



(12) 发明专利

(10) 授权公告号 CN 111340177 B

(45) 授权公告日 2025. 02. 18

(21) 申请号 201910778533.5

G06N 3/084 (2023.01)

(22) 申请日 2019.08.22

(56) 对比文件

(65) 同一申请的已公布的文献号
申请公布号 CN 111340177 A

Sharan Narang et al.Mixed Precision Training.《arXiv: 1710.03740v1》.2017,第1-14页.

(43) 申请公布日 2020.06.26

Xianyan Jia et al.Highly Scalable Deep Learning Training System with Mixed-Precision:Training ImageNet in Four Minutes.《arXiv: 1807.11205v1》.2018,第1-9页.

(30) 优先权数据
10-2018-0165585 2018.12.19 KR

He Wen et al.Training Bit Fully Convolutional Network for Fast Semantic Segmentation.《arXiv: 1612.00212v1》.2016,第1-8页.

(73) 专利权人 三星电子株式会社
地址 韩国京畿道水原市

Sharan Narang et al.Mixed Precision Training.《arXiv: 1710.03740v1》.2017,第1-14页.

(72) 发明人 李曙炯 郭荣竣 孙辰雨 孙昌用
郑相一 崔昌圭 韩在濬

审查员 王闪

(74) 专利代理机构 北京铭硕知识产权代理有限公司 11286

专利代理师 刘灿强 张川绪

权利要求书5页 说明书21页 附图18页

(51) Int.Cl.

G06N 3/045 (2023.01)

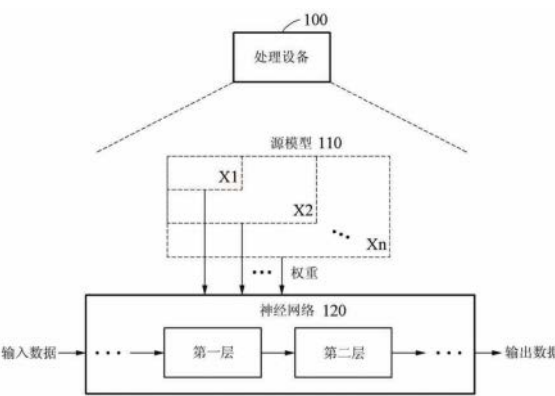
G06N 3/063 (2023.01)

(54) 发明名称

基于嵌套位表示的神经网络处理方法和设备

(57) 摘要

公开基于嵌套位表示的神经网络处理方法和设备。所述方法包括:确定神经网络的第一层的位宽;通过从与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重;以及通过基于获得的第二权重执行神经网络的第一层来处理神经网络的第一层的输入数据。



1. 一种神经网络处理方法,包括:
获得与神经网络的第一层对应的源模型的第一层的第一权重;
确定神经网络的第一层的位宽;
通过从与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重;以及
通过基于获得的第二权重执行神经网络的第一层来处理神经网络的第一层的输入数据,其中,神经网络的第一层的输入数据包括多媒体数据,多媒体数据包括文本、语音、动态图像和静态图像中的至少一个,
其中,源模型基于以下处理被训练:
通过量化源模型的第一层的高位宽的第一权重,确定与高位宽的第一权重对应的低位宽的权重,
通过将训练输入数据应用于源模型的第一层来确定与确定的低位宽的权重对应的损失值,以及
基于确定的损失值更新高位宽的第一权重,
其中,更新高位宽的第一权重的步骤包括:基于与确定的损失值对应的损失梯度的统计信息来更新高位宽的第一权重,
其中,更新高位宽的第一权重的步骤还包括:通过将高加权值分配给与低位宽的权重之中设置了高优先级的权重对应的损失梯度来计算统计信息。
2. 根据权利要求1所述的神经网络处理方法,其中,第一权重具有高于第二权重的位精度的位精度。
3. 根据权利要求1所述的神经网络处理方法,其中,第二权重嵌套在第一权重中。
4. 根据权利要求1所述的神经网络处理方法,其中,基于与神经网络的第一层对应的处理特性来确定神经网络的第一层的位宽,
其中,处理特性包括需要的处理速度、需要的处理准确度、处理难度和终端性能中的至少一个。
5. 根据权利要求1所述的神经网络处理方法,还包括:
确定神经网络的第二层的位宽;
获得与神经网络的第二层对应的源模型的第二层的第三权重;
通过从与神经网络的第二层对应的源模型的第二层的第三权重中的每个提取与确定的神经网络的第二层的位宽对应的至少一个位,获得神经网络的第二层的第四权重;以及
通过基于获得的第四权重执行神经网络的第二层来处理神经网络的第二层的输入数据,其中,第二层的输入数据对应于第一层的输出数据。
6. 根据权利要求5所述的神经网络处理方法,其中,第三权重具有高于第四权重的位精度的位精度。
7. 根据权利要求5所述的神经网络处理方法,其中,第四权重嵌套在第三权重中。
8. 根据权利要求5所述的神经网络处理方法,其中,基于第二权重执行的神经网络的第一层被配置为:基于第一层的输入数据处理与第一处理特性对应的第一任务,并且
基于第四权重执行的神经网络的第二层被配置为:基于第二层的输入数据处理与第一任务不同的与第二处理特性对应的第二任务。

9. 一种神经网络处理方法,包括:
获得第一神经网络的源模型的第一权重;
确定第一神经网络的位置;
通过从源模型的第一权重中的每个提取与确定的位置对应的至少一个位,获得第一神经网络的第二权重;以及
通过基于获得的第二权重执行第一神经网络来处理第一神经网络的输入数据,其中,第一神经网络的输入数据包括多媒体数据,多媒体数据包括文本、语音、动态图像和静态图像中的至少一个,
其中,源模型基于以下处理被训练:
通过量化源模型的高位宽的第一权重,确定与高位宽的第一权重对应的低位宽的权重,
通过将训练输入数据应用于源模型来确定与确定的低位宽的权重对应的损失值,以及基于确定的损失值更新高位宽的第一权重,
其中,更新高位宽的第一权重的步骤包括:基于与确定的损失值对应的损失梯度的统计信息来更新高位宽的第一权重,
其中,更新高位宽的第一权重的步骤还包括:通过将高加权值分配给与低位宽的权重之中设置了高优先级的权重对应的损失梯度来计算统计信息。
10. 根据权利要求9所述的神经网络处理方法,其中,第一权重具有高于第二权重的位精度的位精度。
11. 根据权利要求9所述的神经网络处理方法,其中,第二权重嵌套在第一权重中。
12. 根据权利要求9所述的神经网络处理方法,还包括:
基于由第一神经网络执行的输入数据的结果确定第二神经网络的位置;
通过从第一权重中的每个提取与确定的第二神经网络的位置对应的至少一个位,获得第二神经网络的第三权重;以及
通过基于获得的第三权重执行第二神经网络来处理第二神经网络的输入数据。
13. 根据权利要求12所述的神经网络处理方法,其中,第一权重具有高于第三权重的位精度的位精度。
14. 根据权利要求12所述的神经网络处理方法,其中,第二权重和第三权重嵌套在第一权重中。
15. 根据权利要求12所述的神经网络处理方法,其中,
基于第二权重执行的第一神经网络被配置为:基于第一神经网络的输入数据来处理与第一处理特性对应的第一任务,并且
基于第三权重执行的第二神经网络被配置为:基于第二神经网络的输入数据处理与第一任务不同的与第二处理特性对应的第二任务。
16. 根据权利要求9所述的神经网络处理方法,其中,基于与第一神经网络对应的处理特性来确定第一神经网络的位置,
其中,处理特性包括需要的处理速度、需要的处理准确度、处理难度和终端性能中的至少一个。
17. 一种神经网络训练方法,包括:

通过量化与神经网络的第一层对应的高位宽的权重,确定与神经网络的第一层对应的低位宽的权重;

通过将输入数据应用于第一层来确定与确定的低位宽的权重对应的损失值,其中,输入数据包括多媒体数据,多媒体数据包括文本、语音、动态图像和静态图像中的至少一个;以及

基于确定的损失值更新高位宽的权重,

其中,更新高位宽的权重的步骤包括:基于与确定的损失值对应的损失梯度的统计信息来更新高位宽的权重,

其中,更新高位宽的权重的步骤还包括:通过将高加权值分配给与低位宽的权重之中设置了高优先级的权重对应的损失梯度来计算统计信息。

18. 根据权利要求17所述的神经网络训练方法,还包括:

在与高位宽的权重相关联的训练完成之后,通过量化高位宽的权重来确定与第一层对应的低位宽的权重集。

19. 根据权利要求18所述的神经网络训练方法,其中,低位宽的权重集包括第一位宽的权重集和第二位宽的权重集,第二位宽的权重集具有低于第一位宽的权重集的位精度的位精度,

其中,第二位宽的权重集嵌套在第一位宽的权重集中。

20. 根据权利要求17所述的神经网络训练方法,其中,低位宽的权重包括第一位宽的第一权重和第二位宽的第二权重,其中,第一位宽的第一权重具有低于高位宽的权重的位精度的位精度,第二位宽的第二权重具有低于第一位宽的第一权重的位精度的位精度。

21. 根据权利要求20所述的神经网络训练方法,其中,确定低位宽的权重的步骤包括:

通过量化高位宽的权重来确定第一位宽的第一权重;以及

通过从确定的第一位宽的第一权重中的每个提取至少一个位,确定第二位宽的第二权重。

22. 根据权利要求20所述的神经网络训练方法,其中,确定低位宽的权重的步骤包括:

通过量化高位宽的权重来确定第二位宽的第二权重;

将确定的第二位宽的第二权重确定为第一位宽的第一权重的高位组;以及

通过量化高位宽的权重来确定第一位宽的第一权重的低位组。

23. 一种神经网络处理设备,包括:

处理器;以及

存储器,被配置为存储能够由处理器读取的指令,

其中,当所述指令被处理器执行时,处理器被配置为:

获得与神经网络的第一层对应的源模型的第一层的第一权重;

确定神经网络的第一层的位宽;

通过从与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重;以及

通过基于获得的第二权重执行神经网络的第一层来处理神经网络的第一层的输入数据,其中,神经网络的第一层的输入数据包括多媒体数据,多媒体数据包括文本、语音、动态图像和静态图像中的至少一个,

其中,源模型基于以下处理被训练:

通过量化源模型的第一层的高位宽的第一权重,确定与高位宽的第一权重对应的低位宽的权重,

通过将训练输入数据应用于源模型的第一层来确定与确定的低位宽的权重对应的损失值,以及

基于确定的损失值更新高位宽的第一权重,

其中,更新高位宽的第一权重的步骤包括:基于与确定的损失值对应的损失梯度的统计信息来更新高位宽的第一权重,

其中,更新高位宽的第一权重的步骤还包括:通过将高加权值分配给与低位宽的权重之中设置了高优先级的权重对应的损失梯度来计算统计信息。

24. 根据权利要求23所述的神经网络处理设备,其中,处理器被配置为:

确定神经网络的第二层的位宽;

获得与神经网络的第二层对应的源模型的第二层的第三权重;

通过从与神经网络的第二层对应的源模型的第二层的第三权重中的每个提取与确定的神经网络的第二层的位宽对应的至少一个位,获得神经网络的第二层的第四权重;以及

通过基于获得的第四权重执行神经网络的第二层来处理神经网络的第二层的输入数据,其中,第二层的输入数据对应于第一层的输出数据。

25. 一种识别对象的设备,包括:

处理器;以及

存储器,被配置为存储能够由处理器读取的指令,

其中,当所述指令由处理器执行时,处理器被配置为:

获得第一神经网络的源模型的第一权重;

确定第一神经网络的位宽;

通过从源模型的第一权重中的每个提取与确定的位宽对应的至少一个位,获得第一神经网络的第二权重;以及

通过基于获得的第二权重执行第一神经网络来处理第一神经网络的输入数据,其中,第一神经网络的输入数据包括多媒体数据,多媒体数据包括文本、语音、动态图像和静态图像中的至少一个,

其中,源模型基于以下处理被训练:

通过量化源模型的高位宽的第一权重,确定与高位宽的第一权重对应的低位宽的权重,

通过将训练输入数据应用于源模型来确定与确定的低位宽的权重对应的损失值,以及基于确定的损失值更新高位宽的第一权重,

其中,更新高位宽的第一权重的步骤包括:基于与确定的损失值对应的损失梯度的统计信息来更新高位宽的第一权重,

其中,更新高位宽的第一权重的步骤还包括:通过将高加权值分配给与低位宽的权重之中设置了高优先级的权重对应的损失梯度来计算统计信息。

26. 根据权利要求25所述的设备,其中,处理器被配置为:

基于由第一神经网络处理输入数据的结果确定第二神经网络的位宽;

通过从第一权重中的每个提取与确定的第二神经网络的位宽对应的至少一个位,获得第二神经网络的第三权重;以及

通过基于获得的第三权重执行第二神经网络来处理第二神经网络的输入数据。

27. 一种神经网络训练设备,包括:

处理器;以及

存储器,被配置为存储能够由处理器读取的指令,

其中,当所述指令被处理器执行时,处理器被配置为:

通过量化与神经网络的第一层对应的高位宽的权重,确定与神经网络的第一层对应的低位宽的权重;

通过将输入数据应用于第一层,确定与确定的低位宽的权重对应的损失值,其中,输入数据包括多媒体数据,多媒体数据包括文本、语音、动态图像和静态图像中的至少一个;以及

基于确定的损失值更新高位宽的权重,

其中,更新高位宽的权重的步骤包括:基于与确定的损失值对应的损失梯度的统计信息来更新高位宽的权重,

其中,更新高位宽的权重的步骤还包括:通过将高加权值分配给与低位宽的权重之中设置了高优先级的权重对应的损失梯度来计算统计信息。

28. 根据权利要求27所述的神经网络训练设备,其中,低位宽的权重包括第一位宽的第一权重和第二位宽的第二权重,其中,第一位宽的第一权重具有低于高位宽的权重的位精度的位精度,第二位宽的第二权重具有低于第一位宽的第一权重的位精度的位精度。

29. 根据权利要求28所述的神经网络训练设备,其中,处理器被配置为:

通过量化高位宽的权重来确定第一位宽的第一权重;以及

通过从确定的第一位宽的第一权重中的每个提取至少一个位,确定第二位宽的第二权重。

30. 根据权利要求28所述的神经网络训练设备,其中,处理器被配置为:

通过量化高位宽的权重来确定第二位宽的第二权重;

将确定的第二位宽的第二权重确定为第一位宽的第一权重的高位组;以及

通过量化高位宽的权重来确定第一位宽的第一权重的低位组。

基于嵌套位表示的神经网络处理方法和设备

[0001] 本申请要求于2018年12月19日提交到韩国知识产权局的第10-2018-0165585号韩国专利申请的权益,所述韩国专利申请的全部公开出于所有目的通过引用包含于此。

技术领域

[0002] 以下描述涉及一种基于嵌套位表示的神经网络处理方法和设备。

背景技术

[0003] 例如,识别的技术自动化已经通过处理器实现的神经网络模型来实现,其中,处理器实现的神经网络模型作为专用计算架构在大量训练之后可提供输入模式与输出模式之间的计算上直观的映射。产生这样的映射的训练能力可称为神经网络的学习能力。此外,由于专门的训练,这样的经过专门训练的神经网络因此可具有关于例如神经网络可能尚未被训练的输入模式生成相对准确的输出的泛化能力。

[0004] 在神经网络用于对象识别(诸如,图像识别或语音识别)的情况下,神经网络中的权重会占用较大的存储器空间,这使得存储器效率被降低。

发明内容

[0005] 提供本发明内容,以按照简化的形式介绍下面在具体实施方式中进一步描述的构思的选择。本发明内容不意图确定要求权利的主题的关键特征或必要特征,也不意图作为帮助确定要求权利的主题的范围而被使用。

[0006] 在一个总体方面,一种神经网络处理方法包括:确定神经网络的第一层的位宽;通过与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重;以及通过基于获得的第二权重执行神经网络的第一层来处理神经网络的第一层的输入数据。

[0007] 第一权重可具有高于第二权重的位精度。第二权重可嵌套在第一权重中。

[0008] 可基于与神经网络的第一层对应的处理特性来确定神经网络的第一层的位宽。处理特性可包括需要的处理速度、需要的处理准确度、处理难度或终端性能中的至少一个。

[0009] 所述处理方法还可包括:确定神经网络的第二层的位宽;通过从与神经网络的第二层对应的源模型的第二层的第三权重中的每个提取与确定的神经网络的第二层的位宽对应的至少一个位,获得神经网络的第二层的第四权重;以及通过基于获得的第四权重执行神经网络的第二层来处理神经网络的第二层的输入数据。

[0010] 第三权重可具有高于第四权重的位精度。第四权重可嵌套在第三权重中。

[0011] 基于第二权重执行的神经网络的第一层可被配置为:基于第一层的输入数据处理第一任务,以及基于第四权重执行的神经网络的第二层可被配置为:基于第二层的输入数据处理与第一任务不同的第二任务。

[0012] 在另一总体方面,一种神经网络处理方法包括:确定第一神经网络的位宽;通过从源模型的第一权重中的每个提取与确定的位宽对应的至少一个位,获得第一神经网络的第

二权重;以及通过基于获得的第二权重执行第一神经网络来处理第一神经网络的输入数据。

[0013] 所述处理方法还可包括:基于由第一神经网络处理输入数据的结果确定第二神经网络的位宽;通过从第一权重中的每个提取与确定的第二神经网络的位宽对应的至少一个位,获得第二神经网络的第三权重;以及通过基于获得的第三权重执行第二神经网络来处理第二神经网络的输入数据。

[0014] 基于第二权重执行的第一神经网络可被配置为:基于第一神经网络的输入数据来处理第一任务,以及基于第三权重执行的第二神经网络可被配置为:基于第二神经网络的输入数据处理与第一任务不同的第二任务。

[0015] 在另一总体方面,一种神经网络训练方法包括:通过量化与神经网络的第一层对应的高位宽的权重,确定与神经网络的第一层对应的低位宽的权重;通过将输入数据应用于第一层来确定与确定的低位宽的权重对应的损失值;以及基于确定的损失值更新高位宽的权重。

[0016] 所述训练方法还可包括:在与高位宽的权重相关联的训练完成之后,通过量化高位宽的权重来确定与第一层对应的低位宽的权重集。

[0017] 低位宽的权重集可包括第一位宽的权重集和第二位宽的权重集,第二位宽的权重集具有低于第一位宽的权重集的位精度。第二位宽的权重集可嵌套在第一位宽的权重集中。

