



(21) 申请号 201910822062.3

(22) 申请日 2019.08.30

(65) 同一申请的已公布的文献号
申请公布号 CN 110876062 A

(43) 申请公布日 2020.03.10

(30) 优先权数据

10-2018-0103933 2018.08.31 KR

10-2019-0081288 2019.07.05 KR

(73) 专利权人 三星电子株式会社
地址 韩国京畿道

(72) 发明人 赵仁相 李元宰 黄赞荣

(74) 专利代理机构 中科专利商标代理有限责任
公司 11021

专利代理师 吴晓兵

(51) Int.Cl.

H04N 19/42 (2014.01)

H04N 19/426 (2014.01)

H04N 19/184 (2014.01)

G06N 3/0464 (2023.01)

G06N 3/08 (2023.01)

(56) 对比文件

US 2018189981 A1, 2018.07.05

CN 108062780 A, 2018.05.22

审查员 罗容

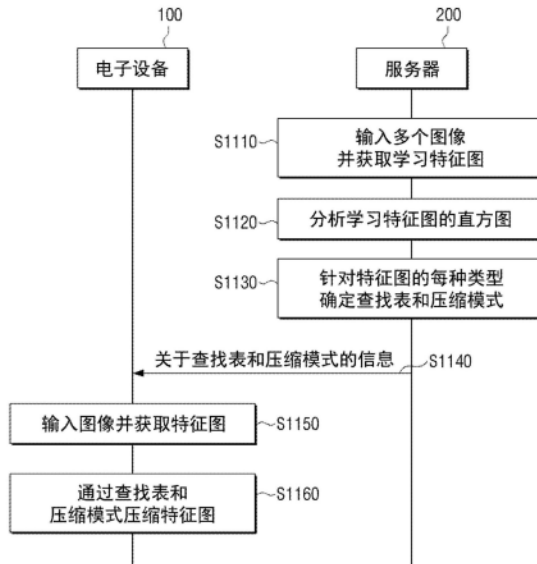
权利要求书2页 说明书19页 附图9页

(54) 发明名称

用于特征图的高速压缩处理的电子设备及其控制方法

(57) 摘要

一种电子设备及其控制方法包括：将输入图像输入到人工智能模型中，获取输入图像的特征图，通过与特征图相对应的查找表转换特征图，以及通过与特征图相对应的压缩模式压缩特征图来存储转换后的特征图。



1. 一种控制电子设备的方法,包括:
将图像作为输入提供到人工智能模型中;
从所述人工智能模型获得所述图像的特征图作为输出;
使用与所述特征图相对应的查找表将所述特征图转换为转换后的特征图,其中所述转换后的特征图的像素的像素值的标准偏差小于所述特征图的像素的像素值的标准偏差;
使用多种压缩模式中与所述转换后的特征图相对应的压缩模式来压缩所述转换后的特征图;以及
在电子设备的存储器中存储经压缩的特征图。
2. 根据权利要求1所述的方法,其中,生成所述查找表,使得对关于与所述特征图的类型相对应的学习特征图的直方图信息进行分析,并且所述特征图中包括的像素值之间的最大残差减小。
3. 根据权利要求2所述的方法,其中,所述查找表是用于将所述特征图转换为所述转换后的特征图的查找表,使得所述特征图中包括的像素的像素值中具有高出现频率的像素值的像素对应于与所述像素的像素值变化范围的中间值接近的值。
4. 根据权利要求2所述的方法,其中,所述压缩包括:
分析与所述特征图的类型相对应的所述学习特征图;
识别多种压缩模式中与所述特征图相对应的压缩模式;以及
根据所述压缩模式压缩所述转换后的特征图,并且
其中,所述存储包括:在报头中存储关于所述压缩模式的信息。
5. 根据权利要求4所述的方法,其中,所述压缩包括:
基于所述压缩模式识别与所述转换后的特征图中包括的多个像素中的每个像素相邻的多个像素中的至少一个相邻像素的值;
通过使用所述至少一个相邻像素的值预测所述多个像素中的每个像素的值;以及
将所述多个像素中的每个像素的值减小与预测值一样的大小。
6. 根据权利要求1所述的方法,还包括:
通过所述压缩模式恢复所述经压缩的特征图的残差;
通过将所述经压缩的特征图的像素的预测值与所述残差相加来获取所述转换后的特征图;
通过所述查找表对所述转换后的特征图进行逆转换来恢复输入图像的特征图;以及
将所恢复的特征图输入到所述人工智能模型中包括的第二层中。
7. 根据权利要求1所述的方法,还包括:
通过以预定数量的像素为单位对所述经压缩的特征图中包括的多个像素进行分组来获取多个像素组;
将像素组中包括的像素被压缩之后比特量最小的像素组识别为报头分组;以及
识别所述报头分组中与像素值的差相对应的比特数量,并且
其中,所述存储包括:基于所述比特数量存储所述经压缩的特征图。
8. 根据权利要求7所述的方法,其中,所述存储包括:
在报头中存储关于所述报头分组的信息。
9. 一种电子设备,包括:

存储器;以及

处理器,被配置为:

基于图像被输入到人工智能模型中,从所述人工智能模型获得所述图像的特征图作为输出;

使用与所述特征图相对应的查找表将所述特征图转换为转换后的特征图,其中所述转换后的特征图的像素的像素值的标准偏差小于所述特征图的像素的像素值的标准偏差;

使用多种压缩模式中与所述转换后的特征图相对应的压缩模式来压缩所述转换后的特征图;以及

在所述存储器中存储经压缩的特征图。

10. 根据权利要求9所述的电子设备,其中,生成所述查找表,使得对关于与所述特征图的类型相对应的学习特征图的直方图信息进行分析,并且所述特征图中包括的像素值之间的最大残差减小。

11. 根据权利要求10所述的电子设备,其中,所述查找表是用于将所述特征图转换为所述转换后的特征图的查找表,使得所述特征图中包括的像素的像素值中具有高出现频率的像素值的像素对应于与所述像素的像素值变化范围的中间值接近的值。

12. 根据权利要求10所述的电子设备,其中,所述处理器还被配置为:

分析与所述特征图的类型相对应的所述学习特征图,

识别多种压缩模式中与所述特征图相对应的压缩模式,

根据所述压缩模式压缩所述转换后的特征图,以及

在报头中存储关于所述压缩模式的信息。

13. 根据权利要求12所述的电子设备,其中,所述处理器还被配置为:

基于所述压缩模式识别与所述转换后的特征图中包括的多个像素中的每个像素相邻的多个像素中的至少一个相邻像素的值,

通过使用所述至少一个相邻像素的值预测所述多个像素中的每个像素的值,以及

将所述多个像素中的每个像素的值减小与预测值一样的大小。

14. 根据权利要求9所述的电子设备,其中,所述处理器还被配置为:

通过所述压缩模式恢复所述存储器中存储的所述经压缩的特征图的残差,

通过将所述经压缩的特征图的像素的预测值与所述残差相加来获取所述转换后的特征图,

通过所述查找表对所述转换后的特征图进行逆转换来恢复输入图像的特征图,以及

将所恢复的特征图输入到所述人工智能模型中包括的第二层中。

15. 根据权利要求9所述的电子设备,其中,所述处理器被配置为:

通过以预定数量的像素为单位对所述经压缩的特征图中包括的多个像素进行分组来获取多个像素组,

将像素组中包括的像素被压缩之后比特量最小的像素组识别为报头分组,

识别所述报头分组中与像素值的差相对应的比特数量,以及

基于所述比特数量在所述存储器中存储所述经压缩的特征图。

用于特征图的高速压缩处理的电子设备及其控制方法

[0001] 相关申请的交叉引用

[0002] 本申请基于并要求于2018年8月31日在韩国知识产权局递交的韩国专利申请No.10-2018-0103933和于2019年7月5日在韩国知识产权局递交的韩国专利申请No.10-2019-0081288的优先权,其全部内容通过引用并入本文。

技术领域

[0003] 本公开涉及一种用于压缩在基于卷积神经网络(CNN)的媒体处理期间生成的多通道特征图图像的电子设备及一种用于控制该电子设备的方法。

背景技术

[0004] 在CNN操作的中间过程中,生成多个通道中的特征图图像。图像具有如下结构:针对每个CNN层级生成多个特征图图像,并且在最后层级处将图像转换为最终结果。当在基于CNN的媒体处理过程期间在存储器中存储或读取特征图图像时,需要较大的传输容量用于图像数据。

[0005] 在诸如图像识别之类的领域中,当实现多层感知器(MLP)或多层神经网络时,MLP中的所有输入具有相同级别的重要性,而不管它们的位置如何。因此,如果通过使用MLP构造全连接神经网络,则存在参数大小变得过大的问题。在常规技术中,通过使用CNN作为解决方案解决了这样的问题。

[0006] 同时,在常规技术中,可以通过将常规的JPEG、JPEG2000、PNG或Lempel-Ziv行程编码方法应用于每个通道的图像以在CNN操作过程期间对各个通道上生成的多个通道的图像进行压缩,来减少用于特征图图像的存储容量。此外,可以通过使用应用于运动图像压缩的基于MPEG的压缩以基于图像通道之间的预测进一步改善压缩性能,来减少用于特征图图像的存储容量,或者可以通过使用3D多级树集合分裂(SPIHT)方法来减少用于特征图图像的存储容量,在该3D SPIHT方法中,用于单个图像的小波压缩方法被广泛应用于多个通道中的图像,以用于压缩卫星图像的多光谱图像。

[0007] 在将常规的图像压缩方法应用于特征图图像的压缩的情况下,可以有效地减少图像的存储容量,但是由于常规压缩方法不是为了在嵌入式系统中操作而实现的算法,因此难以有效地利用常规压缩方法。因此,需要一种在能够在嵌入式系统中实现的级别上的复杂程度内的有效压缩算法。此外,由于常规压缩方法被开发用于有效地压缩一般图像,因此常规压缩方法未被优化用于特征图的压缩。

发明内容

[0008] 本公开提供了一种用于特征图图像的有效压缩方法以及一种用于实现该方法的设备,以减少存储或读取在CNN处理期间生成的特征图图像所需的传输容量。

[0009] 提供了一种例如在嵌入式系统中的特征图的有效压缩处理结构,该结构利用在CNN学习过程期间基于特征图的特征而获取的信息。

[0010] 一种根据本公开的实施例的控制电子设备的方法可以包括:将图像作为输入提供到人工智能模型中;从人工智能模型获得图像的特征图作为输出;使用与特征图相对应的查找表将特征图转换为转换后的特征图,其中转换后的特征图的像素的像素值的标准偏差小于特征图的像素的像素值的标准偏差;使用多种压缩模式中与转换后的特征图相对应的压缩模式来压缩转换后的特征图;以及在电子设备的存储器中存储经压缩的特征图。

[0011] 同时,与特征图相对应的查找表可以是如下生成的查找表,使得对关于与特征图的类型相对应的学习特征图的直方图信息进行分析,并且特征图中包括的像素值之间的最大残差减小。

[0012] 此外,查找表可以是用于将特征图转换为转换后的特征图的查找表,使得特征图中包括的像素的像素值中具有高出现频率的像素值的像素对应于与像素的像素值变化范围的中间值接近的值。

[0013] 同时,压缩和存储特征图的步骤可以包括以下步骤:分析与特征图的类型相对应的学习特征图,识别多种压缩模式中与特征图相对应的压缩模式,根据压缩模式压缩转换后的特征图,并在报头中存储关于压缩模式的信息。

[0014] 此外,压缩特征图的步骤可以包括以下步骤:基于压缩模式识别与转换后的特征图中包括的多个像素中的每个像素相邻的多个像素中的至少一个相邻像素的值,通过使用所述至少一个相邻像素的值来预测所述多个像素中的每个像素的值,并将所述多个像素中的每个像素的值减小与预测值一样的大小。

[0015] 此外,用于控制电子设备的方法还可以包括以下步骤:通过压缩模式恢复经压缩的特征图的残差,通过将经压缩的特征图的像素的预测值与残差相加来获取转换后的特征图,通过查找表对转换后的特征图进行逆转换来恢复输入图像的特征图,并将所恢复的特征图输入到人工智能模型中包括的第二层中。

