

(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
2017 年 10 月 5 日 (05.10.2017)



(10) 国际公布号
WO 2017/167114 A1

- (51) 国际专利分类号:
G06N 3/08 (2006.01)
- (21) 国际申请号: PCT/CN2017/077897
- (22) 国际申请日: 2017 年 3 月 23 日 (23.03.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610201731.1 2016 年 3 月 31 日 (31.03.2016) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。
- (72) 发明人: 及
- (71) 申请人 (仅对美国): 王思宇 (WANG, Siyu) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

- (74) 代理人: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区小关北里甲 2 号渔阳置业大厦 B 座 605, Beijing 100029 (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, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO,

[见续页]

(54) Title: METHOD AND DEVICE FOR TRAINING MODEL OF QUASI-ALEXNET

(54) 发明名称: 一种类 Alexnet 网络的模型训练方法和装置



图 3

- 301 Utilize a first graphical processing unit (GPU) to compute a first gradient and a second gradient in a quasi-Alexnet
- 302 Receive a third gradient in the quasi-Alexnet and transmitted by a second GPU
- 303 Compute, according to the first gradient and the third gradient, a first model parameter of the quasi-Alexnet
- 304 Receive a fourth gradient in the quasi-Alexnet and transmitted by the second GPU
- 305 Compute, according to the second gradient and the fourth gradient, a second model parameter of the quasi-Alexnet
- 306 Utilize the first model parameter and the second model parameter to train a model of the quasi-Alexnet to perform computation and communication processes separately

(57) Abstract: The embodiments of the invention provide a method and device for training a model of a quasi-Alexnet. The method comprises: utilizing a first graphical processing unit (GPU) to compute a first gradient and a second gradient in a quasi-Alexnet; receiving a third gradient in the quasi-Alexnet and transmitted by a second GPU; computing, according to the first gradient and the third gradient, a first model parameter of the quasi-Alexnet; receiving a fourth gradient in the quasi-Alexnet and transmitted by the second GPU; computing, according to the second gradient and the fourth gradient, a second model parameter of the quasi-Alexnet; utilizing the first model parameter and the second model parameter to train a model of the quasi-Alexnet to perform computation and communication processes separately. The embodiment can perform in parallel computation of a convolutional layer in the quasi-Alexnet and communication using fully connected layer parameters, effectively reducing the time consumed for training a model, and increasing operational efficiency of training the model.

(57) 摘要: 本申请实施例提供了一种类 Alexnet 网络的模型训练方法和装置, 所述方法包括: 采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值; 接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值; 依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数; 接收所述第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第四梯度值; 依据所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模型参数; 采用所述第一模型参数和第二模型参数训练所述类 Alexnet 网络的模型, 使计算和通信两个过程分开进行, 进一步使类 Alexnet 网络的卷积层的计算和全连接参数通信并行, 有效地减少了模型训练过程耗费的时间, 提高了模型训练的运行效率。

WO 2017/167114 A1



RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, **本国际公布:**

CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, — 包括国际检索报告(条约第 21 条(3))。
TG)。

一种类 Alexnet 网络的模型训练方法和装置

技术领域

本申请涉及信息技术领域，特别是涉及一种类 Alexnet 网络的模型训练方法和一种类 Alexnet 网络的模型训练装置。

5 背景技术

人工智能 (Artificial Intelligence) 是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器，该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。人工智能从诞生以来，理论和技术日益成熟，应用领域也不断扩大。近年来，深度学习 (Deep Learning) 直接尝试解决抽象认知的难题，并取得了突破性的进展。深度学习引爆的这场革命，将人工智能带上了一个新的台阶，不仅学术意义巨大，而且实用性很强。

深度学习的动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本。通常，深度学习是通过建立相应的网络模型，采用模型训练的方式来进行的。根据不同的学习框架建立的学习模型很是不同，例如，卷积神经网络 (Convolutional neural networks, 简称 CNNs) 就是一种深度的监督学习下的机器学习模型，其中，Alexnet 网络又是开发者经常使用的一种经典的卷积神经网络。