[0018] 低位宽的权重可包括第一位宽的第一权重和第二位宽的第二权重,其中,第一位宽的第一权重具有低于高位宽的权重的位精度,第二位宽的第二权重具有低于第一位宽的第一权重的位精度。

[0019] 确定低位宽的权重的步骤可包括:通过量化高位宽的权重来确定第一位宽的第一权重;以及通过从确定的第一位宽的第一权重中的每个提取至少一个位,确定第二位宽的第二权重。

[0020] 确定低位宽的权重的步骤可包括:通过量化高位宽的权重来确定第二位宽的第二权重;将确定的第二位宽的第二权重确定为第一位宽的第一权重的高位组;以及通过量化高位宽的权重来确定第一位宽的第一权重的低位组。

[0021] 更新高位宽的权重的步骤可包括:基于与确定的损失值对应的损失梯度的统计信息来更新高位宽的权重。更新高位宽的权重的步骤还可包括:通过将高加权值分配给与低位宽的权重之中设置了高优先级的权重对应的损失梯度来计算统计信息。

[0022] 在另一总体方面,一种神经网络处理设备包括:处理器;以及存储器,被配置为存储能够由处理器读取的指令。当所述指令被处理器执行时,处理器可被配置为:确定神经网络的第一层的位宽;通过从与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重;以及通过基于获得的第二权重执行神经网络的第一层来处理神经网络的第一层的输入数据。

[0023] 在一个进一步的总体方面,一种神经网络处理设备包括:处理器;以及存储器,被配置为存储能够由处理器读取的指令。当所述指令由处理器执行时,处理器可被配置为:确定第一神经网络的位宽;通过从源模型的第一权重中的每个提取与确定的位宽对应的至少一个位,获得第一神经网络的第二权重;以及通过基于获得的第二权重执行第一神经网络

来处理第一神经网络的输入数据。

[0024] 在另一总体方面,一种神经网络训练设备包括:处理器;以及存储器,被配置为存储能够由处理器读取的指令。当所述指令被处理器执行时,处理器可被配置为:通过量化与神经网络的第一层对应的高位宽的权重,确定与神经网络的第一层对应的低位宽的权重;通过将输入数据应用于第一层,确定与确定的低位宽的权重对应的损失值;以及基于确定的损失值更新高位宽的权重。

[0025] 在另一总体方面,一种识别对象的方法,包括:接收多媒体数据作为神经网络的第一层的输入数据;基于第一任务资源需求确定神经网络的第一层的位宽,其中,第一层包括至少一个层;通过从与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重;以及通过基于获得的第二权重执行神经网络的第一层来对神经网络的第一层的输入数据执行对象识别。

[0026] 第一权重可具有高于第二权重的位精度的位精度。

[0027] 第二权重可嵌套在第一权重中。

[0028] 第一任务资源需求越高,将神经网络的第一层的位宽确定得越高。

[0029] 第一任务资源需求包括需要的第一处理速度、需要的第一处理准确度、第一处理难度、输入数据的第一质量和第一终端性能中的至少一个。

[0030] 确定神经网络的第一层的位宽的步骤可包括:输入数据的第一质量越高,将神经网络的第一层的位宽确定得越高。

[0031] 输入数据可以是包括第一级输入数据和第二级输入数据的多级多媒体数据,第一级输入数据的第一质量低于第二级输入数据的第一质量。执行对象识别的步骤可包括:基于第一级输入数据将神经网络的第一层的位宽确定为第一位宽,并将第一级输入数据应用于基于第一位宽权重执行的神经网络的第一层,以确定第一级输入数据中是否存在预定对象;当确定第一级输入数据中存在预定对象时,将神经网络的第一层的位宽确定为高于第一位宽的第二位宽,并将第二级输入数据应用于基于第二位宽权重执行的神经网络的第一层,以确定预定对象的位置信息和/或对预定对象进行验证。

[0032] 多媒体数据可以是图像数据,图像数据的分辨率越高,第一质量越高。

[0033] 所述方法还可包括:基于第二任务资源需求确定神经网络的第二层的位宽,其中,第二层包括至少一个层;通过从与神经网络的第二层对应的源模型的第二层的第三权重中的每个提取与确定的神经网络的第二层的位宽对应的至少一个位,获得神经网络的第二层的第四权重;以及通过基于获得的第四权重执行神经网络的第二层来对神经网络的第二层的输入数据执行对象识别,其中,第二层的输入数据对应于第一层的输出数据,其中,第二任务资源需求包括需要的第二处理速度、需要的第二处理准确度、第二处理难度、输入数据的第二质量和第二终端性能中的至少一个,其中,第二任务资源需求与第一任务资源需求相同或不同。

[0034] 第三权重可具有高于第四权重的位精度的位精度。

[0035] 第四权重可嵌套在第三权重中。

[0036] 基于第二权重执行的神经网络的第一层可被配置为:基于第一层的输入数据处理与第一任务资源需求对应的第一任务,并且基于第四权重执行的神经网络的第二层可被配置为:基于第二层的输入数据处理与第一任务不同的与第二任务资源需求对应的第二任

务。

[0037] 多媒体数据可包括文本、语音、动态图像和静态图像中的至少一个。

[0038] 在另一总体方面,一种识别对象的方法,包括:接收多媒体数据作为神经网络的输入数据;基于第一任务资源需求确定第一神经网络的位宽;通过从源模型的第一权重中的每个提取与确定的位宽对应的至少一个位,获得第一神经网络的第二权重;以及通过基于获得的第二权重执行第一神经网络来对第一神经网络的输入数据执行对象识别。

[0039] 第一权重可具有高于第二权重的位精度的位精度。

[0040] 第二权重可嵌套在第一权重中。

[0041] 所述方法还可包括:基于由第一神经网络处理输入数据的结果确定是否需要第二神经网络;响应于确定需要第二神经网络,基于第二任务资源需求确定第二神经网络的位宽;通过从第一权重中的每个提取与确定的第二神经网络的位宽对应的至少一个位,获得第二神经网络的第三权重;以及通过基于获得的第三权重执行第二神经网络来对第二神经网络的输入数据执行对象识别。

[0042] 所述结果可指示对象是否存在于多媒体数据中,其中,基于所述结果指示对象存在于多媒体数据中,确定需要第二神经网络。

[0043] 第一权重可具有高于第三权重的位精度的位精度。

[0044] 第二权重和第三权重可嵌套在第一权重中。

[0045] 基于第二权重执行的第一神经网络可被配置为:基于第一神经网络的输入数据来处理与第一任务资源需求对应的第一任务,并且基于第三权重执行的第二神经网络可被配置为:基于第二神经网络的输入数据处理与第一任务不同的与第二任务资源需求对应的第二任务。

[0046] 第一任务资源需求可包括需要的处理速度、需要的处理准确度、处理难度、输入数据的质量和终端性能中的至少一个。

[0047] 第一神经网络的输入数据的质量低于第二神经网络的输入数据的质量,并且第二权重的位精度低于第三权重的位精度。

[0048] 第一任务是检测输入数据中是否存在预定的对象,第二任务是检测关于对象的位置和分类中的至少一个的信息。

[0049] 第一任务是检测输入数据是否对应于语音,第二任务是从输入数据检测关键词或验证说话者。

[0050] 在另一总体方面,一种用于识别对象的神经网络的训练方法,包括:接收多媒体数据作为神经网络的训练数据;通过量化与神经网络的第一层对应的高位宽的权重,确定与神经网络的第一层对应的低位宽的权重;通过将多媒体数据应用于第一层来确定与确定的低位宽的权重对应的损失值;以及基于确定的损失值更新高位宽的权重。

[0051] 所述训练方法还可包括:在与高位宽的权重相关联的训练完成之后,通过量化高位宽的权重来确定与第一层对应的低位宽的权重集。

[0052] 低位宽的权重集可包括第一位宽的权重集和第二位宽的权重集,第二位宽的权重集具有低于第一位宽的权重集的位精度的位精度,其中,第二位宽的权重集嵌套在第一位宽的权重集中。

[0053] 低位宽的权重可包括第一位宽的第一权重和第二位宽的第二权重,其中,第一位

宽的第一权重具有低于高位宽的权重的位精度的位精度,第二位宽的第二权重具有低于第一位宽的第一权重的位精度的位精度。

[0054] 确定低位宽的权重的步骤可包括:通过量化高位宽的权重来确定第一位宽的第一权重;以及通过从确定的第一位宽的第一权重中的每个提取至少一个位,确定第二位宽的第二权重。

[0055] 确定低位宽的权重的步骤可包括:通过量化高位宽的权重来确定第二位宽的第二权重;将确定的第二位宽的第二权重确定为第一位宽的第一权重的高位组;以及通过量化高位宽的权重来确定第一位宽的第一权重的低位组。

[0056] 更新高位宽的权重的步骤可包括:基于与确定的损失值对应的损失梯度的统计信息来更新高位宽的权重。

[0057] 更新高位宽的权重的步骤还可包括:通过将高加权值分配给与低位宽的权重之中设置了高优先级的权重对应的损失梯度来计算统计信息。

[0058] 在另一总体方面,一种识别对象的设备,包括:处理器;以及存储器,被配置为存储能够由处理器读取的指令,其中,当所述指令被处理器执行时,处理器被配置为:接收多媒体数据作为神经网络的第一层的输入数据;基于第一任务资源需求确定神经网络的第一层的位宽,其中,第一层包括至少一个层;通过从与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重;以及通过基于获得的第二权重执行神经网络的第一层来对神经网络的第一层的输入数据执行对象识别。

[0059] 处理器可被配置为:基于第二任务资源需求确定神经网络的第二层的位宽,其中,第二层包括至少一个层;通过从与神经网络的第二层对应的源模型的第三权重中的每个提取与确定的神经网络的第二层的位宽对应的至少一个位,获得神经网络的第二层的第四权重;以及通过基于获得的第四权重执行神经网络的第二层来对神经网络的第二层的输入数据进行对象识别,其中,第二层的输入数据对应于第一层的输出数据,其中,第二任务资源需求包括需要的第二处理速度、需要的第二处理准确度、第二处理难度和第二终端性能中的至少一个,其中,第二任务资源需求与第一任务资源需求相同或不同。

[0060] 在另一总体方面,一种识别对象的设备,包括:处理器;以及存储器,被配置为存储能够由处理器读取的指令,其中,当所述指令由处理器执行时,处理器被配置为:接收多媒体输入数据作为神经网络的输入数据;基于第一任务资源需求确定第一神经网络的位宽;通过从源模型的第一权重中的每个提取与确定的位宽对应的至少一个位,获得第一神经网络的第二权重;以及通过基于获得的第二权重执行第一神经网络来对第一神经网络的输入数据执行对象识别。

[0061] 处理器可被配置为:基于由第一神经网络处理输入数据的结果确定是否需要第二神经网络;响应于确定需要第二神经网络,基于第二任务资源需求确定第二神经网络的位宽;通过从第一权重中的每个提取与确定的第二神经网络的位宽对应的至少一个位,获得第二神经网络的第三权重;以及通过基于获得的第三权重执行第二神经网络来对第二神经网络的输入数据执行对象识别。

[0062] 所述结果可指示对象是否存在于多媒体数据中,其中,处理器被配置为基于所述结果指示对象存在于多媒体数据中,确定需要第二神经网络。

[0063] 在另一总体方面,一种用于识别对象的神经网络的训练设备,包括:处理器;以及存储器,被配置为存储能够由处理器读取的指令,其中,当所述指令被处理器执行时,处理器被配置为:接收多媒体数据作为神经网络的训练数据;通过量化与神经网络的第一层对应的高位宽的权重,确定与神经网络的第一层对应的低位宽的权重;通过将多媒体数据应用于第一层,确定与确定的低位宽的权重对应的损失值;以及基于确定的损失值更新高位宽的权重。

[0064] 低位宽的权重可包括第一位宽的第一权重和第二位宽的第二权重,其中,第一位宽的第一权重具有低于高位宽的权重的位精度的位精度,第二位宽的第二权重具有低于第一位宽的第一权重的位精度的位精度。

[0065] 处理器可被配置为:通过量化高位宽的权重来确定第一位宽的第一权重;以及通过从确定的第一位宽的第一权重中的每个提取至少一个位,确定第二位宽的第二权重。

[0066] 处理器可被配置为:通过量化高位宽的权重来确定第二位宽的第二权重;将确定的第二位宽的第二权重确定为第一位宽的第一权重的高位组;以及通过量化高位宽的权重来确定第一位宽的第一权重的低位组。

[0067] 从以下的具体实施方式、附图和权利要求书,其他特征和方面将是清楚的。

附图说明

[0068] 图1是示出基于嵌套位表示(nested bit representation-based)的神经网络处理设备的示例的示意图。

[0069] 图2是示出嵌套位表示的示例的示意图。

[0070] 图3是示出基于嵌套位表示的权重的示例的示意图。

[0071] 图4是示出针对每至少一个层调整权重的位宽的示例的示意图。

[0072] 图5是示出针对每个神经网络调整权重的位宽的示例的示意图。

[0073] 图6是示出基于嵌套位表示的训练设备的示例的示意图。

[0074] 图7是示出用于训练的前向处理的示例的示意图。

[0075] 图8是示出用于训练的前向处理的另一示例的示意图。

[0076] 图9是示出用于训练的后向处理的示例的示意图。

[0077] 图10是示出使用嵌套位表示来处理单个任务的示例的示意图。

[0078] 图11是示出使用嵌套位表示来处理多个任务的示例的示意图。

[0079] 图12是示出用于面部验证的多任务处理的示例的示意图。

[0080] 图13是示出用于对象分类的多任务处理的示例的示意图。

[0081] 图14是示出用于使用多任务处理以低功率来感测面部的设备的示例的示意图。

[0082] 图15是示出用于使用多任务处理以低功率来识别语音命令的设备的示例的示意图。

[0083] 图16是示出神经网络处理设备的示例的示意图。

[0084] 图17是示出电子设备的示例的示意图。

[0085] 图18是示出神经网络处理方法的示例的流程图。

[0086] 图19是示出神经网络处理方法的另一示例的流程图。

[0087] 图20是示出神经网络训练方法的示例的流程图。

[0088] 贯穿附图和具体实施方式,除非另有描述或提供,否则相同的附图参考标号将被

理解为表示相同的元件、特征和结构。附图可不按比例绘制,并且为了清楚、说明和方便起见,可夸大附图中的元件的相对尺寸、比例和描绘。

具体实施方式

[0089] 提供以下具体实施方式以帮助读者获得对在此描述的方法、设备和/或系统的全面理解。然而,在理解了本申请的公开之后,在此描述的方法、设备和/或系统的各种变化、修改和等同物将是清楚的。例如,在此描述的操作顺序仅仅是示例,并不限于在此阐述的操作顺序,而是除了必须以特定顺序发生的操作之外,操作顺序可如在理解本申请的公开之后清楚地那样改变。此外,为了更加清楚和简明,可省略在理解了本申请的公开之后已知的特征的描述。

[0090] 在此描述的特征可以以不同的形式来实现,并且不被解释为限于在此描述的示例。相反,已提供在此描述的示例仅仅示出在理解本申请的公开之后将是清楚的实现在此描述的方法、设备和/或系统的许多可能的方式中的一些方式。

[0091] 贯穿说明书,当诸如层、区域或基底的元件被描述为“在”另一元件“上”、“连接到”或“结合到”另一元件时,所述元件可直接“在”所述另一元件“上”、直接“连接到”或直接“结合到”所述另一元件,或者在它们之间可存在一个或多个其他元件。相比之下,当元件被描述为“直接在”另一元件“上”、“直接连接到”或“直接结合到”另一元件时,在它们之间可不存在其他元件。如在此使用的,术语“和/或”包括相关所列项的任意一个或任意两个或更多的任意组合。

[0092] 尽管可在此使用诸如“第一”、“第二”和“第三”的术语来描述各种构件、组件、区域、层或部分,但是这些构件、组件、区域、层或部分不受这些术语限制。相反,这些术语仅用于将一个构件、组件、区域、层或部分与另一构件、组件、区域、层或部分区分开。因此,在不脱离示例的教导的情况下,在此描述的示例中所称的第一构件、第一组件、第一区域、第一层或第一部分也可被称为第二构件、第二组件、第二区域、第二层或第二部分。

[0093] 在此使用的术语仅用于描述各种示例,并不用于限制本公开。除非上下文另外清楚地指示,否则单数形式也意图包括复数形式。术语“包含”、“包括”和“具有”说明存在阐述的特征、数量、操作、构件、元件和/或它们的组合,但不排除存在或添加一个或多个其他特征、数量、操作、构件、元件和/或它们的组合。

[0094] 除非另外定义,否则在此使用的所有术语(包括技术术语和科学术语)具有与基于本申请的公开的理解和本公开所属领域的普通技术人员通常理解的含义相同的含义。除非在此明确地如此定义,否则术语(诸如在通用词典中定义的术语)应被解释为具有与它们的相关领域的上下文中的含义一致的含义,并且不被理想化或过于形式化地解释。

[0095] 此外,在示例实施例的描述中,当认为在理解本申请的公开之后已知的结构或功能的详细描述将导致对示例实施例的模糊解释时,将省略这种描述。

[0096] 在下文中,将参照附图详细描述示例,并且附图中相同的参考标号始终表示相同的元件。

[0097] 图1是示出基于嵌套位表示的神经网络处理设备的示例的示图。神经网络处理设备在下文中将被简称为处理设备。

[0098] 参照图1,处理设备100处理神经网络120的操作。例如,神经网络120的操作可包括

对象识别(例如,识别或验证输入图像中的对象)。将理解,这里的“对象识别”是一个广义的概念,可包括对象感测或检测、对象验证或识别、对象分类、识别语音中的关键词、识别文本中包括的文本的含义等,但不限于此。在这种情况下,处理设备可以是用于对象识别的设备。与将在下文中描述的神经网络120相关联的处理操作的至少一部分可由软件、包括神经处理器的硬件或软件和硬件的组合来实现。

[0099] 神经网络120可包括可基于权重进行操作的所有类型的神经网络。例如,神经网络120可以是包括例如全连接网络、深度卷积网络和递归神经网络(RNN)的深度神经网络(DNN)。尽管下面的描述集中于作为神经网络120的一个示例的卷积神经网络(CNN),但是它也可适用于基于权重进行操作的其他类型的神经网络。

