



(12) 发明专利申请

(10) 申请公布号 CN 115171663 A

(43) 申请公布日 2022. 10. 11

(21) 申请号 202210761612.7

(22) 申请日 2022.06.29

(71) 申请人 美的集团(上海)有限公司

地址 201700 上海市青浦区盈港东路168号
9幢2层205室

申请人 美的集团股份有限公司

(72) 发明人 丁维浩

(74) 专利代理机构 北京路浩知识产权代理有限公司 11002

专利代理师 赵娜

(51) Int. Cl.

G10L 15/02 (2006.01)

G10L 15/22 (2006.01)

G10L 25/03 (2013.01)

G10L 25/51 (2013.01)

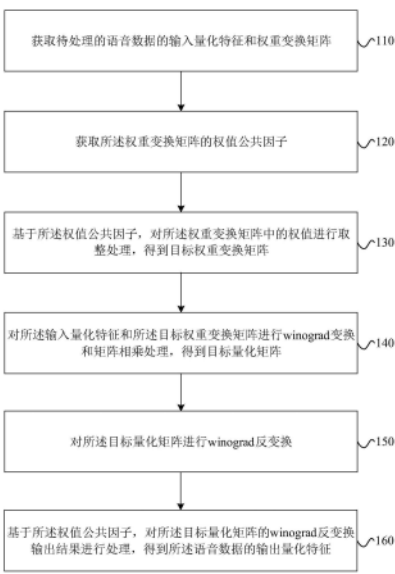
权利要求书2页 说明书11页 附图4页

(54) 发明名称

语音数据的处理方法、电子设备及存储介质

(57) 摘要

本申请涉及计算机领域,提供一种语音数据的处理方法、电子设备及存储介质,该方法包括:获取待处理的语音数据的输入量化特征和权重变换矩阵;获取权重变换矩阵的权值公共因子;基于权值公共因子,对权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵;对输入量化特征和目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵;对目标量化矩阵进行winograd反变换;基于权值公共因子,对目标量化矩阵的winograd反变换输出结果进行处理,得到语音数据的输出量化特征。该方法通过权值公共因子对winograd变换前的权重变换矩阵和winograd反变换输出结果进行处理,有效防止数据计算溢出。



1. 一种语音数据的处理方法,其特征在于,包括:

获取待处理的语音数据的输入量化特征和权重变换矩阵;

获取所述权重变换矩阵的权值公共因子;

基于所述权值公共因子,对所述权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵;

对所述输入量化特征和所述目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵;

对所述目标量化矩阵进行winograd反变换;

基于所述权值公共因子,对所述目标量化矩阵的winograd反变换输出结果进行处理,得到所述语音数据的输出量化特征。

2. 根据权利要求1所述的语音数据的处理方法,其特征在于,所述基于所述权值公共因子,对所述权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵,包括:

将所述权重变换矩阵中的权值乘以所述权值公共因子,得到所述目标权重变换矩阵。

3. 根据权利要求1所述的语音数据的处理方法,其特征在于,所述基于所述权值公共因子,对所述目标量化矩阵的winograd反变换输出结果进行处理,得到所述语音数据的输出量化特征,包括:

将所述目标量化矩阵中最后一列和最后一行的winograd反变换输出结果除所述权值公共因子,得到所述输出量化特征。

4. 根据权利要求3所述的语音数据的处理方法,其特征在于,所述目标量化矩阵的winograd反变换输出结果采用所述winograd反变换输出结果的两倍整数型数据格式存储。

5. 根据权利要求1-4任一项所述的语音数据的处理方法,其特征在于,所述对所述输入量化特征和所述目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵,包括:

对所述输入量化特征和所述目标权重变换矩阵分别进行winograd变换,得到变换后特征矩阵和变换后权重矩阵;

将所述变换后特征矩阵和所述变换后权重矩阵按照批量矩阵乘法进行相乘,得到所述目标量化矩阵。

6. 根据权利要求5所述的语音数据的处理方法,其特征在于,所述将所述变换后特征矩阵和所述变换后权重矩阵按照批量矩阵乘法进行相乘,得到所述目标量化矩阵,包括:

将所述变换后特征矩阵和所述变换后权重矩阵输入至添加第一参数的批量矩阵乘函数接口,得到所述批量矩阵乘函数接口输出的所述目标量化矩阵。

7. 根据权利要求5所述的语音数据的处理方法,其特征在于,所述对所述输入量化特征和所述目标权重变换矩阵分别进行winograd变换,得到变换后特征矩阵和变换后权重矩阵,包括:

基于所述输入量化特征和所述输入量化特征的特征变换矩阵进行winograd变换,得到所述变换后特征矩阵;

基于所述变换后权重矩阵和量化卷积核进行winograd变换,得到所述变换后权重矩阵。

8. 一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算

机程序,其特征在于,所述处理器执行所述程序时实现如权利要求1至7任一项所述语音数据的处理方法。

9.一种非暂态计算机可读存储介质,其上存储有计算机程序,其特征在于,该计算机程序被处理器执行时实现如权利要求1至7任一项所述语音数据的处理方法。

10.一种计算机程序产品,包括计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至7任一项所述语音数据的处理方法。

语音数据的处理方法、电子设备及存储介质

技术领域

[0001] 本申请涉及计算机技术领域,尤其涉及语音数据的处理方法、电子设备及存储介质。

背景技术

[0002] 搭载有处理语音数据的深度学习模型的边缘端侧设备可以进行语音识别或声纹识别等工作,实现机器与人即时的沟通及身份识别等功能,处理语音数据的深度学习模型部署过程中需要较高的算力成本,在算力不充足的边缘设备上经常会遇到推理实时性较差的问题。

[0003] 目前出现了使用winograd快速卷积算法来加速语音数据处理时卷积运算占据大部分时间的问题,通过用增加的加法运算次数替换减少的乘法运算次数,缩短卷积计算所需要的时间。