[0016] 另外,用于控制电子设备的方法还可以包括以下步骤:通过以预定数量的像素为单位对经压缩的特征图中包括的多个像素进行分组来获取多个像素组,将像素组中包括的像素被压缩之后比特量最小的像素组识别为报头分组,确定报头分组中与像素值的差相对应的比特数量,并且基于比特数量存储经压缩的特征图。

[0017] 此外,预定数量的像素可以是 4×2 像素块,并且在存储经压缩的特征图的步骤中,可以将关于报头分组的信息存储在报头中。

[0018] 同时,一种根据本公开的实施例的电子设备的可以包括:存储器;以及处理器,被配置为:基于图像被输入到人工智能模型中,从人工智能模型获得图像的特征图作为输出;使用与特征图相对应的查找表将特征图转换为转换后的特征图,其中转换后的特征图的像素的像素值的标准偏差小于特征图的像素的像素值的标准偏差;使用多种压缩模式中与转换后的特征图相对应的压缩模式来压缩转换后的特征图;以及在存储器中存储经压缩的特征图。

[0019] 另外,处理器可以分析与特征图的类型相对应的学习特征图,识别多种压缩模式中与特征图相对应的压缩模式,根据压缩模式压缩转换后的特征图,并在报头中存储关于压缩模式的信息。

[0020] 此外,处理器可以基于压缩模式识别与转换后的特征图中包括的多个像素中的每个像素相邻的多个像素中的至少一个相邻像素的值,通过使用所述至少一个相邻像素的值

来预测所述多个像素中的每个像素的值,并将所述多个像素中的每个像素的值减小与预测值一样的大小。

[0021] 另外,处理器可以通过压缩模式恢复存储在存储器中的经压缩的特征图的残差,通过将经压缩的特征图的像素的预测值与残差相加来获取转换后的特征图,通过查找表对转换后的特征图进行逆转换来恢复输入图像的特征图,并将所恢复的特征图输入到人工智能模型中包括的第二层中。

[0022] 此外,处理器可以通过以预定数量的像素为单位对经压缩的特征图中包括的多个像素进行分组来获取多个像素组,将像素组中包括的像素被压缩之后比特量最小的像素组识别为报头分组,确定报头分组中与像素值的差相对应的比特数量,并且基于比特数量将经压缩的特征图存储在存储器中。

[0023] 此外,预定数量的像素可以是 4×2 像素块,并且处理器可以在存储器中将关于报头分组的信息存储在报头中。

[0024] 同时,一种根据本公开的实施例的用于服务器训练人工智能模型的方法可以包括以下步骤:将多个学习图像作为输入提供到待训练的人工智能模型中,从人工智能模型获取所述多个学习图像的多个特征图作为输出,通过分析所述多个特征图的直方图来识别所述多个特征图的类型,生成与所述多个特征图的类型相对应的查找表,识别多种压缩模式中与所述多个特征图的类型相对应的压缩模式,并且将关于查找表和压缩模式的信息发送给外部设备。

[0025] 此外,与特征图的类型相对应的查找表可以是与所述多个特征图的类型相对应的查找表,其被如下生成:使得所述多个特征图中包括的多个像素的像素值之间的最大残差减小,并且与特征图的类型相对应的压缩模式可以是被识别用于压缩所述多个特征图的压缩模式,使得对所述多个特征图进行压缩之后的比特量在所述多种压缩模式中最小。

[0026] 另外,根据本公开的实施例的服务器可以包括:通信器;以及处理器,被配置为:基于多个学习图像作为输入被提供给待训练的人工智能模型,从人工智能模型获取所述多个学习图像的多个特征图作为输出,通过分析所述多个特征图的直方图来识别所述多个特征图的类型,生成与所述多个特征图的类型相对应的查找表,识别多种压缩模式中与所述多个特征图的类型相对应的压缩模式,并且控制通信器将关于查找表和压缩模式的信息发送给外部设备。

[0027] 同时,与特征图的类型相对应的查找表可以是如下生成的查找表:使得所述多个特征图中包括的多个像素的像素值之间的最大残差减小,并且与特征图的类型相对应的压缩模式可以是被识别用于压缩所述多个特征图的压缩模式,使得对所述多个特征图进行压缩之后的比特量在所述多种压缩模式中最小。

[0028] 通过本公开,可以有效地压缩电子设备中在利用基于人工智能模型的应用或嵌入式系统的过程期间所生成的大量特征图数据。

[0029] 因此,通过本公开,可以通过压缩数据和减少的操作时间,在诸如用户终端设备之类的对设备内的传输容量具有限制的电子设备中有效地利用人工智能模型。

附图说明

[0030] 图1是示出了根据本公开的实施例的用于使用人工智能模型的包括电子设备和服

务器的系统的示图；

[0031] 图2是示出了根据本公开的实施例的对输入到基于CNN的人工智能模型中的图像进行编码和解码的处理的示图；

[0032] 图3是示出了根据本公开的实施例的电子设备的配置的框图；

[0033] 图4是示出了根据本公开的实施例的电子设备的配置的框图；

[0034] 图5是示出了根据本公开的实施例的服务器的配置的框图；

[0035] 图6是示出了根据本公开的实施例的在训练人工智能模型的过程中生成查找表并确定压缩模式的方法的流程图；

[0036] 图7A是示出了根据本公开的实施例的在通过查找表转换特征图之前的直方图的曲线图；

[0037] 图7B是示出了根据本公开的实施例的在通过查找表转换特征图之后的直方图的曲线图；

[0038] 图8A示出了根据实施例的查找表；

[0039] 图8B示出了根据本公开的实施例的在训练人工智能模型的过程中确定压缩模式的示图；

[0040] 图9是示出了根据本公开的实施例的通过使用查找表和压缩模式转换并压缩输入图像的方法的流程图；

[0041] 图10是示出了根据本公开的实施例的通过使用查找表和压缩模式转换并压缩输入图像的处理的示图；

[0042] 图11是示出了根据本公开的实施例的通过包括服务器和电子设备的系统压缩特征图的方法的序列图；

[0043] 图12A是示出了根据本公开的实施例的对特征图进行编码和解码的方法的流程图；

[0044] 图12B是示出了根据本公开的实施例的对特征图进行编码和解码的处理的示图；

[0045] 图12C是示出了根据本公开的实施例的对特征图进行编码和解码的方法的流程图；

[0046] 图13A是示出了根据本公开的实施例的针对多个图像使用特定查找表和特定压缩模式的压缩率的结果的示图；以及

[0047] 图13B是示出了根据本公开的实施例的针对多个图像使用特定查找表和特定压缩模式的压缩率的结果的示图。

具体实施方式

[0048] 在下文中,将参考附图描述本公开的各种实施例。各种实施例不是用于将本公开中描述的技术限制于特定实施例,而是它们应被解释为包括本公开中描述的实施例的各种修改、等同和/或替代。此外,关于附图的详细描述,类似的组件可以由类似的附图标记指定。

[0049] 此外,本公开中使用的表达“第一”、“第二”等可以用于描述各种元件,而与任何顺序和/或重要程度无关。这些表达仅用于将一个元件与另一个元件区分开,而不旨在对元件进行限制。例如,第一用户设备和第二用户设备可以指代彼此不同的用户设备,而与任何顺

序或重要程度无关。因此,在不脱离本公开的范围的情况下,第一元件可以被称为第二元件,以类似的方式,第二元件可以被称为第一元件。

[0050] 另外,本公开中关于一个元件(例如,第一元件)“(可操作地或通信地)与另一元件(例如,第二元件)耦接”或“连接到”另一元件(例如,第二元件)的描述应被解释为包括以下两种情况:该一个元件直接耦接到该另一元件,以及该一个元件通过另一个中间元件(例如,第三元件)或附加的中间元件耦接到该另一元件。相反,关于一个元件(例如,第一元件)与另一元件(例如,第二元件)“直接耦接”或“直接连接”的描述应被解释为使得所述元件之间不存在或未介入另一个元件(例如,第三元件)。

[0051] 同时,选择本公开中使用的术语以解释本公开的某些实施例,并不旨在限制其他实施例的范围。此外,单数表达可以包括复数表达,除非在上下文中明显不同地限定。本公开中使用的术语,包括技术或科学术语,可以具有与本公开中描述的本领域普通技术人员通常已知的含义相同的含义。本文所使用的术语中在一般词典中定义的术语可以被解释为具有与相关技术中的上下文含义相同的含义或相似的含义。除非另有定义,否则本文所使用的术语不可被解释为具有理想的或过于形式的含义。在一些情况下,即使在本文中定义的术语也不可被解释为排除本文的实施例。

[0052] 在下文中,将参考附图详细描述本公开的各种实施例。

[0053] 图1是示出了根据本公开的实施例的用于使用人工智能模型的包括电子设备和服务器的系统的示意图。

[0054] 参考图1,服务器200可以通过将多个图像作为输入提供到人工智能模型中,来训练基于卷积神经网络(CNN)的人工智能模型。服务器200可以在训练人工智能模型的过程期间获取多个图像中的每个图像的特征图。此外,服务器200可以分析关于所获取的多个特征图的特征(例如,根据特征图类型的规则图案),并且可以确定用于基于特征图的特征或特性有效地压缩特征图的查找表(LUT)和压缩模式(PMODE)。然后,服务器200可以向电子设备100发送与根据多个图像的特征图的每种类型而具有良好压缩效率的LUT和PMODE有关的信息。

[0055] 这里,LUT可以是用于将特征图中包括的多个像素值中的每个像素值转换为不同值所参考的表。作为示例,在8比特像素数据的情况下,LUT可以包括用于将从0到255的每个像素值转换为不同值的信息。这里,LUT可以包括关于转换值的信息,使得从0到255的每个像素值的转换值以1:1的关系对应于从0到255的每个像素值,以使这些值不彼此重叠。

[0056] 作为示例,LUT可以提供如下关联:该关联用于将特征图中包括的多个像素中具有高频率的像素值转换为所述多个像素的像素值变化范围的中间值。然而,LUT不限于此,并且它们可以以各种形式实现以用于有效压缩。例如,LUT可以包括关于以1:1的关系与每个像素值相对应的转换值的信息,或者包括关于用于将像素值转换为不同值的特定数学公式的信息。将参考图7A和图7B在这方面进行详细解释。

[0057] 同时,PMODE可以用于分析与特征图的类型相对应的学习特征图,并且用于压缩特征图,使得特征图上的每个像素的值在多种压缩模式中最小化。然而,PMODE不限于此,并且它们可以以各种形式确定以用于有效压缩。电子设备100可以是用于驱动基于CNN的人工智能模型的电子设备。电子设备100中基于CNN的人工智能模型可以通过对在服务器200处输入的多个图像进行学习而确定的人工智能模型。

[0058] 作为示例,电子设备100可以驱动样式传递应用,该样式传递应用通过一系列图像处理将输入图像转换为类似于名作绘画样式的图像。这里,当用于图像识别的图像被输入到电子设备100中基于CNN的人工智能模型(应用)中时,电子设备100可以通过人工智能模型的第一层获得输入图像的特征图。此外,这里,电子设备100可以转换和压缩特征图,并将特征图存储在存储器中,从而可以防止在将特征图存储在存储器中以及读取特征图的过程中瓶颈现象的发生。

[0059] 根据本公开的实施例的电子设备100可以从服务器200接收关于LUT和PMODE的信息,并且将该信息存储在存储器中,其中所述LUT和PMODE根据在学习多个图像的过程期间所获取的特征图的每种类型。此外,电子设备100可以确定输入图像的特征图的类型,并基于存储在存储器中的LUT和PMODE的信息来转换和压缩输入图像的特征图,如图2所示。电子设备100可以将针对设备通过自学习而获取的每种类型的特征图的LUT和PMODE的信息存储在存储器中。