[0100] 神经网络120可通过基于深度学习映射处于非线性关系的输入数据和输出数据来执行对象识别和对对象验证。深度学习是指用于从大数据集识别图像或话音(或语音)的机器学习方法。深度学习可被解释为优化的问题解决处理,其中,问题解决处理在使用准备好的训练数据训练神经网络120的同时定位能量最小化的点。深度学习可分为有监督学习和无监督学习,其中,通过有监督学习和无监督学习,与神经网络120的架构或模型相应的权重被获得。通过这样获得的权重,输入数据和输出数据可被映射。在CNN中,可基于在卷积层中使用的权重核的权重来执行深度学习。

[0101] 神经网络120可包括多个层,其中,多个层包括输入层、至少一个隐藏层和输出层。例如,如示出的,包括在神经网络120中的第一层和第二层可以是多个层中的至少一部分。尽管在图1的示例中示出两个层包括在神经网络120中,但是提供这个示例是为了便于描述,因此更多数量的层可包括在神经网络120中。在一个示例中,输入数据可以是多媒体数据。例如,多媒体数据可包括文本、语音、动态图像和静态图像中的至少一个。

[0102] 参照图1,源模型110包括第一位宽 X_1 的权重至第 n 位宽 X_n 的权重。在这个示例中,当 X_n 的 n 增大时,位宽增大。处理设备100通过从源模型110选择期望的位宽的权重来执行神经网络120。也就是说,选择性的可变位宽可被应用于神经网络120。源模型110基于神经网络120的处理特性来选择期望的位宽的权重。例如,处理特性可包括需要的处理速度、需要的处理准确度、处理难度和终端性能中的至少一个。在这个示例中,准确度可对应于错误接受率(FAR)或验证率(VR)。这里,处理特性也可被称为任务资源需求。也就是说,任务资源需求可包括需要的处理任务的速度、需要的处理任务的准确度、处理任务的难度、输入数据的质量和终端的性能中的至少一个。例如,神经网络120的第一层的位宽可与第一层的任务资源需求(例如,第一任务资源需求)有关,神经网络120的第二层的位宽可与第二层的任务资源需求(例如,第二任务资源需求)有关。任务资源需求越高,神经网络的对应层的位宽越高。

[0103] 例如,在需要高处理速度或者在资源不足的移动终端中处理数据的情况下,源模型110可以以与低位宽对应的第一位宽 X_1 执行神经网络120。又例如,在需要高处理准确度或需要高处理难度的情况下,源模型110可以以与高位宽对应的第 n 位宽 X_n 执行神经网络120。例如,在输入数据的质量低(例如,输入图像的分辨率低)的情况下,源模型110可以以与低位宽对应的第一位宽 X_1 执行神经网络120。又例如,在输入数据的质量高(例如,输入图像的分辨率高)的情况下,源模型110可以以与高位宽对应的第 n 位宽 X_n 执行神经网络120。位精度对应于位宽,因此可变的位宽可指示可变的位精度。

[0104] 在一个示例中,可针对每至少一个层调整权重的位宽。例如,与第一位宽X1对应的权重可施加到第一层,并且与第二位宽X2对应的权重可施加到神经网络120的第二层,因此,如前所述,权重的位宽可针对每个层被调整。可选地,与第一位宽X1对应的权重可施加到神经网络120的第一层至第三层,并且与第二位宽X2对应的权重可施加到神经网络120的第四层和第五层,因此,如前所述,权重的位宽可针对每至少两个层被调整。

[0105] 在这个示例中,可基于层的处理特性来确定将施加到至少一个层的权重的位宽。例如,低位宽的权重可施加到需要高处理速度的层或者在资源不足的移动终端中处理数据的层或者低质量数据被输入的层。也就是说,低位宽的权重可施加到执行需要高处理速度的任务的层或者在资源不足的移动终端中执行任务的层。相反,高位宽的权重可施加到需要高处理准确度的层或者具有高处理难度的层。也就是说,高位宽的权重可施加到执行需要高处理准确度的任务的层或者执行具有高处理难度的任务的层。

[0106] 在另一示例中,可针对每个神经网络调整权重的位宽。例如,与第一位宽X1对应的权重可施加到第一神经网络,并且与第二位宽X2对应的权重可施加到第二神经网络。在这个示例中,可基于相应的神经网络的处理特性来确定将施加到每个神经网络的权重的位宽。例如,低位宽的权重可施加到处理具有相对低难度的操作(诸如,以面部感测或检测为例)的神经网络。相反,高位宽的权重可施加到处理具有相对高难度的操作(诸如,以面部验证或识别为例)的神经网络。

[0107] 然而,当包括在源模型110中的第一位宽X1的权重至第n位宽Xn的权重占用独立的存储器空间时,存储器效率可降低。因此,与第一位宽X1的权重至第n位宽Xn的权重对应的位可被嵌套,因此存储器空间可被有效地使用。例如,包括在源模型110中的第一位宽X1的权重至第n位宽Xn的权重可被嵌套。在这个示例中,第一位宽X1的权重可被嵌套在第二位宽X2的权重和第n位宽Xn的权重中,第二位宽X2的权重可被嵌套在第n位宽Xn的权重中。将参照图2详细描述这样的嵌套位表示。

[0108] 图2是示出嵌套位表示的示例的示图。在图2的示例中,CNN的权重核的权重由嵌套位表示。这样的嵌套位表示可适用于其他类型的神经网络。参照图2,存在 3×3 权重核210、220和230。权重核的每个元素可对应于一个权重或权重值。

[0109] 在这个示例中,权重核210的权重由X1位表示,权重核220的权重由X2位表示,权重核230的权重由X3位表示。在这个示例中,X2是大于X1的值,X3是大于X2的值。例如,X1位可以是2位,X2位可以是4位,X3位可以是8位。然而,X1至X3的值不限于前面描述的示例值,并且除了X1至X3之外,权重可由更多类型或更少类型的位宽来表示。

[0110] 权重核210、220和230可共享位值。例如,在权重核210的特定位置处的第一元素可具有权重值 $01_{(2)}$,在权重核220的相应位置处的第二元素可具有权重值 $0111_{(2)}$,在权重核230的相应位置处的第三元素可具有权重值 $01110100_{(2)}$ 。在这个示例中, $_{(2)}$ 指示二进制表示。因此,第一元素至第三元素可共享位值 $01_{(2)}$,并且第二元素和第三元素可共享位值 $0111_{(2)}$,因此第一元素至第三元素可通过被嵌套来表示。

[0111] 如上所述,权重核210、220和230可在相应的元素之中共享位值,因此可被合并以定义嵌套位模型250。为了存储嵌套位模型250,可仅需要用于权重核230的存储器空间,因此通过嵌套位模型250,存储器效率可被提高。

[0112] 返回参照图1,包括在源模型110中的第一位宽X1的权重至第n位宽Xn的权重可对

应于上面参照图2描述的嵌套位模型250。例如,第一位宽X1的权重可对应于权重核210,第二位宽X2的权重可对应于权重核220,第n位宽Xn的权重可对应于权重核230。因此,第一位宽X1至第n位宽Xn的所有权重可存储在用于第n位宽Xn的权重的存储器空间中。

[0113] 神经网络120可在训练处理中基于训练数据被训练,并且在推断处理中执行关于输入数据的推断操作(诸如,以分类、识别和检测为例)。包括在源模型110中的第一位宽X1的权重至第n位宽Xn的权重可通过训练处理确定。训练可离线或在线执行。由于最近的具有加速训练的硬件(诸如,以神经处理器为例)的引入,因此在线训练可以是可行的。

[0114] 权重核可预先确定,这指示权重核在神经网络120启动之前被确定。神经网络120被启动可指示神经网络120准备好推断。例如,神经网络120被启动可指示神经网络120被加载到存储器中,或者在神经网络120被加载到存储器中之后,用于推断的输入数据被输入到神经网络120。

[0115] 图3是示出基于嵌套位表示的权重的示例的示图。

[0116] 参照图3,输出数据330基于输入数据310与权重核320之间的运算来确定。在这个示例中,输入数据310具有 $H \times W \times C$ 尺寸(dimension),权重核320中的每个具有 $K \times K \times C$ 尺寸,并且权重核320被分类为D个组。输出数据330具有 $H \times W \times D$ 尺寸。在这个示例中,输入数据310、权重核320和输出数据330可与第一层相关联。

[0117] 权重核320包括具有权重 $116_{(10)}$ 的元素(在下文中被简称为第一元素)。在这个示例中, $116_{(10)}$ 被表示为 $01110100_{(2)}$,其中, $_{(10)}$ 指示十进制表示, $_{(2)}$ 指示二进制表示。源模型包括与第一元素的 $01110100_{(2)}$ 对应的权重值350。

[0118] 处理设备确定第一层的位宽(在下文中被称为第一位宽)。例如,处理设备基于对象识别的任务资源需求来确定第一位宽。此外,处理设备通过从权重值350提取与第一位宽对应的至少一个位来获得第一元素的权重值。例如,如示出的,在确定的位宽是2位的情况下,处理设备通过从权重值350提取 $01_{(2)}$ 来获得权重值351。在确定的位宽是4位的情况下,处理设备通过从权重值350提取 $0111_{(2)}$ 来获得权重值352。在确定的位宽是8位的情况下,处理设备通过从权重值350提取 $01110100_{(2)}$ 来获得权重值353。在这个示例中,2位、4位和8位被提供作为可由源模型表示的示例位宽,因此源模型可表示更大数量或更少数量的位宽。

[0119] 处理设备还通过与与权重核320的剩余元素对应的源模型的权重值提取与第一位宽对应的至少一个位,获得权重核320的剩余元素的权重值。处理设备通过基于获得的权重值执行第一层来处理输入数据310。例如,如示出的,处理设备通过执行第一层来执行输入数据310与权重核320之间的运算,并将输出数据330确定为执行运算的结果。尽管图3的示例集中于单个层,但是通过与上面参照图3描述的处理对应的处理,全部的单个神经网络的权重可被设置。

[0120] 在图3的示例中, $01_{(2)}$ 对应于 $1_{(10)}$, $0111_{(2)}$ 对应于 $7_{(10)}$, $01110100_{(2)}$ 对应于 $116_{(10)}$ 。这指示权重值基于位提取(bit extraction)而变化。因此,嵌套位表示的模型可需要被训练,以基于位宽的变化来对应于每个权重值的变化。在下文中,将详细描述训练嵌套位表示的模型或源模型。

[0121] 图4是示出针对每至少一个层调整权重的位宽的示例的示图。

[0122] 参照图4,处理设备确定第一层的位宽。在图4的示例中,第一层的位宽被确定为4位。处理设备通过从与第一层对应的源模型的第一层的权重410中的每个提取与4位的位宽

对应的至少一个位,获得第一层的权重420。例如,如示出的,处理设备通过从权重410中的元素01110100₍₂₎提取高4位0111₍₂₎来获得权重420中的对应元素。

[0123] 处理设备通过从权重410中的剩余元素中的每个提取4位来获得第一层的权重420。通过这样的位提取,01110100₍₂₎和0111₍₂₎可共享位。处理设备通过基于获得的权重420执行第一层来处理第一层的输入数据430。

[0124] 可对第二层执行与对第一层执行的处理对应的处理。处理设备确定第二层的位宽。处理设备可针对第二层确定与针对第一层确定的位宽相同的位宽或者与针对第一层确定的位宽不同的位宽。在图4的示例中,第二层的位宽被确定为2位。

[0125] 处理设备通过从与第二层对应的源模型的第二层的权重440中的每个提取与2位的位宽对应的至少一个位,来获得第二层的权重450。例如,如示出的,处理设备通过从权重440中的元素01000101₍₂₎提取高2位01₍₂₎来获得权重450中的对应元素。处理设备通过从权重440中的剩余元素中的每个提取2位来获得第二层的权重450。在这个示例中,01000101₍₂₎和01₍₂₎可共享位。处理设备通过基于获得的权重450执行第二层来处理第二层的输入数据460。

[0126] 可对第二层的后续层执行与对第一层和第二层执行的处理对应的处理。如上所述,处理设备可针对每个层调整权重的位宽,或者针对每至少两个层调整权重的位宽。可基于层的处理特性来确定将被应用于至少一个层的权重的位宽。处理特性可包括需要的处理速度、需要的处理准确度、处理难度和终端性能中的至少一个。换句话说,任务资源需求可包括需要的处理速度、需要的处理准确度、处理难度、输入数据的质量和终端的性能中的至少一个。

[0127] 图5是示出针对每个神经网络调整权重的位宽的示例的示意图。

[0128] 参照图5,处理设备确定神经网络510的位宽。在图5的示例中,神经网络510的位宽被确定为2位。处理设备通过从源模型500的第一权重中的每个提取与2位的位宽对应的至少一个位,获得神经网络510的第二权重。处理设备通过基于获得的第二权重执行神经网络510来处理神经网络510的输入数据。例如,用于对象识别的处理设备接收输入数据(例如,多媒体数据,其中,多媒体数据可包括文本、语音、动态图像和静态图像中的至少一种)。处理设备基于对象识别的任务资源需求将神经网络510的位宽确定为2位。处理设备通过从源模型500的第一权重中的每个提取与2位的位宽对应的至少一个位,获得神经网络510的第二权重。处理设备通过基于获得的第二权重执行神经网络510来对神经网络510的输入数据执行对象识别(例如,面部感测)。

[0129] 还可对神经网络520执行与对神经网络510执行的前述处理对应的处理。处理设备确定神经网络520的位宽。在图5的示例中,处理设备基于由神经网络510处理输入数据的结果来确定神经网络520的位宽。例如,处理设备基于由神经网络510处理输入数据的结果来确定是否需要执行神经网络520。响应于确定需要执行神经网络520,处理设备基于神经网络520的任务资源需求确定神经网络520的位宽。例如,神经网络510可处理面部感测,神经网络520可处理面部检测。在这个示例中,当神经网络510成功地感测到面部时,处理设备可确定需要执行神经网络520,并确定神经网络520的位宽以执行神经网络520。例如,针对具有相对较小难度的面部感测,神经网络510的位宽可被确定为2位,针对具有相对较大难度的面部检测,神经网络520的位宽可被确定为4位。

[0130] 处理设备通过从源模型500的第一权重中的每个提取与4位的位宽对应的至少一个位,获得神经网络520的第三权重。处理设备通过基于获得的第三权重执行神经网络520来处理神经网络520的输入数据。神经网络520的输入数据可具有与神经网络510的输入数据的质量或分辨率相比相对较高的质量或分辨率。例如,低分辨率图像可在低分辨率模式下作为输入数据被提供给神经网络510。在这个示例中,当神经网络510成功地感测到面部时,高分辨率图像可在高分辨率模式下作为输入数据被提供给神经网络520。

[0131] 图6是示出基于嵌套位表示的训练设备的示例的示图。基于嵌套位表示的训练设备也可被称为神经网络处理设备,并在下文将被简称为训练设备。

[0132] 参照图6,训练设备600包括存储器610和处理器620。存储器610存储神经网络615。处理器620从存储器610获得神经网络615,并基于训练数据训练神经网络615。存储器610可存储可由处理器620读取的指令,当存储在存储器610中的指令被处理器620执行时,处理器620可执行用于训练神经网络615的操作。

[0133] 训练数据可包括训练输入和训练输出。训练输入可以是将被输入到神经网络615的输入数据,输入数据可以是多媒体数据,例如文本、动态图像、静态图像和语音等,但不限于此。作为映射到训练输入的数据,训练输出可以是例如将响应于训练输入被输入到神经网络615而从神经网络615输出的标签。训练数据也可称为输入数据。

[0134] 处理器620训练神经网络615,使得训练输出从训练输入被获得。神经网络615的训练可包括:训练神经网络615的参数、更新神经网络615或更新神经网络615的参数。

[0135] 例如,处理器620可通过量化与神经网络615的第一层对应的高位宽的权重(也被称为高位宽权重),确定与神经网络615的第一层对应的低位宽的权重(也被称为低位宽权重),并通过将训练输入应用于第一层来确定与确定的低位宽权重对应的损失值。处理器620可基于确定的损失值来更新高位宽权重。可针对神经网络615的所有层重复这样的前述处理。

[0136] 在与高位宽权重相关联的训练完成之后,处理器620可通过量化高位宽权重来确定低位宽权重集。例如,低位宽权重集可对应于图1的源模型110。例如,低位宽权重集可包括表示为被嵌套的第一位宽权重至第n位宽权重。

[0137] 对神经网络615的这样的训练的处理可包括用于确定损失值的前向处理和用于通过反向传播来更新权重的后向处理。根据一个示例,各种位宽的权重可通过单个训练处理被训练。在下文中将详细描述用于训练神经网络615的前向处理和后向处理。

[0138] 图7是示出用于训练的前向处理的示例的示图。

[0139] 在图7的示例中,示出了输入数据710、高位宽权重720和低位宽权重730至750。权重720具有高于权重730至750的位精度的位精度。例如,如示出的,权重730、740和750分别由8位的位宽、4位的位宽和2位的位宽表示。权重720由32位的位宽(例如,浮点类型)表示。示出的位数仅作为示例提供,因此权重720至750可由其他的各种位数来表示。权重730具有高于权重740的位精度的位精度,权重740具有高于权重750的位精度的位精度。在图7的这个示例中,输入数据710、高位宽权重720和低位宽权重730至750可对应于神经网络的第一层。

[0140] 训练设备通过量化权重720来确定权重730至750。训练设备通过量化权重720来确定权重730,通过从权重730中的每个提取至少一个位来确定权重740,并通过从权重740中

的每个提取至少一个位来确定权重750。通过这样的处理,权重730至750可共享位。例如,如示出的,权重730和权重740共享4位,权重740和权重750共享2位。

[0141] 训练设备通过将输入数据710应用于第一层来确定分别与权重730、740和750对应的损失值(例如,如示出的 $Loss_8$ 、 $Loss_4$ 和 $Loss_2$)。损失值 $Loss_8$ 、 $Loss_4$ 和 $Loss_2$ 可对应于由第一层和第一层的后续层输出的神经网络的输出。在前向处理中,权重730、740和750用于确定损失值 $Loss_8$ 、 $Loss_4$ 和 $Loss_2$,在后向处理中,权重720基于损失值 $Loss_8$ 、 $Loss_4$ 和 $Loss_2$ 被更新。例如,输入数据710可以是多媒体数据,神经网络的输出可以是对象识别的结果。在下文中将详细描述后向处理。

[0142] 当通过这样的前述处理来完成训练权重720时,通过将完成训练的权重720量化为与权重730至750对应的位宽来确定源模型。例如,可通过量化完成训练的权重720来确定具有8位的位宽的第一权重集,可通过从第一权重集的权重中的每个提取至少一个位来确定具有4位的位宽的第二权重集,可通过从第二权重集的权重中的每个提取至少一个位来确定具有2位的位宽的第三权重集。前面描述的位数仅作为示例提供,因此权重集可由与权重730至750对应的各种位数来表示。确定的权重集可对应于图1的源模型110。

