

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2021 年 12 月 30 日 (30.12.2021)



(10) 国际公布号
WO 2021/258752 A1

- (51) 国际专利分类号:
G06N 3/08 (2006.01) **G06N 3/04** (2006.01)
- (21) 国际申请号: PCT/CN2021/076982
- (22) 国际申请日: 2021 年 2 月 20 日 (20.02.2021)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
202010589233.5 2020年6月24日 (24.06.2020) CN
- (71) 申请人: 苏州浪潮智能科技有限公司
(**INSPUR SUZHOU INTELLIGENT TECHNOLOGY CO., LTD.**) [CN/CN]; 中国江苏省苏州市吴中区吴中经济开发区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。
- (72) 发明人: 王曦辉 (**WANG, Xihui**); 中国江苏省苏州市吴中区郭巷街道官浦路1号9幢, Jiangsu 215000 (CN)。

- (74) 代理人: 北京集佳知识产权代理有限公司
(**UNITALEN ATTORNEYS AT LAW**); 中国北京市朝阳区建国门外大街22号赛特广场7层, Beijing 100004 (CN)。
- (81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。
- (84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,

(54) **Title:** 4-BIT QUANTIZATION METHOD AND SYSTEM FOR NEURAL NETWORK

(54) 发明名称: 一种神经网络的4比特量化方法及系统

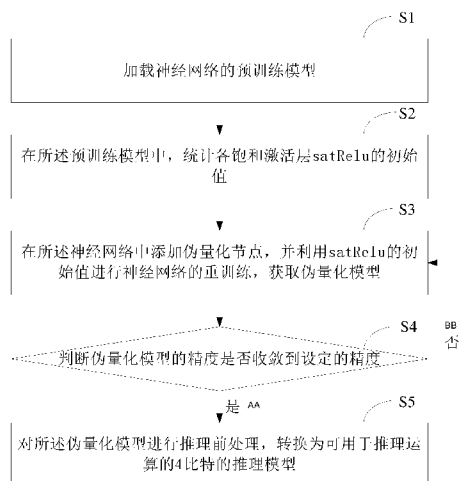


图 1

- S1 LOAD A PRE-TRAINING MODEL OF A NEURAL NETWORK
S2 IN THE PRE-TRAINING MODEL, COLLECT STATISTICS ABOUT INITIAL VALUES OF SATURATION ACTIVATION LAYERS SATRELU
S3 ADD PSEUDO QUANTIZATION NODES TO THE NEURAL NETWORK AND RE-TRAIN THE NEURAL NETWORK BY USING THE INITIAL VALUES OF THE SATRELU TO OBTAIN A PSEUDO QUANTIZATION MODEL
S4 DETERMINE WHETHER THE PRECISION OF THE PSEUDO QUANTIZATION MODEL CONVERGES TO A SET PRECISION
S5 PERFORM REASONING PREPROCESSING ON THE PSEUDO QUANTIZATION MODEL, AND CONVERT SAME INTO A 4-BIT REASONING MODEL THAT CAN BE USED FOR REASONING OPERATION
AA YES
BB NO

(57) **Abstract:** A 4-bit quantization method and system for a neural network. The method comprises: loading a pre-training model of the neural network (S1); in the pre-training model, collecting statistics about initial values of saturation activation layers satRelu (S2); adding pseudo quantization nodes to the neural network, and re-training the neural network by using the initial values of the satReLU to obtain a pseudo quantization model (S3); determining whether the precision of the pseudo quantization model converges to a set precision; (S4) if yes, performing reasoning preprocessing on the pseudo quantization model, and converting same into a 4-bit reasoning model that can be used for reasoning operation (S5); otherwise, returning to carry out re-training of the neural network. The system mainly comprises: a loading module, a statistics module, a re-training module, a determination module, and a conversion module. According to the method and system, the training efficiency can be effectively improved on the basis of ensuring the accuracy of the training result.

IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：一种神经网络的4比特量化方法及系统，该方法包括：加载神经网络的预训练模型(S1)；在预训练模型中，统计各饱和激活层satRelu的初始值(S2)；在神经网络中添加伪量化节点，并利用satRelu的初始值进行神经网络的重训练，获取伪量化模型(S3)；判断伪量化模型的精度是否收敛到设定的精度(S4)；如果是，对伪量化模型进行推理前处理，转换为可用于推理运算的4比特的推理模型(S5)；否则返回进行神经网络的重训练。该系统主要包括：加载模块、统计模块、重训练模块、判断模块和转换模块。通过该方法及系统能够在确保训练结果准确性的基础上，有效提高训练效率。

一种神经网络的 4 比特量化方法及系统

本申请要求于 2020 年 06 月 24 日提交中国专利局、申请号为 202010589233.5、发明名称为“一种神经网络的 4 比特量化方法及系统”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本申请涉及神经网络模型压缩技术领域，特别是涉及一种神经网络的 4 比特量化方法及系统。

10 背景技术

在神经网络中，Neural Network模型一般会占用很大的磁盘空间，比如 AlexNet的模型文件超过了200 MB（MByte，兆字节）。模型包含了数百万的参数，绝大部分的磁盘空间都用于存储模型参数。由于模型参数是浮点数类型的，普通的压缩算法很难压缩其空间，因此，引入模型量化，通过减少表示每个权重所需的比特数来压缩原始网络，能够大大提高网络的运行速度。因此，如何对神经网络进行量化是个重要的技术问题。

目前神经网络量化的主流方法是8bit量化，大多数训练和推理框架都支持8比特量化。然而相比于8比特量化，4比特量化可以在8比特的基础上将模型的体积继续压缩1倍，运行速度提升50%。因此，4比特量化也逐渐引起人们重视。