[0004] 但winograd快速卷积算法在一些量化模型处理语音数据时,经常会发生数据计算溢出,影响深度学习模型的部署和语音数据的处理。

发明内容

[0005] 本申请旨在至少解决现有技术中存在的技术问题之一。为此,本申请提出一种语音数据的处理方法,防止语音数据的winograd快速卷积发生数据计算溢出。

[0006] 根据本申请第一方面实施例的语音数据的处理方法,包括:

[0007] 获取待处理的语音数据的输入量化特征和权重变换矩阵;

[0008] 获取所述权重变换矩阵的权值公共因子;

[0009] 基于所述权值公共因子,对所述权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵;

[0010] 对所述输入量化特征和所述目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵;

[0011] 对所述目标量化矩阵进行winograd反变换;

[0012] 基于所述权值公共因子,对所述目标量化矩阵的winograd反变换输出结果进行处理,得到所述语音数据的输出量化特征。

[0013] 根据本申请实施例的语音数据的处理方法,通过权值公共因子对Winograd变换前的权重变换矩阵和winograd反变换输出结果进行处理,调整量化模型的整型数据动态范围,有效防止数据计算溢出,提升语音数据的处理速度,加快语音数据处理相关深度学习模型的部署。

[0014] 根据本申请的一个实施例,所述基于所述权值公共因子,对所述权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵,包括:

[0015] 将所述权重变换矩阵中的权值乘以所述权值公共因子,得到所述目标权重变换矩阵。

[0016] 根据本申请的一个实施例,所述基于所述权值公共因子,对所述目标量化矩阵的winograd反变换输出结果进行处理,得到所述语音数据的输出量化特征,包括:

[0017] 将所述目标量化矩阵中最后一列和最后一行的winograd反变换输出结果除所述权值公共因子,得到所述输出量化特征。

[0018] 根据本申请的一个实施例,所述目标量化矩阵的winograd反变换输出结果采用所述winograd反变换输出结果的两倍整数型数据格式存储。

[0019] 根据本申请的一个实施例,所述对所述输入量化特征和所述目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵,包括:

[0020] 对所述输入量化特征和所述目标权重变换矩阵分别进行winograd变换,得到变换后特征矩阵和变换后权重矩阵;

[0021] 将所述变换后特征矩阵和所述变换后权重矩阵按照批量矩阵乘法进行相乘,得到所述目标量化矩阵。

[0022] 根据本申请的一个实施例,所述将所述变换后特征矩阵和所述变换后权重矩阵按照批量矩阵乘法进行相乘,得到所述目标量化矩阵,包括:

[0023] 将所述变换后特征矩阵和所述变换后权重矩阵输入至添加第一参数的批量矩阵乘函数接口,得到所述批量矩阵乘函数接口输出的所述目标量化矩阵。

[0024] 根据本申请的一个实施例,所述对所述输入量化特征和所述目标权重变换矩阵分别进行winograd变换,得到变换后特征矩阵和变换后权重矩阵,包括:

[0025] 基于所述输入量化特征和所述输入量化特征的特征变换矩阵进行winograd变换,得到所述变换后特征矩阵;

[0026] 基于所述变换后权重矩阵和量化卷积核进行winograd变换,得到所述变换后权重矩阵。

[0027] 根据本申请第二方面实施例的语音数据的处理装置,包括:

[0028] 第一获取模块,用于获取待处理的语音数据的输入量化特征和权重变换矩阵;

[0029] 第二获取模块,用于获取所述权重变换矩阵的权值公共因子;

[0030] 第一处理模块,用于基于所述权值公共因子,对所述权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵;

[0031] 第二处理模块,用于对所述输入量化特征和所述目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵;

[0032] 第三处理模块,用于对所述目标量化矩阵进行winograd反变换;

[0033] 第四处理模块,用于基于所述权值公共因子,对所述目标量化矩阵的winograd反变换输出结果进行处理,得到所述语音数据的输出量化特征。

[0034] 根据本申请第三方面实施例的电子设备,包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序,所述处理器执行所述计算机程序时实现如上述任一种所述语音数据的处理方法。

[0035] 根据本申请第四方面实施例的非暂态计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现如上述任一种所述语音数据的处理方法。

[0036] 根据本申请第五方面实施例的计算机程序产品,包括计算机程序,所述计算机程序被处理器执行时实现如上述任一种所述语音数据的处理方法。

[0037] 本申请实施例中的上述一个或多个技术方案,至少具有如下技术效果之一:

[0038] 通过权值公共因子对winograd变换前的权重变换矩阵和winograd反变换输出结果进行处理,调整量化模型的整型数据动态范围,有效防止数据计算溢出,提升语音数据的处理速度,加快语音数据处理相关深度学习模型的部署。

[0039] 进一步的,将目标量化矩阵中最后一列和最后一行的winograd反变换输出结果除权值公共因子,调整目标量化矩阵的整型数据动态范围,防止winograd反变换输出的数据计算溢出。

[0040] 更进一步的,目标量化矩阵的winograd反变换输出结果采用winograd反变换输出结果的两倍整型数据格式存储,可以保持权值矩阵转换过程计算的正确性。

[0041] 本申请的附加方面和优点将在下面的描述中部分给出,部分将从下面的描述中变得明显,或通过本申请的实践了解到。

附图说明

[0042] 为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0043] 图1是本申请实施例提供的语音数据的处理方法的流程示意图之一;

[0044] 图2是本申请实施例提供的语音数据的处理方法的流程示意图之二;

[0045] 图3是本申请实施例提供的语音数据的处理方法的流程示意图之三;

[0046] 图4是本申请实施例提供的语音数据的处理装置的结构示意图;