[0143] 权重720可基于权重730至750被优化,并且权重集可具有与权重730至750对应的位宽。因此,尽管权重值由于位提取而变化,但是每个权重值可保持在优化状态下。此外,尽管各种位宽的权重通过单个训练处理被训练,但是具有不同位宽的低位宽权重可彼此影响,因此权重可保持优化状态。

[0144] 图8是示出用于训练的前向处理的另一示例的示意图。

[0145] 在图8的示例中,示出了输入数据810、高位宽权重820和低位宽权重830至850。权重820具有高于权重830至850的位精度的位精度。例如,如示出的,权重830、840和850分别由2位的位宽、4位的位宽和8位的位宽表示。权重820由32位的位宽(例如,浮点类型)表示。示出的位数仅作为示例提供,因此权重820至850可由其他的各种位数来表示。权重830具有低于权重840的位精度的位精度,权重840具有低于权重850的位精度的位精度。在图8的这个示例中,输入数据810、高位宽权重820和低位宽权重830至850可对应于神经网络的第一层。

[0146] 训练设备通过量化权重820来确定权重830至850。训练设备通过量化权重820来确定权重830,并基于权重820和830来确定权重840。例如,训练设备可将权重830确定为权重840的高位组,并且通过量化权重820来确定权重840的低位组。训练装置可执行量化,使得在权重840的高位组被固定的同时,权重840具有与权重820对应的值。

[0147] 通过与前述确定权重840的处理对应的处理,训练设备基于权重820和权重840确定权重850。例如,训练设备可将权重840确定为权重850的高位组,并通过量化权重820来确定权重850的低位组。在图8的示例中,基于包括在权重830中的权重值 $01_{(2)}$ 确定包括在权重840中的权重值 $0111_{(2)}$,基于包括在权重840中的权重值 $0111_{(2)}$ 来确定包括在权重850中的权重值 $01110100_{(2)}$ 。

[0148] 训练设备通过将输入数据810应用于第一层来确定分别与权重830、840和850对应的损失值(例如,如示出的 $Loss_2$ 、 $Loss_4$ 和 $Loss_8$)。损失值 $Loss_2$ 、 $Loss_4$ 和 $Loss_8$ 可对应于由第一层和第一层的后续层输出的神经网络的输出。在后向处理中,权重820基于损失值 $Loss_2$ 、 $Loss_4$ 和 $Loss_8$ 被更新。当通过这样的前述处理来完成训练权重820时,通过将完成训练的权

重820量化为与权重830至850对应的位宽来确定源模型。此外,上面参照图7描述的前向处理可应用于参照图8描述的前向处理。

[0149] 在参照图7描述的前向处理中,由于位提取,低位宽权重的位值可取决于高位宽权重的位值,因此线性量化可基于参照图7描述的前向处理来执行。在参照图8描述的前向处理中,可从高位宽权重的位值仅提取低位宽权重的位值的一部分,并且可基于高位宽权重经由量化来确定低位宽权重的位值的剩余部分。因此,通过经由量化确定的位值,可基于参照图8描述的前向处理来执行非线性量化。因此,图8的示例中的权重830至850之间的位共享可小于图7的示例中的权重730至750之间的位共享。

[0150] 图9是示出用于训练的后向处理的示例的示意图。

[0151] 在图9的示例中,示出了高位宽权重910和低位宽权重920至940。权重910具有高于权重920至940的位精度的位精度。权重920具有n位的位宽,权重930具有n-1位的位宽,权重940具有1位的位宽。在图9的示例中,省略了具有2位的位宽的权重和具有n-2位的位宽的权重等。尽管在图9的示例中位宽之间的每个间隔被示出为1位,但是位宽之间的间隔可被定义为各种位数。此外,间隔可彼此相同或不同。在图9的示例中,高位宽权重910和低位宽权重920至940可对应于神经网络的第一层。

[0152] 训练设备基于损失值 $Loss_n$ 、 $Loss_{n-1}$ 和 $Loss_1$ 来更新权重910。训练设备在此确定与损失值 $Loss_n$ 、 $Loss_{n-1}$ 和 $Loss_1$ 对应的损失的梯度或损失梯度,并基于确定的损失梯度的统计信息更新权重910。在此,损失梯度可包括由 $\frac{\partial L}{\partial w_i^q}$ 表示的第一损失梯度和由 $\frac{\partial L}{\partial w_i}$ 表示的第二

损失梯度,其中,i表示1与n之间的整数值,L、w和 w^q 分别表示损失、权重和量化的权重。例如,损失值 $Loss_n$ 、 $Loss_{n-1}$ 和 $Loss_1$ 可对应于对象识别的结果。

[0153] 训练设备基于第二损失梯度的统计信息更新权重910。例如,统计信息可以由等式1表示的第二损失梯度的平均值,或者是由等式2表示的第二损失梯度的加权平均值。

[0154] [等式1]

$$[0155] \quad \frac{1}{n} \sum_{i=1}^n \frac{\partial L}{\partial w_i}$$

[0156] [等式2]

$$[0157] \quad \frac{1}{n} \sum_{i=1}^n \alpha_i \frac{\partial L}{\partial w_i}$$

[0158] 在等式2中, α 表示加权值。训练设备可通过将高加权值分配给与权重920至940之中设置了高优先级的权重对应的损失梯度来计算统计信息。例如,可关于具有较高位精度的权重分配较高的加权值,或者可关于具有较低位精度的权重分配较高的加权值。在此,可基于训练的结果或训练的目的来确定加权值。

[0159] 当通过这样的前述处理来完成训练权重910时,通过将完成训练的权重910量化为与权重920至940对应的位宽来确定源模型。

[0160] 图10是示出使用嵌套位表示来处理单个任务的示例的示意图。

[0161] 参照图10,基于具有2位的位宽的权重1021、具有4位的位宽的权重1022以及具有8

位的位宽的权重1023来执行神经网络1020。权重1021至1023可对应于嵌套位模型。示出的位数仅作为示例提供,因此权重1021至1023可由其他的各种位数来表示。尽管在图10的示例中示出三种类型的位宽,但是还可基于具有两种或四种或者更多种类型的位宽的权重来执行神经网络1020。

[0162] 神经网络1020通过处理输入数据1010生成输出数据1030。当输入数据1010被输入时,神经网络1020从权重1021至1023选择与至少一个位宽对应的权重来处理输入数据1010。例如,用于对象识别的神经网络1020通过处理输入数据1010生成输出数据1030。输入数据1010可以是文本、语音、动态图像、静态图像等,但不限于此。当输入数据1010被输入时,神经网络1020基于对象识别的任务资源需求,从权重1021至1023选择与至少一个位宽对应的权重,以对输入数据1010执行对象识别。

[0163] 神经网络1020基于针对每至少一个层具有不同位宽的权重来执行。例如,神经网络1020的第一层可基于第一位宽的权重来执行,神经网络1020的第二层可基于第二位宽的权重来执行。可选地,神经网络1020可基于针对每个网络具有不同位宽的权重来执行。例如,在神经网络1020基于第一位宽的权重被执行之后,神经网络1020可基于第二位宽的权重再次被执行。在这个示例中,基于第一位宽的权重执行的神经网络1020可被称为第一神经网络,基于第二位宽的权重执行的神经网络1020可被称为第二神经网络。

[0164] 在图10的示例中,神经网络1020基于各种位宽的权重1021至1023来处理单个任务。在这个示例中,与权重1021至1023的位宽对应的位精度可影响处理单个任务的准确度。例如,基于与相对高的位宽的权重1023生成的输出数据1030可比基于与相对较低的位宽对应的权重1021生成的输出数据1030更准确。例如,这样的处理准确度可对应于FAR或VR。因此,可基于单个任务所需的处理准确度来确定位宽。

[0165] 图11是示出使用嵌套位表示来处理多个任务的示例的示意图。

[0166] 参照图11,当多级输入数据1110应用于神经网络1120时,神经网络1120处理多级输入数据1110以生成多级输出数据1130。多级输入数据1110可以是多级图像数据或多级语音数据。例如,作为第一级输入数据的输入数据1111可对应于低分辨率图像,作为第二级输入数据的输入数据1112可对应于中分辨率图像,作为第三级输入数据的输入数据1113可对应于高分辨率图像。又例如,输入数据1111可对应于低分辨率图像,输入数据1112可对应于高分辨率图像,输入数据1113可对应于裁剪的图像。

[0167] 神经网络1120基于具有第一位宽(例如,2位的位宽)的权重1121、具有第二位宽(例如,4位的位宽)的权重1122和具有第三位宽(例如,8位的位宽)的权重1123来执行。权重1121至1123可被训练以执行不同的任务。例如,权重1121可被训练以感测图像中的面部,权重1122可被训练以检测图像中的面部的位置,权重1123可被训练以验证图像中的面部。在这个示例中,输出数据1131可指示面部是否被感测到,输出数据1132可指示关于面部的位置的信息,输出数据1133可指示面部是否被验证。

[0168] 又例如,权重1121可被训练以感测图像中的对象,权重1122可被训练以检测图像中的对象的位置并识别对象的超类(superclass)或粗粒度类(coarse-grained class),权重1123可被训练以识别对象的子类或细粒度类(fine-grained class)。在这个示例中,输出数据1131可指示对象是否被感测到,输出数据1132可指示关于对象的位置和对象的超类的信息,输出数据1133可指示对象的子类。多级输出数据1130可用于后续任务。例如,基于

与第一任务对应的权重1121生成的输出数据1131可用于触发第二任务。

[0169] 权重1121至1123可对应于嵌套位模型。权重1121至1123可包括共享位和独特位(individual bit)。例如,2位的权重01₍₂₎和4位的权重0111₍₂₎可具有共享位01和独特位11。神经网络1120可被训练以通过独特位执行多个任务。例如,当权重1121被训练以感测面部时,权重1122和权重1123可被训练,以通过独特位来检测面部的位置。

[0170] 神经网络1120可基于针对每至少一个层具有不同位宽的权重来执行,或者基于针对每个网络具有不同位宽的权重来执行。例如,在需要在一个网络中处理多个任务的情况下,神经网络1120可基于针对每至少一个层具有不同位宽的权重来执行。在这个示例中,神经网络1120的第一层组可基于权重1121执行面部感测,神经网络1120的第二层组可基于权重1122执行面部位置检测,神经网络1120的第三层组可基于权重1123执行面部验证。在这个示例中,每个层组可包括至少一个层。

[0171] 又例如,在使用多个网络来处理多个任务的情况下,神经网络1120可基于针对每个网络具有不同位宽的权重来执行。例如,神经网络1120可通过基于权重1121被执行来执行对象感测,通过基于权重1122被执行来执行对象位置检测和超类识别,并通过基于权重1123被执行来执行子类识别。在这个示例中,神经网络1120可被称为分别与这些任务对应的第一神经网络、第二神经网络和第三神经网络。

[0172] 对于神经网络1120的更详细描述,可参照上面参照图10提供的神经网络1020的描述。

[0173] 图12是示出用于面部验证的多任务处理的示例的示图。

[0174] 参照图12,神经网络1220处理多级输入数据1210。处理设备基于具有2位的位宽的权重1221执行神经网络1220的至少一个层。基于权重1221执行的神经网络1220的至少一个层输出指示低质量图像1211中是否存在面部的面部感测信息。处理设备基于由神经网络1220的至少一个层基于权重1221的产生的输出执行操作1230。如果面部不存在,则处理设备再次将低质量图像1211应用于基于权重1221执行的神经网络1220的至少一个层。

[0175] 当面部存在时,处理设备基于具有4位的位宽的权重1222执行神经网络1220的至少一个层,并且将高质量图像1212应用于基于权重1222执行的神经网络1220的至少一个层。基于权重1222执行的神经网络1220的至少一个层检测高质量图像1212中的面部的位置,以在操作1240中输出指示面部的位置的信息。

[0176] 在一个示例中,处理设备基于面部的位置从高质量图像1212生成裁剪的图像1213。处理设备基于具有8位的位宽的权重1223执行神经网络1220的至少一个层,并且将裁剪的图像1213应用于基于权重1223执行的神经网络1220的至少一个层。基于权重1223执行的神经网络1220的至少一个层通过裁剪的图像1213执行面部验证,并在操作1250中输出面部验证的结果。

[0177] 对于神经网络1220的更详细描述,可参照上面参照图10提供的神经网络1020的描述和上面参照图11提供的神经网络1120的描述。

[0178] 图13是示出用于对象分类的多任务处理的示例的示图。参照图13,多级输入数据1310包括低分辨率图像1311、高分辨率图像1312和裁剪的图像1313。基于具有2位的位宽的权重1321执行的神经网络1320的至少一个层输出指示低分辨率图像1311中是否存在对象的对象感测信息,低分辨率图像1311中是否存在对象在操作1330中基于对象感测信息来确

定。在操作1340中,基于具有4位的位宽的权重1322执行的神经网络1320的至少一个层从高分辨率图像1312检测对象的超类和对象的位置,并输出指示对象的超类和位置的信息。在操作1350中,基于具有8位的位宽的权重1323执行的神经网络1320的至少一个层通过裁剪的图像1313对对象进行分类,并输出关于对象的子类或细粒度类的信息。对于神经网络1320的更详细描述,可参照上面参照图10提供的神经网络1020的描述、上面参照图11提供的神经网络1120的描述以及上面参照图12提供的神经网络1220的描述。

[0179] 图14是示出用于使用多任务处理以低功率来感测面部的设备的示例的示意图。

[0180] 参照图14,图像传感器1400包括像素1410、逻辑1420和接口(I/F)1430。图像传感器1400可始终进行操作。为了执行这样的始终进行操作,可能需要功耗低。因此,图像传感器1400可通过使用嵌套位模型基于面部是否被感测到,在低分辨率模式或高分辨率模式下操作来减少功耗。

[0181] 在低分辨率模式下,像素1410仅激活像素1410的一部分以生成低分辨率图像,逻辑1420基于与低分辨率模式对应的低位宽权重来执行低位宽网络1421,以感测低分辨率图像中是否存在面部。例如,低位宽可以是2位或4位。

[0182] 当低分辨率图像中存在面部时,图像传感器1400的操作模式从低分辨率模式改变为高分辨率模式。在高分辨率模式下,像素1410激活像素1410的部分大于在低分辨率模式下激活的像素1410的部分,以生成高分辨率图像。例如,在高分辨率模式下,像素1410可激活像素1410的整个区域。逻辑1420基于与高分辨率模式对应的高位宽权重来执行高位宽网络1422,以从高分辨率图像检测与面部对应的感兴趣区域(ROI)。例如,高位宽可以是4位或8位。

[0183] 接口1430向应用1450发送唤醒中断和裁剪的ROI图像。应用1450响应于唤醒中断而唤醒,然后使用裁剪的ROI图像执行用户验证。

[0184] 低位宽网络1421和高位宽网络1422可基于嵌套位模型来执行。嵌套位模型可占用较少的存储器空间,因此即使当逻辑1420具有低容量的存储器时嵌套位模型也可在逻辑1420中运行,并且图像传感器1400因此可提供低分辨率模式和高分辨率模式。在低分辨率模式下,仅像素1410的一部分可被激活,具有小处理能力的低分辨率图像可在逻辑1420中处理,因此在低分辨率模式下将被消耗的功率可被降低。因此,通过基于嵌套位模型的图像传感器1400,面部感测或检测可始终以低功率执行。此外,应用1450可在面部ROI被检测到之后运行,因此应用1450的运行所消耗的功率可被最小化。

[0185] 图15是示出用于使用多任务处理以低功率来识别语音命令的设备的示例的示意图。

[0186] 参照图15,麦克风模块1500包括麦克风1510和数字信号处理器(DSP)1520。麦克风模块1500可始终进行操作。麦克风模块1500可使用嵌套位模型来降低用于这样的始终进行操作的功耗。

[0187] 麦克风1510在这样的始终工作状态下输出音频信号。DSP 1520基于低位宽权重执行低位宽网络1521,并使用低位宽网络1521确定音频信号是否对应于语音或语音。例如,低位宽可以是2位。当音频信号被确定为对应于语音或语音时,DSP 1520基于中位宽权重执行中位宽网络1522。例如,中位宽可以是4位。DSP 1520使用中位宽网络1522从音频信号检测关键词。关键词可以是用于唤醒应用1550的命令(例如“你好,XXX(Hi,XXX)”和“好,XXX(Okay,XXX)”)。

[0188] 当关键词从音频信号被检测到时,DSP 1520基于高位宽权重执行高位宽网络1523。例如,高位宽可以是8位。DSP 1520使用高位宽网络1523验证说话者。当说话者被验证时,或者当说话者对应于注册用户时,唤醒中断和语音间隔中的音频信号被发送到应用1550。应用1550响应于唤醒中断而唤醒,然后使用语音间隔中的音频信号执行话音或语音识别。

[0189] 低位宽网络1521、中位宽网络1522和高位宽网络1523可基于嵌套位模型来执行。嵌套位模型可占用较少的存储器空间,因此即使当DSP 1520具有低容量的存储器时,嵌套位模型也可在DSP 1520中运行。DSP 1520可正常操作低位宽网络1521,并且可仅在音频信号对应于话音或语音时以及当关键词被检测到时操作中位宽网络1522和高位宽网络1523,以降低功耗。此外,应用1550可在说话者被验证之后运行,因此应用1550的运行所消耗的功率可被最小化。

[0190] 图16是示出处理设备的示例的示意图。

[0191] 参照图16,处理设备1600接收输入数据。处理设备1600使用神经网络来处理输入数据。例如,输入数据可以是多媒体数据。例如,处理设备1600可使用神经网络处理关于输入数据的识别或验证。数据库(DB) 1630可存储源模型。

[0192] 处理设备1600可执行在此描述的用于面部验证的操作中的至少一个,并向用户提供面部验证的结果。在图16的示例中,处理设备1600包括至少一个处理器1610和存储器1620。存储器1620可连接到处理器1610,并且存储可由处理器1610执行的指令,以及由处理器1610处理的数据或将由处理器1610处理的数据。存储器1620可包括非暂时性计算机可读介质(例如,高速随机存取存储器(RAM))和/或非易失性计算机可读存储介质(例如,至少一个磁盘存储装置、闪存装置和其他非易失性固态存储器装置)。