[0060] 图1示出了实现为电视的电子设备100的示例。然而,电子设备100不限于此,并且电子设备100可以被实现为各种类型的电子设备,包括服务器200和能够与外部设备发送和接收数据的通信装置。例如,电子设备100可以被实现为诸如智能电话和智能电视之类的包括显示器的电子设备,或者实现为包括在另一电子设备中并执行功能的芯片的形式。

[0061] 图2是示出了根据本公开的实施例的对输入到基于CNN的人工智能模型中的图像进行编码和解码的处理的示意图。

[0062] 参考图2,当图像被输入到电子设备100中基于CNN的应用中时,电子设备100可以通过CNN中包括的第一层101获得输入图像的特征图102。这里,第一层101可以是如下卷积层,该卷积层通过使用基于对多个图像进行学习的结果而设置的权重或卷积滤波器来生成输入图像的特征图。

[0063] 根据本公开的实施例,在人工智能模型是包括10个层级的CNN并且针对一个层级生成128个特征图的情况下,对于一个输入图像,电子设备100需要存储或读取1280(10个层级 $\times$ 128个通道)个特征图。如上所述,在电子设备100中的人工智能模型被构造为嵌入式系统的情况下,需要通过压缩特征图来减少传输容量和存储容量。

[0064] 为了通过压缩来存储多个特征图,电子设备100可以通过从服务器200发送的LUT和PMODE来转换和压缩(编码)所生成的特征图。参考图2,电子设备100可以在存储器中存储通过编码而具有减小的数据大小的特征图,并且如果需要,可以对存储在存储器中的特征图进行解码。这里,存储器可以是电子设备100中包括的存储器,或者是外部设备或外部服务器中包括的存储器。

[0065] 同时,电子设备100可以对所存储的特征图进行解码,所述特征图在通过在编码期间使用的LUT和PMODE进行压缩的同时被存储,并且因此,电子设备100可以恢复编码之前的特征图103。此外,电子设备100可以将特征图103输入到人工智能模型中包括的第二层104中,并且基于CNN执行处理。第二层104可以是用于对特征图进行子采样的池化层、或ReLU层、或另一卷积层,然而第二层104不限于此。通过将输入图像输入到第一层101中所获取的特征图102和通过对存储在存储器中的经编码的特征图进行解码所获取的特征图103可以是相同的。根据本公开的实施例,由于基于LUT和PMODE对特征图进行编码和解码,电子设备100可以执行无损压缩。

[0066] 根据前述实施例,即使生成了大量特征图,电子设备100也可以在基于CNN的应用或基于人工智能的嵌入式系统中以相对小的容量执行操作。

[0067] 图3是示出了根据本公开的实施例的电子设备的配置的框图。参考图3,电子设备100可以包括存储器110和处理器120。

[0068] 存储器110可以在电子设备100中存储基于CNN的应用或基于人工智能的嵌入式系统。此外,存储器110可以存储各种类型的信息,例如关于在通过CNN执行操作时生成的特征图的信息、关于用于压缩特征图的多种压缩模式的信息、关于先前输入的图像的信息等。另外,存储器110可以存储从外部服务器200接收的关于与特征图的每种类型相对应的LUT和PMode的信息。同时,这仅是示例,并且电子设备100可以在存储器110中存储关于如下LUT和PMode的信息,所述LUT和PMode与设备通过基于多个样本图像的学习而获取的各种类型的特征图的每种类型相对应。

[0069] 此外,存储器110可以存储用于电子设备100的整体操作的各种类型的数据,例如用于处理器120的处理或控制的程序等。此外,存储器110可以存储用于在电子设备100处驱动的一个或多个应用程序(应用程序或应用)的操作以及电子设备100的操作的数据和指令。可以通过无线通信从外部服务器下载这样的应用程序中的至少一些。此外,从制造商发布产品的时间起,电子设备100中可以存在这样的应用程序中的至少一些,用于电子设备100的基本功能。此外,这样的应用程序可以存储在存储器110中,并且可以在处理器120的控制下被执行以执行电子设备100的操作(或功能)。

[0070] 存储器110可以被实现为非易失性存储器、易失性存储器、闪存、硬盘驱动器(HDD)或固态硬盘(SSD)等。存储器110可以由处理器120访问,并且可以执行处理器120对数据的读取/记录/校正/删除/更新等。在本公开中,术语“存储器”可以包括处理器120内部的存储器110、ROM和RAM、或者安装在电子设备100上的存储卡(例如,微型SD卡、记忆棒、USB存储器等)。

[0071] 处理器120通过执行存储在存储器110中的各种类型的程序来控制电子设备100的整体操作。

[0072] 当输入图像被输入到人工智能模型中时,处理器120可以通过人工智能模型中包括的第一层获取图像的特征图,并通过与特征图相对应的查找表转换特征图。

[0073] 此外,根据本公开的实施例的处理器120可以通过多种压缩模式中与特征图相对应的压缩模式来压缩转换后的特征图,并将特征图存储在存储器110中。例如,处理器120可以基于从服务器200接收的关于与特征图相对应的压缩模式的信息,对通过查找表转换的特征图执行压缩。作为另一示例,处理器120可以基于多种压缩模式中的每种压缩模式对转换后的特征图执行压缩,并且基于具有最佳压缩率的压缩模式来压缩转换后的特征图,并将特征图存储在存储器110中。作为又一示例,从服务器200接收的关于与特征图相对应的压缩模式的信息可以指示多种压缩模式中的第一压缩模式和第二压缩模式与特征图相对应。处理器120可以识别第一压缩模式和第二压缩模式之间具有相对高压缩率的压缩模式,并且基于所识别的压缩模式压缩转换后的特征图。

[0074] 根据本公开的实施例的处理器120可以基于为根据压缩模式压缩转换后的特征图而转换的特征图所包含的多个像素中与一像素相邻的像素中的至少一个像素的值来预测该像素的值,并用排除了预测值的残差替换该像素的值。这里,较小的残差可以意味着较

高的压缩率。

[0075] 此外,根据本公开的实施例的处理器120可以通过基于用于对存储在存储器110中的经压缩的特征图进行解码的压缩模式将预测值与残差相加,来恢复该像素的值,并通过经由反向查找表对转换后的特征图进行逆转换来释放压缩。另外,处理器120可以将恢复的特征图输入到人工智能模型中包括的第二层中。下面将详细描述该处理。

[0076] 此外,根据本公开的实施例的处理器120可以在报头中存储关于用于压缩特征图的压缩模式的信息。例如,处理器120可以通过以预定数量的像素为单位对经压缩的特征图中包括的多个像素进行分组来获取多个像素组,识别所述多个像素组中在组内压缩之后比特量最小的像素组作为报头分组,确定报头分组中与像素值的差相对应的比特数量,并且基于该比特数量将经压缩的特征图存储在存储器110中。下面也将详细描述该处理。

[0077] 图4是示出了根据本公开的实施例的电子设备的配置的框图。

[0078] 参考图4,电子设备100可以包括存储器110、处理器120、通信器130、显示器140、接口150和相机160。然而,本公开不限于此,电子设备100可以包括各种组件。将省略关于存储器110和处理器120的冗余解释。

[0079] 处理器120包括RAM 121、ROM 122、图形处理器123、主CPU 124以及第一接口125-1至第n接口125-n。这里,RAM 121、ROM 122、图形处理器123(图形处理单元GPU)、主CPU 124、第一接口125-1至第n接口125-n等可以通过至少一条总线126彼此连接。

[0080] ROM 122存储用于系统启动的指令集等。当向电子设备100供电时,主CPU 124根据存储在ROM 122中的指令将存储在存储器110中的操作系统(O/S)复制到RAM 121中,并通过执行O/S来启动系统。当启动完成时,主CPU 124将存储在存储器110中的各种类型的应用程序复制到RAM 121中,并通过执行复制到RAM 121中的应用程序来执行各种类型的操作。

[0081] 根据本公开的实施例,处理器120可以被实现为处理数字信号的数字信号处理器(DSP)和微处理器。然而,本公开不限于此,处理器120可以包括以下项中的一个或多个:中央处理单元(CPU)、微控制器单元(MCU)、微处理单元(MPU)、控制器、应用处理器(AP)或通信处理器(CP)以及ARM处理器。此外,处理器120可以被实现为其中存储有处理算法的片上系统(SoC)或大规模集成电路(LSI),或者以现场可编程门阵列(FPGA)的形式实现。

[0082] 通信器130是用于与外部服务器200通信的通信接口。通信器130可以从外部服务器200接收与根据多个图像的特征图的每种类型而具有良好压缩效率的LUT和PMODE有关的信息。然而,这仅是示例,通信器130可以从外部服务器200接收各种类型的信息,并且可以与除外部服务器200之外的各种类型的外部设备通信。

[0083] 同时,通信器130可以包括用于以下至少一种通信方法的通信模块:无线保真(WiFi)、蓝牙(BT)、近场通信(NFC)、全球定位系统(GPS)、以及通过载波侦听多址接入/冲突检测(CSMA/CD)方法的以太网或蜂窝通信(例如,4G(第四代)、LTE、LTE-A、5G、CDMA、WCDMA、UMTS、Wibro或GSM等)。此外,通信器130可以被实现为与如上所述的至少一个通信模块相对应的多个通信部件。

[0084] 具体地,WiFi芯片和蓝牙芯片可以分别通过WiFi方法和蓝牙方法执行通信。在使用WiFi芯片或蓝牙芯片的情况下,可以预先发送和接收各种类型的连接信息,例如SSID或会话密钥,并且通过使用该信息建立通信,然后可以发送和接收各种类型的信息。无线通信芯片意指根据各种通信标准(例如,IEEE、Zigbee、第三代(3G)、第三代合作伙伴计划

(3GPP)、长期演进 (LTE) 等) 执行通信的芯片。同时, NFC芯片意指在近场通信 (NFC) 方法中使用诸如135kHz、13.56MHz、433MHz、860~960MHz和2.45GHz之类的各种RF-ID频带中的13.56MHz频带进行操作的芯片。

[0085] 显示器140可以显示在图形处理器123处生成的用于根据处理器120的控制来执行功能的各种屏幕。显示器140可以显示要输入到电子设备100中包括的人工智能模型中的图像, 并且显示器140还可以显示通过人工智能模型向图像添加各种效果的结果。例如, 显示器140可以显示关于对输入图像执行样式传递的结果的图像。

[0086] 显示器140可以被实现为液晶显示器 (LCD) 面板、有机发光二极管 (OLED) 等, 但是显示器140不限于此。另外, 根据电子设备100的设计和实现方式, 显示器140可以被实现为柔性显示器、透明显示器等。

[0087] 此外, 根据本公开的实施例的显示器140不仅可以包括输出图像的显示面板, 还可以包括容纳显示面板的边框。具体地, 根据本公开的实施例的边框可以包括用于检测用户交互的触摸检测器。

[0088] 显示器140可以具有各种尺寸。例如, 显示器140可以具有用于指示显示器140的可视区域的尺寸, 诸如3英寸、4英寸、4.65英寸、5英寸、6.5英寸、8.4英寸、32英寸、45英寸等, 如通常测量的那样。此外, 显示器140可以包括多个像素。这里, 宽度×长度形式的多个像素的数量可以表示为显示器140的分辨率。

[0089] 此外, 显示器140可以被实现为各种形式的显示面板。例如, 显示面板可以实现为各种类型的显示技术, 例如液晶显示器 (LCD)、有机发光二极管 (OLED)、有源矩阵有机发光二极管 (AM-OLED)、硅上液晶 (LcoS) 或数字光处理 (DLP) 等。另外, 显示器140可以以柔性显示器的形式耦接到电子设备100的前表面区域、侧表面区域或后表面区域中的至少一个。

[0090] 显示器140可以耦接到触摸检测器, 并且可以被实现为分层结构的触摸屏。触摸屏可以具有显示功能, 并且还可以具有检测触摸输入的压力或电容变化以及触摸输入和触摸区域的位置的功能。此外, 显示器140可以具有检测接近触摸以及真实触摸的功能。