目前的4比特量化算法通常是从头开始训练一个网络，直到整个网络训练完毕，以imagenet这种大型数据集一般需要训练100个以上的周期数才能训练完毕。且采用非线性量化的方式提升模型精度。

然而，目前的4比特量化算法中，由于量化周期数较多，训练时间较长，使得量化效率较低。

发明内容

-2-

本申请提供了一种神经网络的4比特量化方法及系统,以解决现有技术中的神经网络量化方法量化效率较低的问题。

为了解决上述技术问题,本申请实施例公开了如下技术方案:

一种神经网络的4比特量化方法,所述方法包括:

5 加载神经网络的预训练模型;

在所述预训练模型中,统计各饱和激活层satRelu的初始值;

在所述神经网络中添加伪量化节点,并利用satRelu的初始值进行神经网络的重训练,获取伪量化模型;

判断所述伪量化模型的精度是否收敛到设定的精度;

10 如果是,对所述伪量化模型进行推理前处理,转换为可用于推理运算的4比特的推理模型,所述推理前处理过程包括:常量折叠、二次量化以及激活等价变换;

如果否,继续进行神经网络的重训练。

15 可选地,在所述预训练模型中,统计各饱和激活层satRelu的初始值的方法,包括:

将所述神经网络中所有的激活层relu替换为饱和激活层satRelu;

根据所获取的命令,获取各饱和激活层satRelu的激活值;

根据所述激活值,利用直方图统计分布数据;

20 选取所述直方图中位于99.999%点位的激活值作为饱和激活层satRelu中参数max的初始值,其中,satRelu定义为:

$$y = \text{satRelu}(x) = \begin{cases} 0, & x \in (-\infty, 0) \\ x, & x \in [0, \max) \\ \max, & x \in [\max, +\infty) \end{cases},$$

且反向传播中,satRelu对于参数max的梯度为: $\frac{\partial y}{\partial \max} = \begin{cases} 0, & x \in (-\infty, \max) \\ 1, & x \in [\max, +\infty) \end{cases}$,

satRelu对于输入x的梯度为 $\frac{\partial y}{\partial x} = \begin{cases} 0, & x \in (-\infty, 0) \cup [\max, +\infty) \\ 1, & x \in [0, \max) \end{cases}$, max为饱和激活

层satRelu的最大值。

25 可选地,在神经网络重训练过程中,采用L2正则化的方法对参数max进行压缩。

—3—

可选地，重训练周期 ≤ 10 ，参数max的取值 ≤ 1 。

可选地，在所述神经网络中添加伪量化节点，并利用satRelu的初始值进行神经网络的重训练，获取伪量化模型的方法，包括：

在神经网络的权重层之前插入权重伪量化层，且在激活层之前插入激活伪量化层；

利用公式 $y = quant(w) = clip(round(w * scale)) / scale$ ，对权重伪量化层进行神经网络的重训练，其中，w为权重值，n取值为4，比例系数scale为：

$$scale = \frac{2^{n-1} - 1}{\max(w) - \min(w)} ;$$

利用公式 $y = quant(x) = clip(round(x * scale)) / scale$ ，对激活伪量化层进行神经网络的重训练，其中，x为每层的激活值，max为satRelu的最大值，n取

值为4，比例系数scale为： $sclae = \frac{2^n - 1}{max}$ ；

可选地，进行神经网络的重训练时，反向传播过程采用直通估计器计算梯度。

一种神经网络的4比特量化系统，所述系统包括：

15 加载模块，用于加载神经网络的预训练模型；

统计模块，用于在所述预训练模型中，统计各饱和激活层satRelu的初始值；

重训练模块，用于在所述神经网络中添加伪量化节点，并利用satRelu的初始值进行神经网络的重训练，获取伪量化模型；

20 判断模块，用于判断所述伪量化模型的精度是否收敛到设定的精度；

转换模块，用于当所述伪量化模型的精度收敛到设定的精度时，对所述伪量化模型进行推理前处理，转换为可用于推理运算的4比特的推理模型，所述推理前处理过程包括：常量折叠、二次量化以及激活等价变换。

可选地，所述统计模块包括：

25 替换单元，用于将所述神经网络中所有的激活层relu替换为饱和激活层satRelu；

—4—

激活值获取单元, 用于根据所获取的命令, 获取各饱和激活层satRelu的激活值;

统计单元, 用于根据所述激活值, 利用直方图统计分布数据;

初始值选取单元, 用于选取所述直方图中位于99.999%点位的激活值

5 作为饱和激活层satRelu中参数max的初始值, 其中, satRelu定义为:

$$y = \text{satRelu}(x) = \begin{cases} 0, & x \in (-\infty, 0) \\ x, & x \in [0, \max) \\ \max, & x \in [\max, +\infty) \end{cases},$$

且反向传播中, satRelu对于参数max的梯度为: $\frac{\partial y}{\partial \max} = \begin{cases} 0, & x \in (-\infty, \max) \\ 1, & x \in [\max, +\infty) \end{cases}$,

satRelu对于输入x的梯度为 $\frac{\partial y}{\partial x} = \begin{cases} 0, & x \in (-\infty, 0) \cup [\max, +\infty) \\ 1, & x \in [0, \max) \end{cases}$, max为饱和激活层satRelu的最大值。

10 可选地, 所述重训练模块包括:

伪量化层插入单元, 用于在神经网络的权重层之前插入权重伪量化层, 且在激活层之前插入激活伪量化层;

第一重训练单元, 用于利用公式 $y = \text{quant}(w) = \text{clip}(\text{round}(w * \text{scale})) / \text{scale}$,

对权重伪量化层进行神经网络的重训练, 其中, w为权重值, n取值为4,

15 比例系数scale为: $\text{scale} = \frac{2^{n-1} - 1}{\max(w) - \min(w)}$;

第二重训练单元, 用于利用公式 $y = \text{quant}(x) = \text{clip}(\text{round}(x * \text{scale})) / \text{scale}$, 对

