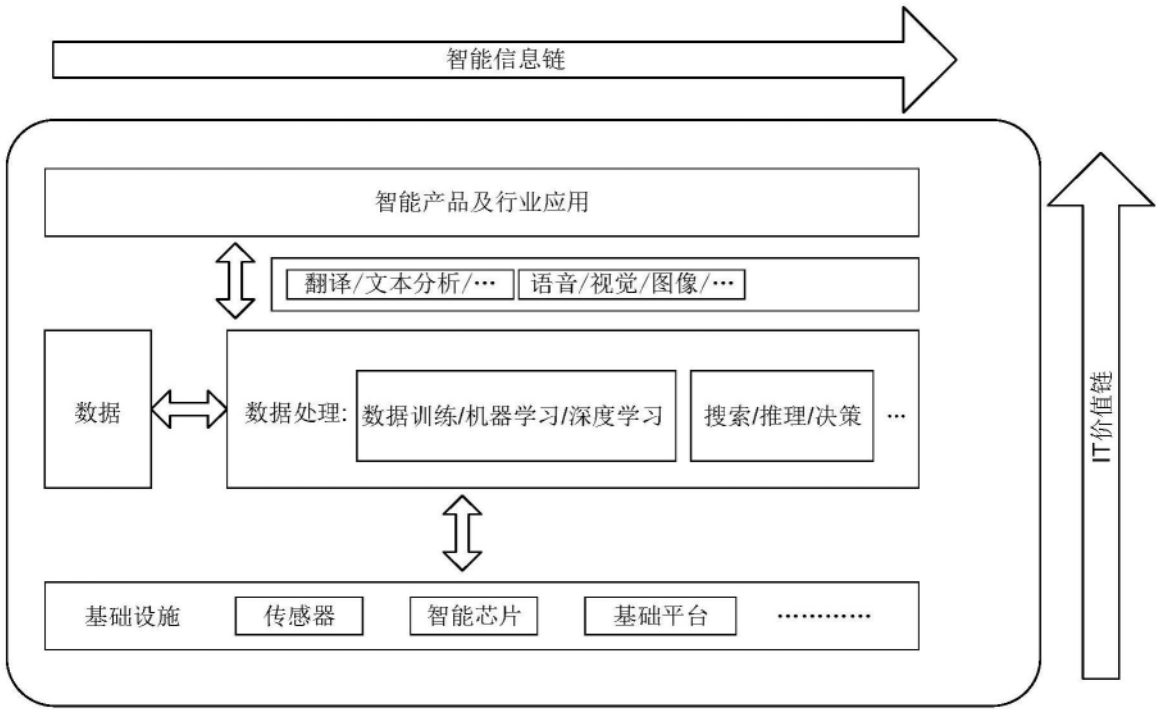

智能信息链
智能产品及行业应用
翻译/文本分析/…
语音/视觉/图像/…
数据
数据处理:
数据训练/机器学习/深度学习
搜索/推理/决策
…
基础设施
传感器
智能芯片
基础平台
……
IT价值链

图1

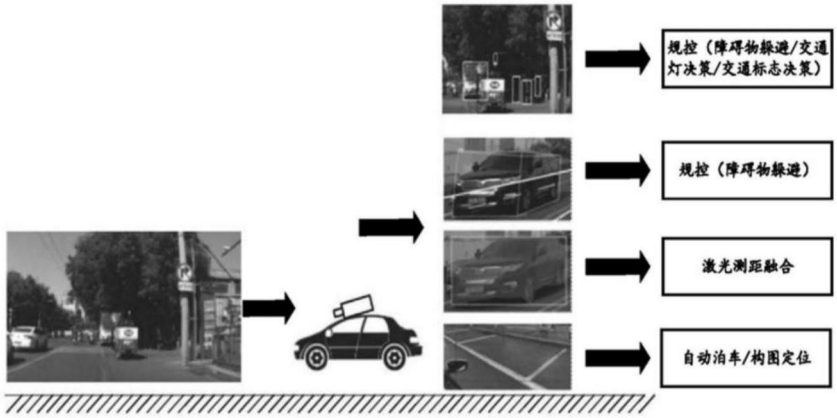

规控（障碍物躲避/交通灯决策/交通标志决策）
规控（障碍物躲避）
激光测距融合
自动泊车/构图定位

图2a

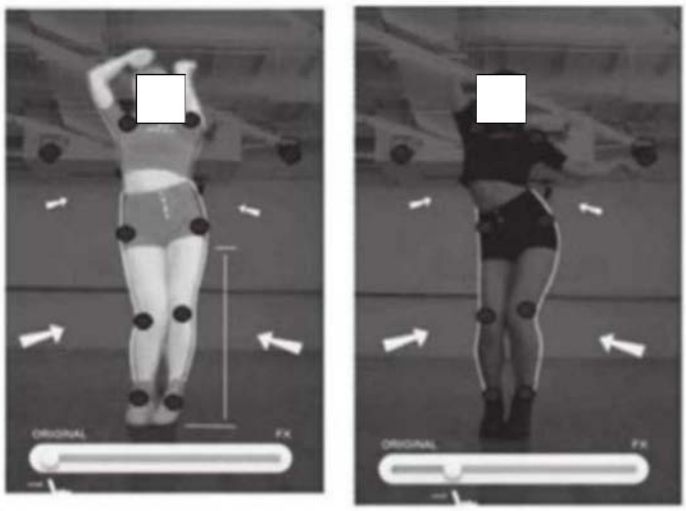


图2b

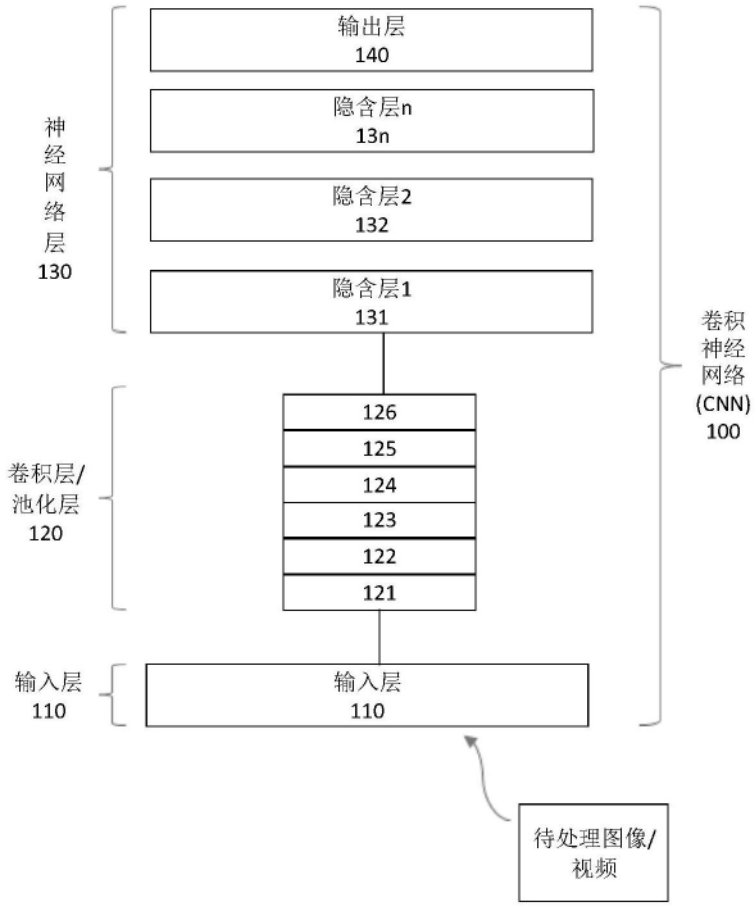

输出层
140
隐含层n
13n
隐含层2
132
隐含层1
131
神经网络层
130
126
125
124
123
122
121
卷积层/池化层
120
输入层
110
输入层
110
卷积神经网络(CNN)
100
待处理图像/视频