[0047] 图5是本申请实施例提供的电子设备的结构示意图。

具体实施方式

[0048] 下面结合附图和实施例对本申请的实施方式作进一步详细描述。以下实施例用于说明本申请,但不能用来限制本申请的范围。

[0049] 在本申请实施例的描述中,需要说明的是,术语“第一”、“第二”、“第三”仅用于描述目的,而不能理解为指示或暗示相对重要性。

[0050] 在本说明书的描述中,参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本申请实施例的至少一个实施例或示例中。在本说明书中,对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且,描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外,在不相互矛盾的情况下,本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0051] 下面结合图1至图3描述本申请实施例的语音数据的处理方法,该方法的执行主体可以为终端设备的控制器,或者云端,或者边缘服务器。

[0052] 如图1所示,本申请实施例的语音数据的处理方法包括步骤110至步骤160。

[0053] 步骤110、获取待处理的语音数据的输入量化特征和权重变换矩阵。

[0054] 在该步骤中,可以将待处理的语音数据输入至语音处理模型的量化卷积层进行特征提取,获得待处理的语音数据所对应的输入量化特征和权重变换矩阵。

[0055] 需要说明的是,一般模型内部的计算采用浮点数计算,浮点数的计算会消耗比较大的计算资源,在该实施例,语音处理模型可以是进行语音数据处理的量化模型,采用整数型(int)的数据计算,大大减小消耗的计算资源。

[0056] 在该实施例中,待处理的语音数据的输入量化特征是整数型的特征数据,在实际执行中,输入量化特征可以是8位整数型的数据,也即输入量化特征是int8量化后的数据。

[0057] 可以理解的是,输入量化特征进行winograd快速卷积处理所使用的权重变换矩阵也可以是8位整数型的数据。

[0058] 步骤120、获取权重变换矩阵的权值公共因子。

[0059] 可以理解的是,输入量化特征进行卷积处理所需的多个权值按照对应的行和列排列,得到权重变换矩阵。

[0060] 在该步骤中,获取权重变换矩阵的权值公共因子,也即获取权重变换矩阵中多个权值的公共因子。

[0061] 需要说明的是,权重变换矩阵的权值可以为带小数的权值。

[0062] 在实际执行中,可以采用最大似然法、主轴迭代法、加权最小二乘法、广义加权最小二乘法及最小残差法等提取公共因子的方法,获取权重变换矩阵的权值公共因子。

[0063] 例如,权重变换矩阵G如下式

$$[0064] \quad G = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{24} & \frac{1}{12} & \frac{1}{6} \\ \frac{1}{24} & -\frac{1}{12} & \frac{1}{6} \\ 0 & 0 & 1 \end{pmatrix}$$

[0065] 在该实施例中,获取的权重变换矩阵G的权值公共因子为24。

[0066] 步骤130、基于权值公共因子,对权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵。

[0067] 在该步骤中,根据权重变换矩阵的权值公共因子,将权重变换矩阵的进行取整处理,得到一个新的权重变换矩阵,也即目标权重变换矩阵。

[0068] 需要说明的,利用权值公共因子对权重变换矩阵的权值进行取整处理时,对权重变换矩阵内的所有权值进行相同的处理,仅将原先的权重变换矩阵的小数权值处理为整数权值,权重变换矩阵的权重分布情况不变。

[0069] 例如,利用权值公共因子24对权重变换矩阵G进行取整处理,得到目标权重变换矩阵G',如下所示:

$$[0070] \quad G' = \begin{pmatrix} 6 & 0 & 0 \\ -4 & -4 & -4 \\ -4 & 4 & -4 \\ 1 & 2 & 4 \\ 1 & -2 & 4 \\ 0 & 0 & 24 \end{pmatrix}$$

[0071] 其中,权重变换矩阵G和目标权重变换矩阵G'均为6行3列的矩阵,权重变换矩阵G和目标权重变换矩阵G'的权重分布情况不变。

[0072] 在实际执行中,在语音处理模型的量化卷积层对输入量化特征和权重变换矩阵进行winograd快速卷积处理之前,设置数据预处理模块,利用权值公共因子实现对权重变换矩阵中的权值的取整处理。

[0073] 步骤140、对输入量化特征和目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵。

[0074] 在该步骤中,对输入量化特征进行winograd变换,进行矩阵变换和数据重排,实现winograd快速卷积算法中的输入变换(Input transform);对目标权重变换矩阵进行winograd变换,进行矩阵变换和数据重排,实现winograd快速卷积算法中的权值变换(Filter transform)。

[0075] 将输入量化特征和目标权重变换矩阵进行winograd变换得到的矩阵进行矩阵相乘处理,可以通过批量矩阵乘法(Batched-GEMM)进行矩阵相乘,得到目标量化矩阵。

[0076] 步骤150、对目标量化矩阵进行winograd反变换。

[0077] 在该步骤中,对目标量化矩阵进行winograd反变换,进行数据重排和矩阵变换,实现winograd快速卷积算法中的输出变换(Output transform)。

[0078] 步骤160、基于权值公共因子,对目标量化矩阵的winograd反变换输出结果进行处理,得到语音数据的输出量化特征。

[0079] 在实际执行中,在语音处理模型的量化卷积层对输入量化特征和目标权重变换矩阵进行批量矩阵相乘之后,设置数据后处理模块,利用权值公共因子实现winograd快速卷积算法中的输出变换时防止数据计算溢出。

[0080] 例如,输入量化特征可以是8位整数型的数据(int8),目标权重变换矩阵也是8位整数型的数据(int8)。

[0081] 将输入量化特征和目标权重变换矩阵进行winograd变换得到的矩阵进行矩阵相乘处理,执行int16*int16计算,对目标量化矩阵进行winograd反变换,是将目标量化矩阵中的数据转换为int16数据输出的过程。