激活伪量化层进行神经网络的重训练, 其中, x为每层的激活值, max为

satRelu的最大值, n取值为4, 比例系数scale为: $\text{sclae} = \frac{2^n - 1}{\max}$ 。

可选地, 所述推理前处理模块包括:

20 常量折叠单元, 用于将batchNorm层融合进卷积中, 其中, 卷积的计算

公式为: $z = w * x + b$, batchNorm的计算公式为: $y = \frac{z - \text{mean}}{\sqrt{\text{var}}} * \beta + \gamma$, 合并

-5-

后的新卷积的计算公式为： $w' = \frac{w}{\sqrt{var}} * \beta$ ， $b' = (b - mean) \sqrt{var} * \beta + \gamma$ ；

二次量化单元，用于对权重进行二次量化，获取二次量化后权重的量化比例系数scale；

激活等价变换单元，用于利用公式 $y = \frac{W_q * x_q + W_q * scale_a * 8 + b_q}{scale_w * scale_a}$ ，对

5 激活做等价变换。

本申请的实施例提供的技术方案可以包括以下有益效果：

本申请提供一种神经网络的4比特量化方法，该量化方法首先加载神经网络的预训练模型，在预训练模型中统计各satRelu层的初始值，在神经网络中添加伪量化节点，并利用satRelu的初始值进行神经网络的重训练，
10 通过几个周期的重训练获取伪量化模型，并利用推理方法将伪量化模型转换为4比特的推理模型，通过推理算法使在推理前完成全4比特推理过程，有利于提高计算速度，并确保最终的推理模型能够直接应用于4比特的GPU上，支持4比特的GPU运算，有利于提高本发明的实用性。该方法在重训练时，采用伪量化的方式进行，通过在权重层和激活层之前插入相
15 应的伪量化层，用于模拟模型量化对整个神经网络的影响，借由训练来使模型学习并适应对神经网络的影响，从而能够大大提高量化模型的准确率，获取到的模型精度损失能够控制在1%以内，确保4比特量化结果满足精度要求。而且，本实施例中重训练的周期为10，即：对神经网络进行10
20 次重训练，大大节省训练时间，能够在确保精度的基础上有效提高训练效率，进而提高神经网络4比特量化效率。

本实施例中对神经网络进行重训练，对权重层和激活层采用线性伪量化，在确保训练精度的基础上提高计算速度，从而有效提高神经网络量化效率。

本申请还提供一种神经网络的4比特量化系统。该系统主要包括：加载模块、统计模块、重训练模块、判断模块和转换模块。通过加载模块和重训练模块，能够通过重训练的方式加载一预训练模型，对其进行几个周
25 期的重训练，获取伪量化模型，当判断模块判定伪量化模型的精度收敛到

—6—

设定精度时，启动转换模块将对伪量化模型进行推理前处理，转换为可用于推理运算的 4 比特的推理模型。本实施例采用较少的重训练次数，通常为 10 个周期，能够有效节省训练时间，从而大大提高神经网络量化效率。而且，利用推理前处理模块在推理态对伪量化模型进行处理，能够在推理前完成全 4 比特推理，有利于进一步提高计算速度，从而提高量化效率。

应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。

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，对于本领域普通技术人员而言，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。

图 1 为本申请实施例所提供的一种神经网络的 4 比特量化方法的流程示意图；

图 2 为本申请实施例所提供的一种神经网络的 4 比特量化系统的结构示意图。

具体实施方式

为了使本技术领域的人员更好地理解本申请中的技术方案，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本申请保护的范围。

为了更好地理解本申请，下面结合附图来详细解释本申请的实施方式。

实施例一

参见图 1，图 1 为本申请实施例所提供的一种神经网络的 4 比特量化方法的流程示意图。由图 1 可知，本实施例中神经网络的 4 比特量化方法，

-7-

主要包括如下过程:

S1: 加载神经网络的预训练模型。

S2: 在预训练模型中, 统计各饱和激活层 satRelu 的初始值。

具体地, 步骤 S2 又包括:

5 S21: 将神经网络中所有的激活层 relu 替换为饱和激活层 satRelu。

S22: 根据所获取的命令, 获取各饱和激活层 satRelu 的激活值。

S23: 根据激活值, 利用直方图统计分布数据。

S24: 选取直方图中位于 99.999% 点位的激活值作为饱和激活层 satRelu 中参数 max 的初始值。其中, satRelu 定义为:

$$10 \quad y = \text{satRelu}(x) = \begin{cases} 0, & x \in (-\infty, 0) \\ x, & x \in [0, \max) \\ \max, & x \in [\max, +\infty) \end{cases},$$

且反向传播中, satRelu 对于参数 max 的梯度为: $\frac{\partial y}{\partial \max} = \begin{cases} 0, & x \in (-\infty, \max) \\ 1, & x \in [\max, +\infty) \end{cases}$,

satRelu 对于输入 x 的梯度为 $\frac{\partial y}{\partial x} = \begin{cases} 0, & x \in (-\infty, 0) \cup [\max, +\infty) \\ 1, & x \in [0, \max) \end{cases}$ 。

由步骤 S1 和 S2 可知, 本实施例在进行训练前, 首先加载神经网络的预训练模型, 可以通过运行网络脚本的形式实现。通过运行网络脚本, 统计各层激活值的分布情况, 对于每层激活值, 可以采用 4096 个采样点统计直方图分布, 选取直方图分布中位于 99.999% 的点位的激活值, 作为 satRelu 中唯一参数 max 的初始值, 用于后续的迭代训练。

20 本实施例中 satRelu 的下限值为 0, satRelu 的下限值为 max, 也就是 satRelu 的最大值为 max, max 是一个变量, 在神经网络训练的过程逐渐变小, 此处需要求取 max 的初始值。