[0193] 处理器1610可执行指令,以执行上面参照图1至图15描述的操作中的至少一个。例如,处理器1610可确定神经网络的第一层的位宽,通过与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重,并通过基于获得的第二权重执行神经网络的第一层来处理神经网络的第一层的输入数据。例如,处理器1610可接收多媒体数据作为输入数据,基于对象识别的任务资源需求确定神经网络的第一层的位宽,通过与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重,并通过基于获得的第二权重执行神经网络的第一层来处理神经网络的第一层的输入数据以执行对象识别。可选地,处理器1610可确定第一神经网络的位宽,通过从源模型的第一权重中的每个提取与确定的位宽对应的至少一个位,获得第一神经网络的第二权重,并通过基于获得的第二权重执行第一神经网络来处理第一神经网络的输入数据。

[0194] 图17是示出电子设备的示例的示意图。

[0195] 电子设备1700可接收输入图像,并处理与输入图像相关联的神经网络的操作。例如,神经网络的操作可包括识别或验证输入图像中的对象。电子设备1700可使用上述源模型来处理神经网络的操作。电子设备1700可包括上面参照图1至图16描述的处理设备,或者执行上面参照图1至图16描述的处理设备的功能。

[0196] 参照图17,电子设备1700包括处理器1710、存储器1720、相机1730、存储装置1740、输入装置1750、输出装置1760和网络接口1770。处理器1710、存储器1720、相机1730、存储装

置1740、输入装置1750、输出装置1760和网络接口1770可通过通信总线1780彼此通信。

[0197] 处理器1710可执行电子设备1700中的功能和指令。例如,处理器1710可处理存储在存储器1720或存储装置1740中的指令。处理器1710可执行上面参照图1至图16描述的操作中的至少一个。

[0198] 存储器1720可存储将用于处理神经网络的操作的信息。存储器1720可包括计算机可读存储介质或计算机可读存储装置。存储器1720可存储将由处理器1710执行的指令,并且在电子设备1700正在执行软件程序或应用的同时存储相关信息。

[0199] 相机1730可捕获静止图像、视频图像或者这两种图像。相机1730可捕获由用户输入的用于面部验证的面部区域的图像。相机1730还可提供包括对象的深度信息的三维(3D)图像。

[0200] 存储装置1740可包括计算机可读存储介质或计算机可读存储装置。存储装置1740可存储包括将用于处理神经网络的信息(诸如,以源模型、权重核和权重值为例)的DB。例如,与存储器1720相比,存储装置1740可将更大量的信息存储更长的时间段。例如,存储装置1740可包括磁性硬盘、光盘、闪存、软盘以及相关技术领域中公知的其他类型的非易失性存储器。

[0201] 输入装置1750可通过传统的输入方法(例如,键盘和鼠标)和/或新的输入方法(例如,触摸输入、语音输入和图像输入)从用户接收输入。例如,输入装置1750可包括键盘、鼠标、触摸屏、麦克风和可从用户检测输入并将检测到的输入发送到电子设备1700的其他装置。

[0202] 输出装置1760可通过视觉、听觉或触觉通道向用户提供电子设备1700的输出。例如,输出装置1760可包括显示器、触摸屏、扬声器、振动发生器以及可向用户提供输出的其他装置。网络接口1770可通过有线网络或无线网络与外部装置通信。

[0203] 图18是示出神经网络处理方法的示例的流程图。在一个示例中,神经网络处理方法可以是识别对象的方法。参照图18,在操作1810中,处理设备确定神经网络的第一层的位宽。在操作1820中,处理设备通过从与神经网络的第一层对应的源模型的第一层的第一权重中的每个提取与确定的位宽对应的至少一个位,获得神经网络的第一层的第二权重。在操作1830中,处理设备通过基于获得的第二权重执行神经网络的第一层来处理神经网络的第一层的输入数据。对于神经网络处理方法的操作的详细描述,可参照上面参照图1至图17提供的描述。

[0204] 图19是示出神经网络处理方法的另一示例的流程图。在一个示例中,神经网络处理方法可以是识别对象的方法。参照图19,在操作1910中,处理设备确定第一神经网络的位宽。在操作1920中,处理设备通过从源模型的第一权重中的每个提取与确定的位宽对应的至少一个位,获得第一神经网络的第二权重。在操作1930中,处理设备通过基于获得的第二权重执行第一神经网络来处理第一神经网络的输入数据。对于神经网络处理方法的操作的详细描述,可参照上面参照图1至图17提供的描述。

[0205] 图20是示出神经网络训练方法的示例的流程图。参照图20,在操作2010中,训练设备通过量化与神经网络的第一层对应的高位宽权重,确定与神经网络的第一层对应的低位宽权重。在操作2020中,训练设备通过将输入数据应用于第一层来确定与确定的低位宽权重对应的损失值。在操作2030中,训练设备基于确定的损失值来更新高位宽权重。对于神经

网络训练方法的操作的详细描述,可参照上面参照图1至图17提供的描述。

[0206] 在此针对图1、图6、图16和图17描述的处理设备、训练设备以及其他设备、单元、模块、装置和其他组件通过硬件组件实现。可用于执行在本申请中描述的操作的硬件组件的示例适当地包括:控制器、传感器、生成器、驱动器、存储器、比较器、算术逻辑单元、加法器、减法器、乘法器、除法器、积分器以及被配置为执行在本申请中描述的操作的任何其他电子组件。在其他示例中,通过计算硬件(例如,通过一个或多个处理器或计算机)来实现执行在本申请中描述的操作的硬件组件中的一个或多个硬件组件。可通过一个或多个处理元件(诸如,逻辑门阵列、控制器和算术逻辑单元、数字信号处理器、微型计算机、可编程逻辑控制器、现场可编程门阵列、可编程逻辑阵列、微处理器或被配置为以限定的方式响应并执行指令以实现期望的结果的任何其他装置或装置的组合)来实现处理器或计算机。在一个示例中,处理器或计算机包括或连接到存储由处理器或计算机执行的指令或软件的一个或多个存储器。由处理器或计算机实现的硬件组件可执行用于执行在本申请中描述的操作的指令或软件(诸如,操作系统(OS)和在OS上运行的一个或多个软件应用)。硬件组件还可响应于指令或软件的执行来访问、操控、处理、创建和存储数据。为了简单起见,单数术语“处理器”或“计算机”可用于本申请中描述的示例的描述,但是在其他的示例中,多个处理器或计算机可被使用,或者处理器或计算机可包括多个处理元件或多种类型的处理元件或二者。例如,单个硬件组件或者两个或更多个硬件组件可通过单个处理器、或者两个或更多个处理器、或者处理器和控制器来实现。一个或多个硬件组件可通过一个或多个处理器、或者处理器和控制器来实现,一个或多个其他硬件组件可通过一个或多个其他处理器,或者另外的处理器和另外的控制器来实现。一个或多个处理器、或者处理器和控制器可实现单个硬件组件或者两个或更多个硬件组件。硬件组件可具有不同的处理配置中的任何一个或多个,其中,不同的处理配置的示例包括:单个处理器、独立处理器、并行处理器、单指令单数据(SISD)多处理、单指令多数据(SIMD)多处理、多指令单数据(MISD)多处理以及多指令多数据(MIMD)多处理。

[0207] 在图1至图5、图7至图15以及图18至图20中示出的执行在本申请中描述的操作的方法通过计算硬件(例如,通过一个或多个处理器或计算机)来执行,其中,计算硬件被如上所述实现为执行指令或软件,以执行在本申请中描述的通过该方法执行的操作。例如,单个操作或者两个或更多个操作可通过单个处理器或者两个或更多个处理器、或者处理器和控制器来执行。一个或多个操作可通过一个或多个处理器、或者处理器和控制器来执行,并且一个或多个其他操作可通过一个或多个其他处理器、或者另外的处理器和另外的控制器来执行。一个或多个处理器、或者处理器和控制器可执行单个操作或者两个或更多个操作。

[0208] 用于控制处理器或计算机实现硬件组件并执行如上所述的方法的指令或软件被编写为计算机程序、代码段、指令或它们的任何组合,以单独地或共同地指示或配置处理器或者计算机作为机器或专用计算机进行操作,以执行由如上所述的硬件组件和方法执行的操作。在一个示例中,指令或软件包括由处理器或者计算机直接执行的机器代码(诸如,由编译器产生的机器代码)。在另一个示例中,指令或软件包括由处理器或者计算机使用解释器执行的高级代码。本领域的普通技术编程人员可基于附图中所示的框图和流程图以及说明书中的对应描述,容易地编写指令或软件,其中,附图中所示的框图和流程图以及说明书中的对应描述公开了用于执行由如上所述的硬件组件和方法执行的操作的算法。

[0209] 用于控制处理器或计算机以实现硬件组件并执行如上所述的方法的指令或软件以及任何相关联的数据、数据文件和数据结构被记录、存储或固定在一个或多个非暂时性计算机可读存储介质中,或者被记录、存储或固定在一个或多个非暂时性计算机可读存储介质上。非暂时性计算机可读存储介质的示例包括:只读存储器(ROM)、随机存取可编程只读存储器(PROM)、电可擦除可编程只读存储器(EEPROM)、随机存取存储器(RAM)、动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM)、闪存、非易失性存储器、CD-ROM、CD-R、CD+R、CD-RW、CD+RW、DVD-ROM、DVD-R、DVD+R、DVD-RW、DVD+RW、DVD-RAM、BD-ROM、BD-R、BD-R LTH、BD-RE、蓝光或光盘存储器、硬盘驱动器(HDD)、固态硬盘(SSD)、卡式存储器(诸如,多媒体微型卡或卡(例如,安全数字(SD)或极速数字(XD))、磁带、软盘、磁光数据存储装置、光学数据存储装置、硬盘、固态硬盘以及任何其他装置,其中,任何其他装置被配置为以非暂时性方式存储指令或软件以及任何相关联的数据、数据文件和数据结构并将所述指令或软件以及任何相关联的数据、数据文件和数据结构提供给处理器或计算机使得处理器或计算机能执行指令。

[0210] 虽然本公开包括具体的示例,但是本领域普通技术人员将清楚,在不脱离权利要求及其等同物的精神和范围的情况下,可在这些示例中进行形式和细节上的各种改变。在此描述的示例仅被认为是描述性的,而不是出于限制的目的。每个示例中的特征或方面的描述将被认为可适用于其他示例中的类似特征或方面。如果描述的技术以不同的顺序被执行,和/或如果描述的系统、架构、装置或电路中的组件以不同的方式被组合,和/或由其他组件或它们的等同物替换或补充,则可实现合适的结果。因此,本公开的范围不是由具体实施方式限定,而是由权利要求及其等同物限定,并且在权利要求及其等同物的范围内的所有变化应被解释为包括在本公开中。