[0091] 接口150是用于将外部设备连接到电子设备100的组件, 用于输入和输出视频和音频。例如, 接口150可以包括以下至少一项: 通用串行总线 (USB)、高清多媒体接口 (HDMI)、推荐标准232 (RS-232)、普通老式电话服务 (POTS) DVI、显示器端口 (DP) 或雷电接口 (thunderbolt), 它们是有用于有线通信的组件。HDMI是能够为输出音频和视频的AV设备传输高性能数据的接口。DP是能够实现具有超高分辨率 (例如, 2560×1600 或 3840×2160) 的屏幕以及 1920×1080 的全高清级别的屏幕和3D立体图像的接口, 并且其还能够传输数字声音。雷电接口是用于高速传输和连接数据并且能够将PC、显示器、存储设备等与一个端口并联连接的输入和输出接口。

[0092] 然而, 前述输入和输出端子仅是示例, 并且除了前述端子之外, 接口150还可以包括仅用于输出音频信号的端口或者仅用于输出视频信号的端口。

[0093] 相机160是用于根据用户的控制拍摄图像的成像设备。具体地, 相机160可以拍摄各种类型的图像, 用于拍摄要输入到电子设备100中包括的基于CNN的应用或基于人工智能的嵌入式系统中的图像。

[0094] 图5是示出了根据本公开的实施例的服务器的配置的框图。参考图5, 服务器200可以包括存储器210、通信器220和处理器230。

[0095] 存储器210可以存储待训练的人工智能模型(例如,CNN等)。此外,存储器210可以存储各种类型的信息,例如关于在执行通过CNN的操作时生成的特征图的信息、关于用于压缩特征图的多种压缩模式的信息、关于先前输入的图像的信息等。另外,存储器210可以存储关于与对应于学习图像的学习特征图的每种类型相对应的LUT和PMODE的信息。

[0096] 此外,存储器210可以存储用于服务器200的整体操作的各种类型的数据,例如用于处理器230的处理或控制的程序等。此外,存储器210可以被实现为非易失性存储器、易失性存储器、闪存、硬盘驱动器(HDD)或固态驱动器(SSD)等。

[0097] 通信器220是用于与电子设备100通信的通信接口。通信器220可以向电子设备100发送与根据多个图像的特征图的每种类型而具有良好压缩效率的LUT和PMODE有关的信息。然而,这仅是示例,通信器220可以向电子设备100发送各种类型的信息,并且可以与除电子设备100之外的各种类型的外部设备通信。

[0098] 同时,通信器220可以包括用于以下至少一种通信方法的通信模块:无线保真(WiFi)、蓝牙(BT)、近场通信(NFC)、全球定位系统(GPS)、以及通过载波侦听多址接入/冲突检测(CSMA/CD)方法的以太网或蜂窝通信(例如,LTE、LTE-A、CDMA、WCDMA、UMTS、Wibro或GSM等)。此外,通信器220可以被实现为与如上所述的至少一个通信模块相对应的多个通信组件。

[0099] 处理器230通过执行存储在存储器210中的各种类型的程序来控制服务器200的整体操作。

[0100] 处理器230可以生成用于转换特征图的查找表,以有效地压缩在多个学习图像被输入人工智能模型中时获取的特征图。此外,处理器230可以确定多种压缩模式中用于有效地压缩每种类型的转换后的特征图的压缩模式。

[0101] 具体而言,处理器230可以分析学习特征图的直方图信息,并根据特征图中包括的多个像素来识别像素值的频率数。例如,在8比特像素数据的情况下,处理器230可以根据特征图中包括的多个像素来识别从0到255的每个像素值的频率数。

[0102] 根据本公开的实施例的处理器230可以生成如下查找表,该查找表用于将具有最高频率的像素值转换为0到255之间的特定值。例如,如果具有最高频率的像素值是0,则像素值0可以被转换为127。然后,处理器230可以生成用于将其余像素值转换为其他值的查找表。根据本公开的实施例的查找表可以包括关于以1:1的关系与从0到255的每个像素值相对应的转换值的信息。例如,查找表可以包括关于1:1对应关系的信息,使得像素值0被转换为127,像素值1被转换为126,并且像素值255被转换为128。同时,这仅是示例,并且查找表不限于此。例如,查找表可以包括关于用于将像素值转换为另一值的特定数学公式的信息,并且处理器230可以基于数学公式计算与像素值相对应的转换值,并基于所获取的转换值转换特征图。

[0103] 此外,根据本公开的实施例的处理器230可以生成用于将像素值之间的最大残差转换为较低值的查找表。例如,处理器230可以生成用于执行转换的查找表,使得特征图中包括的多个像素中具有最高频率的像素值的像素对应于所述多个像素的像素值变化范围(例如,0到255)的中间值。此外,处理器230可以生成用于执行转换的查找表,使得具有第二最高频率的像素值的像素对应于与中间值相邻或接近中间值的值。例如,如果具有最高频率的像素值是0,则处理器230可以将像素值0转换为像素值变化范围中的0到255之间的

127,并且如果具有第二最高频率的像素值是255,则处理器230可以将像素值255转换为像素值变化范围中的0到255之间的126或128。此外,查找表可以包括关于1:1对应关系的信息,使得其余像素值具有基于频率数趋于远离中间值(例如,127)的值。将参考图7A和图7B在这方面进行详细解释。

[0104] 另外,处理器230可以分析与特征图的类型相对应的学习特征图,并且确定用于压缩特征图的压缩模式,使得压缩之后特征图的比特量在多种压缩模式中变为最小值。

[0105] 同时,处理器230包括RAM、ROM、图形处理器、主CPU和第一接口至第n接口。这里,RAM、ROM、图形处理器、主CPU、第一接口至第n接口等可以通过至少一条总线彼此连接。同时,为了便于解释,基于多个学习图像获取与特征图相对应的查找表,并且在服务器200处执行用于识别压缩模式的一系列操作。然而,这仅是示例,可以在电子设备100处执行操作。作为另一示例,在服务器200处执行基于多个学习图像获取与特征图相对应的查找表的操作,并且可以在电子设备100处执行用于识别多种压缩模式中对于特征图具有最高压缩率的压缩模式的操作。

[0106] 图6至图8B描述了根据本公开的实施例的在训练人工智能模型的过程中生成查找表并确定压缩模式。

[0107] 参考图6,服务器200可以将多个图像输入到人工智能模型中,以训练人工智能模型(例如,CNN)。这里,在人工智能模型中,可以调节每个节点的权重或卷积滤波器的值,使得多个图像的特征图形成特定图案。同时,在操作S610处,基于通过这样的训练过程所设置的滤波器或权重,人工智能模型可以获取多个输入图像中每一个的特征图。

[0108] 在操作S620处,服务器200可以分析多个图像的特征图的直方图。这里,直方图可以包括与特征图中包括的多个像素中的每个像素的像素值的分布有关的信息。此外,服务器200可以基于多个像素值的分布来确定特征图的类型。也就是说,服务器200可以通过获取多个特征图的像素值分布的数据库来确定特征图的类型,确定多个特征图中每一个的分布,并且将多个特征图中每一个的像素值分布与数据库进行比较。

[0109] 例如,在特征图中包括的像素值用“8比特灰阶”表示的情况下,每个像素值可以由0~255之间的值表示。服务器200可以考虑各种因素来确定像素值的分布,例如在特征图中包括的像素中像素值在0~50之间的像素的比率、像素值在200~255之间的像素的比率,并将具有特定比率的特征图定义为第一类型、第二类型等。作为示例,服务器200可以将其中像素值在0~30之间的像素的比率为30%并且像素值在225~255之间的像素的比率为40%的特征图定义为“第一类型”。因此,在确定特征图的像素值分布类似于先前“第一类型”的像素值分布的情况下,服务器200可以将特征图识别为第一类型。然而,前述实施例仅是示例,并且服务器200可以考虑各种因素来识别各种类型。

[0110] 另外,在操作S630处,服务器200可以生成查找表,并确定压缩模式,以根据多个特征图的每种类型有效地压缩多个特征图。具体而言,查找表是用于转换特征图的每个像素值以有效地压缩特征图的查找表。这样的查找表可以被构造为以与多个像素一一对应的方式转换每个像素。

[0111] 图7A和图7B以及图8A和图8B是示出了用于转换特征图的查找表和压缩模式的示图。

[0112] 具体而言,图7A示出了在通过查找表转换特征图之前的直方图,图7B示出了根据

本公开的实施例的转换之后的直方图。参考图7A,在通过查找表转换之前的特征图的直方图中,像素值在0~30之间的像素的比率以及像素值在225~255之间的像素的比率在所有值0~255的像素中为高。换句话说,与值在0~30和225~255范围之外的像素相比,值在0~30之间的像素的数量和值在225~255之间的像素的数量相对较高。服务器200可以将具有这种像素值分布的特征图识别为特定类型(例如,“第一类型”)。像素值的分布可以是特征图的特征,通过该特征,可以将特征图与具有不同像素值分布特征的其他类型的特征图相对识别出来。

[0113] 同时,在服务器200通过与图7A中所示的直方图相对应的“第一查找表”转换特征图的情况下,特征图的像素值在0~30之间的像素可以转换为具有在120~150之间的像素值,并且像素值在225~255之间的像素可以转换为具有在150~180之间的像素值。也就是说,特征图的像素值分布可以从其中像素值分散于直方图两侧的分布(图7A)转换为其中像素值集中在直方图中心的分布(图7B)。

[0114] 同时,根据本公开的实施例,基于如下假设进行了说明:服务器200基于多个学习图像识别特征图的多种类型,并获取用于每种类型的特征图的查找表。然而,本公开不限于此。

[0115] 根据本公开的实施例,服务器200可以通过如下方法获取查找表:基于特征图的直方图,将具有最高频率的像素值转换为像素值范围(例如,0到255)中的中间值,并且将与第二最高频率相对应的像素值转换为与中间值相邻的值。此外,服务器200可以通过如下方法获取查找表:基于直方图,将中间值指派给具有最高频率的像素值,将与中间值相邻且大于中间值的值指派给具有第二最高频率的像素值,并将与中间值相邻且小于中间值的值指派给具有第三最高频率的像素值。

[0116] 作为另一示例,应用了激活函数的特征图可以包括具有从0到255的像素值的多个像素。因此,服务器200可以基于数学公式转换特征图,如下所示。作为示例,对于从0到127的像素值,可以基于数学公式1执行特征图的转换,对于从128到255的像素值,可以基于数学公式2执行特征图的转换。如图7B所示,在转换后的特征图中,0和255的像素值可以分别转换为127和128,并且标准偏差可以减小。

[0117] [数学公式1]

[0118] $pv1 = pv + (127 - 2 * pv)$

[0119] [数学公式2]

[0120] $pv1 = pv + 383 - 2 * pv$

[0121] 这里,pv1表示转换后的像素值,pv表示原像素值。同时,如上所述,获取查找表的操作可以由电子设备100而非服务器200执行。

[0122] 如图7B所示,在直方图上的像素值分布的标准偏差减小的情况下(即,在像素值集中在曲线图中心的情况下),每个像素的值与每个像素的预测值之间的差(或残差)总体上减小。参考图7A,与其余像素值相比,像素值0和像素值255的频率数相对较高。如果预测值是中间值(例如,127),则在像素值与预测值之间的差(或残差)中127(像素值0和中间值127之间的差)或128(像素值255和中间值127之间的差)所占据的比例高,相应地,压缩率减小。这是因为编码较大值(例如,大的残差值)的有效性可能很差。相反,在通过查找表转换的特征图的情况下,如果预测值是中间值(例如,127),则在像素值和预测值之间的差(或残差)

中127或128所占据的比例低,并且0(像素值0的转换值127与中间值127之间的差)或1(像素值255的转换值128与中间值127之间的差)所占据的比例高。相应地,压缩率变得更高。这是因为编码较小值(例如,小的残差值,特别是值0和1)的有效性可以很高。

[0123] 根据本公开的实施例,电子设备100可以根据压缩模式对特征图执行压缩,其中所述特征图基于查找表被转换为指示具有较小标准偏差的曲线图的形式。在使用如上所述经过转换以使标准偏差减小的特征图的情况下,存在电子设备100能够以低存储容量存储特征图的效果。也就是说,经编码的特征图的大小减小。将参考图8A和图8B对这方面进行详细解释。