本实施例在训练时, 需要将原网络中的所有 relu 层替换为 satRelu 层, satRelu 的定义如下:

$$y = \text{satRelu}(x) = \begin{cases} 0, & x \in (-\infty, 0) \\ x, & x \in [0, \max) \\ \max, & x \in [\max, +\infty) \end{cases}$$

其中, max 为直方图统计出的各层的初始值。且反向传播中, satRelu

—8—

对于参数max的梯度为： $\frac{\partial y}{\partial \max} = \begin{cases} 0, & x \in (-\infty, \max) \\ 1, & x \in [\max, +\infty) \end{cases}$ ，对于输入x的梯度为 $\frac{\partial y}{\partial x} = \begin{cases} 0, & x \in (-\infty, 0) \cup [\max, +\infty) \\ 1, & x \in [0, \max) \end{cases}$ 。

进一步地，本实施例在神经网络重训练过程中，对于max参数，采用L2正则化对其进行压缩，能够有效减小量化误差，增加网络精度。

5 本实施例中重训练周期 ≤ 10 ，参数max的取值小于等于1，其中优选取值为：重训练周期为10，参数max为1。本实施例中参数max的取值设置，能够确保每层输出适当，有利于提高网络精度。在训练过程中，需要观察max值，采用合适的正则化参数，保证在网络收敛时，max取值在1附近。

10 继续参见图1可知，统计各饱和激活层satRelu的初始值之后，执行步骤S3：在神经网络中添加伪量化节点，并利用satRelu的初始值进行神经网络的重训练，获取伪量化模型。

本实施例在神经网络中添加伪量化节点，具体是：对神经网络的卷积的输入和权重，以及，神经网络的全连接的输入和权重添加伪量化节点。

15 S31：在神经网络的权重层之前插入权重伪量化层，且在激活层之前插入激活伪量化层。

本实施例在模型训练中，整个神经网络采用伪量化的方式进行训练，即：在常规神经网络中，在权重层和激活层之前插入伪量化层，通过权重伪量化层和激活伪量化层，能够模拟模型量化对整个神经网络的影响，并通过训练使得模型学习并适应这种影响，从而能够有效提高最种量化模型的准确率。

20 S32：利用公式 $y = \text{quant}(w) = \text{clip}(\text{round}(w * \text{scale})) / \text{scale}$ ，对权重伪量化层进行神经网络的重训练。

其中，w为权重值，n取值为4，比例系数scale为：

$$25 \quad \text{scale} = \frac{2^{n-1} - 1}{\max(w) - \min(w)}。$$

本实施例中对权重伪量化层进行神经网络的重训练之后，获取到伪量化模型，伪量化模型中的权重值因为训而发生变化。另外，本实施例为神

—9—

神经网络 4 比特量化，量化后的取值范围为-8~7，因此，clip 的上限为 7，下限为-8，round 函数用于对输入进行四舍五入取整，scale 的取值采用每个输出通道共享一个 scale 的方式。

5 S33: 利用公式 $y = \text{quant}(x) = \text{clip}(\text{round}(x * \text{scale})) / \text{scale}$ ，对激活伪量化层进行神经网络的重训练。

其中，x 为每层的激活值，max 为 satRelu 的最大值，n 取值为 4，比例系数 scale 为： $\text{scale} = \frac{2^n - 1}{\text{max}}$ 。

10 由于每层激活后面会有一个 relu 层，因此激活值不会出现负值，量化后的取值范围在 0~15 之间，因此 clip 的上限为 15，下限为 0，后续在推理时需要继续将激活映射到-8~7 之间，从而便于 GPU (Graphics Processing Unit, 图形处理器) 计算，scale 的取值是每个层共享一个 scale 的方式。

本实施例中进行神经网络的重训练时，反向传播过程采用直通估计器计算梯度，即 STE (straight-through estimator)。具体地，由于量化函数为离散不可导函数，采用直通估计器计算梯度，利用网络量化的 bit 数逐渐降低到 4 比特的方式，量化的 bit 数 n 取值为 8, 6, 5, 4，即先从 8bit 量化开始训练，在训练过程中逐渐将 bit 数降低到 4bit。使用该方式可以减小梯度不匹配的问题，使网络逐渐适应由于量化带来的误差，从而有效提高运行精度和量化效率。

20 继续参见图 1 可知，本实施例中进行神经网络的重训练，获取伪量化模型之后，执行步骤 S4: 判断伪量化模型的精度是否收敛到设定的精度。

如果伪量化模型的精度收敛到设定的精度，执行步骤 S5: 对伪量化模型进行推理前处理，转换为可用于推理运算的 4 比特的推理模型。

25 本实施例先重新加载训练模型，将 satRelu 层重新更改为普通的 Relu 层，从而确保网络结构不改变。推理前处理过程包括：常量折叠、二次量化以及激活等价变换。

具体地，在训练过程中，batchNorm 层不做任何处理，不会影响网络的精度。在训练后，需将 batchNorm 层融合进卷积中，常量折叠过程的计算公式包括：

-10-

1) 卷积的计算过程: $z = w * x + b$;

2) batchNorm 的计算过程: $y = \frac{z - \text{mean}}{\sqrt{\text{var}}} * \beta + \gamma$;

3) 合并后的新卷积的计算过程采用公式:

$$w' = \frac{w}{\sqrt{\text{var}}} * \beta \text{ 和 } b' = (b - \text{mean})\sqrt{\text{var}} * \beta + \gamma。$$

5 由于 batchnorm 层折叠后, 权重参数发生变化, 本实施例通过对权重做二次量化, 能够提高计算的精度。此处量化方法与训练时权重的量化方式一致, 重新得到权重的量化 scale, 而激活的 scale 不变, 仍采用训练获得的 scale。