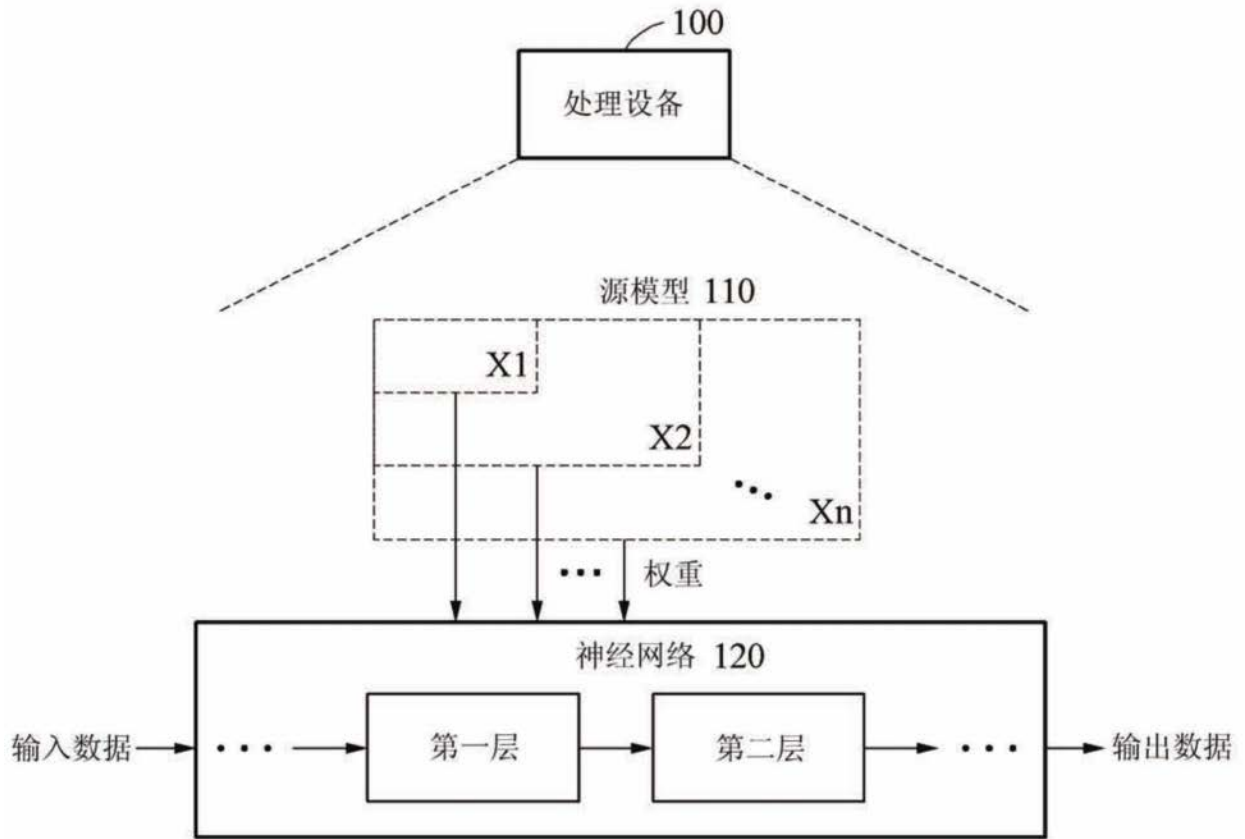


图1

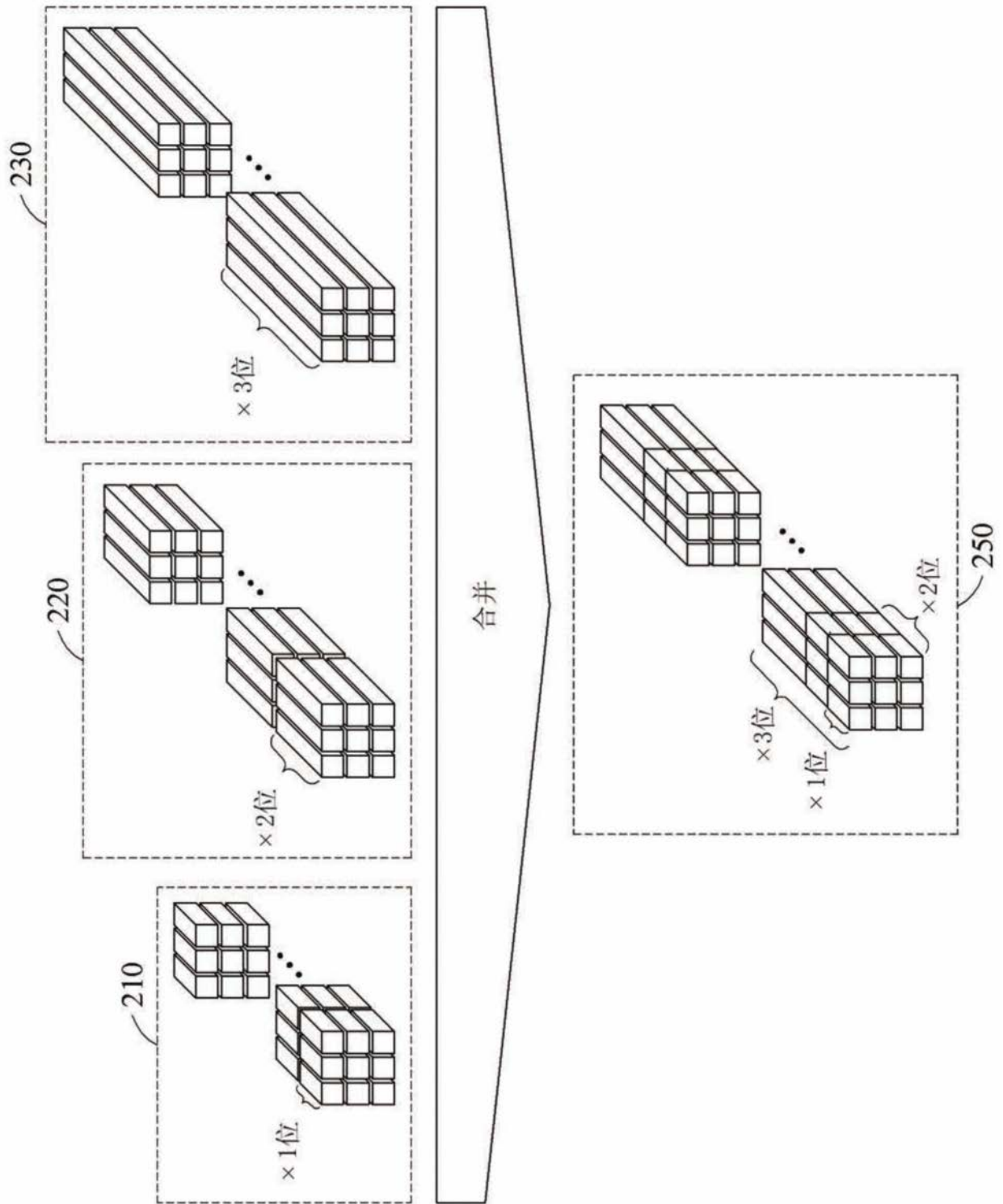


图2

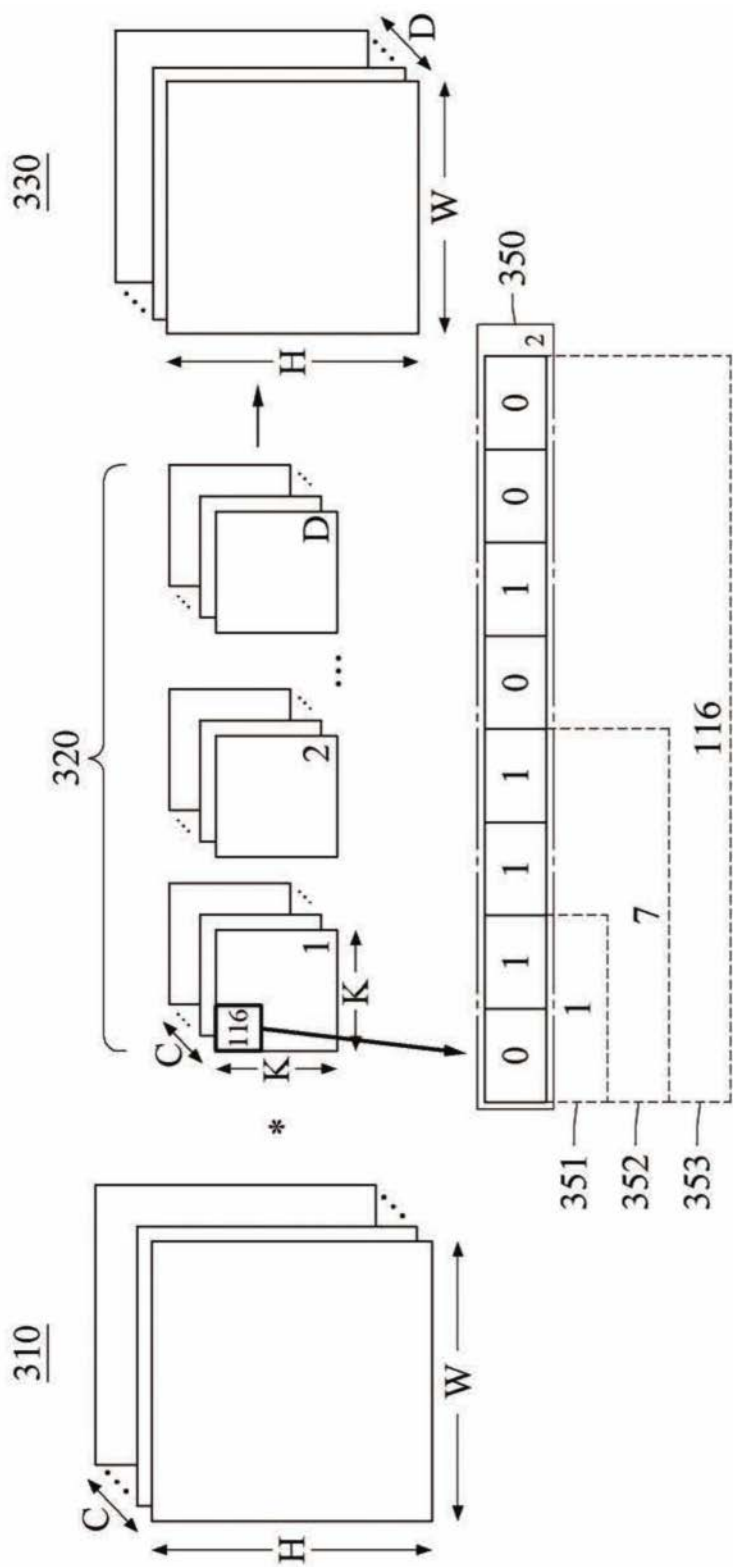


图3

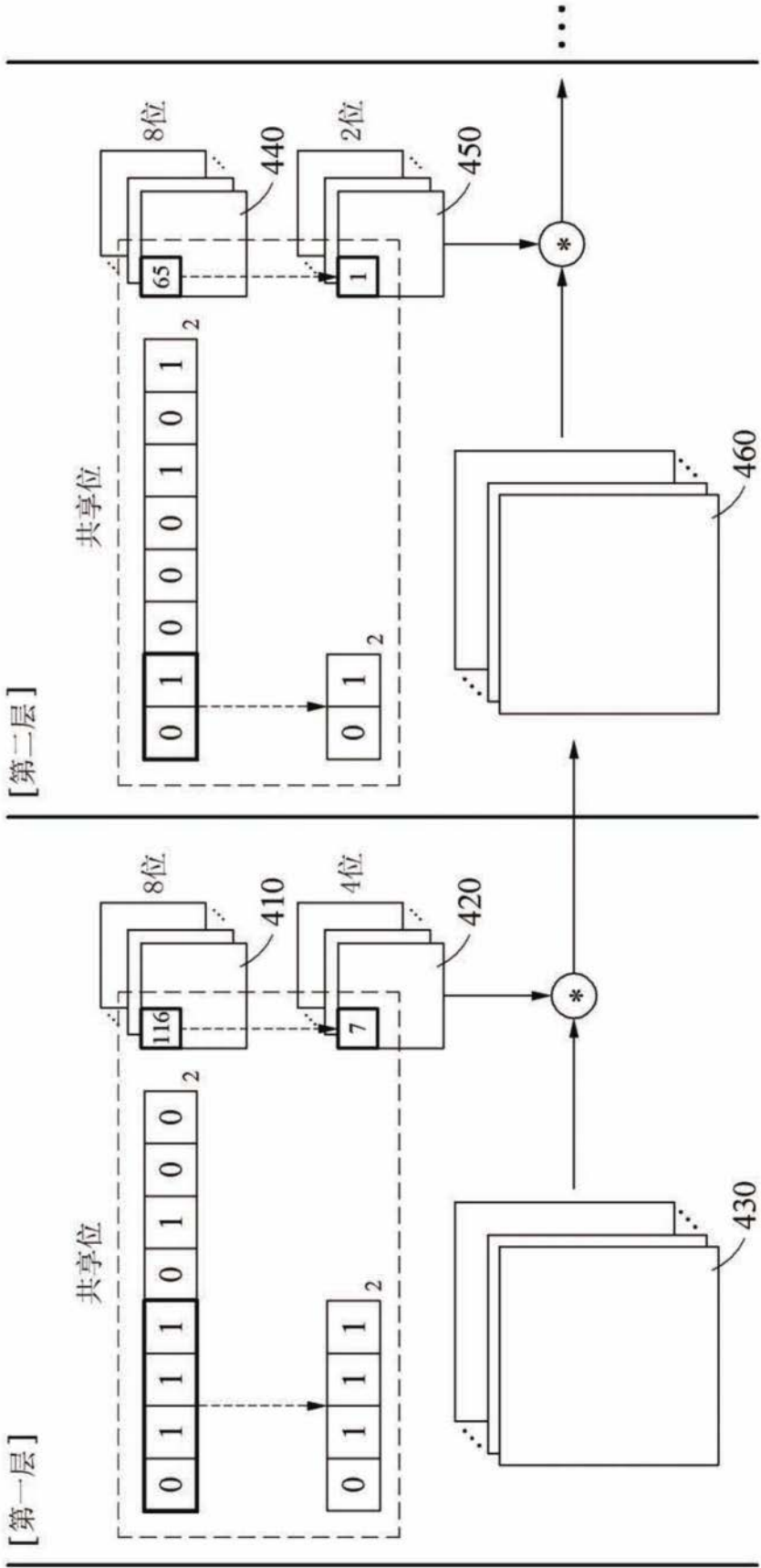


图4

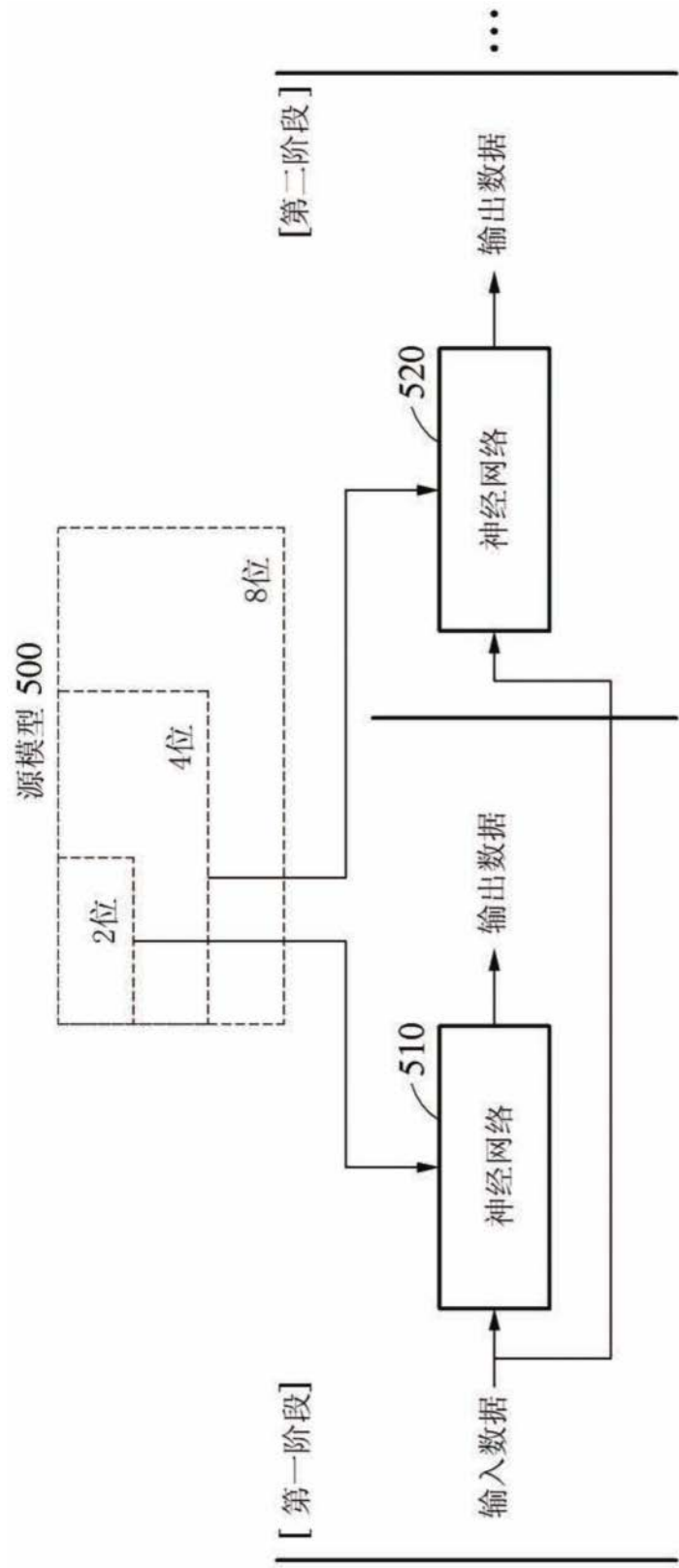


图5

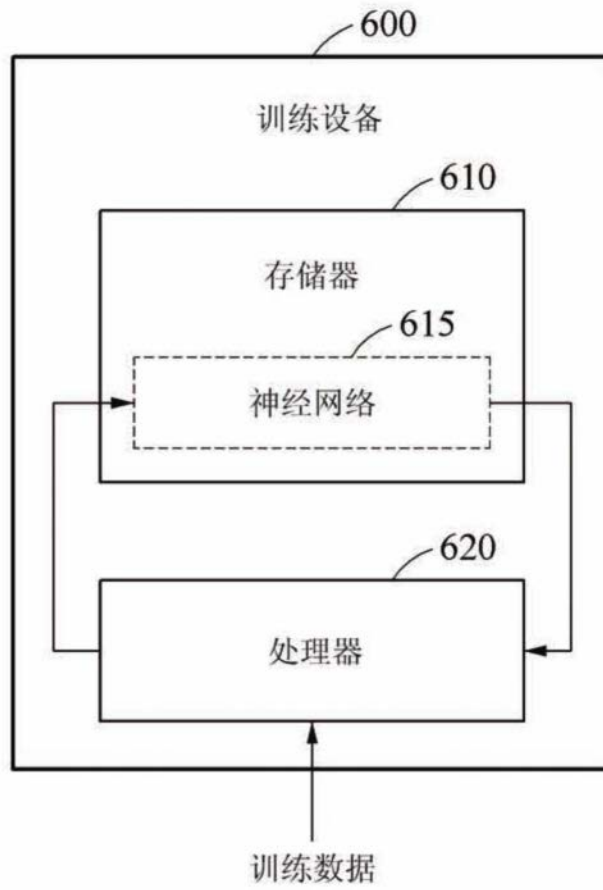


图6

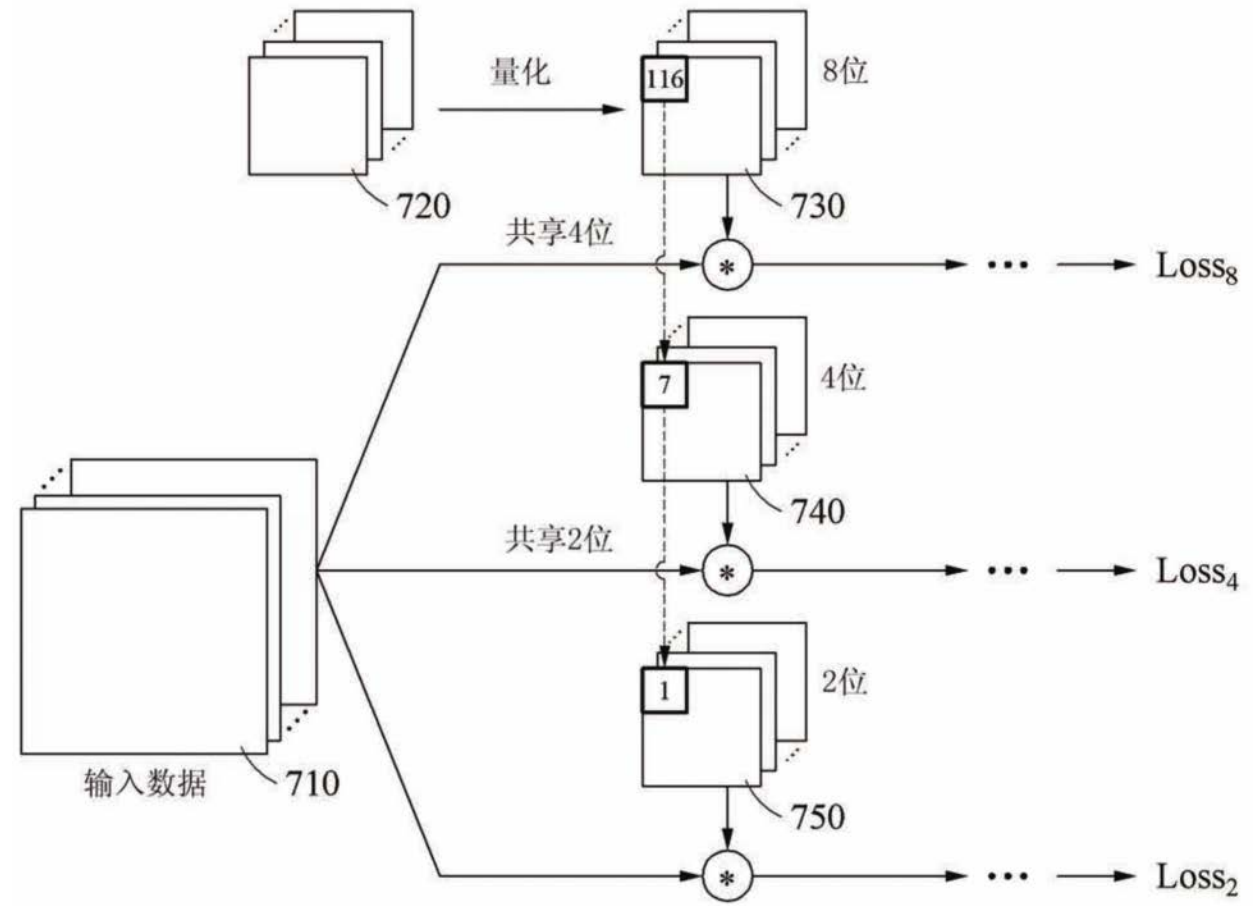