[0124] 图8A和图8B是示出了根据本公开的实施例的用于压缩特征图的压缩模式的示意图。

[0125] 服务器200可以预测每个像素的值。服务器200可以通过多种压缩模式(或PMODE)预测像素值。这里,压缩模式可以表示像素和相邻像素之间的关系。参考图8A,PMODE 0可以通过使用待预测像素x附近的三个像素a、b、c来计算待预测像素x的预测值Pred_x。同样地,PMODE 1可以使用待预测像素x附近的像素a。同时,PMODE 2可以使用待预测像素x附近的像素b,PMODE 3可以使用待预测像素x的相邻像素a、b的平均值,PMODE 4可以使用待预测像素x附近的像素c。在这种情况下,附近像素a、b、c可以是图像中从待预测像素x的左侧的像素开始沿像素的顺时针顺序布置的三个像素。将参考图8B对这方面进行详细解释。

[0126] 参考图8B,服务器200可以使用与对象像素10的左侧相邻的像素“a”、与对象像素10相邻的左上对角线方向上的像素“b”和与对象像素10的上侧相邻的像素“c”中的至少一个像素来获取预测值。作为示例,压缩模式PMODE 1可以是基于与对象像素的左侧相邻的像素a获取预测值(Pred_x=a)的压缩模式。在这种情况下,电子设备100可以在从对象像素10的像素值减去像素“a”的值之后获取残差,并且用获取的残差值替换对象像素10的像素值。这里,如果像素“a”的值大于对象像素10的值,则残差值可以是0,并且对象像素10的值可以被值0替换。

[0127] 同时,为了便于解释,基于如下假设描述了其中电子设备100利用根据压缩模式基于与对象像素10相邻的三个像素中的至少一个像素的组合来获取预测值的实施例:以预定大小的像素块为单位使用从左侧到右侧的扫描顺序对特征图执行压缩。然而,本公开不限于此。例如,电子设备100可以利用根据压缩模式基于与对象像素10相邻的五个像素中的至少一个像素的组合来获取预测值。

[0128] 同时,图8A中所示的压缩模式PMODE和像素a、b、c仅是示例,并且可以通过各种压缩模式执行压缩。例如,PMODE 5可以使用待预测像素x的相邻像素a、b、c的平均值。服务器200可以将每个像素值减小与通过压缩模式预测的预测值(Pred_x)一样的大小。例如,在特定像素具有8比特灰阶值150,并且通过特定压缩模式预测的该像素的预测值是145的情况下,服务器200可以将像素值减小与预测值一样的大小。在这种情况下,像素值150可以被值5替换。因此,尽管值150需要8比特的存储容量,但是像素值为5的像素可以仅以4比特的存储容量来存储。

[0129] 同时,服务器200可以评估通过用于压缩特征图的多种压缩模式(例如,如图8A中的5种压缩模式)中的每种压缩模式进行的压缩。在作为压缩的结果确定PMODE 1总体上具有最佳压缩效率的情况下,服务器200可以确定与特征图相对应的压缩模式是PMODE 1,并且进一步地,确定与特征图所属的类型相对应的压缩模式是PMODE 1。作为示例,服务器200

可以利用PMODE 1将多个像素中的每个像素的像素值指示为排除了预测值的残差,并且确定是否可以以4比特的存储容量存储残差。例如,如果像素值中排除了预测值的残差是0到15,则可以以4比特的存储容量存储像素值,因此压缩率可以变得更高。

[0130] 此外,服务器200可以评估通过多种压缩模式中的每种压缩模式对通过如图7A至图7B所示确定的查找表而转换的特征图进行的压缩。同样地,服务器200可以确定针对转换后的特征图表现出最佳压缩效率的压缩模式与查找表相对应。

[0131] 也就是说,服务器200可以确定如下查找表和压缩模式,其中输入图像的特征图通过该查找表和压缩模式被最大程度地压缩。此外,服务器200可以确定查找表和压缩模式并存储所确定的查找表和压缩模式以与特征图所属的类型相对应。例如,在输入的第一图像的特征图属于“第一类型”,并且通过“第一查找表”和“PMODE 1”的转换和压缩表现出最大压缩效率的情况下,服务器200可以将特征图的“第一类型”、“第一查找表”和“PMODE 1”存储为彼此对应。这里,在稍后输入第二图像的情况下,如果确定第二图像的特征图被分类为第一类型,则服务器200可以通过“第一查找表”和“PMODE1”来压缩第二图像。

[0132] 同时,服务器200可以存储压缩模式,使得针对每种类型的特征图的至少两种压缩模式彼此对应。作为示例,如果特征图是第一类型,则服务器200可以存储压缩模式,使得PMODE 1和PMODE 2彼此对应。此外,根据本公开的实施例,电子设备100可以基于与特征图相对应的第一查找表来转换特征图,并且基于PMODE 1和PMODE 2来执行压缩。然后,电子设备100可以识别PMODE 1和PMODE 2之间具有相对高压缩率的压缩模式,并且基于所识别的压缩模式对转换后的特征图执行压缩。同时,这仅是示例,电子设备100可以基于任意数量的多种压缩模式(例如,PMODE 0至PMODE 4)中的每种压缩模式的压缩率来识别具有最高压缩率的压缩模式。

[0133] 在前述实施例中,仅描述了在服务器200处压缩图像的情况,但是电子设备100可以通过在服务器200处获取的查找表和压缩模式来压缩图像。将参考图9和图10对此进行详细的描述。

[0134] 图9和图10描述了根据本公开的实施例的通过使用查找表和压缩模式转换并压缩输入图像的示例。

[0135] 具体而言,图9是示出了根据本公开的实施例的电子设备压缩图像的方法的流程图。

[0136] 在操作S910处,电子设备100可以将图像输入到电子设备100中的人工智能模型中,并获取与图像相对应的特征图。此外,电子设备100可以通过特征图的像素值分布来确定特征图所属的类型。作为示例,电子设备100可以将像素值的变化范围划分为预定数量,并且识别第一范围中包括的像素的数量、第二范围中包括的像素的数量等。然后,电子设备100可以基于识别结果来识别特征图的类型。作为示例,如果最大量的像素包括在从0到30的像素值中,并且第二最大量的像素包括在从195到255的像素值中,则电子设备100可以将特征图识别为第一类型。然而,这仅是示例,电子设备100可以根据各种标准识别特征图的类型。例如,电子设备100可以基于全部像素的平均值和具有最高频率的像素值来识别特征图的类型。

[0137] 然后,在操作S920处,电子设备100可以通过与特征图的类型相对应的查找表来转换特征图。如上所述,通过查找表转换的特征图可以以一对一的对应关系被转换,以便改进

或优化像素值的分布以进行压缩。此外,在操作S930处,电子设备100可以通过用于最佳地压缩转换后的特征图的压缩模式来压缩特征图。

[0138] 在这种情况下,查找表和压缩模式可以通过在电子设备100内部训练人工智能模型而已被确定,或者在服务器200处被确定后通过通信器已被接收。将参考图11描述通过服务器200接收查找表和压缩模式的方法。

[0139] 同时,图10是示出了根据本公开的实施例的在通过查找表转换图像之前的图像和转换之后的图像的示意图。

[0140] 参考图10,在通过查找表转换图像之前,在针对图10的(a)中的输入图像的图10的(b)中的特征图中,图像内像素的灰阶分布的标准偏差总体上很大。然而,在电子设备100通过与所获取的特征图的类型相对应的查找表转换特征图的情况下,可以获取图10的(c)中的转换后的特征图,其中图像内的灰阶差异总体上减小。

[0141] 图11是示出了根据本公开的实施例的通过包括服务器和电子设备的系统压缩特征图的方法的序列图。

[0142] 参考图11,服务器200可以通过多个输入图像训练人工智能模型。通过服务器200训练人工智能模型的过程可以是设置权重或卷积滤波器的过程,使得当特定图像被输入到人工智能模型中时可以输出期望的结果。也就是说,服务器200可以通过输入多个输入图像逐渐地调整能够如上所述导出期望结果的权重或卷积滤波器。此外,在操作S1110处,服务器200可以获取用于训练人工智能模型的多个输入图像中的每个输入图像的学习特征图。在操作S1120处,服务器200可以分析与所获取的学习特征图相对应的直方图,并且基于分析的结果获取用于确定学习特征图的类型的数据库。也就是说,服务器200可以获取多个图像的多个学习特征图的直方图信息以及用于多个学习特征图的数据库。

[0143] 服务器200可以基于数据库确定特征图的类型。此外,在操作S1130处,服务器200可以确定用于根据特征图的每种类型有效地压缩特征图的查找表和压缩模式。也就是说,服务器200可以分析多个学习图像的直方图并确定多个特征图的类型,并且可以确定针对特征图的每种类型表现出最大压缩效率的查找表和压缩模式。

[0144] 同时,在操作S1140处,服务器200可以向电子设备100发送与基于在训练人工智能模型的过程中获取的信息而确定的“针对特征图的每种类型的查找表和压缩模式”有关的信息。这里,在服务器200处确定的查找表可以是如下生成的查找表:通过分析学习特征图的直方图信息,使得特征图中包括的多个像素的像素值之间的最大残差减小。同时,服务器200可以向电子设备100发送与人工智能模型的权重值和滤波器值有关的信息,其中该人工智能模型被训练为当输入图像时输出期望的结果。

[0145] 电子设备100可以将将从服务器200接收的与针对特征图的每种类型的查找表和压缩模式有关的信息存储在存储器110中。之后,在操作S1150处,当图像被输入到在服务器200处训练的人工智能模型中时,电子设备100可以通过将图像输入到第一层来获取图像的特征图,并确定所获取的特征图的类型。这里,第一层可以是如下卷积层,其将输入图像的相邻值填补为特定值,并且通过使用在服务器200处训练的人工智能模型的卷积滤波器以预定步幅执行卷积。

[0146] 电子设备100可以确定与作为对输入图像的卷积结果而获取的特征图的类型相对应的查找表和压缩模式。这里,在操作S1160处,电子设备100可以从存储器110读取关于与

所确定的特征图的类型相对应的查找表和压缩模式的信息,并且基于读取的查找表和压缩模式来压缩特征图。

[0147] 也就是说,电子设备100可以通过从服务器200接收的与特征图的类型相对应的查找表转换特征图,使得特征图中包括的像素值的整个分布的标准偏差减小,来获取其中多个像素值之间的最大残差减小的特征图。

[0148] 根据本公开的实施例,电子设备100可以通过查找表转换特征图,使得特征图中包括的多个像素中具有高频率的像素值的像素对应于与所述多个像素的像素值变化范围的中间值接近的值。然而,这仅是示例,并且电子设备100可以转换特征图,使得其他像素对应于与特征图中包括的多个像素中具有高频率的像素值接近的值。

[0149] 另外,电子设备100可以基于转换后的特征图中包括的多个像素中与每个像素相邻的至少一个像素的值来预测每个像素的值,并将每个像素的值减小与预测值一样的大小。具体而言,电子设备100可以基于与用于最佳地压缩转换后的特征图的压缩模式有关的信息,根据压缩模式中包括的计算方法来预测每个像素的值。之后,电子设备100可以仅存储每个像素的值与预测值之间的差值或残差,由此可以压缩转换后的特征图,同时减小用于特征图的存储容量。

[0150] 同时,在前述实施例中,描述了电子设备100通过特定压缩模式压缩通过查找表转换的特征图。然而,电子设备100还可以通过与特征图相对应的压缩模式压缩未通过查找表转换的特征图。也就是说,在电子设备100内部不存在关于与特征图相对应的查找表的信息的情况下,电子设备100可以仅利用与特征图相对应的压缩模式来压缩特征图。

[0151] 此外,在不存在关于与特征图相对应的压缩模式的信息的情况下,电子设备100可以通过与类似特征图相对应的压缩模式来压缩特征图。将参考图13对此进行描述。