图2c

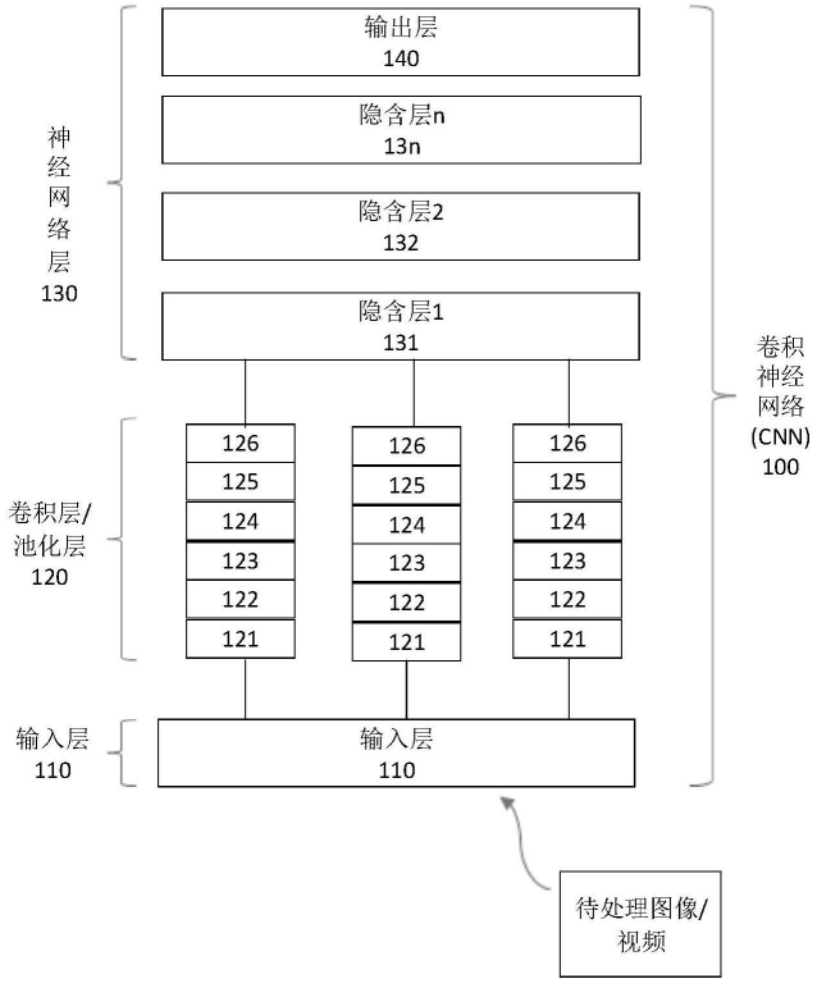

输出层
140
隐含层n
13n
隐含层2
132
隐含层1
131
神经网络层
130
卷积神经网络(CNN)
100
126
125
124
123
122
121
卷积层/池化层
120
输入层
110
输入层
110
待处理图像/视频

图2d

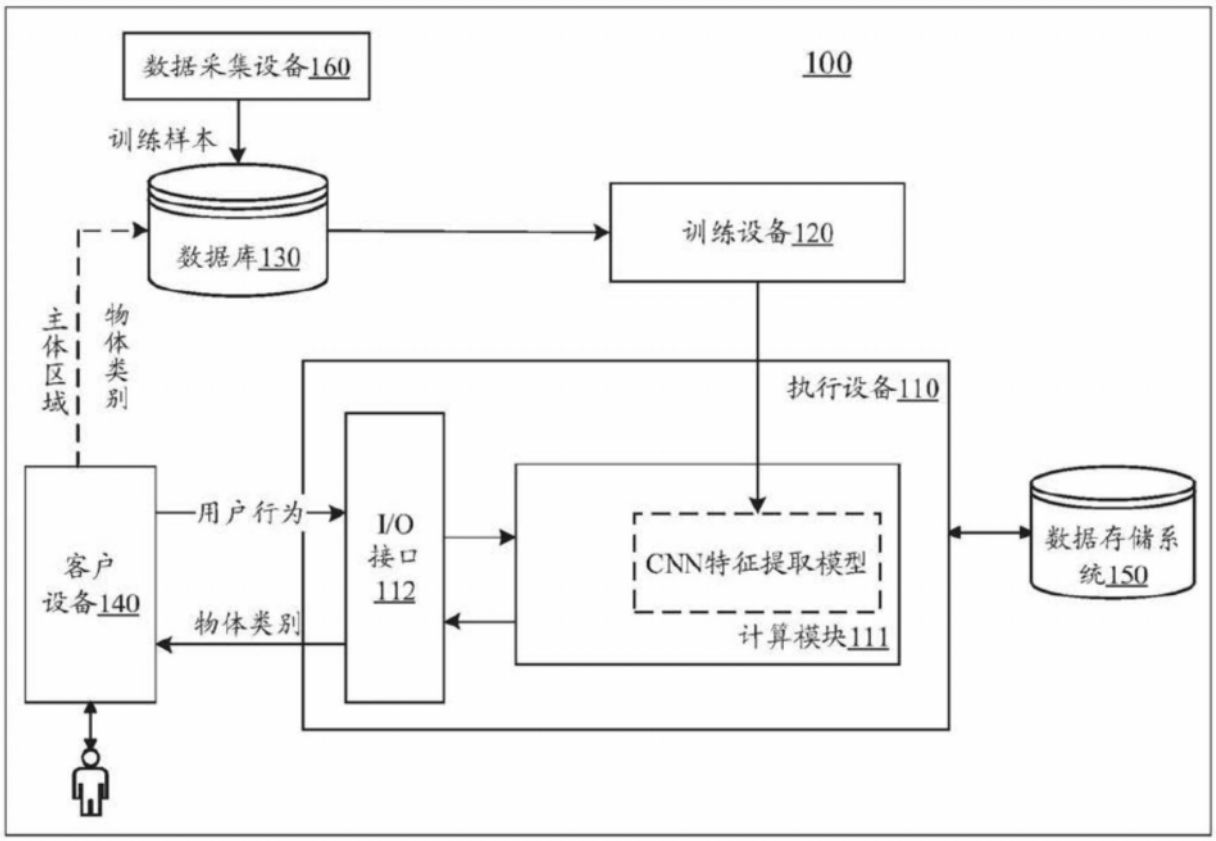

数据采集设备160
100
训练样本
数据库130
训练设备120
主体区域
物体类别
执行设备110
客户设备140
用户行为
I/O接口112
CNN特征提取模型
计算模块111
数据存储系统150
物体类别