如图 1 所示，是一种 Alexnet 网络的结构示例图。在 Alexnet 网络中，比较重要的两种层类型为卷积层 Convolution (即图 1 中 Convolution1 至 pool5 部分) 和全连接层 Inner Product (即图 1 中 Inner Product6 至 loss 层部分)。在 Alexnet 网络中进行一次模型训练的过程可以描述如下：

(1) 先将数据从 Data 层正向传播到 Top 层，此过程途中先经过卷积层部分，再经过全连接层部分；

(2) 在传播到 Top 层后计算损失；

(3) 将损失从 Top 层依次反向传播到 Data 层，并在传播过程中计算梯度值，最后完成连接权重的更新，这一过程途中先经过全连接层部分，再经过卷积层部分。

在 Alexnet 网络中，无论是正向传播过程还是反向传播过程，卷积层部分都会拥有非常大的计算量，几乎占了整个网络的计算时间 80% 以上，但卷积层需要更新的参数量却非常小，只占整个网络参数的 10%；而全连

接层部分的情况则与卷积层完全相反，全连接层部分拥有整个网络 90% 的待更新参数，但计算时间却只占了整个网络的 20%。

在单机多卡（即一台装有多个图形处理单元 GPU 的计算机）环境下，在进行模型训练时，为了能够得到无损的训练结果，必须在每个 GPU 上都保持一份全量的模型，并在两个模型上同时进行训练。以两卡（两个图形处理单元 GPU）为例，可以将两张卡分为主卡和从卡，如图 2 所示，是已有技术中主卡与从卡的工作原理图。在每一轮训练结束后，需要将主卡上的模型计算出来的梯度值发送到主卡模型上，并由主卡在计算梯度值的平均值后更新参数，最后将主卡上最新的模型广播发送到从卡上，才能继续进行下一次的训练。已有技术中一般是先对所有层计算出全部的梯度值之后，将得到的所有层的梯度值发送到主卡上求和平均并更新模型，即必须先进行全部的计算之后才能进行通信，在时间上，计算和通信具有严格的先后顺序。

因此，按照已有技术首先计算出全连接层的梯度值，并在将全连接层的梯度值汇总到主卡上之后，再计算卷积层的梯度值，则整个过程所耗费的时间将会非常多，严重影响模型训练的运行效率。

发明内容

鉴于上述问题，提出了本申请实施例以便提供一种克服上述问题或者至少部分地解决上述问题的一种类 Alexnet 网络的模型训练方法和相应的一种类 Alexnet 网络的模型训练装置。

为了解决上述问题，本申请公开了一种类 Alexnet 网络的模型训练方法，包括：

采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值；

接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值；

依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数；

接收所述第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第四梯度值；

依据所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模

型参数;

采用所述第一模型参数和第二模型参数训练所述类 Alexnet 网络的模型。

可选地, 所述类 Alexnet 网络由全连接层和卷积层组成, 所述采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值的步骤包括:

采用第一图形处理单元 GPU 计算在所述全连接层下的第一梯度值和
在所述卷积层下的第二梯度值。

可选地, 所述第一图形处理单元 GPU 包括第一计算队列, 所述采用第一图形处理单元 GPU 计算在所述全连接层下的第一梯度值和
在所述卷积层下的第二梯度值的步骤包括:

采用第一计算队列计算在所述全连接层下的第一梯度值和
在所述卷积层下的第二梯度值。

可选地, 所述第一图形处理单元 GPU 还包括第一通信队列, 所述第二图形处理单元 GPU 包括第二通信队列, 所述接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值的步骤包括:

采用第一通信队列接收第二通信队列发送的第三梯度值;

所述接收所述第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第四梯度值的步骤包括:

采用第一通信队列接收第二通信队列发送的第四梯度值。

可选地, 所述第二图形处理单元还包括第二计算队列, 所述第三梯度值和所述第四梯度值分别通过如下步骤获得:

采用第二计算队列计算在所述全连接层下的第三梯度值; 以及,

采用第二计算队列计算在所述卷积层下的第四梯度值。

可选地, 所述依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数的步骤包括:

计算所述第一梯度值和第三梯度值的平均值, 获得所述类 Alexnet 网络的第一模型参数。

可选地, 所述采用所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模型参数的步骤包括:

计算所述第二梯度值和第四梯度值的平均值, 获得所述类 Alexnet 网络的第二模型参数。

可选地, 在采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值的步骤前, 还包括:

判断网络是否为类 Alexnet 网络。

可选地，所述网络包括 m 个结构层，所述判断网络是否为类 Alexnet 网络的步骤包括：

对所述网络进行预训练，获得每个结构层的计算时间和参数量；

5 根据所述计算时间和参数量，获得所述网络的计算总时间和总参数量；

按照预设传输顺序，逐层累加所述 m 个结构层的计算时间，分别获得截至第 n 层的计算时间之和；

10 当所述截至第 p 层的计算时间之和与所述计算总时间的比值满足第一预设条件时，累加剩余 $m-p$ 层的参数量，获得所述剩余 $m-p$ 层的参数量之和；

判断所述剩余 $m-p$ 层的参数量之和与所述总参数量的比值是否满足第二预设条件；

若是，则将所述网络划分为类 Alexnet 网络。

15 可选地，所述将所述网络划分为类 Alexnet 网络的步骤包括：

将所述网络的前 p 层划分为类 Alexnet 网络的全连接层；

将所述剩余 $m-p$ 层划分为类 Alexnet 网络的卷积层。

20 为了解决上述问题，本申请还公开了一种类 Alexnet 网络的模型训练装置，包括：

第一计算模块，用于采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值；

第一接收模块，用于接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值；

25 第二计算模块，用于依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数；

第二接收模块，用于接收所述第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第四梯度值；

30 第三计算模块，用于依据所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模型参数；

训练模块，用于采用所述第一模型参数和第二模型参数训练所述类 Alexnet 网络的模型。

可选地，所述类 Alexnet 网络由全连接层和卷积层组成，所述第一计算模块包括：

第一计算子模块，用于采用第一图形处理单元 GPU 计算在所述全连接层下的第一梯度值和在所述卷积层下的第二梯度值。

可选地，所述第一图形处理单元 GPU 包括第一计算队列，所述第一计算子模块包括：

- 5 第一计算单元，用于采用第一计算队列计算在所述全连接层下的第一梯度值和在所述卷积层下的第二梯度值。

可选地，所述第一图形处理单元 GPU 还包括第一通信队列，所述第二图形处理单元 GPU 包括第二通信队列，所述第一接收模块包括：

- 10 第一接收子模块，用于采用第一通信队列接收第二通信队列发送的第三梯度值；

所述第二接收模块包括：

第二接收子模块，用于采用第一通信队列接收第二通信队列发送的第四梯度值。

- 15 可选地，所述第二图形处理单元还包括第二计算队列，所述第三梯度值和所述第四梯度值分别通过如下模块获得：

第四计算模块，用于采用第二计算队列计算在所述全连接层下的第三梯度值；以及，

第五计算模块，用于采用第二计算队列计算在所述卷积层下的第四梯度值。

- 20 可选地，所述第二计算模块包括：

第一模型参数计算子模块，用于计算所述第一梯度值和第三梯度值的平均值，获得所述类 Alexnet 网络的第一模型参数。