[0152] 同时,根据前述实施例,在接收图像输入之前,电子设备100可能已经存储了“针对特征图的每种类型的查找表和压缩模式”。然而,这仅是示例,电子设备100可以在接收图像输入之后实时地请求服务器200发送与所获取的特征图相对应的查找表和压缩模式。

[0153] 人工智能模型的训练过程相对复杂,并且与应用过程相比需要大量操作。同时,即使应用过程相对简单并且需要少量操作,该过程也应由用户每次执行。因此,如图11所示,可以通过其中对处理能力的限制较小的服务器200执行人工智能模型的训练过程,并且可以通过能够容易地与用户进行交互的电子设备100执行人工智能模型的应用过程。也就是说,通过前述实施例,具有的效果是,可以分别通过适当的环境来执行人工智能模型的训练过程和应用过程。

[0154] 同时,在前述实施例中,描述了服务器200执行人工智能模型的训练过程。然而,这仅是示例,可以通过将多个图像输入到另一外部电子设备或电子设备100来执行人工智能模型的训练过程。

[0155] 图12A至图12C是示出了根据本公开的实施例的电子设备对特征图进行编码和解码的处理的示意图。

[0156] 具体而言,图12A是示出了电子设备对所获取的特征图进行编码的过程的流程图。如上所述,电子设备100可以通过将图像输入到电子设备中基于人工智能模型的应用的第一层来获取图像的特征图。然后,在操作S1210处,电子设备100可以通过使用存储在存储器110中的与特征图的类型相对应的查找表(LUT)来转换获取的多个特征图。当通过与特征图

相对应的查找表转换特征图时,电子设备100可以获取转换到如下状态的特征图,在该状态下,特征图中包括的多个像素的像素值的整体分布的标准偏差已减小或被最小化。

[0157] 在操作S1220处,电子设备100可以通过存储在存储器110中的与特征图的类型相对应的压缩模式(PMODE)来压缩转换后的特征图。作为另一示例,电子设备100可以基于多种压缩模式中的每种压缩模式对转换后的特征图执行压缩,并识别与最高压缩率相对应的压缩模式。然后,电子设备100可以根据与PMODE相对应的计算公式来预测每个像素的值的预测值,并用残差替换像素值,其中像素的像素值减小了与预测值一样的大小。如上所述,当通过特定压缩模式压缩特征图时,在特征图中包括的多个像素中的每个像素的像素值已减小的状态下,电子设备100可以获取其数据容量已被压缩的特征图。将参考图12B在这方面进行详细解释。

[0158] 根据图12B,可以用残差替换特定像素的像素值,其中像素值减小了与根据同该特定像素相邻的像素的像素值所获取的预测值一样的大小。例如,可以假设如下情况:压缩模式PMODE 1是其中位于像素左侧的像素的像素值是预测值的模式。如果压缩模式PMODE 1被识别为与特征图的类型相对应的压缩模式,则电子设备100可以用从像素值减去与位于对象像素左侧的像素的像素值一样的大小之后的残差替换对象像素的像素值。根据本公开的实施例,像素值的变化范围是从0到255,但是在用残差替换像素值的情况下,其中像素值被减去与预测值一样的大小,像素值可以是0到16的值。

[0159] 返回到图12A,电子设备100可以将经压缩的特征图存储在存储器中。这里,存储器可以是电子设备100中包括的存储器110、外部设备中包括的存储器或者外部服务器200中包括的存储器。当存储经压缩的特征图时,电子设备100可以直接存储与特征图相对应的数据,但是也可以在以特定数量为单位对像素进行分组的同时存储特征图,以便防止存储容量的不必要浪费。

[0160] 具体而言,根据本公开的实施例的电子设备100可以以 4×2 大小的块为单位对根据压缩模式压缩的特征图中包括的多个像素进行分组。这里,电子设备100可以识别以 4×2 为单位分组的块内部的像素(在该实施例中为8个像素)之间的残差。作为示例,电子设备100可以确定多个块分组中残差最大的块分组。也就是说,在操作S1230处,电子设备100可以确定以 4×2 像素为单位的多个块分组中具有最大长度的报头分组。

[0161] 当存储经压缩的特征图时,电子设备100可以将关于压缩所使用的压缩模式的信息和关于报头分组的信息添加到报头。这里,电子设备100可以基于报头分组的最大长度来确定存储单元。具体而言,在操作S1240处,电子设备100可以确定能够存储与最大长度一样大小的最小存储单元,并且可以在按顺序打包基于确定的存储单元所分组的特征图的残差的同时存储特征图。此外,根据本公开的实施例的电子设备100可以以1比特的存储容量存储关于压缩模式的信息,并且以4比特的存储容量存储与根据最大长度的报头分组有关的信息。

[0162] 根据前述实施例,电子设备100可以通过存储以能够存储与最大长度一样大小的最小单元所压缩的特征图,使存储器的不必要浪费最小化。例如,在报头分组中像素的像素值之间的差(或残差)存在多达7灰阶的情况下,电子设备100可以在存储分组时仅分配3比特。作为另一示例,如果报头分组中像素的像素值之间的差(或残差)是15,则电子设备100可以通过分配关于压缩模式的1比特的信息和关于分组的4比特的信息,将特征图存储在报

头中。

[0163] 同时,前述实施例仅是示例,电子设备100可以在以各种像素单位(例如,2×2单位、4×4单位等)对像素分组的同时存储特征图,并且也可以通过各种方法确定存储单元。

[0164] 图12C是示出了电子设备对经编码的特征图进行解码以将特征图发送给下一层的过程的流程图。

[0165] 在电子设备100从存储器读取与特征图相对应的数据之后,在操作S1250处,电子设备100可以分析包括在数据中的报头。然后,电子设备100可以基于关于报头中包括的报头分组的信息来确定关于存储单元的信息。在操作S1260处,电子设备100可以基于关于存储单元的信息(在所述实施例中为3比特)以4×2像素为单位恢复特征图的结构和像素值。

[0166] 此外,在操作S1270处,电子设备100可以基于关于编码中使用的压缩模式的信息来恢复特征图中包括的像素值的残差。作为示例,电子设备100可以基于关于压缩模式的信息来识别编码中使用的压缩模式。电子设备100可以通过基于与识别的压缩模式相对应的数学公式、相邻像素之间的关系等将像素的预测值与残差相加,来获取转换后的特征图。这里,转换后的特征图可以表示由于查找表已应用于原始特征图而处于已被转换状态下的特征图。然后,在操作S1280处,电子设备100可以通过基于关于编码中使用的查找表的信息执行逆转换来恢复输入图像的特征图。作为示例,在像素值0根据编码中使用的查找表已转换为127的情况下,电子设备100可以将转换后的特征图中包括的127重新转换为像素值0,并执行恢复。另外,电子设备100可以通过将恢复的特征图发送给下一阶段的层来继续执行人工智能模型的处理。根据前述实施例,电子设备100可以在存储器中有效地存储和读取其容量已被压缩的特征图,使得没有浪费的空间。

[0167] 图13A和图13B示出了根据本公开的实施例的针对多个图像使用特定查找表和特定压缩模式的压缩率的结果。

[0168] 参考图13A,电子设备100可以通过使用与芝加哥图像相对应的查找表和压缩模式来压缩多个图像。图13B比较了在人工智能模型的每个层中压缩之前的特征图的数据量与压缩之后的特征图的数据量。也就是说,作为利用查找表和压缩模式压缩在每个层中生成的特征图以压缩芝加哥图像的结果,可以确定在层0中数据量被压缩到4.6%。同时,数据量在层1中被压缩到9.53%,在层2中被压缩到36.14%,在层3中被压缩到17.56%,在层4中被压缩到18.85%,并且总体上,数据量被压缩到17.34%。

[0169] 同时,在通过使用用于最大程度地压缩芝加哥图像的查找表和压缩模式来压缩其他图像(即,菊花、绣球花和企鹅图像)的情况下,总体上,数据量分别被压缩到17.39%、17.84%和18.04%。因此,该查找表和压缩模式对于芝加哥图像表现出最佳压缩率,因为它们是用以压缩芝加哥图像的查找表和压缩模式。

[0170] 然而,即使对于不同的图像,如果应用通过使用人工智能模型所生成和确定的查找表和压缩模式,也可以进行显著量的压缩。因此,即使在没有用于确定输入图像的特征图的类型的充足数据库的情况下,电子设备100也可以通过使用经由不同图像训练的查找表和压缩模式来执行压缩。

[0171] 虽然已经参考本公开的优选实施例示出和描述了本公开,但是本公开不限于上述特定实施例,并且显然的是,在不脱离所附权利要求所要求保护的本公开的主旨的情况下,本公开所属技术领域的普通技术人员可以进行各种修改。此外,旨在不应独立于本公开的