图3

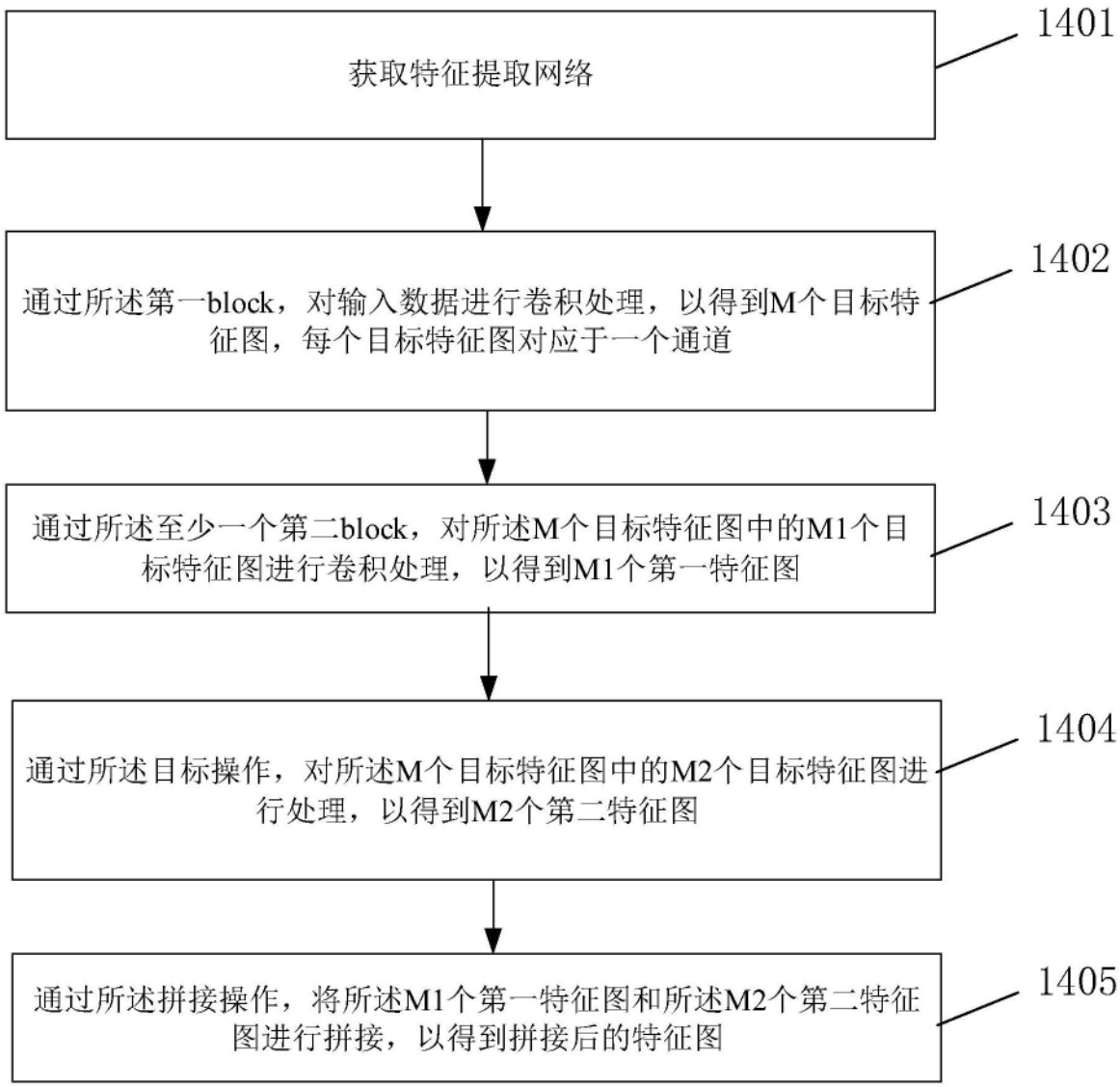

获取特征提取网络
1401
通过所述第一block，对输入数据进行卷积处理，以得到M个目标特征图，每个目标特征图对应于一个通道
1402
通过所述至少一个第二block，对所述M个目标特征图中的M1个目标特征图进行卷积处理，以得到M1个第一特征图
1403
通过所述目标操作，对所述M个目标特征图中的M2个目标特征图进行处理，以得到M2个第二特征图
1404
通过所述拼接操作，将所述M1个第一特征图和所述M2个第二特征图进行拼接，以得到拼接后的特征图
1405

图4

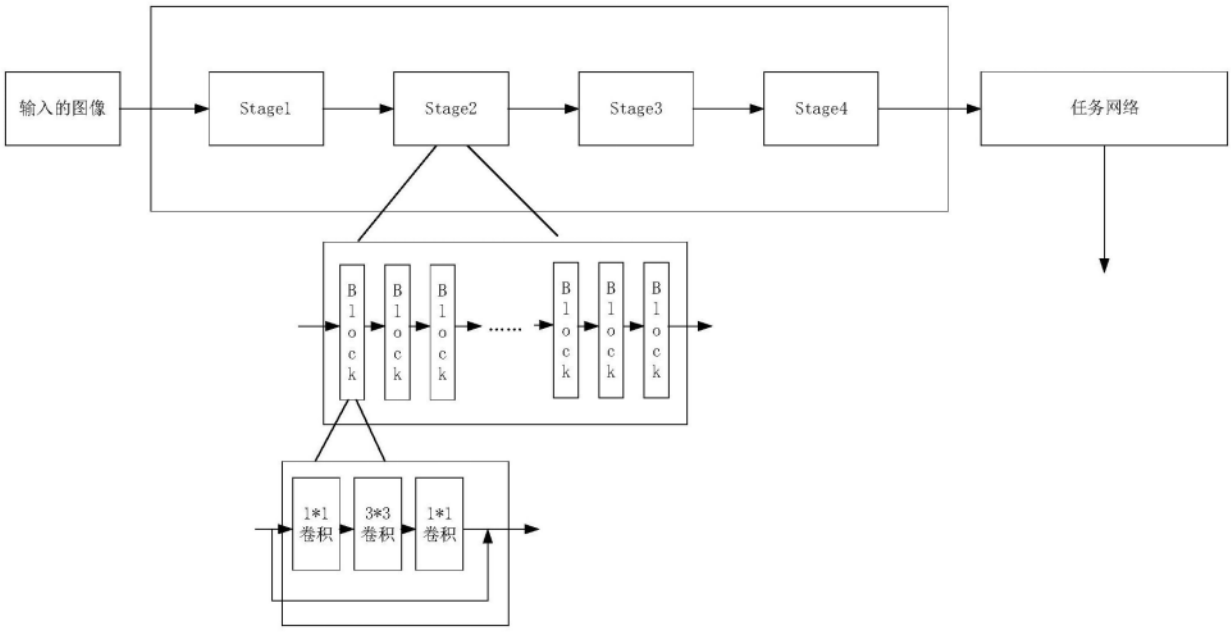

输入的图像
Stage1
Stage2
Stage3
Stage4
任务网络
Block
Block
Block
.....
Block
Block
Block
1*1
卷积
3*3
卷积
1*1
卷积