10 本实施例中 TensorCore 可以支持 GPU 上的 4 比特的运算, 因此可用 TensorCore 代码进行推理计算, 但 TensorCore 仅支持有符号的 4bit 与 4bit 或者无符号 4bit 与无符号 4bit 的运算, 但训练后的网络激活为无符号 4bit 数, 权重为有符号 4bit 数, 通过将激活做等价变换, 能够支持 GPU 运算, 提高本方法的实用性。具体的激活等价变换为: 对激活减掉中间值 8, 变换为有符号 4bit 数据后进行推理。卷积的等价变换为

$$\begin{aligned} 15 \quad y &= W * x + b \\ &= \frac{\text{scale}_w W * \text{scale}_a x + \text{scale}_w * \text{scale}_a * b}{\text{scale}_w * \text{scale}_a} \\ &= \frac{\text{scale}_w W * \text{scale}_a (x - 8) + \text{scale}_w W * \text{scale}_a * 8 + \text{scale}_w * \text{scale}_a * b}{\text{scale}_w * \text{scale}_a} \\ &= \frac{W_q * x_q + W_q * \text{scale}_a * 8 + b_q}{\text{scale}_w * \text{scale}_a} \end{aligned}$$

根据上述卷积等价变换公式, 卷积的推理计算过程分为以下几步:

20 1) 首先对卷积的输入做 pad 操作。

2) 对输入和权重分别乘以各自的 scale, 得到 4bit 的量化值。即 $W_q = \text{scale}_w W$, $x_q = \text{scale}_a (x - 8)$ 。

3) 此处的权重和输入均为 int32 格式, 但其中的数据范围均为 int4, 即数据范围在 -8~7 之间, 而 cpu 中没有 int4 这种数据类型, 因此需分别将

-11-

权重和激活取低 4bit 的数据，通过移位操作将 8 个数据重新拼接为一个 int32，减少数据存储占用的空间，方便 GPU 取数据进行运算。

4) 进行 4bit 卷积运算，即 $W_q * x_q$ ，然后加上偏置 $W_q * scale_a * 8 + b_q$ 。

5) 进行反量化，即除以 $scale_w * scale_a$ ，由于反量化的操作可与下一层的卷积输入做常量折叠合并，即可隐藏掉该运算。

最后，需要一些网络的整体检查：如常量折叠，合并一些常量计算，去除冗余算子后，将该模型保存下来，直接保存出来即是全 int4 推理模型，即可用该模型进行推理计算。

如果伪量化模型的精度没有收敛到设定的精度，返回步骤 S3，重新进行神经网络的重训练，并获取新的伪量化模型，直到新的伪量化模型的精度收敛到设定的精度为止。

实施例二

在图 1 所示实施例的基础之上参见图 2，图 2 为本申请实施例所提供的一种神经网络的 4 比特量化系统的结构示意图。由图 2 可知，本实施例中神经网络的 4 比特量化系统中主要包括：加载模块、统计模块、重训练模块、判断模块和转换模块。

其中，加载模块，用于加载神经网络的预训练模型；统计模块，用于在预训练模型中，统计各饱和激活层 satRelu 的初始值；重训练模块，用于在神经网络中添加伪量化节点，并利用 satRelu 的初始值进行神经网络的重训练，获取伪量化模型；判断模块，用于判断伪量化模型的精度是否收敛到设定的精度；转换模块，用于当伪量化模型的精度收敛到设定的精度时，对伪量化模型进行推理前处理，转换为可用于推理运算的 4 比特的推理模型，其中，推理前处理过程包括：常量折叠、二次量化以及激活等价变换。

进一步地，统计模块包括：替换单元、激活值获取单元、统计单元和初始值选取单元。其中，替换单元，用于将所述神经网络中所有的激活层 relu 替换为饱和激活层 satRelu；激活值获取单元，用于根据所获取的命令，获取各饱和激活层 satRelu 的激活值；统计单元，用于根据所述激活值，利用直方图统计分布数据；初始值选取单元，用于选取所述直方图中位于 99.999% 点位的激活值作为饱和激活层 satRelu 中参数 max 的初始值，其中，

-12-

satRelu 定义为:

$$y = \text{satRelu}(x) = \begin{cases} 0, & x \in (-\infty, 0) \\ x, & x \in [0, \max) \\ \max, & x \in [\max, +\infty) \end{cases},$$

且反向传播中, satRelu 对于参数 max 的梯度为: $\frac{\partial y}{\partial \max} = \begin{cases} 0, & x \in (-\infty, \max) \\ 1, & x \in [\max, +\infty) \end{cases}$,

satRelu 对于输入 x 的梯度为 $\frac{\partial y}{\partial x} = \begin{cases} 0, & x \in (-\infty, 0) \cup [\max, +\infty) \\ 1, & x \in [0, \max) \end{cases}$ 。

- 5 重训练模块包括: 伪量化层插入单元、第一重训练单元和第二重训练单元。其中, 伪量化层插入单元, 用于在神经网络的权重层之前插入权重伪量化层, 且在激活层之前插入激活伪量化层; 第一重训练单元, 用于利用公式 $y = \text{quant}(w) = \text{clip}(\text{round}(w * \text{scale})) / \text{scale}$, 对权重伪量化层进行神经网络的重训练, 其中, w 为权重值, n 取值为 4, 比例系数 scale 为:

$$10 \quad \text{scale} = \frac{2^{n-1} - 1}{\max(w) - \min(w)}; \text{第二重训练单元, 用于利用公式}$$

$y = \text{quant}(x) = \text{clip}(\text{round}(x * \text{scale})) / \text{scale}$, 对激活伪量化层进行神经网络的重训练, 其中, x 为每层的激活值, max 为 satRelu 的最大值, n 取值为 4, 比例系数 scale 为: $\text{sclae} = \frac{2^n - 1}{\max}$ 。