技术构思或前景来解释这些修改。

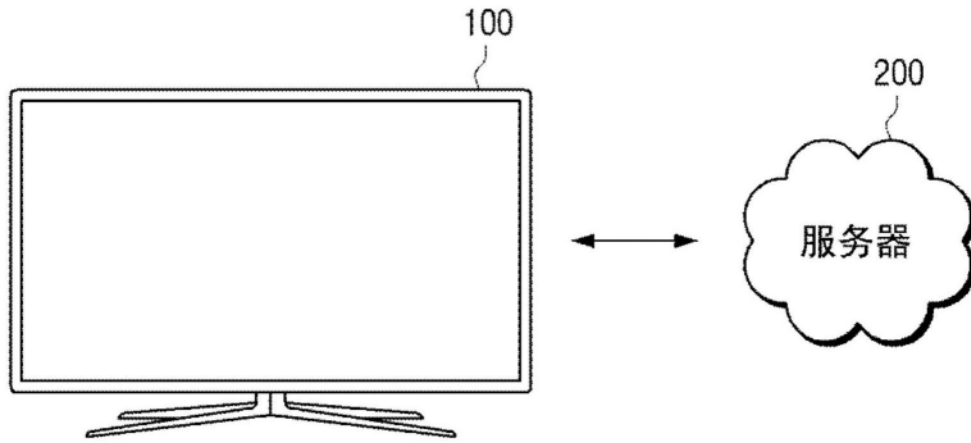


图1

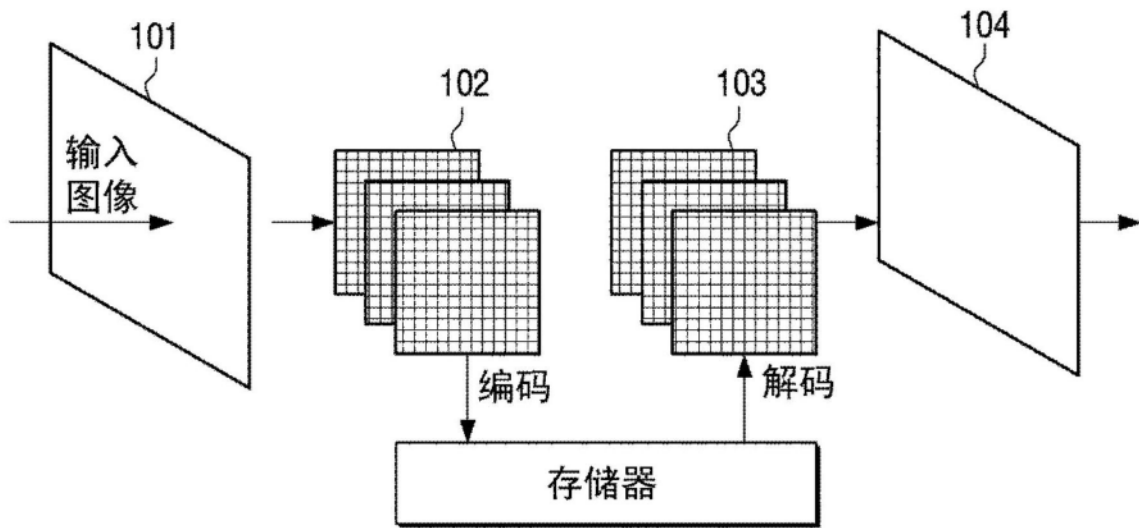


图2

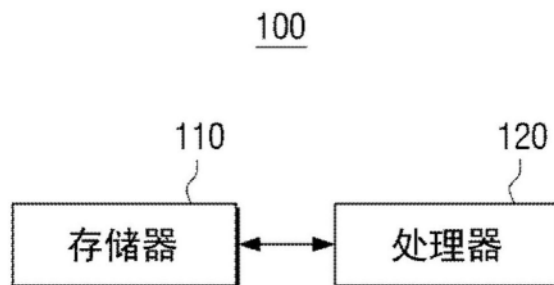


图3

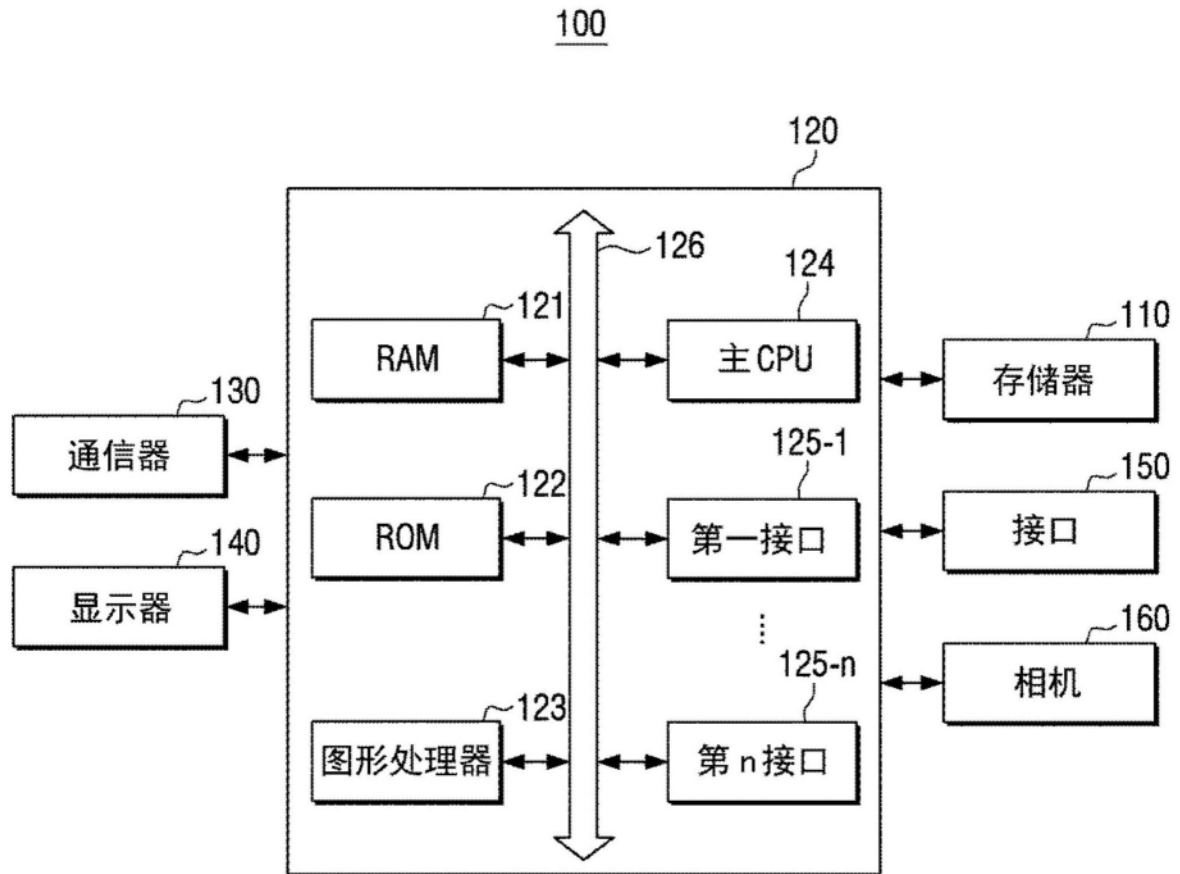


图4

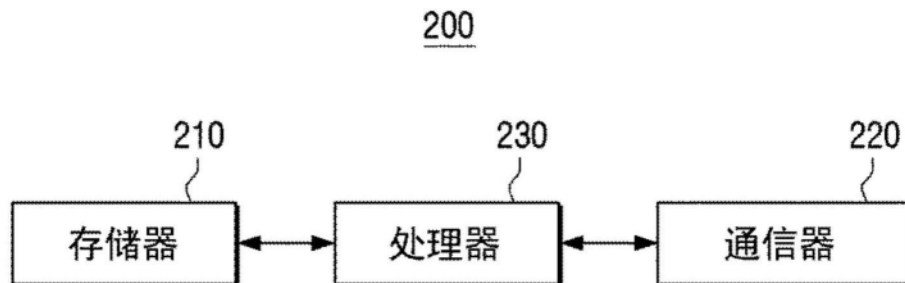


图5

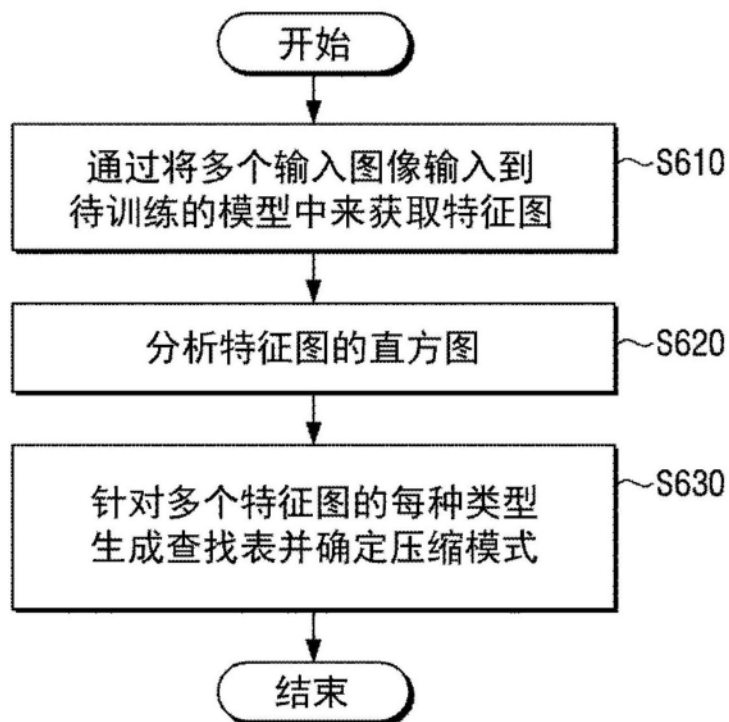
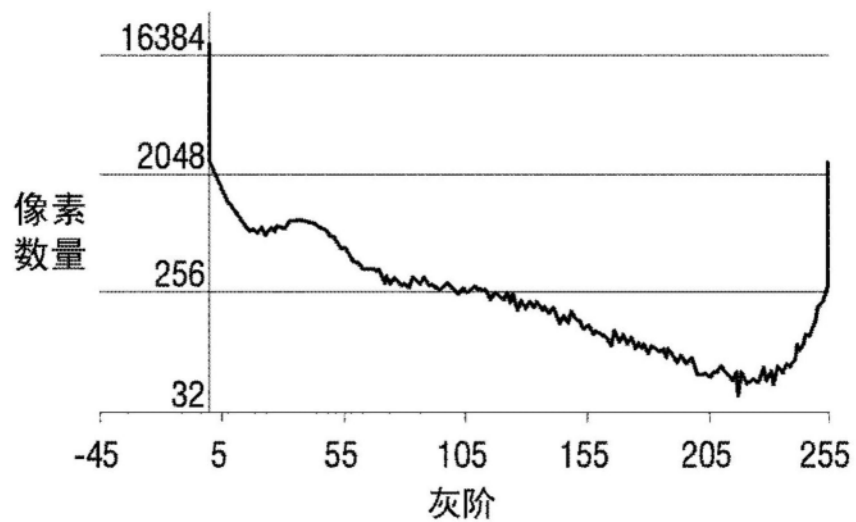
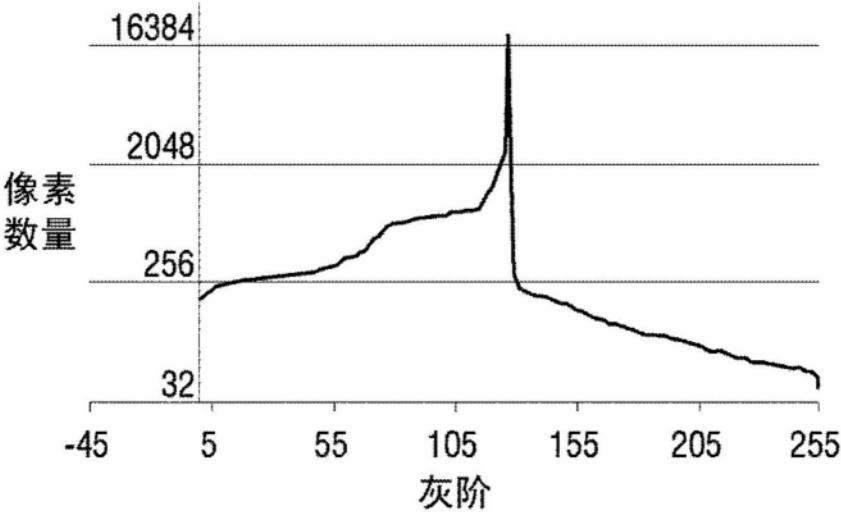


图6



<转换之前>

图7A



<转换之后>

图7B

PMODE	描述
0	$\text{Pred_x} = a + b - c$
1	$\text{Pred_x} = a$
2	$\text{Pred_x} = b$
3	$\text{Pred_x} = (a + b + 1) >> 1$
4	$\text{Pred_x} = c$