图5a

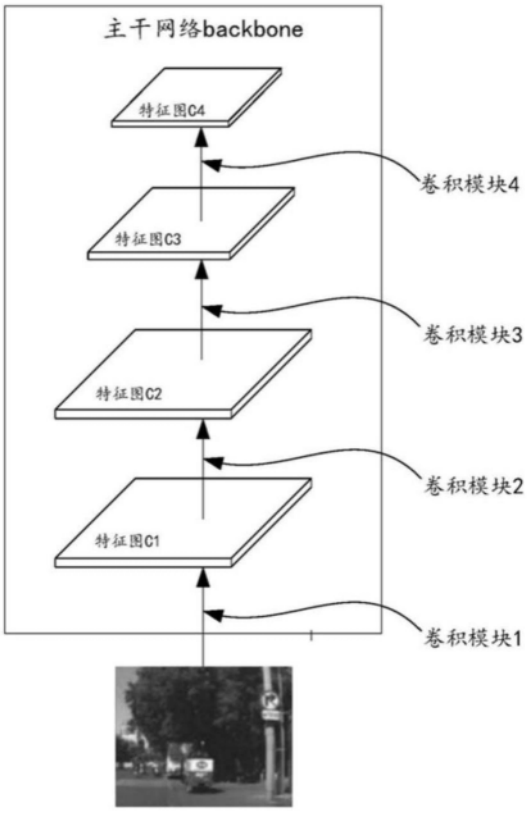

主干网络backbone
特征图C4
特征图C3
特征图C2
特征图C1
卷积模块4
卷积模块3
卷积模块2
卷积模块1

图5b

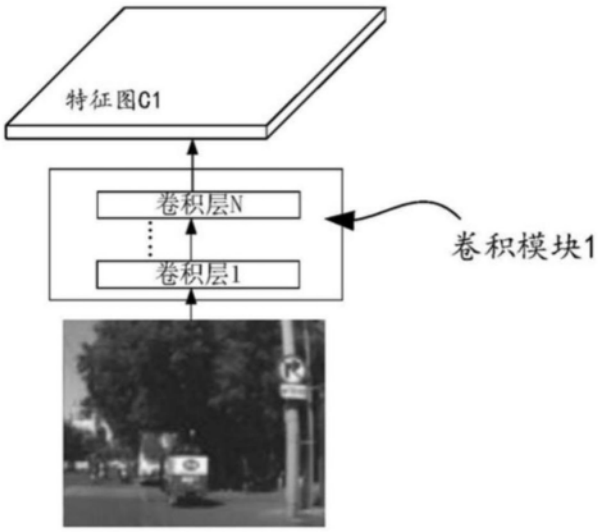

特征图C1
卷积层N
卷积层1
卷积模块1

图5c

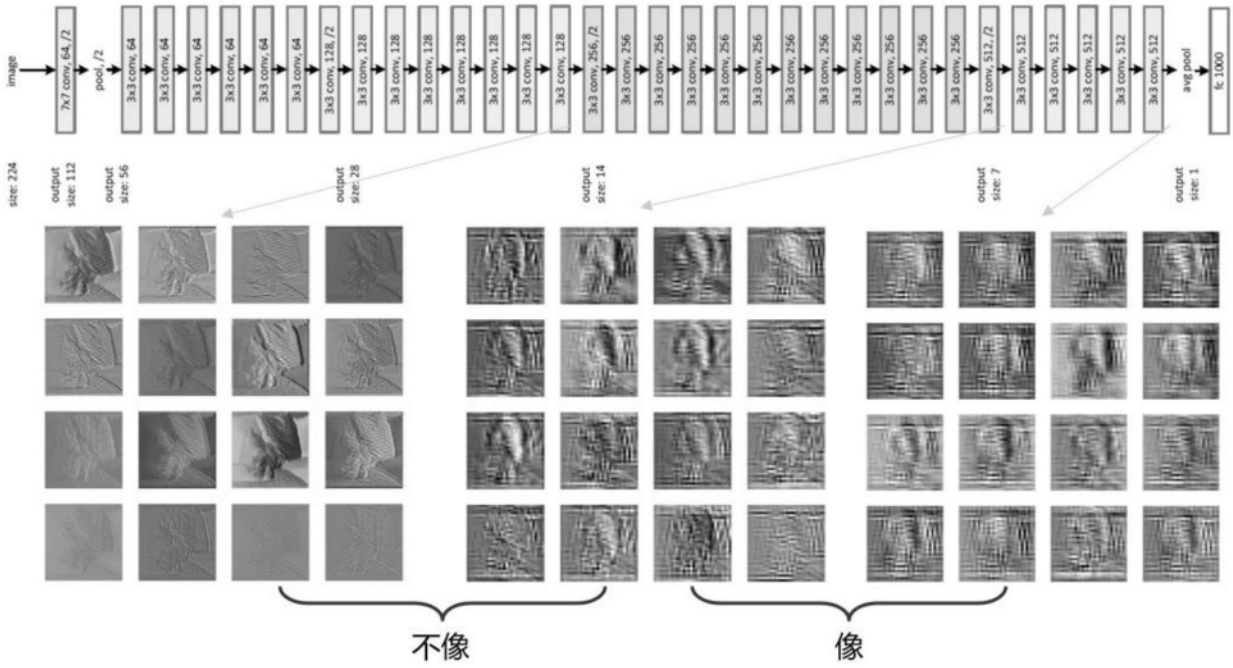

image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
avg pool
fc 1000
size: 224
output size: 112
output size: 56
output size: 28
output size: 14
output size: 7
output size: 1
不像
像