图7

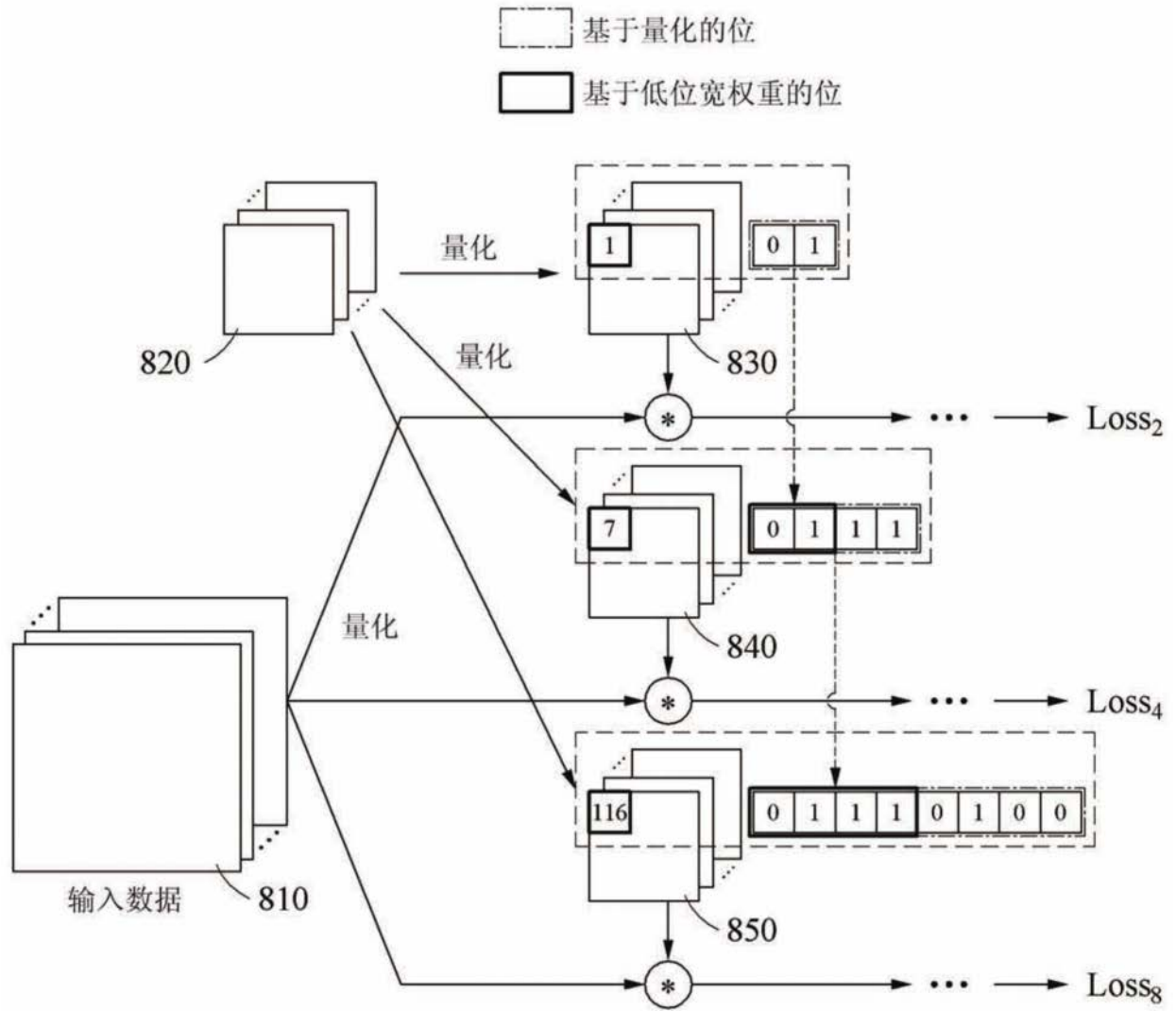


图8

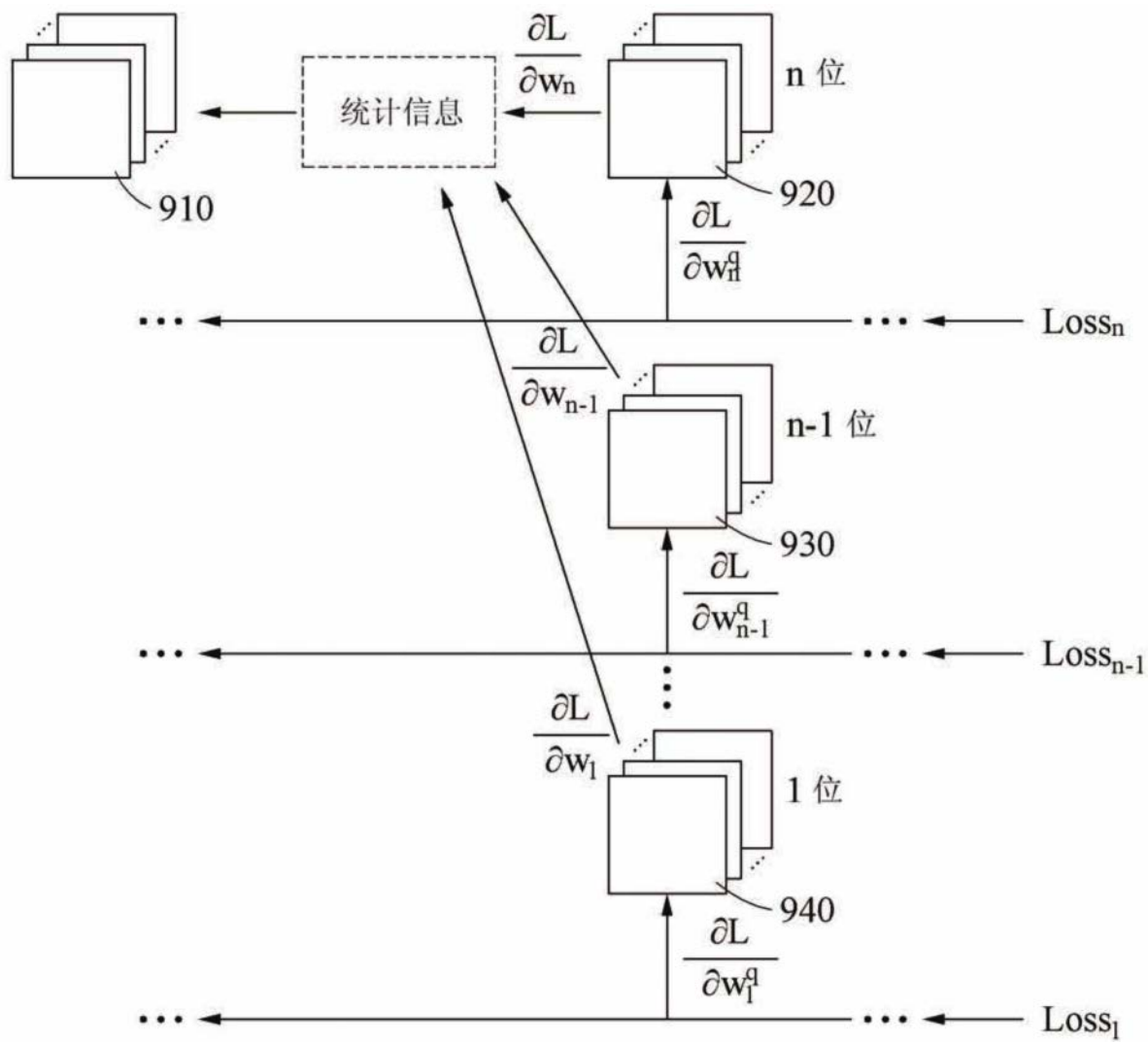


图9

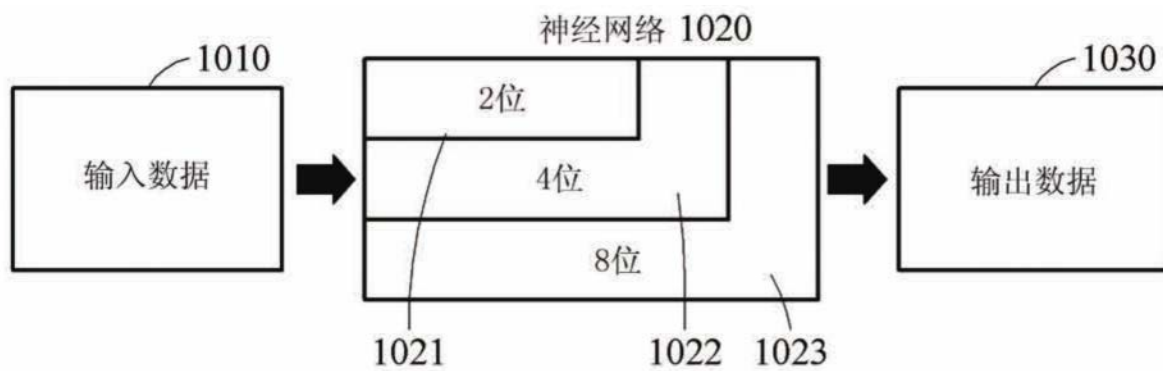


图10

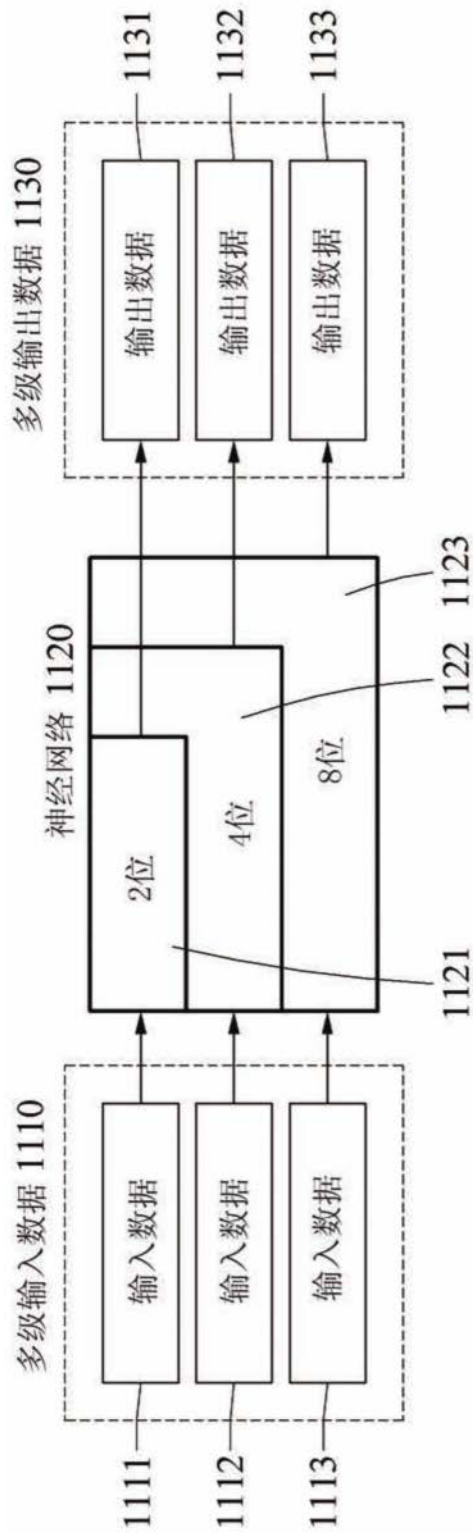


图11

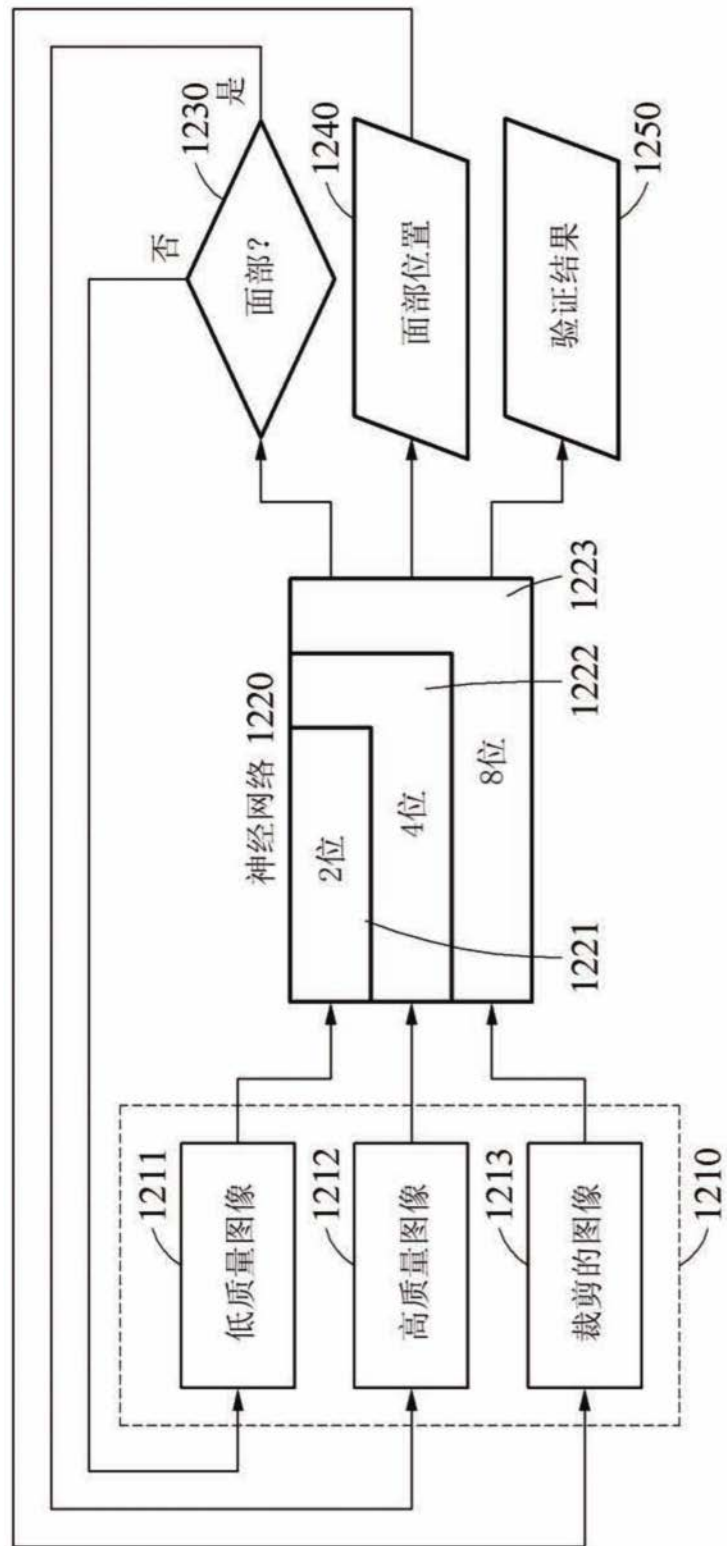


图12

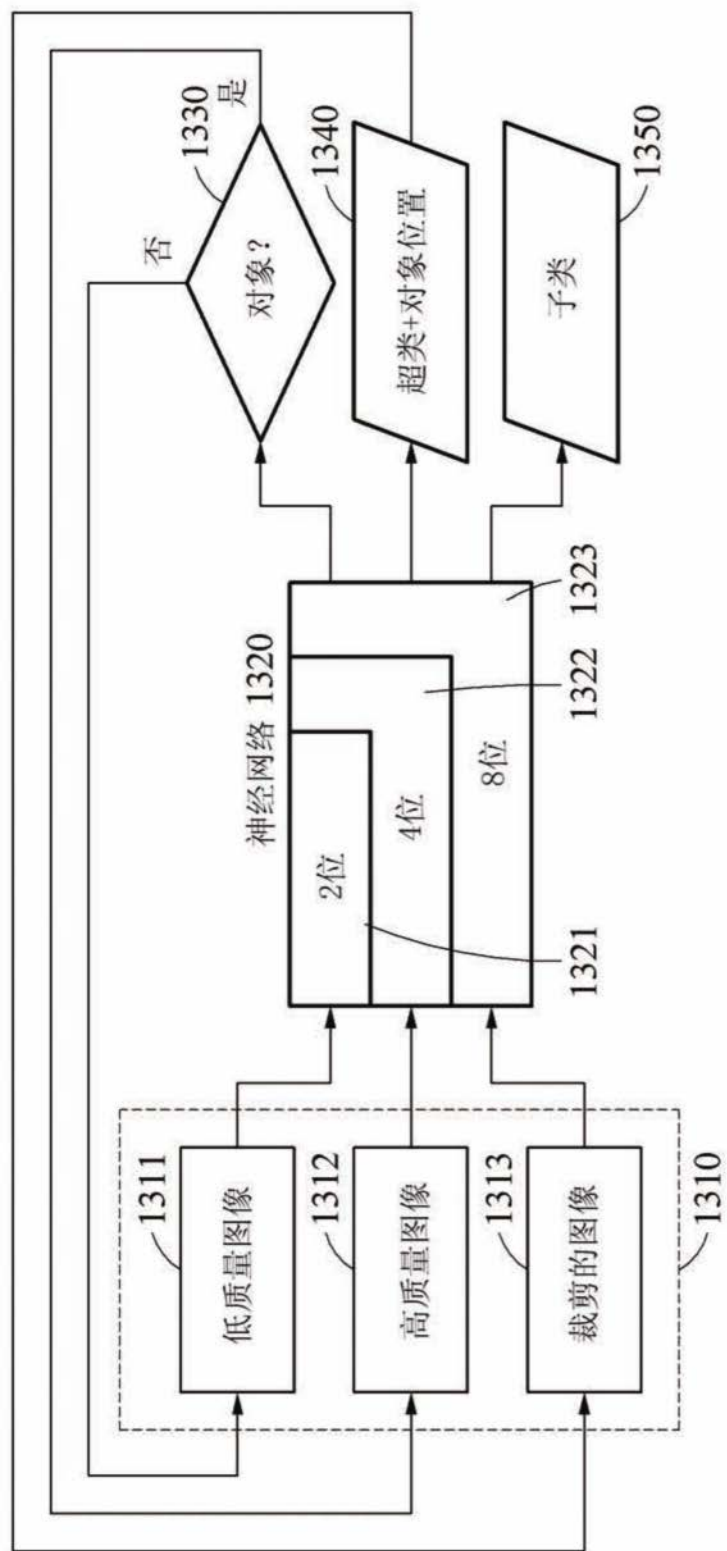


图13