[0082] 语音信号是幅度随时间连续变化的模拟信号,语音数据是由语音信号转换得到的时间和幅度均离散的数字信号,快速、高效地对语音数据进行操作,获得预期的结果,是语音数据处理的关键。

[0083] 在实际执行中,通过整数型的数据计算,可以减小消耗的计算资源,加快语音数据的处理速度。

[0084] 相关技术中,使用winograd快速卷积算法在量化模型处理语音数据时,语音数据中时间和幅度对应的整型数据的变化范围较大,量化模型的整型数据动态范围有限,经常会发生数据计算溢出,影响深度学习模型的部署和语音数据的处理。在该实施例,通过权

值公共因子,对目标量化矩阵的winograd反变换的int16数据输出进行处理,调整量化模型的整型数据动态范围,有效防止数据计算溢出,得到最终输出的语音数据的输出量化特征。通过提取公共因子,实现防止数据溢出的winograd快速卷积算法,可有效的直接部署在实际生产环境中的语音处理模型,具体的加速部署的步骤如下所述:

[0085] 如图3所示,步骤一、部署加载语音处理模型。

[0086] 步骤二、判断语音处理模型的卷积层是否为量化卷积层,也即判断卷积层是否可以处理整数型的特征数据。

[0087] 步骤三、判断是否满足winograd快速卷积算法的加速条件。

[0088] 步骤四、根据输入量化特征的宽高,判断winograd快速卷积的输入量化特征的特征变换矩阵。

[0089] 步骤五、根据输入量化特征的宽高,确定输出结果的数量,选择对应的F (4x4,3x3) winograd int8的快速卷积或F (2x2,3x3) winograd int8的快速卷积。

[0090] 其中,F (4x4,3x3) winograd int8表示8位整数型数据使用winograd快速卷积算法,4x4表示输出结果为16个,3x3表示卷积层的卷积核为3x3卷积核。

[0091] F (2x2,3x3) winograd int8表示8位整数型数据使用winograd快速卷积算法,2x2表示输出结果为4个,3x3表示卷积层的卷积核为3x3卷积核。

[0092] 下面介绍一个具体的实施例。

[0093] 在实现语音唤醒的语音处理模型的量化卷积层上设置数据预处理模块和数据后处理模块,部署于终端设备。

[0094] 进行winograd的快速卷积处理,可有效地解决语音唤醒推理时间过长,无法实时处理而导致语音数据丢帧的问题,例如,从发出“小美小美”的唤醒关键词指令到终端设备反馈“你说”,整个语音链路时延从18毫秒降低到9毫秒以内,语音处理模型整体推理性能提升了一倍。根据本申请实施例提供的语音数据的处理方法,通过权值公共因子对winograd变换前的权重变换矩阵和winograd反变换输出结果进行处理,调整量化模型的整型数据动态范围,有效防止数据计算溢出,提升语音数据的处理速度,加快语音数据处理相关深度学习模型的部署。

[0095] 上述语音数据的处理方法步骤130包括:

[0096] 将权重变换矩阵中的权值乘以权值公共因子,得到目标权重变换矩阵。

[0097] 在该实施例中,将权重变换矩阵中的所有权值均乘以该权重变换矩阵的权值公共因子,得到目标权重变换矩阵,目标权重变换矩阵中所有的权值均为整数。

[0098] 例如,权重变换矩阵G如下式

$$[0099] \quad G = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{1}{6} & -\frac{1}{6} \\ \frac{1}{24} & \frac{1}{12} & \frac{1}{6} \\ \frac{1}{24} & -\frac{1}{12} & \frac{1}{6} \\ 0 & 0 & 1 \end{pmatrix}$$

[0100] 在该实施例中,获取的权重变换矩阵G的权值公共因子为24。

[0101] 将权重变换矩阵G的所有权值乘以权值公共因子24,得到目标权重变换矩阵G',如下所示:

$$[0102] \quad G' = \begin{pmatrix} 6 & 0 & 0 \\ -4 & -4 & -4 \\ -4 & 4 & -4 \\ 1 & 2 & 4 \\ 1 & -2 & 4 \\ 0 & 0 & 24 \end{pmatrix}$$

[0103] 其中,0权值乘以权值公共因子后,仍然为0权值。

[0104] 相关技术中,通过将winograd变换前的权重变换矩阵的小数权值通过四舍五入转换为整数权值,这种方式会使权重变换矩阵的权重分布情况改变,降低winograd快速卷积处理的计算精度。

[0105] 在该实施例中,通过在进行winograd变换之前,将权重变换矩阵中不可整除的小数权值,通过乘以权值公共因子将小数权值预处理转换成相应的整数权值,权重变换矩阵中的所有权值均乘以权值公共因子,目标权重变换矩阵和权重变换矩阵的权重分布情况不变,保证winograd快速卷积处理的计算精度。

[0106] 上述语音数据的处理方法步骤160包括:

[0107] 将目标量化矩阵中最后一列和最后一行的winograd反变换输出结果除权值公共因子,得到输出量化特征。

[0108] 输入量化特征和目标权重变换矩阵进行批量矩阵相乘得到目标量化矩阵,目标量化矩阵内的数据进行winograd反变换,得到对应的winograd反变换输出结果。

[0109] 在目标量化矩阵中最后一列和最后一行输出winograd反变换输出结果前,将目标量化矩阵中最后一列和最后一行的winograd反变换输出结果除权值公共因子,调整目标量化矩阵的整型数据动态范围,防止winograd反变换输出的数据计算溢出。

[0110] 例如,输入量化特征可以是8位整数型的数据(int8),目标权重变换矩阵也是8位整数型的数据(int8)。

[0111] 将输入量化特征和目标权重变换矩阵进行winograd变换得到的矩阵进行矩阵相乘处理,执行int16*int16计算,对目标量化矩阵进行winograd反变换,是将目标量化矩阵中的数据转换为int16数据输出的过程。