图5d

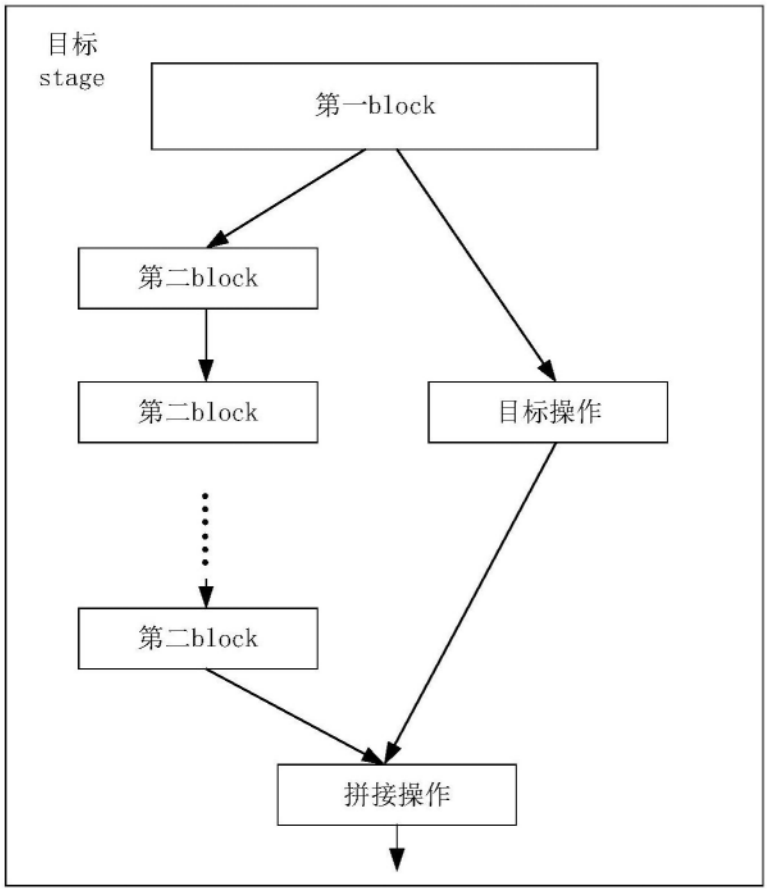


图6

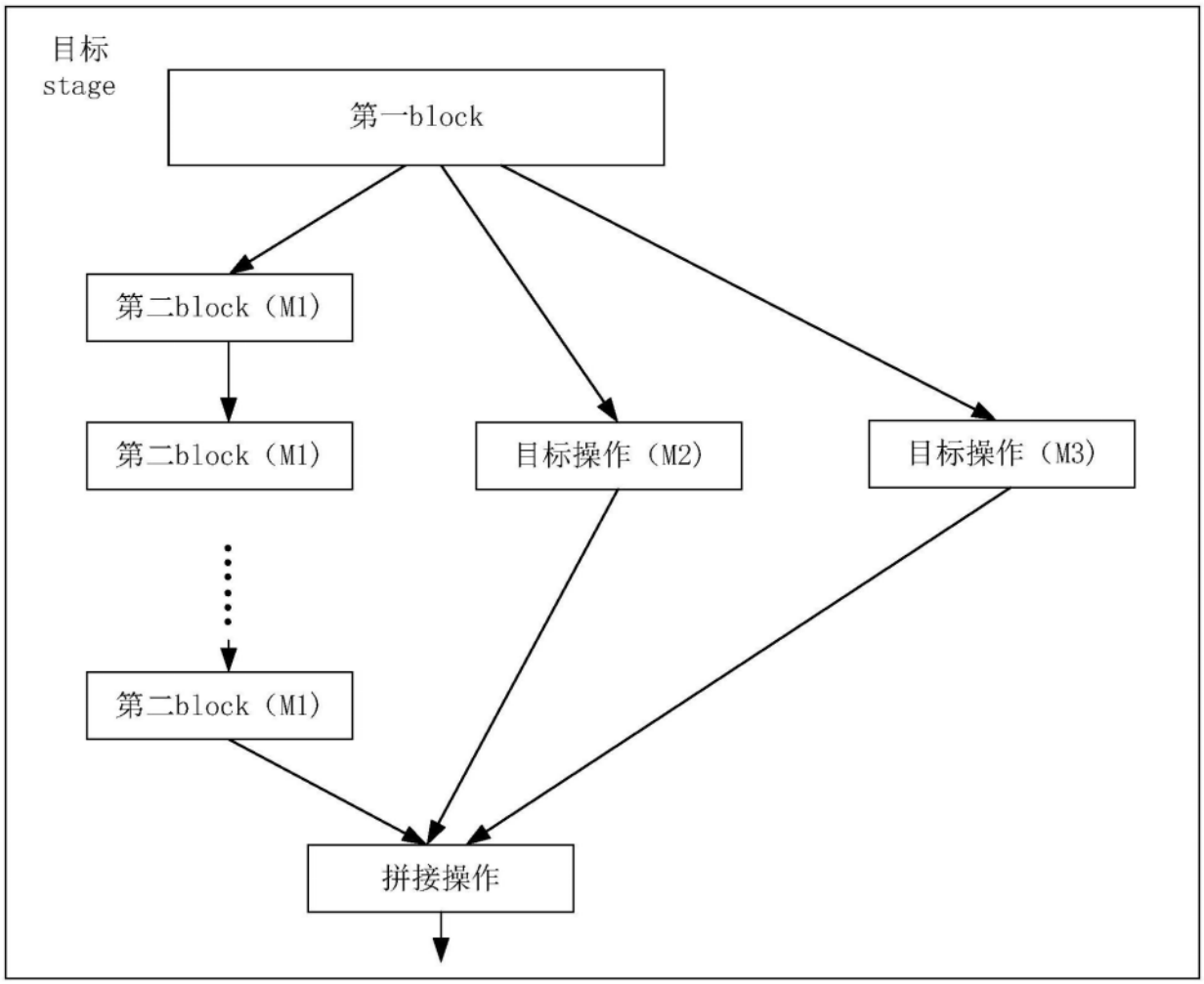

目标
stage
第一block
第二block（M1）
第二block（M1）
目标操作（M2）
目标操作（M3）
第二block（M1）
拼接操作

图7

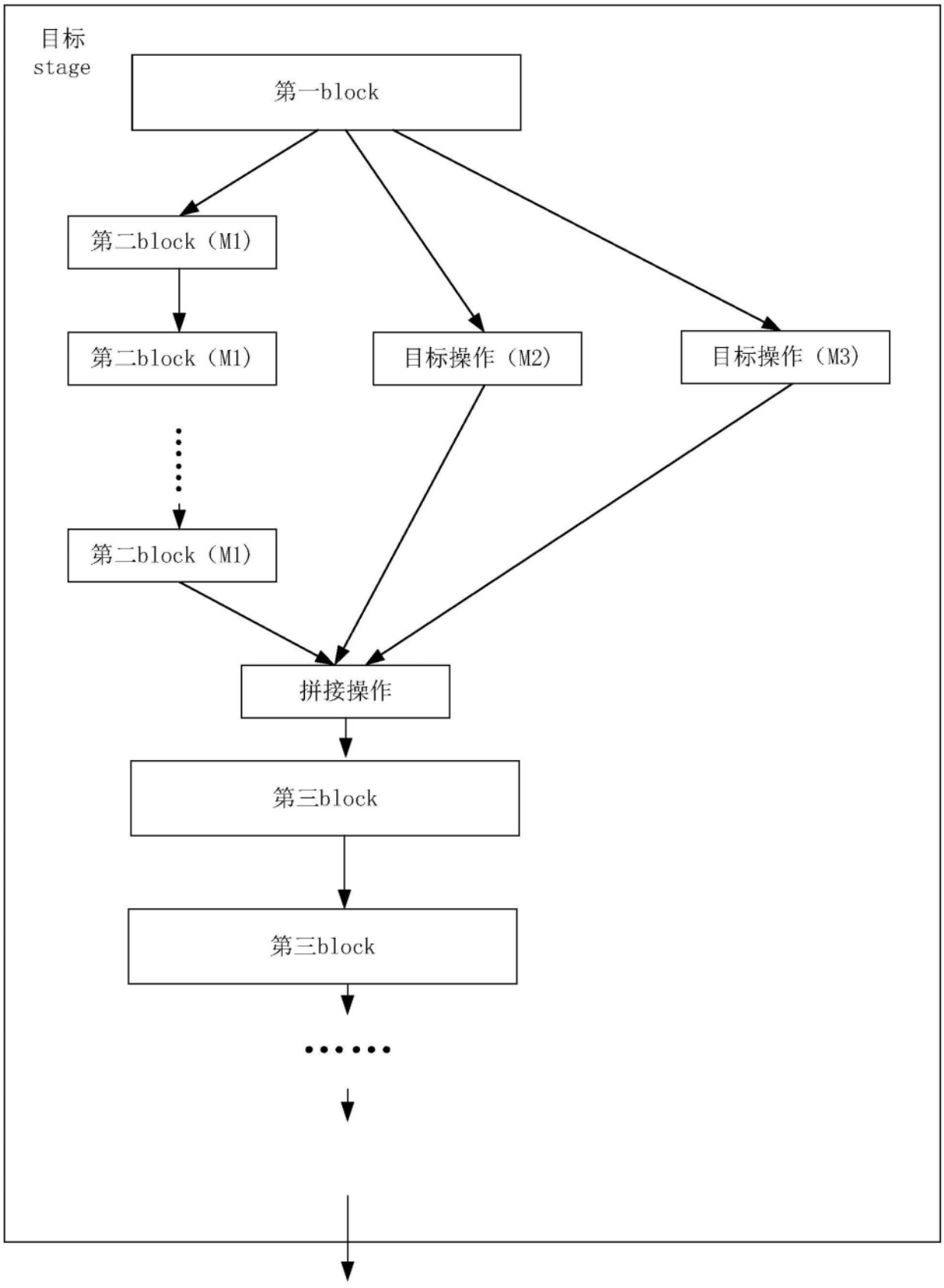

目标
stage
第一block
第二block（M1）
第二block（M1）
目标操作（M2）
目标操作（M3）
第二block（M1）
拼接操作
第三block
第三block