可选地，所述第三计算模块包括：

- 25 第二模型参数计算子模块，用于计算所述第二梯度值和第四梯度值的平均值，获得第二模型参数。

可选地，所述装置还包括：

判断模块，用于判断网络是否为类 Alexnet 网络。

可选地，所述网络包括 m 个结构层，所述判断模块包括：

- 30 计算时间和参数量获得子模块，用于对所述网络进行预训练，获得每个结构层的计算时间和参数量；

计算总时间和总参数量获得子模块，用于根据所述计算时间和参数量，获得所述网络的计算总时间和总参数量；

计算时间之和获得子模块，用于按照预设传输顺序，逐层累加所述 m 个结构层的计算时间，分别获得截至第 p 层的计算时间之和；

参数量之和获得子模块，用于在所述截至第 p 层的计算时间之和与所述计算总时间的比值满足第一预设条件时，累加剩余 $m-p$ 层的参数量，获得所述剩余 $m-p$ 层的参数量之和；

判断子模块，用于判断所述剩余 $m-p$ 层的参数量之和与所述总参数量的比值是否满足第二预设条件；

划分子模块，用于在满足第二预设条件时，将所述网络划分为类 Alexnet 网络。

可选地，所述划分子模块包括：

全连接层划分单元，用于将所述网络的前 p 层划分为类 Alexnet 网络的全连接层；

卷积层划分单元，用于将所述剩余 $m-p$ 层划分为类 Alexnet 网络的卷积层。

与背景技术相比，本申请实施例包括以下优点：

本申请实施例通过分别在第一图形单元 GPU（主卡）和第二图形单元 GPU（从卡）上构建出相应的计算队列和通信队列，采用计算队列执行计算过程，采用通信队列进行数据通信，使计算和通信两个过程分开进行，并进一步使类 Alexnet 网络的卷积层的计算和全连接参数通信并行，有效地减少了模型训练过程耗费的时间，提高了模型训练的运行效率。

其次，在本申请实施例中，在采用某一网络进行模型训练前，还可以对所述网络进行预训练，通过对获得时间参数进行分析，以判断所述网络是否属于类 Alexnet 网络。

附图说明

图 1 是一种 Alexnet 网络的结构示例图；

图 2 是已有技术中主卡与从卡的工作原理图；

图 3 是本申请的一种类 Alexnet 网络的模型训练方法实施例一的步骤流程图；

图 4 是本申请的一种类 Alexnet 网络的模型训练方法实施例一的工作原理图；

图 5 是本申请的一种类 Alexnet 网络的模型训练方法实施例二的步骤流程图；

图 6 是本申请的一种类 Alexnet 网络的模型训练方法实施例二的数据反向传播计算时间和参数量统计图；

图 7 是本申请的判断网络是否为类 Alexnet 网络的算法流程图；

图 8 是本申请的一种类 Alexnet 网络的模型训练装置实施例的结构框图。

具体实施方式

为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本申请作进一步详细的说明。

参照图 3，示出了本申请的一种类 Alexnet 网络的模型训练方法实施例一的步骤流程图，具体可以包括如下步骤：

步骤 301，采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值；

在 Alexnet 网络中，无论是正向传播过程还是反向传播过程，卷积层部分都会拥有非常大的计算量，几乎占了整个网络的计算时间 80%以上，但卷积层需要更新的参数量却非常小，只占整个网络参数的 10%；而全连接层部分的情况则与卷积层完全相反，全连接层部分拥有整个网络 90%的待更新参数，但计算时间却只占了整个网络的 20%。在本申请实施例中，可以将具有上述特点，并且在数据的正向传播过程中先经过卷积层部分，然后才经过全连接层部分的网络，称为类 Alexnet 网络。所述类 Alexnet 网络可以由全连接层和卷积层组成。

图形处理单元 GPU (Graphics Processing Unit) 又称显示核心、视觉处理器、显示芯片等，是一种专门在个人电脑、工作站、游戏机和一些移动设备（如平板电脑、智能手机等）上进行图像运算工作的微处理器，常用于高性能计算，具有高并发处理数据的特性。在本申请实施例中，第一图形处理单元 GPU 可以看做是单机多卡环境下中主卡，第二图形处理单元 GPU 可以看做是单机多卡环境下中从卡。