[0112] 在目标量化矩阵的最后一列和最后一行转int16数据输出前,将最后一列和最后

一行的数据除权值公共因子,防止数据计算溢出,得到语音数据的输出量化特征。

[0113] 在该实施例中,目标量化矩阵的winograd反变换输出结果采用winograd反变换输出结果的两倍整数型数据格式存储。

[0114] 例如,输入量化特征和目标权重变换矩阵进行winograd变换得到的矩阵进行矩阵相乘处理,执行int16*int16计算,目标量化矩阵的winograd反变换输出结果采用int32的数据格式存储。

[0115] 其中,int32是int16的两倍整数型数据格式。

[0116] 在该实施例中,目标量化矩阵的winograd反变换输出结果是int16的计算结果,int16的计算结果使用int32数据格式来存储,可以保持权值矩阵转换过程计算的正确性。

[0117] 上述语音数据的处理方法步骤140包括:

[0118] 对输入量化特征和目标权重变换矩阵分别进行winograd变换,得到变换后特征矩阵和变换后权重矩阵;

[0119] 将变换后特征矩阵和变换后权重矩阵按照批量矩阵乘法进行相乘,得到目标量化矩阵。

[0120] 在该实施例中,对输入量化特征进行winograd变换,进行矩阵变换和数据重排,得到变换后特征矩阵,实现winograd快速卷积算法中的输入变换。

[0121] 在实际执行中,基于输入量化特征和输入量化特征的特征变换矩阵进行winograd变换,得到变换后特征矩阵。

[0122] 在该实施例中,对目标权重变换矩阵进行winograd变换,进行矩阵变换和数据重排,得到变换后权重矩阵,实现winograd快速卷积算法中的权值变换。

[0123] 在实际执行中,基于变换后权重矩阵和量化卷积核进行winograd变换,得到变换后权重矩阵。

[0124] winograd快速卷积算法的计算公式如下:

$$[0125] \quad Y = A^T [[GgG^T] \odot [B^T dB]] A$$

[0126] 其中,d为输入量化特征,g为量化卷积核,B为特征变换矩阵,G为权重变换矩阵,G为输出特征逆变换矩阵,Y为经过winograd快速卷积计算之后的输出量化特征。

[0127] $\odot$ 为哈达玛积,表示矩阵对应位置相乘,变换后特征矩阵和变换后权重矩阵基于哈达玛积的矩阵运算方式实现批量矩阵相乘。

[0128] 在该实施例中, $[GgG^T]$ 表示变换后权重矩阵, $[B^T dB]$ 表示变换后特征矩阵;A为输出特征逆变换矩阵, $[[GgG^T] \odot [B^T dB]]$ 和A以及 A^T 的计算过程为winograd反变换的输出过程。

[0129] 下面介绍一个具体的实施例。

[0130] 如图2所示,d为输入量化特征,g为量化卷积核,B为特征变换矩阵,G为权重变换矩阵。

[0131] 通过提取权重变换矩阵G的权值公共因子,进行矩阵参数预处理,将权重变换矩阵G中不可整除的小数权值,通过乘以权值公共因子将小数权值预处理转换成相应的整数权值,得到目标权重变换矩阵 G' 。

[0132] 基于输入量化特征d和输入量化特征d的特征变换矩阵B进行winograd变换,进行矩阵变化和數據重排,得到变换后特征矩阵。

[0133] 基于目标权重变换矩阵 G' 和为量化卷积核g进行winograd变换,进行矩阵变化和

数据重排,得到变换后权重矩阵。

[0134] 将变换后特征矩阵和变换后权重矩阵进行批量矩阵相乘,得到目标量化矩阵。

[0135] 对目标量化矩阵进行winograd反变换,进行数据重排,将目标量化矩阵中最后一列和最后一行的winograd反变换输出结果除权值公共因子,调整目标量化矩阵的整型数据动态范围,进行矩阵变换,得到输出量化特征Y。

[0136] 在一些实施例中,将变换后特征矩阵和变换后权重矩阵输入至添加第一参数的批量矩阵乘函数接口,得到批量矩阵乘函数接口输出的目标量化矩阵。

[0137] 在该实施例中,进行变换后特征矩阵和变换后权重矩阵的批量矩阵相乘的处理时,将通常使用的批量矩阵乘函数接口中加入第一参数,也即在批量矩阵乘法 (Batched-GEMM) 的函数内部,把变换后特征矩阵和变换后权重矩阵的矩阵对应位置相乘的计算结果保存出去之前,先乘第一参数。

[0138] 可以理解的是,可在批量矩阵乘函数外部调用处,通过第一参数对矩阵索引进行判断,其中,对矩阵索引进行判断第一参数可以传入1或者权值公共因子,将权重变换矩阵进行取整的预处理过程中通过权值公共因子除掉的矩阵系数,在批量矩阵乘法计算时做逆运算添加回来。

[0139] 例如,输入量化特征可以是8位整数型的数据 (int8),目标权重变换矩阵也是8位整数型的数据 (int8)。

[0140] 将输入量化特征和目标权重变换矩阵进行winograd变换得到的矩阵进行矩阵相乘处理,执行int16*int16计算,在标准int16*int16批量矩阵乘函数接口添加第一参数,在批量矩阵乘法函数中,把int32的计算结果保存出去之前,先乘第一参数。

[0141] 下面对本申请实施例提供的语音数据的处理装置进行描述,下文描述的语音数据的处理装置与上文描述的语音数据的处理方法可相互对应参照。

[0142] 如图4所示,本申请实施例提供的语音数据的处理装置包括:

[0143] 第一获取模块410,用于获取待处理的语音数据的输入量化特征和权重变换矩阵;

[0144] 第二获取模块420,用于获取权重变换矩阵的权值公共因子;