图8

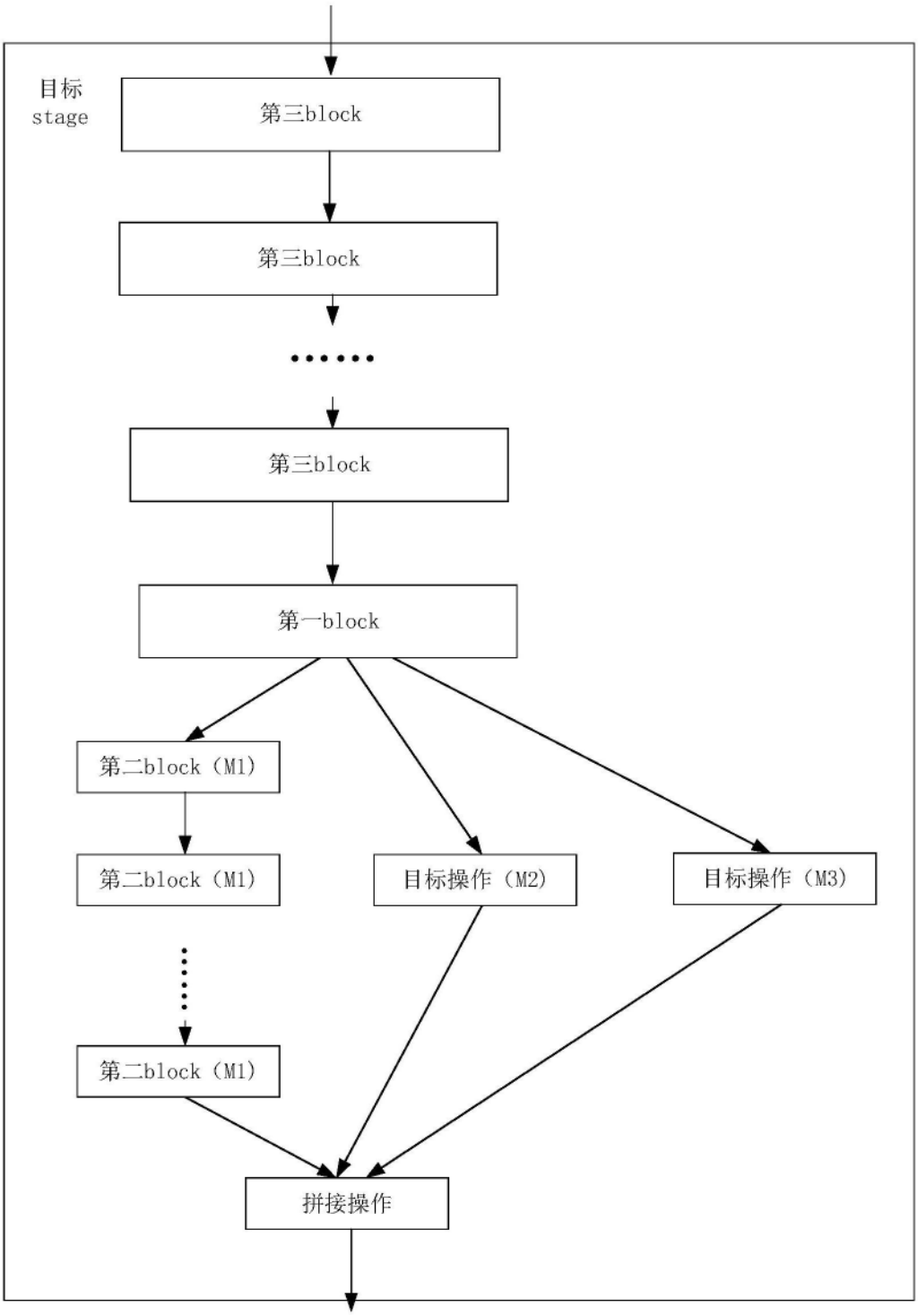

目标
stage
第三block
第三block
第三block
第一block
第二block（M1）
第二block（M1）
目标操作（M2）
目标操作（M3）
第二block（M1）
拼接操作