图8A

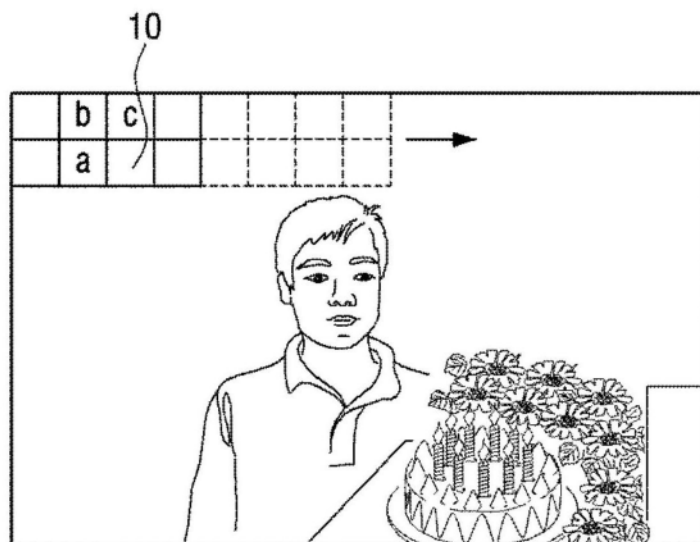


图8B

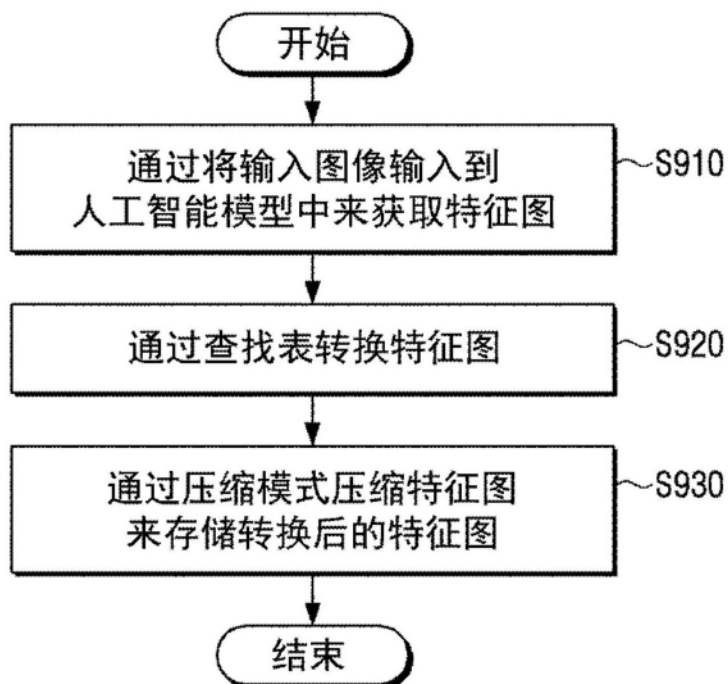


图9

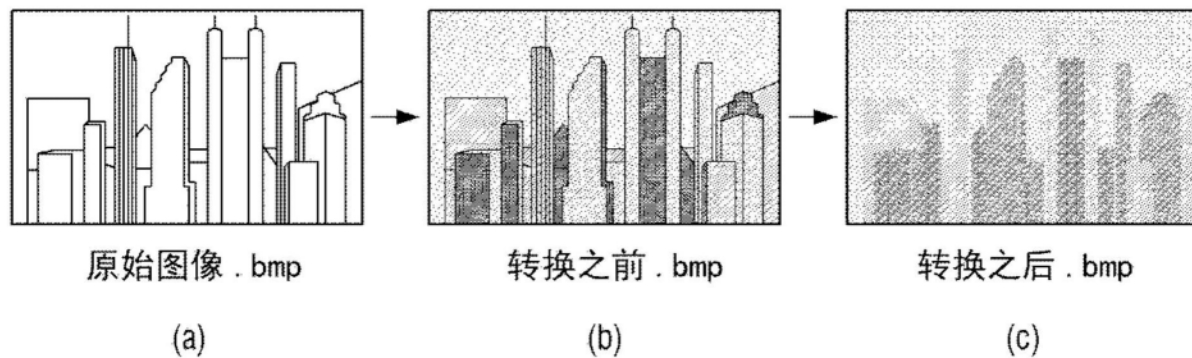


图10

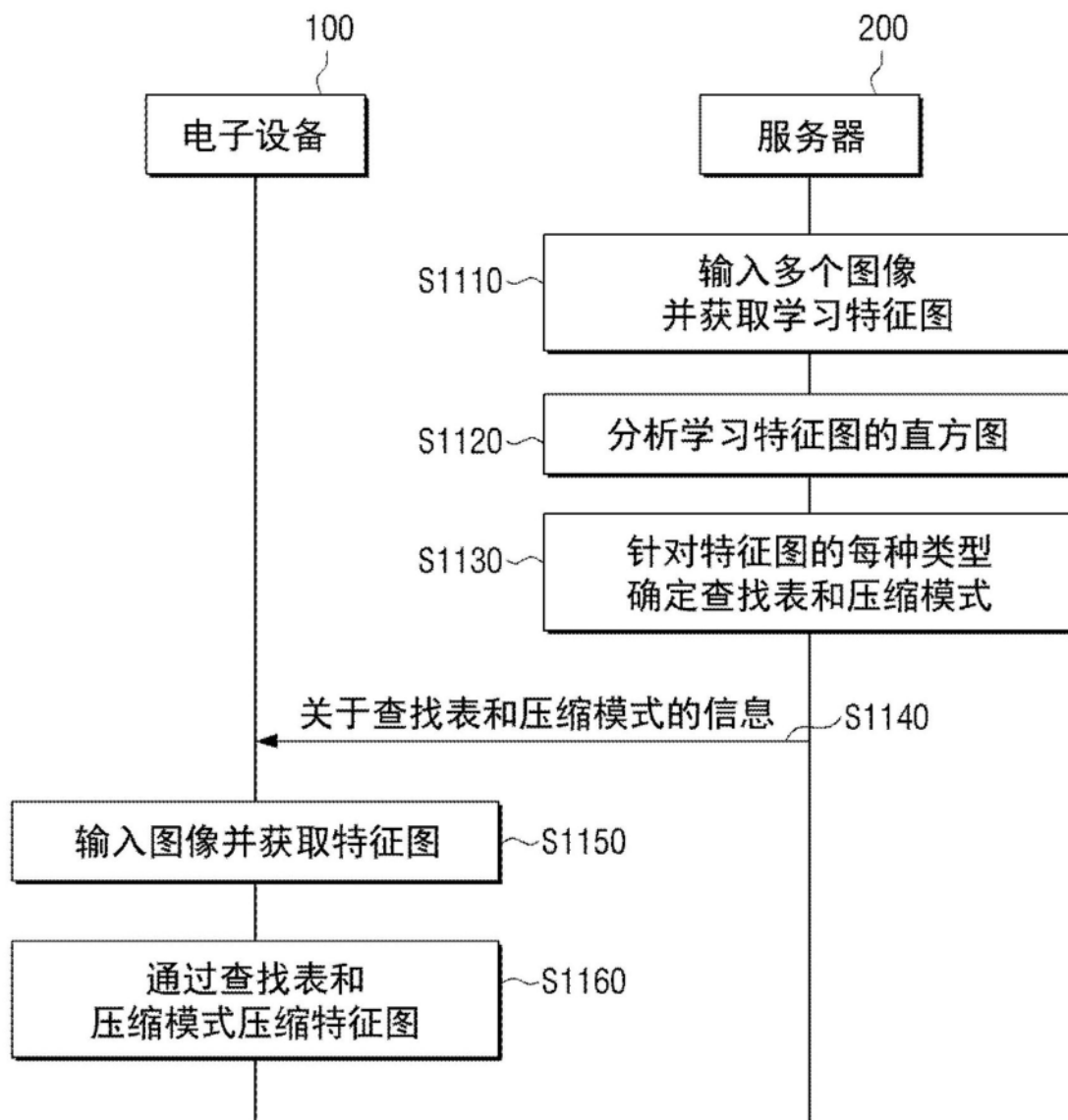


图11

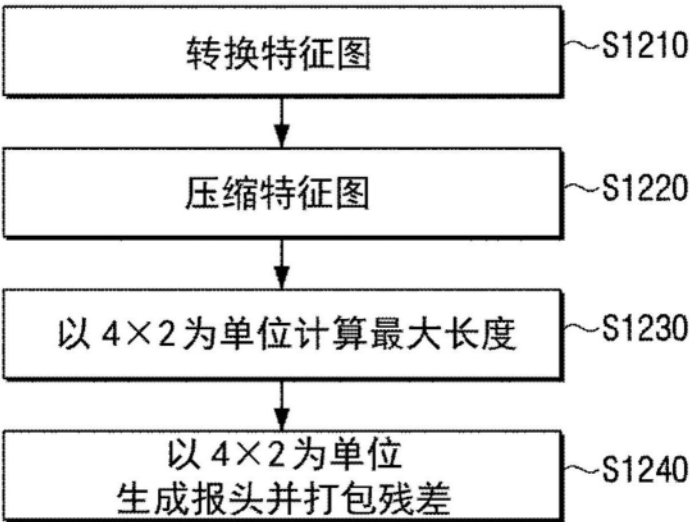


图12A

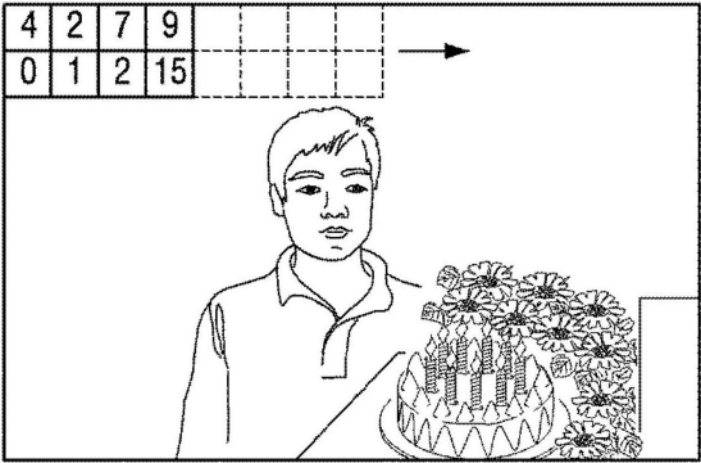


图12B

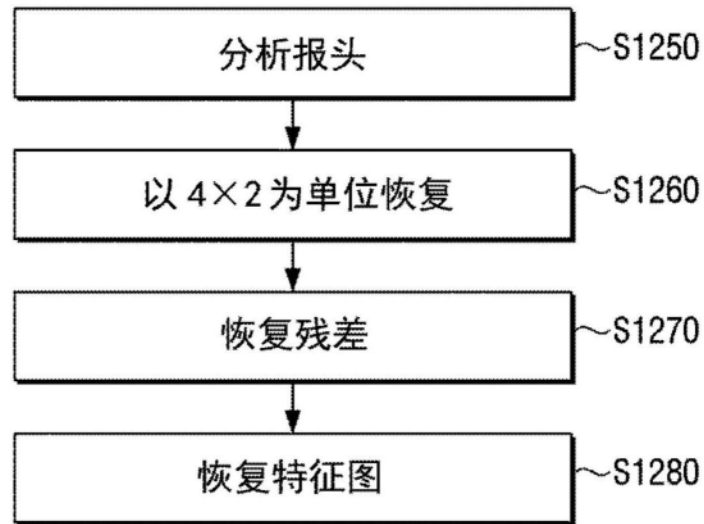


图12C

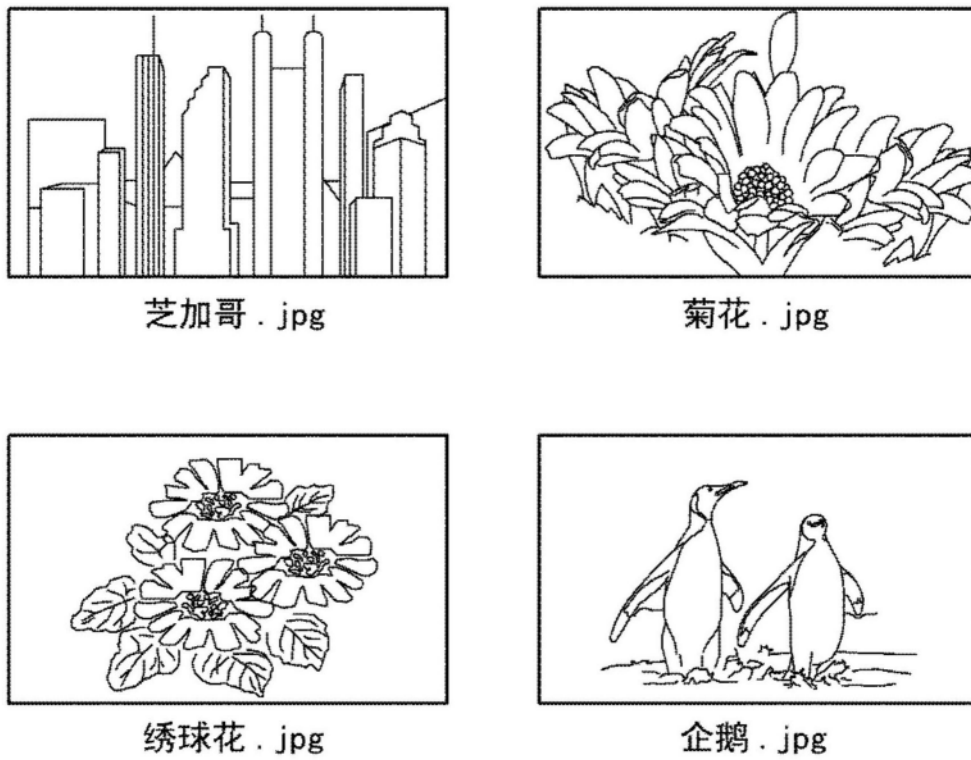


图13A

(应用之后 / 应用之前) [%]				
层	芝加哥	菊花	绣球花	企鹅
0	4.60	3.87	4.73	4.94
1	9.53	9.58	10.05	10.74
2	36.14	36.39	37.06	37.29
3	17.56	18.08	18.25	18.29
4	18.85	19.02	19.10	18.93
Avr.	17.34	17.39	17.84	18.04

图13B