- 推理前处理模块包括: 常量折叠单元、二次量化单元和激活等价变换单元。其中, 常量折叠单元, 用于将 batchNorm 层融合进卷积中, 其中, 卷积的计算公式为: $z = w * x + b$, batchNorm 的计算公式为:

$$y = \frac{z - \text{mean}}{\sqrt{\text{var}}} * \beta + \gamma, \text{合并后的新卷积的计算公式为: } w' = \frac{w}{\sqrt{\text{var}}} * \beta,$$

$b' = (b - \text{mean})\sqrt{\text{var}} * \beta + \gamma$, 二次量化单元, 用于对权重进行二次量化, 获取二次量化后权重的量化比例系数 scale; 激活等价变换单元, 用于利用公

— 13 —

$$\text{式 } y = \frac{W_q * x_q + W_q * scale_a * 8 + b_q}{scale_w * scale_a}, \text{ 对激活做等价变换。}$$

该实施例中神经网络的 4 比特量化系统的工作原理和工作方法，在图 1 所示的实施例一种已经详细阐述，在此不再赘述。

5 以上所述仅是本申请的具体实施方式，使本领域技术人员能够理解或实现本申请。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

权 利 要 求

1、一种神经网络的 4 比特量化方法，其特征在于，所述方法包括：
加载神经网络的预训练模型；

在所述预训练模型中，统计各饱和激活层 satRelu 的初始值；

5 在所述神经网络中添加伪量化节点，并利用 satRelu 的初始值进行神经网络的重训练，获取伪量化模型；

判断所述伪量化模型的精度是否收敛到设定的精度；

如果是，对所述伪量化模型进行推理前处理，转换为可用于推理运算的 4 比特的推理模型，所述推理前处理过程包括：常量折叠、二次量化以及
10 及激活等价变换；

如果否，继续进行神经网络的重训练。

2、根据权利要求 1 所述的一种神经网络的 4 比特量化方法，其特征在于，在所述预训练模型中，统计各饱和激活层 satRelu 的初始值的方法，包括：

15 将所述神经网络中所有的激活层 relu 替换为饱和激活层 satRelu；

根据所获取的命令，获取各饱和激活层 satRelu 的激活值；

根据所述激活值，利用直方图统计分布数据；

选取所述直方图中位于 99.999% 点位的激活值作为饱和激活层 satRelu 中参数 max 的初始值，其中，satRelu 定义为：

$$20 \quad y = \text{satRelu}(x) = \begin{cases} 0, & x \in (-\infty, 0) \\ x, & x \in [0, \text{max}) \\ \text{max}, & x \in [\text{max}, +\infty) \end{cases},$$

且反向传播中，satRelu 对于参数 max 的梯度为： $\frac{\partial y}{\partial \text{max}} = \begin{cases} 0, & x \in (-\infty, \text{max}) \\ 1, & x \in [\text{max}, +\infty) \end{cases}$ ，

satRelu 对于输入 x 的梯度为 $\frac{\partial y}{\partial x} = \begin{cases} 0, & x \in (-\infty, 0) \cup [\text{max}, +\infty) \\ 1, & x \in [0, \text{max}) \end{cases}$ ，max 为饱和激活

层 satRelu 的最大值。

3、根据权利要求 2 所述的一种神经网络的 4 比特量化方法，其特征在于，在神经网络的重训练过程中，采用 L2 正则化的方法对参数 max 进行
25 压缩。

-15-

4、根据权利要求2所述的一种神经网络的4比特量化方法，其特征在于，重训练周期 ≤ 10 ，参数max的取值 ≤ 1 。

5、根据权利要求1所述的一种神经网络的4比特量化方法，其特征在于，在所述神经网络中添加伪量化节点，并利用satRelu的初始值进行神经网络的重训练，获取伪量化模型的方法，包括：

在神经网络的权重层之前插入权重伪量化层，且在激活层之前插入激活伪量化层；

利用公式 $y = \text{quant}(w) = \text{clip}(\text{round}(w * \text{scale})) / \text{scale}$ ，对权重伪量化层进行神经网络的重训练，其中，w为权重值，n取值为4，比例系数scale为：

$$10 \quad \text{scale} = \frac{2^{n-1} - 1}{\max(w) - \min(w)};$$

利用公式 $y = \text{quant}(x) = \text{clip}(\text{round}(x * \text{scale})) / \text{scale}$ ，对激活伪量化层进行神经网络的重训练，其中，x为每层的激活值，max为satRelu的最大值，n取值为4，比例系数scale为： $\text{sclae} = \frac{2^n - 1}{\max}$ 。

6、根据权利要求2-4中任一所述的一种神经网络的4比特量化方法，其特征在于，进行神经网络的重训练时，反向传播过程采用直通估计器计算梯度。

7、一种神经网络的4比特量化系统，其特征在于，所述系统包括：

加载模块，用于加载神经网络的预训练模型；

统计模块，用于在所述预训练模型中，统计各饱和激活层satRelu的初始值；

重训练模块，用于在所述神经网络中添加伪量化节点，并利用satRelu的初始值进行神经网络的重训练，获取伪量化模型；

判断模块，用于判断所述伪量化模型的精度是否收敛到设定的精度；

转换模块，用于当所述伪量化模型的精度收敛到设定的精度时，对所述伪量化模型进行推理前处理，转换为可用于推理运算的4比特的推理模型，所述推理前处理过程包括：常量折叠、二次量化以及激活等价变换。