图9

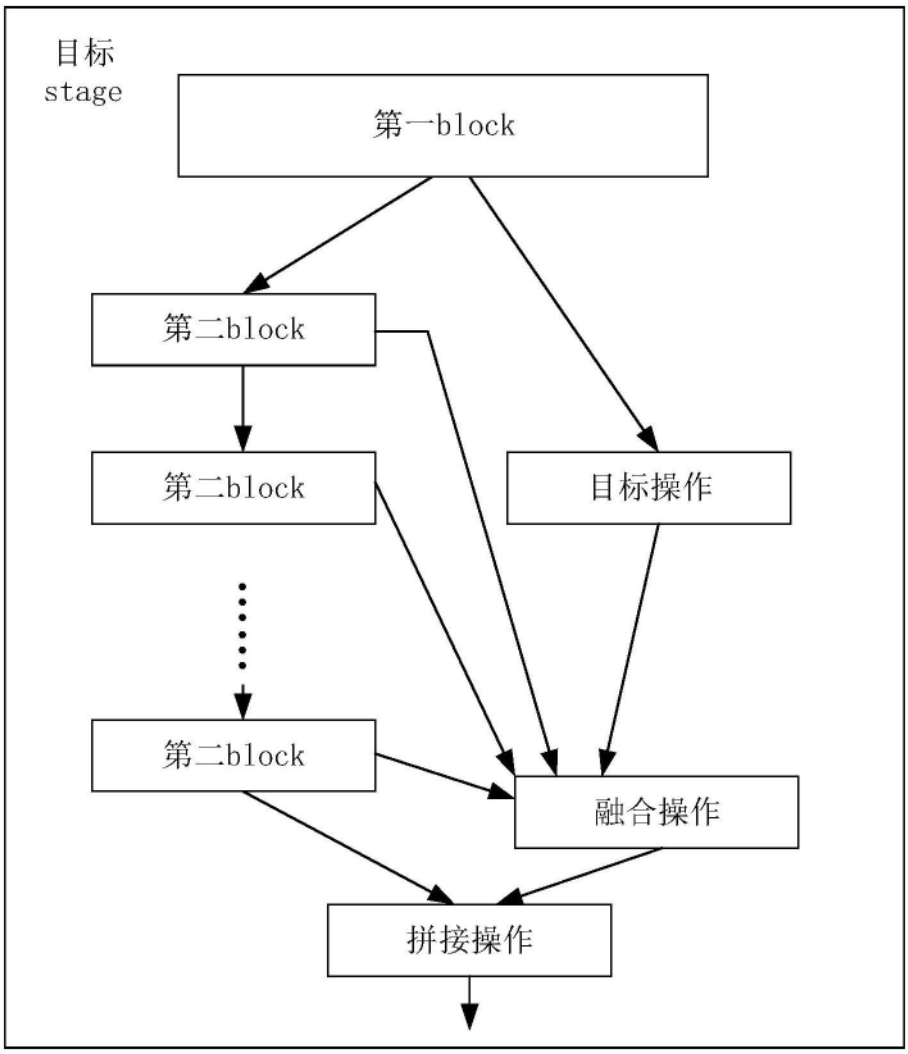


图10

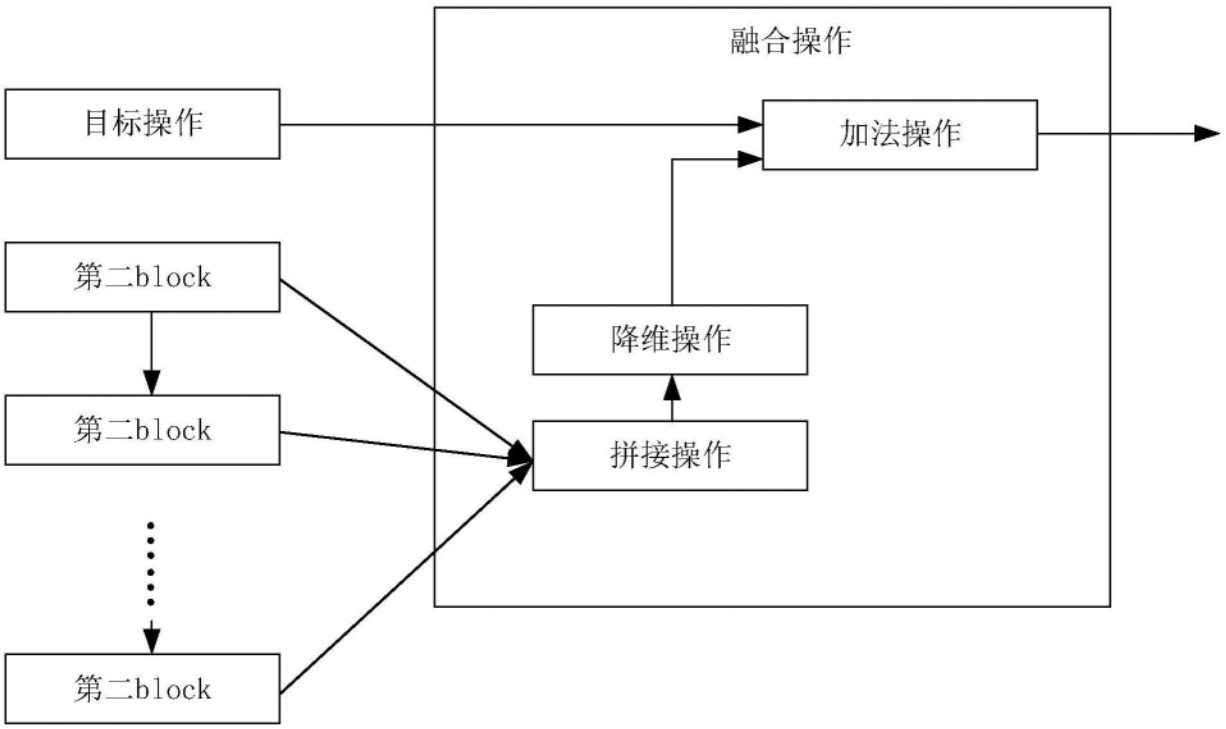

融合操作
目标操作
加法操作
第二block
降维操作
第二block
拼接操作
第二block

图11

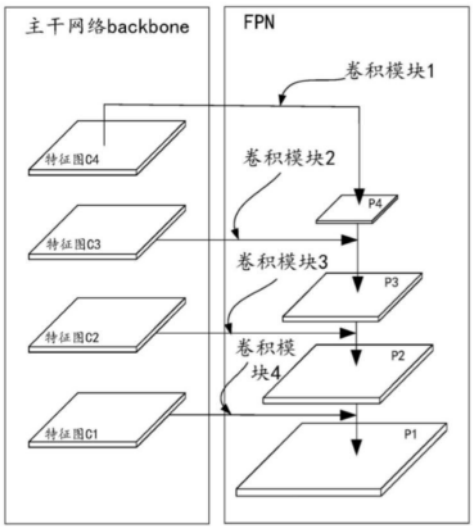

主干网络backbone
FPN
卷积模块1
特征图C4
卷积模块2
P4
特征图C3
卷积模块3
P3
特征图C2
卷积模块4
P2
特征图C1
P1

图12

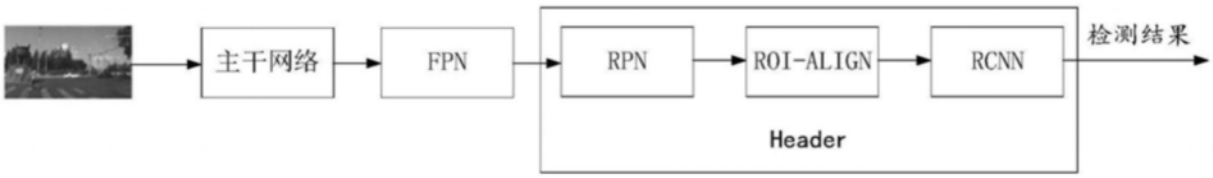

主干网络
FPN
RPN
ROI-ALIGN
RCNN
检测结果
Header

图13

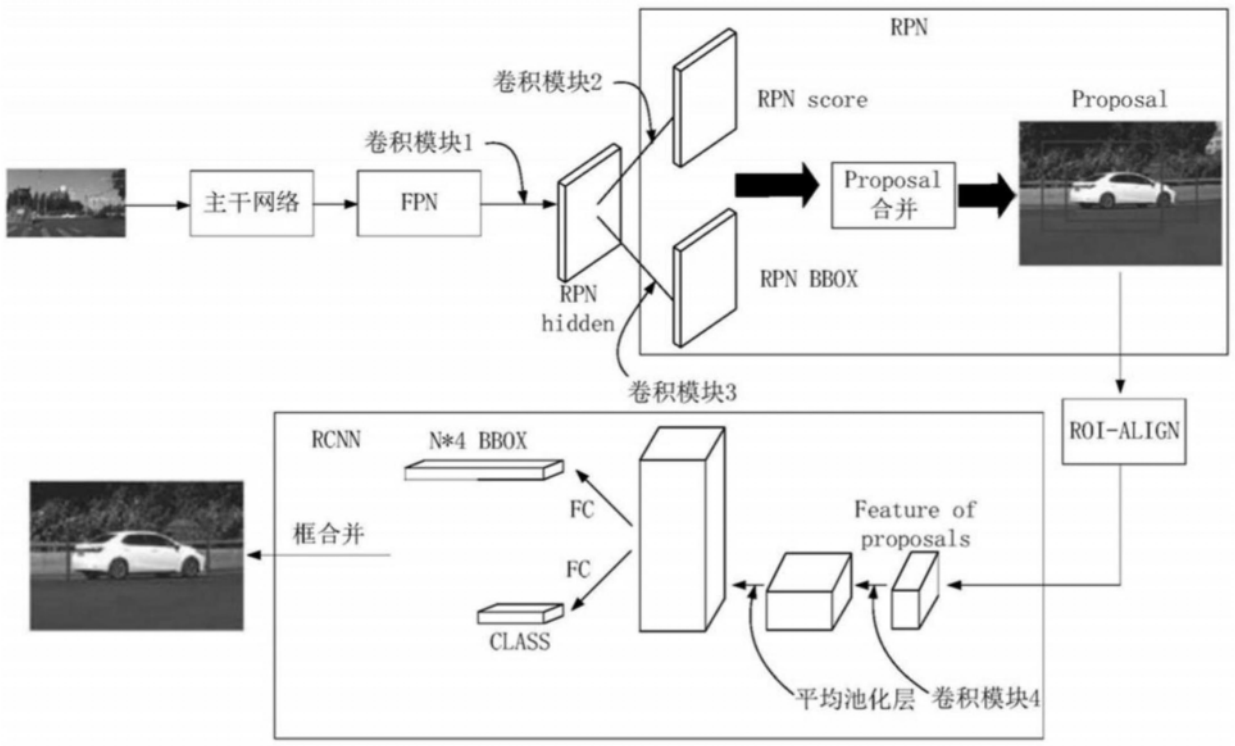

RPN
卷积模块2
RPN score
Proposal
卷积模块1
主干网络
FPN
Proposal
合并
RPN BBOX
RPN
hidden
卷积模块3
ROI-ALIGN
RCNN
N*4 BBOX
FC
Feature of
proposals
框合并
FC
CLASS
平均池化层
卷积模块4