[0145] 第一处理模块430,用于基于权值公共因子,对权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵;

[0146] 第二处理模块440,用于对输入量化特征和目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵;

[0147] 第三处理模块450,用于对目标量化矩阵进行winograd反变换;

[0148] 第四处理模块460,用于基于权值公共因子,对目标量化矩阵的winograd反变换输出结果进行处理,得到语音数据的输出量化特征。

[0149] 根据本申请实施例提供的语音数据的处理装置,通过权值公共因子对Winograd变换前的权重变换矩阵和winograd反变换输出结果进行处理,调整量化模型的整型数据动态范围,有效防止数据计算溢出,提升语音数据的处理速度,加快语音数据处理相关深度学习模型的部署。

[0150] 在一些实施例中,第一处理模块430用于将权重变换矩阵中的权值乘以权值公共因子,得到目标权重变换矩阵。

[0151] 在一些实施例中,第四处理模块460用于基将目标量化矩阵中最后一列和最后一

行的winograd反变换输出结果除权值公共因子,得到输出量化特征。

[0152] 在一些实施例中,目标量化矩阵的winograd反变换输出结果采用winograd反变换输出结果的两倍整数型数据格式存储。

[0153] 在一些实施例中,第二处理模块440用于对输入量化特征和目标权重变换矩阵分别进行winograd变换,得到变换后特征矩阵和变换后权重矩阵;

[0154] 将变换后特征矩阵和变换后权重矩阵按照批量矩阵乘法进行相乘,得到目标量化矩阵。

[0155] 在一些实施例中,第二处理模块440用于将变换后特征矩阵和变换后权重矩阵输入至添加第一参数的批量矩阵乘函数接口,得到批量矩阵乘函数接口输出的目标量化矩阵。

[0156] 在一些实施例中,第二处理模块440用于基于输入量化特征和输入量化特征的特征变换矩阵进行winograd变换,得到变换后特征矩阵;

[0157] 基于变换后权重矩阵和量化卷积核进行winograd变换,得到变换后权重矩阵。

[0158] 图5示例了一种电子设备的实体结构示意图,如图5所示,该电子设备可以包括:处理器(processor) 510、通信接口(Communications Interface) 520、存储器(memory) 530和通信总线540,其中,处理器510,通信接口520,存储器530通过通信总线540完成相互间的通信。处理器510可以调用存储器530中的逻辑指令,以执行语音数据的处理方法,该方法包括:获取待处理的语音数据的输入量化特征和权重变换矩阵;获取权重变换矩阵的权值公共因子;基于权值公共因子,对权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵;对输入量化特征和目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵;对目标量化矩阵进行winograd反变换;基于权值公共因子,对目标量化矩阵的winograd反变换输出结果进行处理,得到语音数据的输出量化特征。

[0159] 此外,上述的存储器530中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用,可以存储在一个计算机可读取存储介质中。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或者该技术方案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一个存储介质中,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,服务器,或者网络设备等)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM, Read-Only Memory)、随机存取存储器(RAM, Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0160] 进一步地,本申请还提供一种计算机程序产品,所述计算机程序产品包括计算机程序,计算机程序可存储在非暂态计算机可读存储介质上,所述计算机程序被处理器执行时,计算机能够执行上述各方法实施例所提供的语音数据的处理方法,该方法包括:获取待处理的语音数据的输入量化特征和权重变换矩阵;获取权重变换矩阵的权值公共因子;基于权值公共因子,对权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵;对输入量化特征和目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵;对目标量化矩阵进行winograd反变换;基于权值公共因子,对目标量化矩阵的winograd反变换输出结果进行处理,得到语音数据的输出量化特征。

[0161] 另一方面,本申请实施例还提供一种非暂态计算机可读存储介质,其上存储有计

算机程序,该计算机程序被处理器执行时实现以执行上述各实施例提供的语音数据的处理方法,该方法包括:获取待处理的语音数据的输入量化特征和权重变换矩阵;获取权重变换矩阵的权值公共因子;基于权值公共因子,对权重变换矩阵中的权值进行取整处理,得到目标权重变换矩阵;对输入量化特征和目标权重变换矩阵进行winograd变换和矩阵相乘处理,得到目标量化矩阵;对目标量化矩阵进行winograd反变换;基于权值公共因子,对目标量化矩阵的winograd反变换输出结果进行处理,得到语音数据的输出量化特征。

[0162] 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。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0163] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件。基于这样的理解,上述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,该计算机软件产品可以存储在计算机可读存储介质中,如ROM/RAM、磁碟、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,服务器,或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0164] 最后应说明的是:以上实施例仅用以说明本申请的技术方案,而非对其限制;尽管参照前述实施例对本申请进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

[0165] 以上实施方式仅用于说明本申请,而非对本申请的限制。尽管参照实施例对本申请进行了详细说明,本领域的普通技术人员应当理解,对本申请的技术方案进行各种组合、修改或者等同替换,都不脱离本申请技术方案的精神和范围,均应涵盖在本申请的权利要求范围中。