8、根据权利要求7所述的一种神经网络的4比特量化系统，其特征在于，所述统计模块包括：

-16-

替换单元，用于将所述神经网络中所有的激活层 relu 替换为饱和激活层 satRelu；

激活值获取单元，用于根据所获取的命令，获取各饱和激活层 satRelu 的激活值；

5 统计单元，用于根据所述激活值，利用直方图统计分布数据；

初始值选取单元，用于选取所述直方图中位于 99.999% 点位的激活值作为饱和激活层 satRelu 中参数 max 的初始值，其中，satRelu 定义为：

$$y = \text{satRelu}(x) = \begin{cases} 0, & x \in (-\infty, 0) \\ x, & x \in [0, \max) \\ \max, & x \in [\max, +\infty) \end{cases},$$

且反向传播中，satRelu 对于参数 max 的梯度为： $\frac{\partial y}{\partial \max} = \begin{cases} 0, & x \in (-\infty, \max) \\ 1, & x \in [\max, +\infty) \end{cases}$ ，

10 satRelu 对于输入 x 的梯度为 $\frac{\partial y}{\partial x} = \begin{cases} 0, & x \in (-\infty, 0) \cup [\max, +\infty) \\ 1, & x \in [0, \max) \end{cases}$ ，max 为饱和激活层 satRelu 的最大值。

9、根据权利要求 7 所述的一种神经网络的 4 比特量化系统，其特征在于，所述重训练模块包括：

15 伪量化层插入单元，用于在神经网络的权重层之前插入权重伪量化层，且在激活层之前插入激活伪量化层；

第一重训练单元，用于利用公式 $y = \text{quant}(w) = \text{clip}(\text{round}(w * \text{scale})) / \text{scale}$ ，对权重伪量化层进行神经网络的重训练，其中，w 为权重值，n 取值为 4，

比例系数 scale 为： $\text{scale} = \frac{2^{n-1} - 1}{\max(w) - \min(w)}$ ；

第二重训练单元，用于利用公式 $y = \text{quant}(x) = \text{clip}(\text{round}(x * \text{scale})) / \text{scale}$ ，

20 对激活伪量化层进行神经网络的重训练，其中，x 为每层的激活值，max 为 satRelu 的最大值，n 取值为 4，比例系数 scale 为： $\text{scale} = \frac{2^n - 1}{\max}$ 。

10、根据权利要求 7 所述的一种神经网络的 4 比特量化系统，其特征在于，所述推理前处理模块包括：

— 17 —

常量折叠单元，用于将 batchNorm 层融合进卷积中，其中，卷积的计算公式为： $z = w * x + b$ ，batchNorm 的计算公式为： $y = \frac{z - \text{mean}}{\sqrt{\text{var}}} * \beta + \gamma$ ，

合并后的新卷积的计算公式为： $w' = \frac{w}{\sqrt{\text{var}}} * \beta$ ， $b' = (b - \text{mean})\sqrt{\text{var}} * \beta + \gamma$ ；

二次量化单元，用于对权重进行二次量化，获取二次量化后权重的量

5 化比例系数 scale；

激活等价变换单元，用于利用公式 $y = \frac{W_q * x_q + W_q * \text{scale}_a * 8 + b_q}{\text{scale}_w * \text{scale}_a}$ ，对