图14

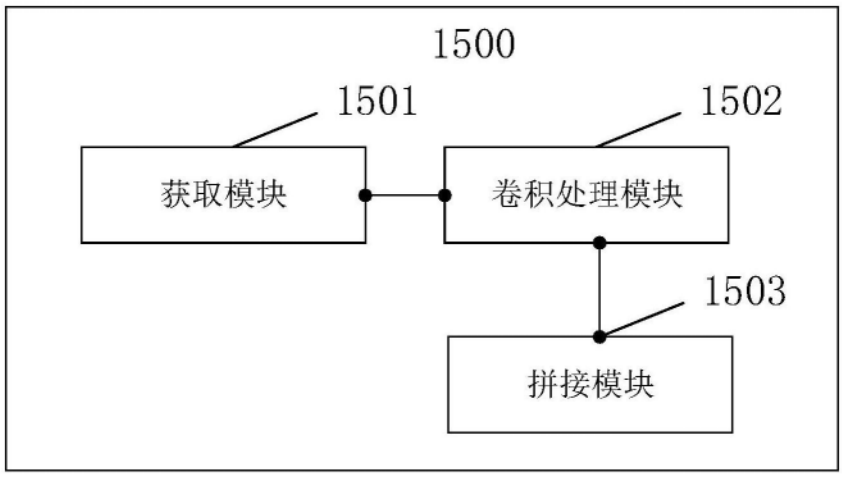

1500
1501
1502
获取模块
卷积处理模块
1503
拼接模块

图15

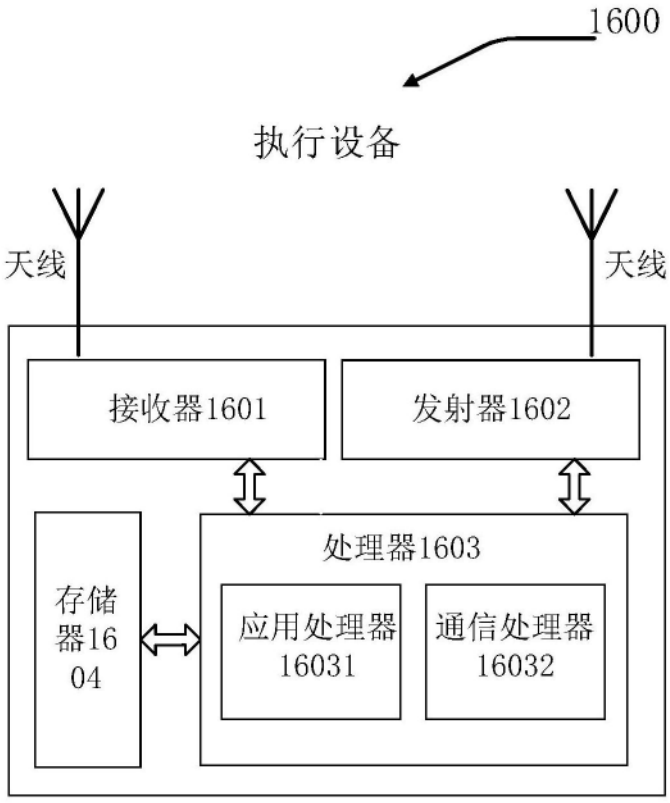

1600
执行设备
天线
天线
接收器1601
发射器1602
处理器1603
存储器1604
应用处理器16031
通信处理器16032

图16

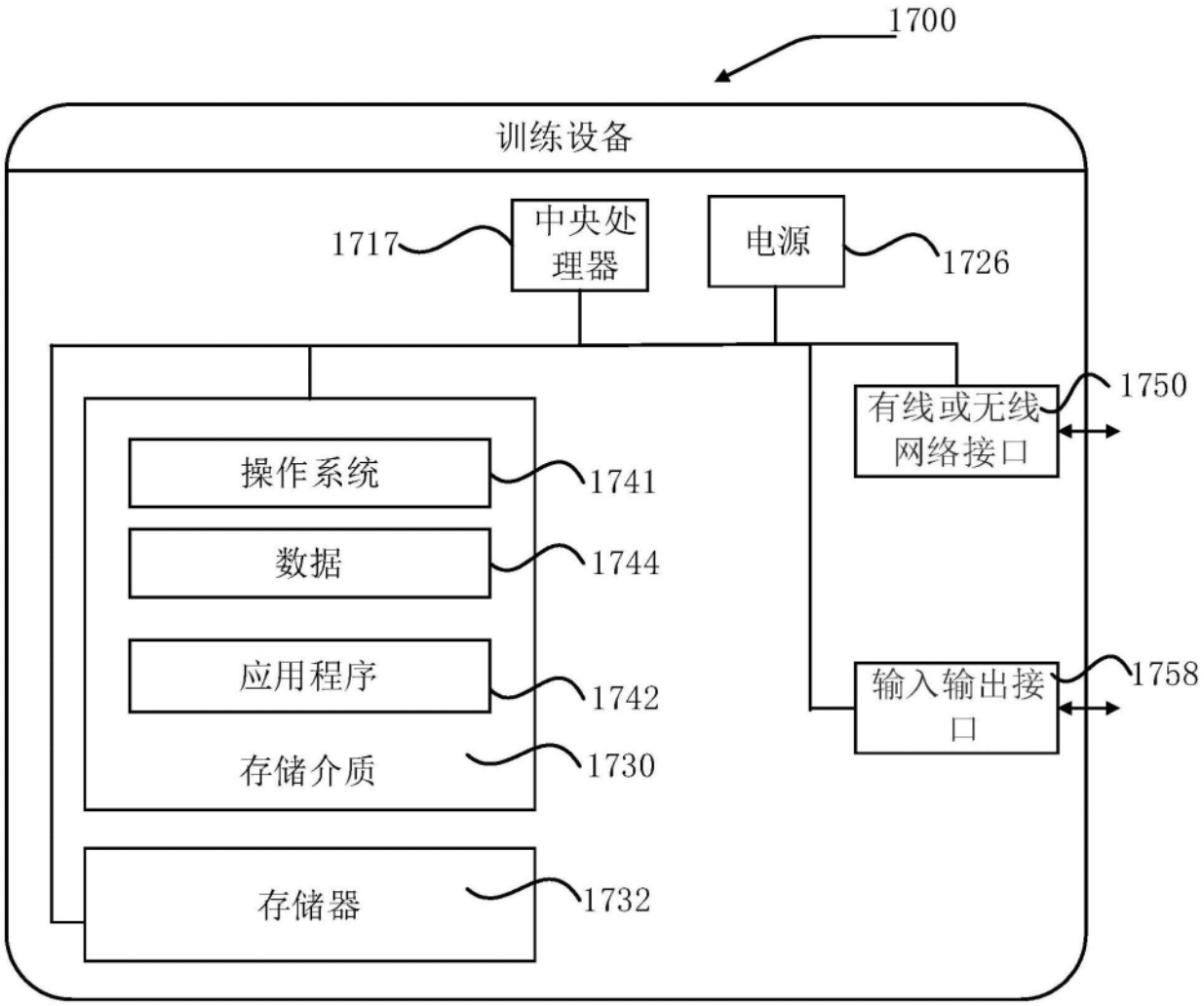

1700
训练设备
1717
中央处理器
电源
1726
有线或无线网络接口
1750
操作系统
1741
数据
1744
应用程序
1742
存储介质
1730
输入输出接口
1758
存储器
1732

图17

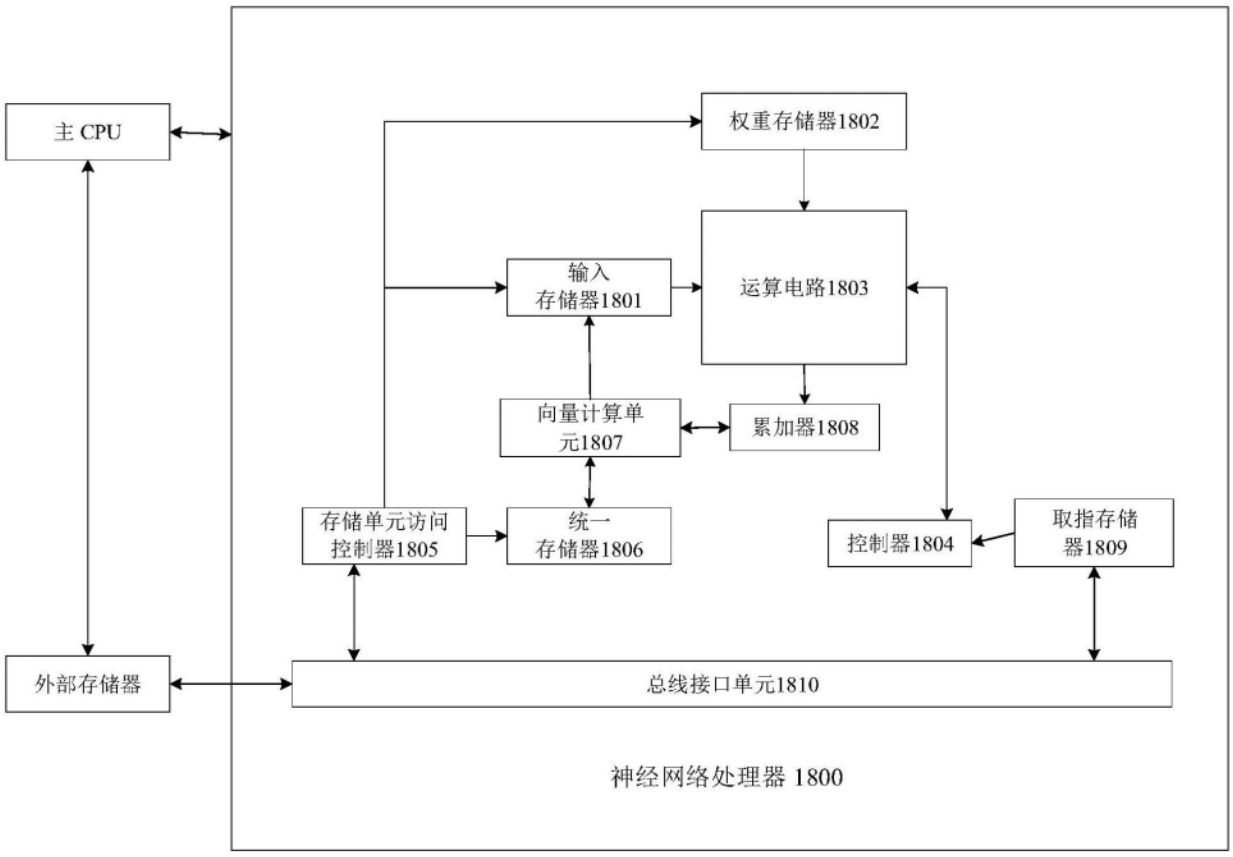

主 CPU
权重存储器1802
输入
存储器1801
运算电路1803
向量计算单
元1807
累加器1808
存储单元访问
控制器1805
统一
存储器1806
控制器1804
取指存储
器1809
外部存储器
总线接口单元1810
神经网络处理器 1800

图18