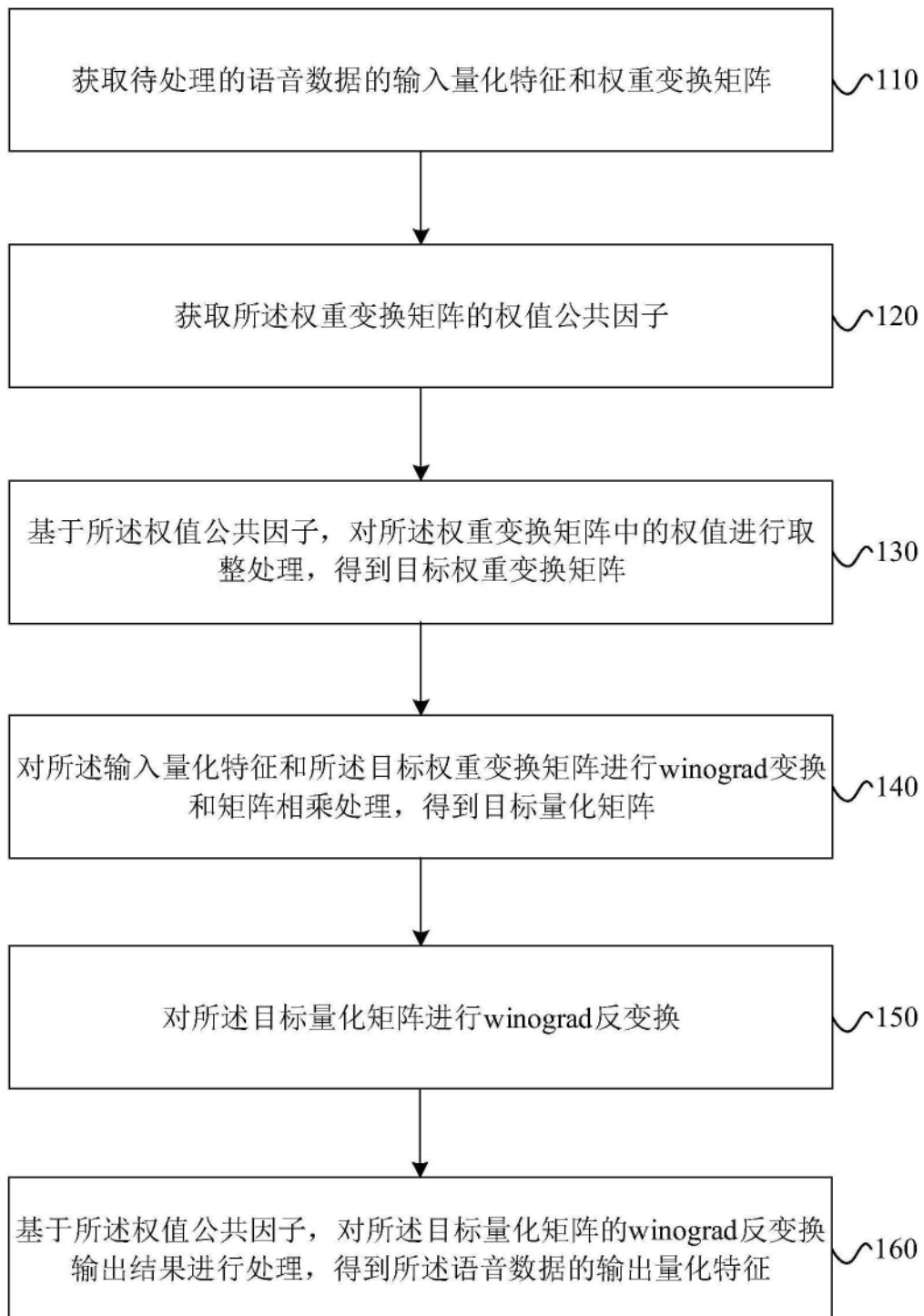


图1

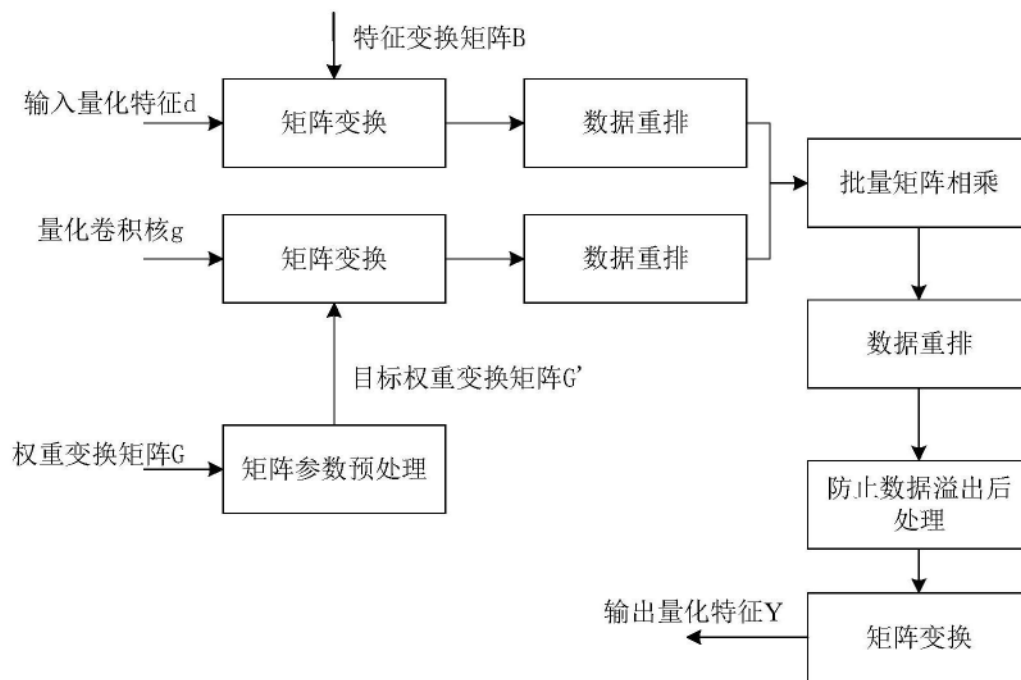


图2

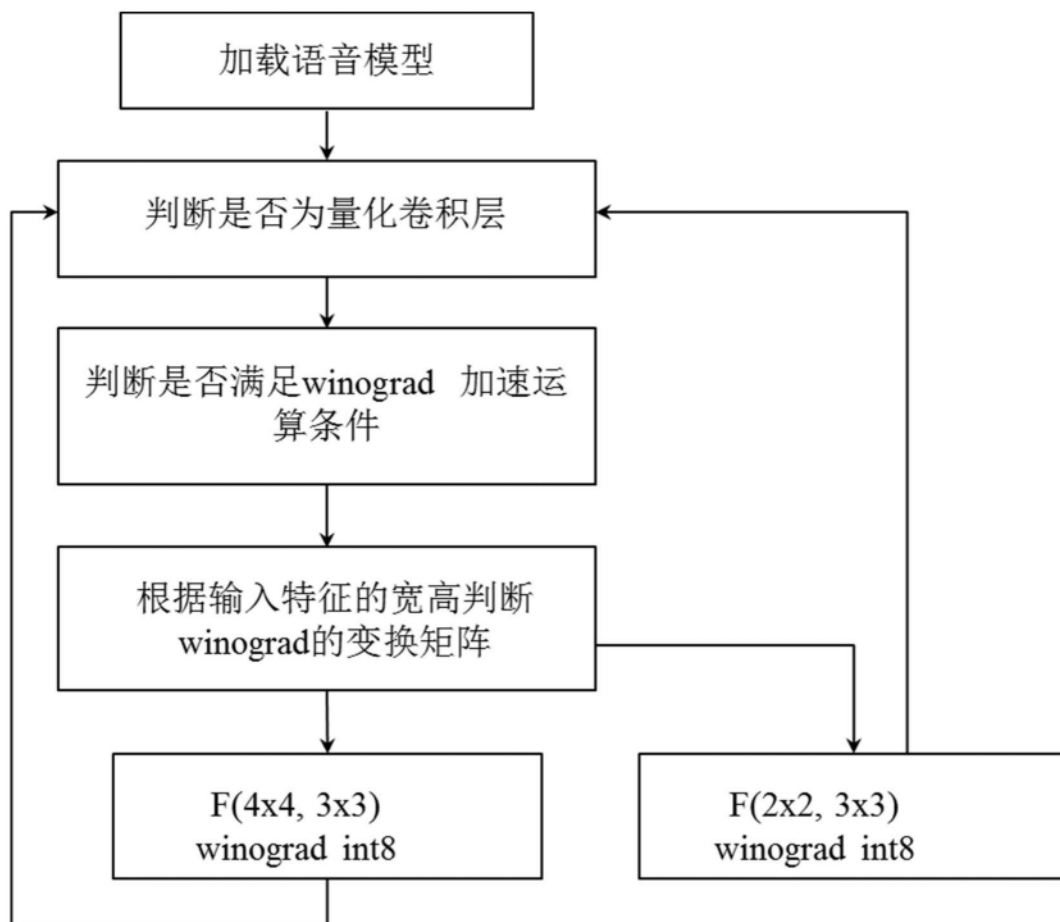