激活做等价变换。

— 1/2 —

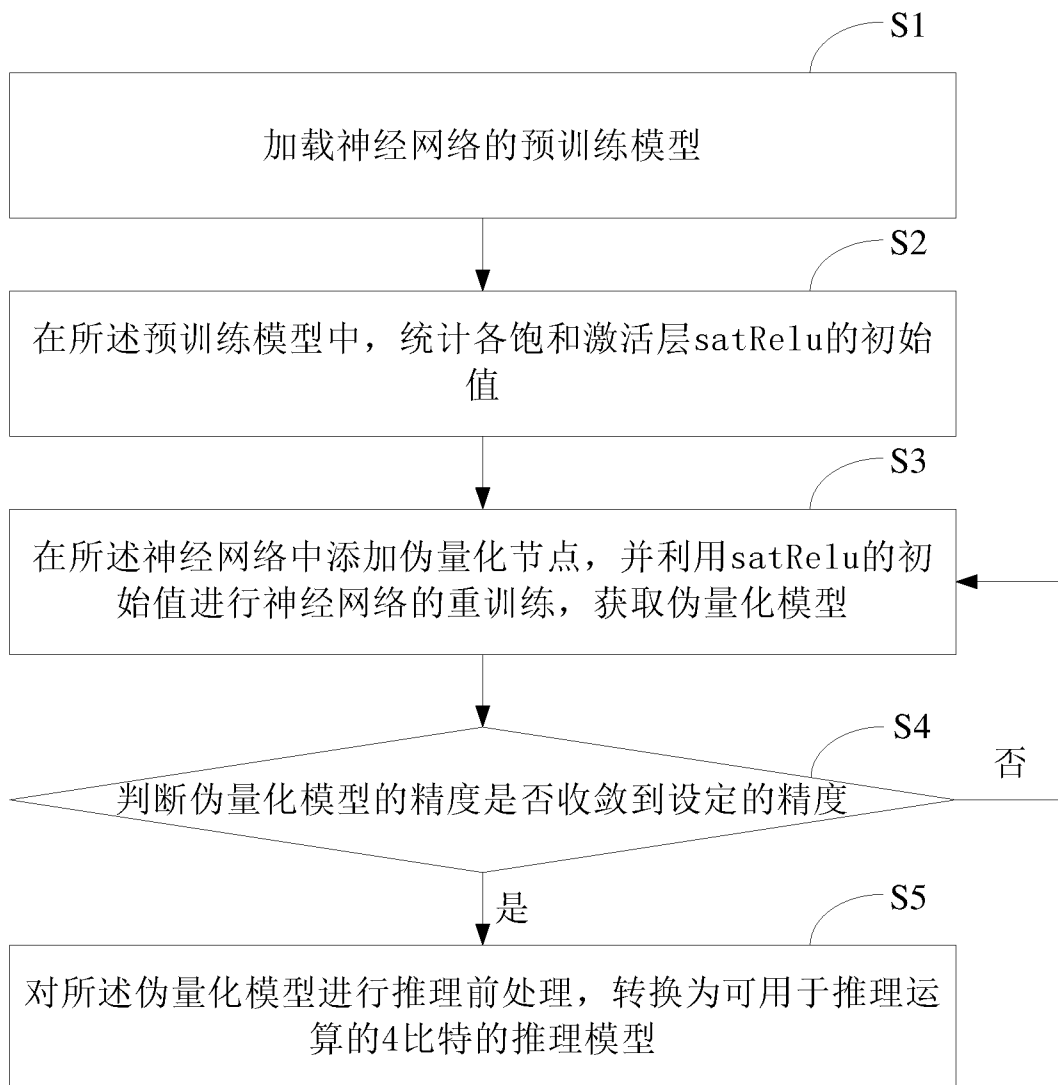


图 1

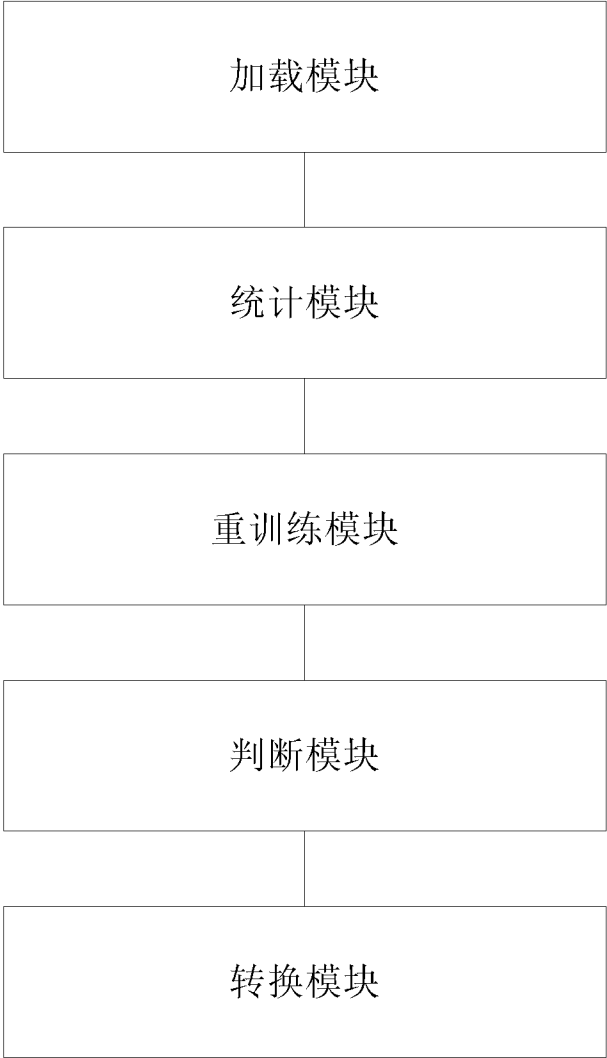


图 2

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/076982

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08(2006.01)i; G06N 3/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N; G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI; EPODOC; CNPAT; CNKI; IEEE: 神经网络, 伪, 量化, 比特, 训练, 推理, 层, 精度, 收敛, neural network, NN, pseudo, fake, quantization, bit?, train+, reason+, layer?, accuracy, convergence

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 111882058 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 03 November 2020 (2020-11-03) claims 1-10	1-10
X	CN 110334802 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 15 October 2019 (2019-10-15) description, paragraphs [0047]-[0085], and figures 1-11	1-4, 6-8, 10
A	CN 110837890 A (XI'AN JIAOTONG UNIVERSITY) 25 February 2020 (2020-02-25) entire document	1-10
A	US 20200097823 A1 (SAMSUNG ELECTRONICS CO., LTD.) 26 March 2020 (2020-03-26) entire document	1-10



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

08 May 2021

Date of mailing of the international search report

20 May 2021

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/076982

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	111882058	A	03 November 2020	None			
CN	110334802	A	15 October 2019	None			
CN	110837890	A	25 February 2020	None			
US	20200097823	A1	26 March 2020	CN	110942136	A	31 March 2020
				KR	2020035198	A	02 April 2020

国际检索报告

国际申请号

PCT/CN2021/076982

A. 主题的分类

G06N 3/08(2006.01)i; G06N 3/04(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06N; G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

WPI;EPDOC;CNPAT;CNKI;IEEE: 神经网络, 伪, 量化, 比特, 训练, 推理, 层, 精度, 收敛, neural network, NN, pseudo, fake, quantization, bit?, train+, reason+, layer?, accuracy, convergence

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 111882058 A (苏州浪潮智能科技有限公司) 2020年 11月 3日 (2020 - 11 - 03) 权利要求1-10	1-10
X	CN 110334802 A (腾讯科技深圳有限公司) 2019年 10月 15日 (2019 - 10 - 15) 说明书第[0047]-[0085]段、附图1-11	1-4, 6-8, 10
A	CN 110837890 A (西安交通大学) 2020年 2月 25日 (2020 - 02 - 25) 全文	1-10
A	US 20200097823 A1 (SAMSUNG ELECTRONICS CO., LTD.) 2020年 3月 26日 (2020 - 03 - 26) 全文	1-10

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2021年 5月 8日

国际检索报告邮寄日期

2021年 5月 20日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

于白

电话号码 86-(10)-53961381

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/076982

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	111882058	A	2020年 11月 3日	无	
CN	110334802	A	2019年 10月 15日	无	
CN	110837890	A	2020年 2月 25日	无	
US	20200097823	A1	2020年 3月 26日	CN 110942136 A	2020年 3月 31日
				KR 2020035198 A	2020年 4月 2日