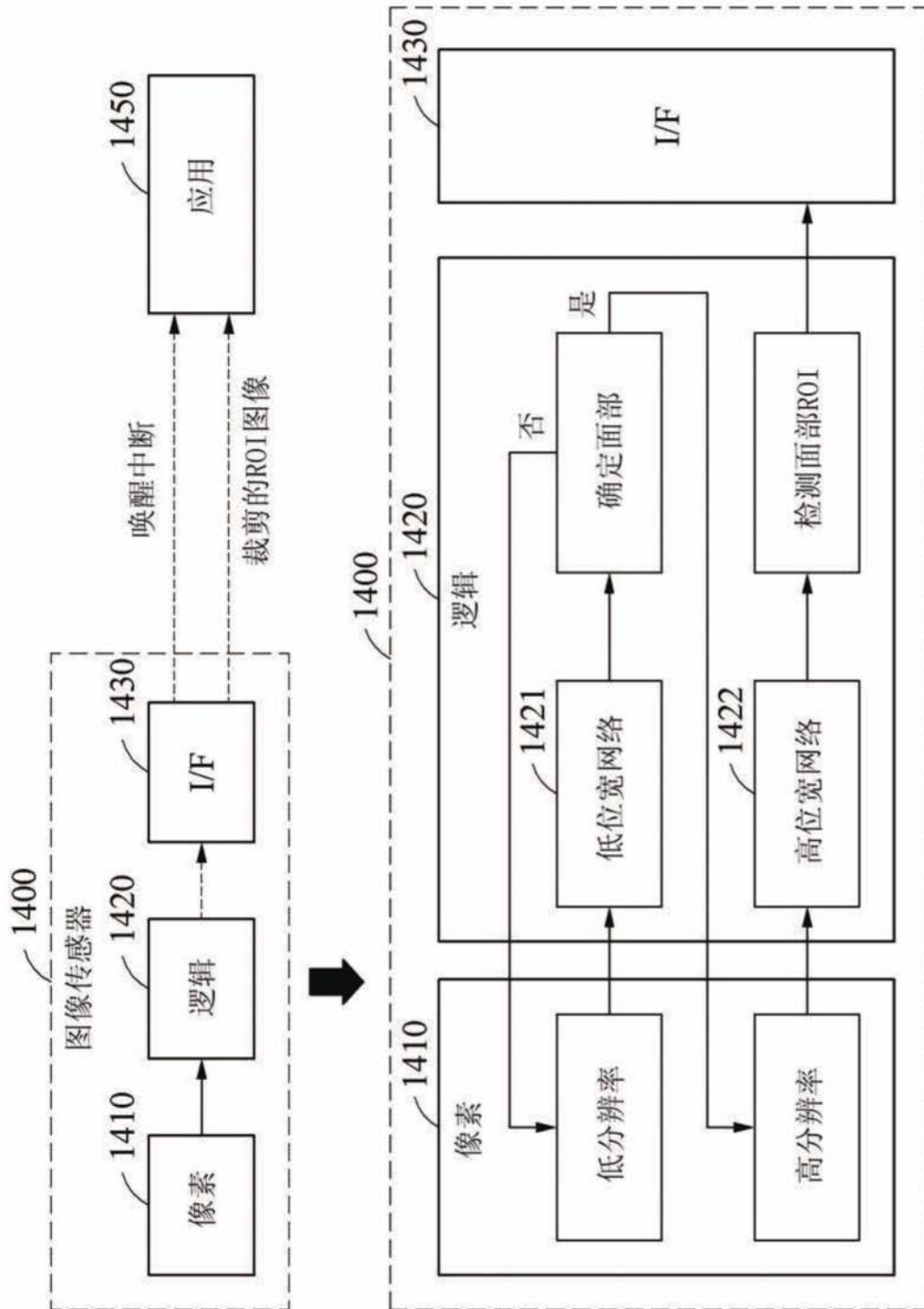


图14

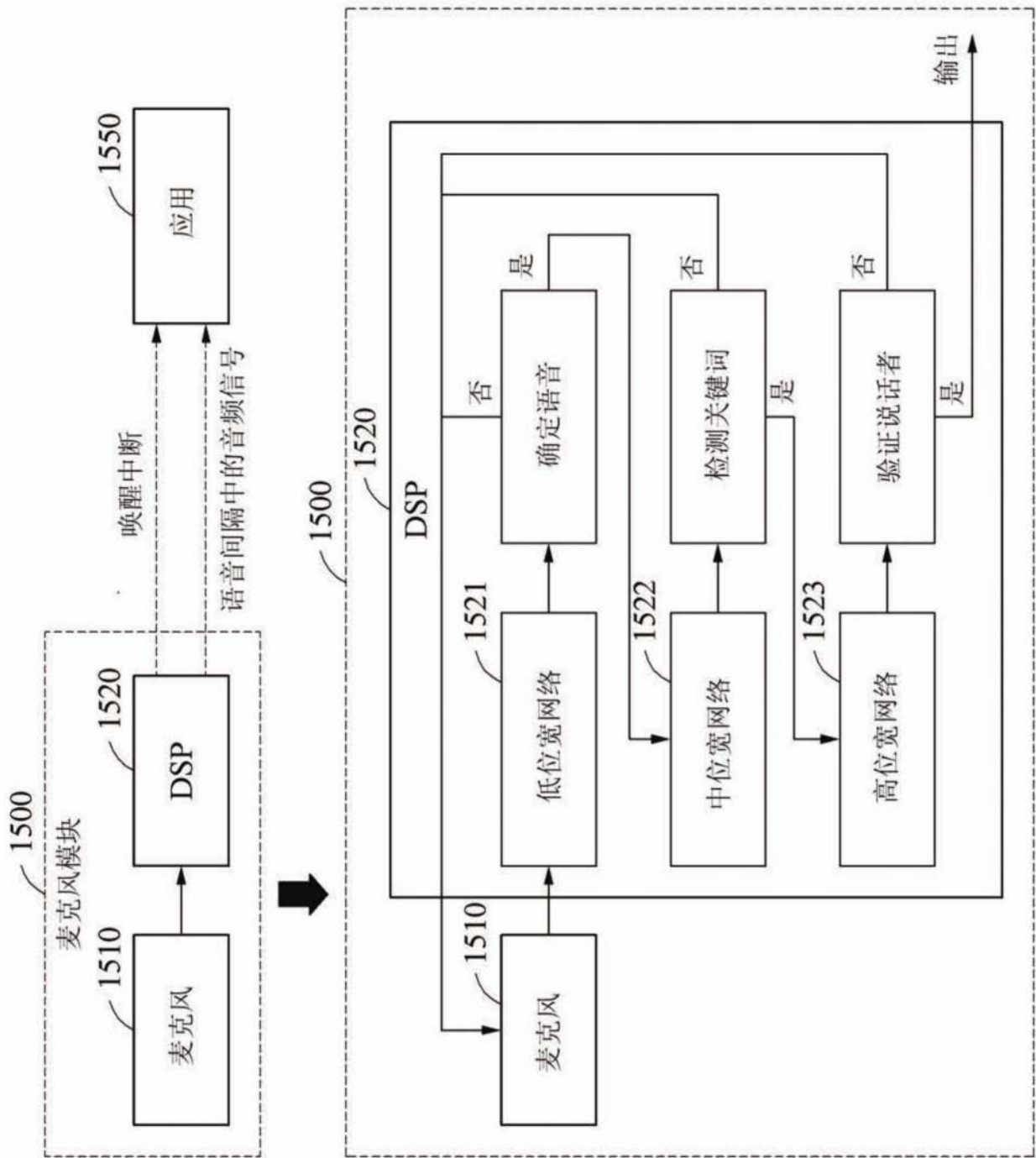


图15

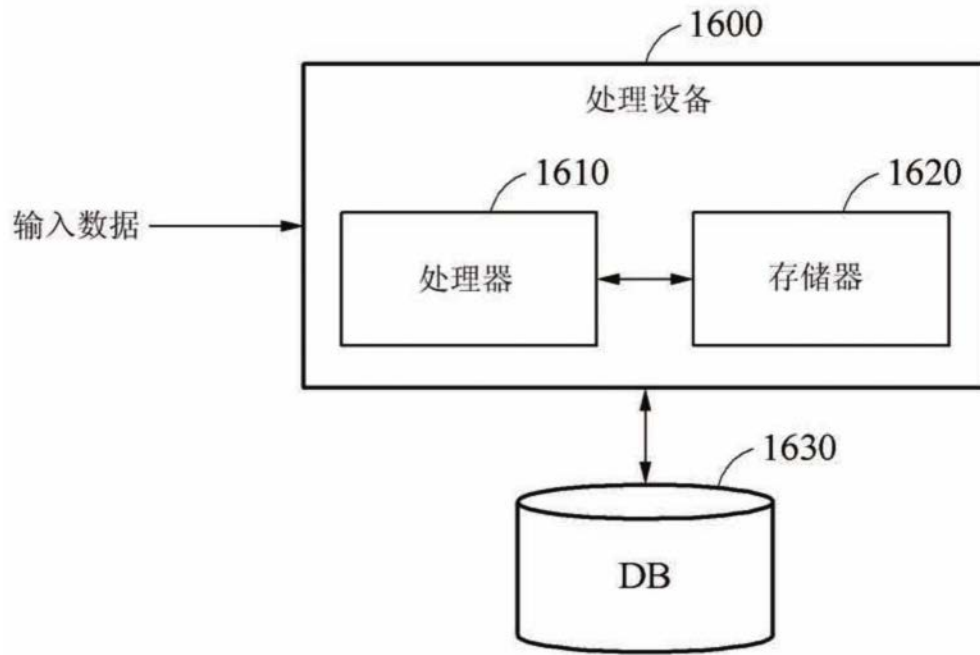


图16

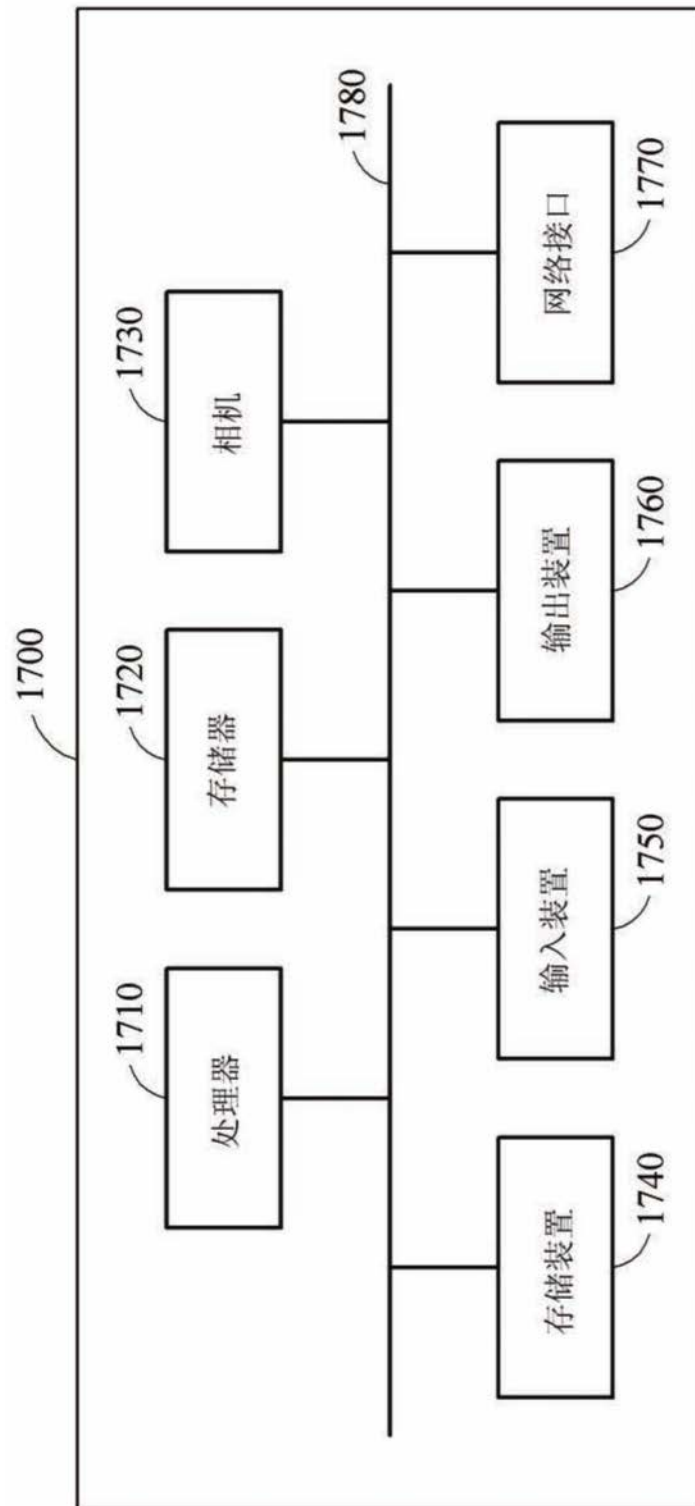


图17

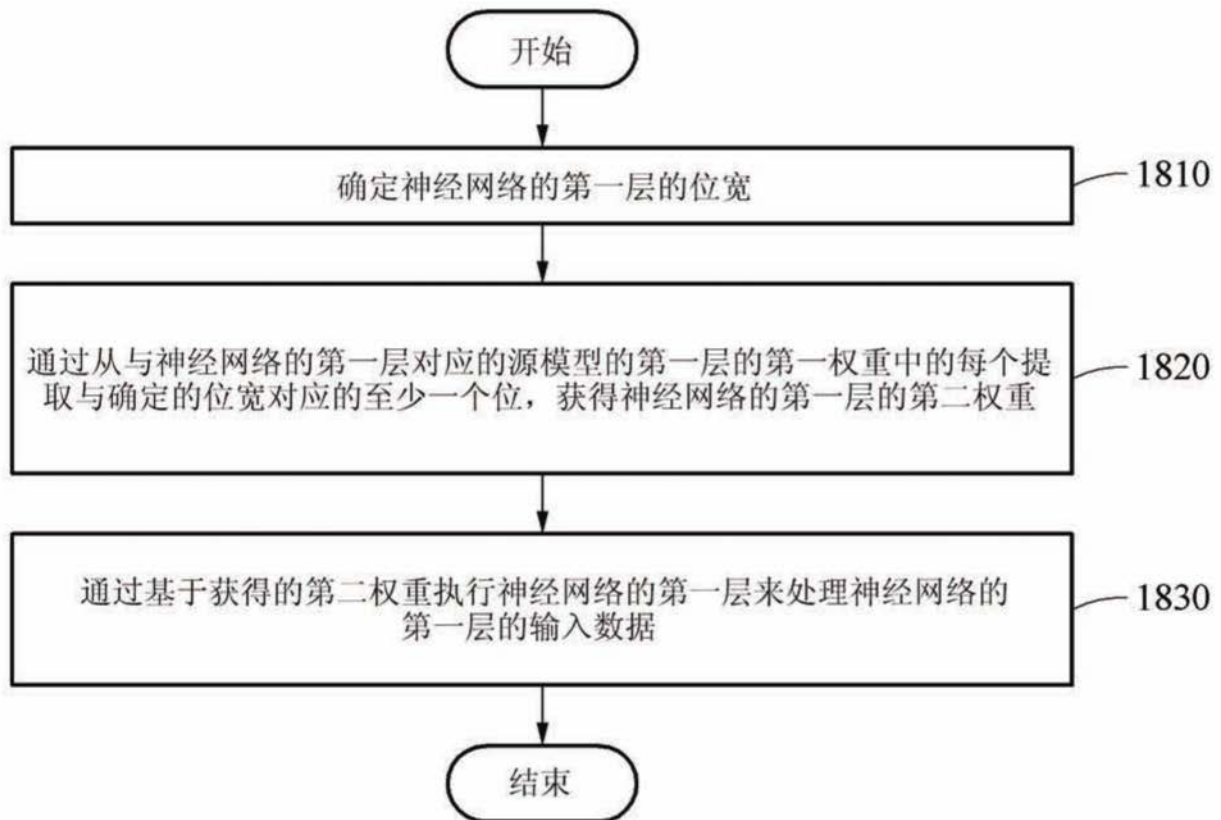


图18

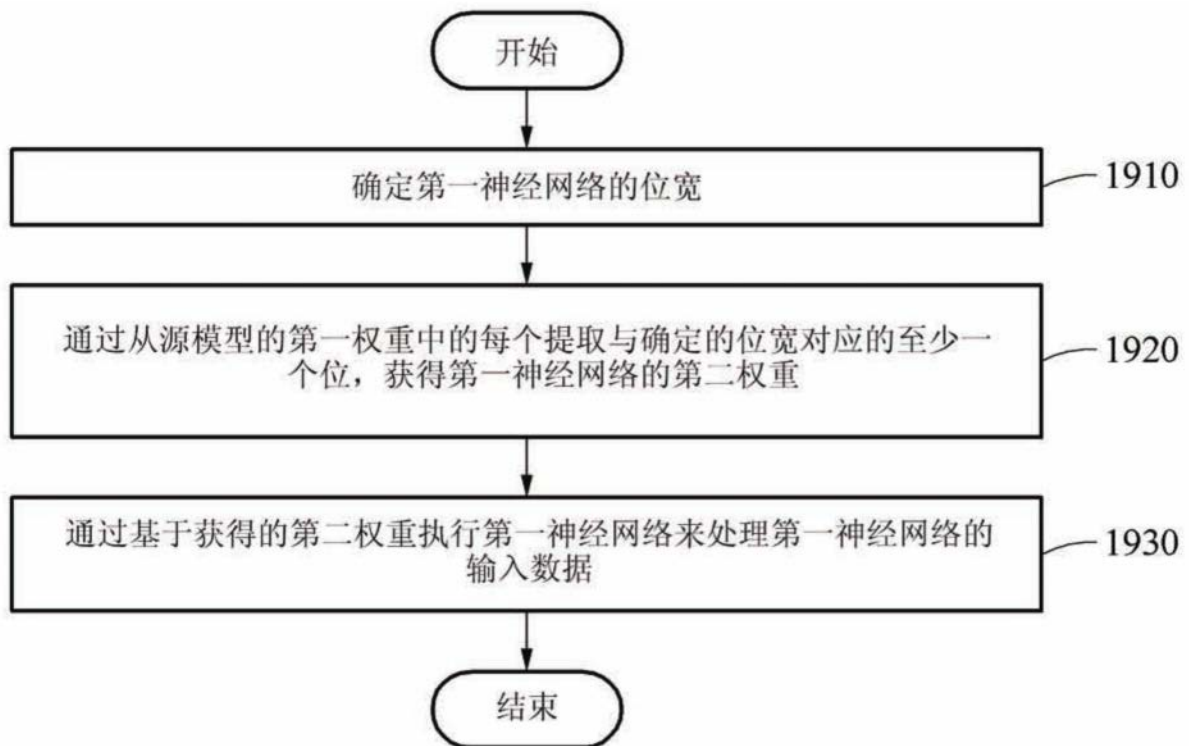


图19

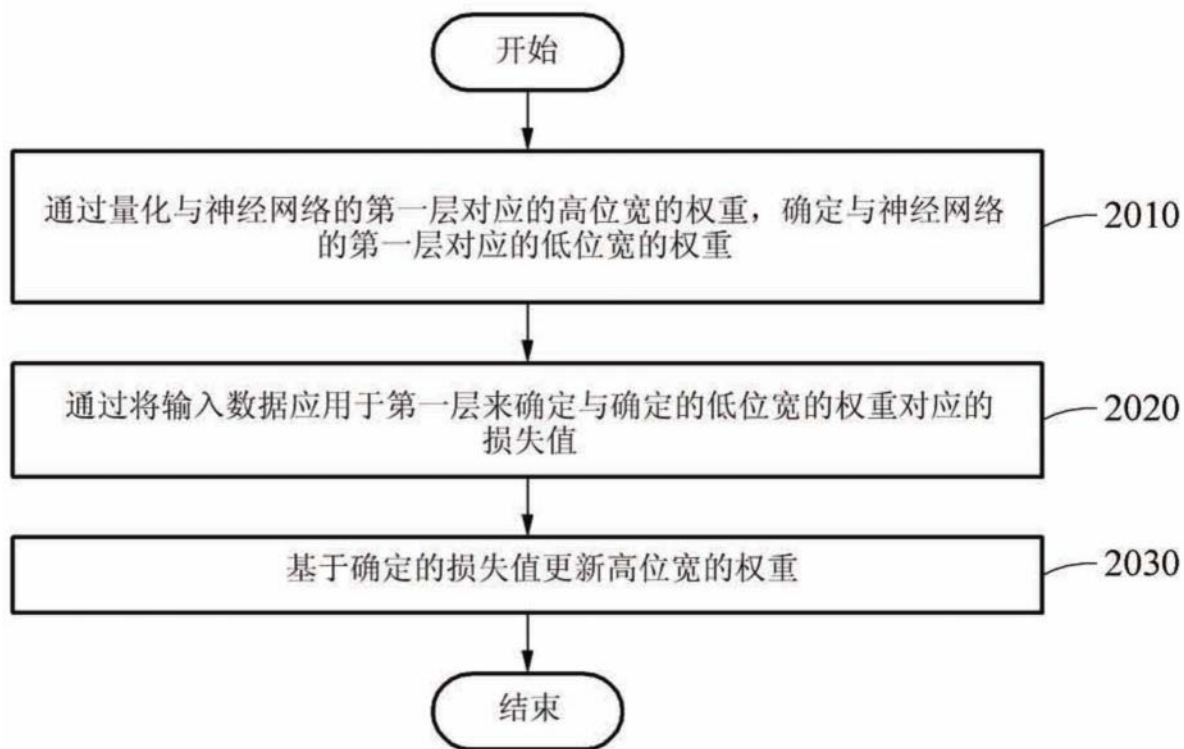


图20