图3

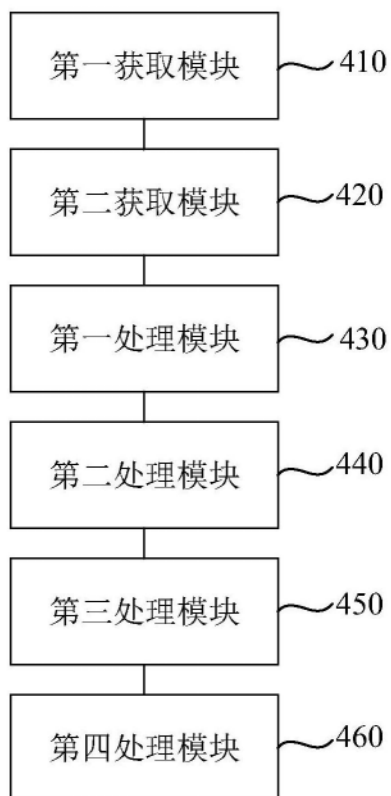


图4

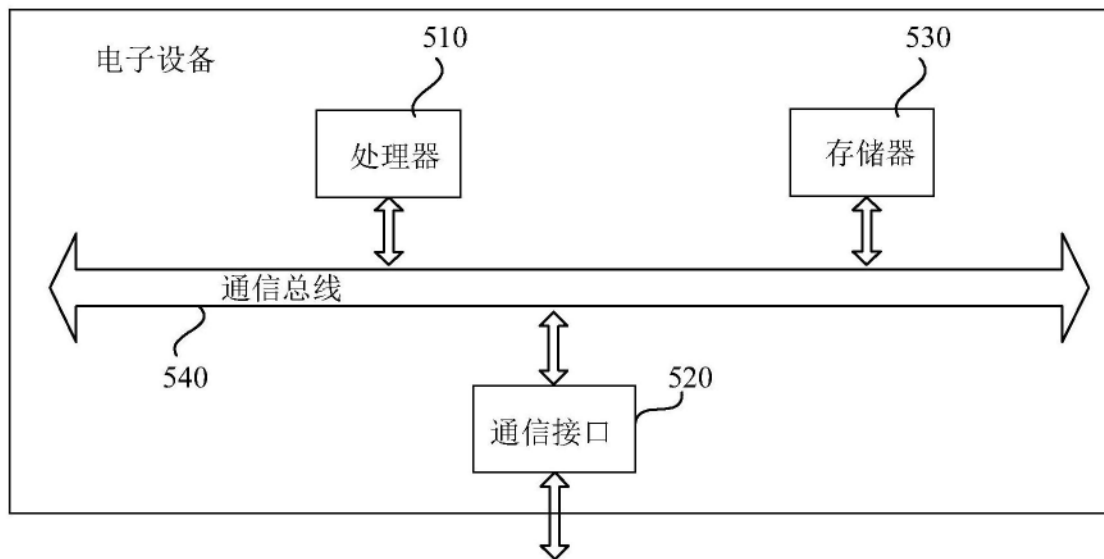


图5