在初始化时，主卡与从卡两张卡必须持有相同的网络结构，因此在 Start 之后，主卡需要将该卡的网络结构广播到从卡上，而从卡通过 Receive Model 过程接收网络结构，使得两张卡保持一致。然后两张卡开始执行相同的行为，目的是进行前向传播，计算 Loss 值。前向传播顾名思义就是从第一层向最后一层计算的过程。

具体地，前向传播的过程是按照图 1 中 Alexnet 网络中每层和每层之间的箭头所指方向进行的。例如从 Data 传播到 Convolution1，再到 relu1...一直到最后的 loss 层。这时 loss 层会得出一个 Loss 值，该值被称为损失值，而后一过程反向传播能够进行的先决条件是需要得出 Loss 值。对于前向传播过程来说，先经过卷积层，后经过全连接层。

然后，进行后向传播，先经过全连接层，后经过卷积层，并相应地计算各层的梯度值。

梯度是一个数学概念，在处理分类问题或回归问题时，在模型训练的过程中，可以用损失值函数 Loss 来作为评估分类是否精准或者回归是否准确。一般情况下，训练得比较好的模型的损失值 Loss 都比较低，而所述 Loss 值又与神经网络的参数有关，如果所述参数符合应用场景的要求，那么 Loss 值就会比较低。如果将网络的所有模型参数组成 w 向量，可以得到 Loss 值是 w 向量有关的，通常，好的 w 向量能够使 Loss 值降低。因此，问题可以归结为如何寻找到好的 w 向量？这需要进行训练，让模型自己去寻找。模型必须找到能够使得 Loss 值下降的正确的方向，而梯度这个数学量就是代表了 Loss 值下降的最快的方向。只要每次让 w 向量按照梯度这个方向更新一步，那么 Loss 值就会减少一些。这就是梯度的作用。

具体地，梯度的计算是根据 Loss 值关于各个 w 向量的偏微分求出来的，而求偏微分的过程就是在数据的反向传播的过程中进行的。

在本申请实施例中，所述第一梯度值即为全连接层梯度，所述第二梯度值即为卷积层梯度。

因此，所述采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值的步骤具体可以包括如下子步骤：

子步骤 3011，采用第一图形处理单元 GPU 计算在所述全连接层下的第一梯度值和所述卷积层下的第二梯度值。

通常，在 GPU 中可以有多组不同的操作队列，即 CUDA 流，并且该队列中的操作可以按照添加到队列的先后顺序执行，不同流间的操作可以并行执行。CUDA 是一种由 NVIDIA 推出的通用并行计算架构，该架构使 GPU 能够解决复杂的计算问题，并使得在计算机上实现 GPU 编程成为可能。

在本申请实施例中，所述第一图形处理单元 GPU 即主卡上可以包括有第一计算队列和第一通信队列，所述第二图形处理单元 GPU 即从卡上可以包括有第二计算队列和第二通信队列，所述第一计算队列、第一通信队列、第二计算队列和第二通信队列均是 CUDA 流，其中，第一计算队列和第二计算队列可以用于计算，而第一通信队列和第二通信队列可以用于通信，以使主卡和从卡的计算和通信分开，实现并行处理。

在本申请的一种优选实施例中，所述采用第一图形处理单元 GPU 计算在所述全连接层下的第一梯度值和所述卷积层下的第二梯度值的子步骤可以进一步包括：

采用第一计算队列计算在所述全连接层下的第一梯度值和在所述卷积层下的第二梯度值。

在具体实现中，在计算全连接层下的第一梯度值时，该过程只是反向传播的前半部分。反向传播的完整过程是从 loss 层开始（最后一层），按照箭头相反的方向逐层传播。例如从 loss 层传播到 inner produce8，再到 drop7...，一直到 convolution1。而计算第一梯度值的过程只是包含图中从 loss 层传播到 inner product6 的过程。在全连接层相关层的反向传播的过程中，每反向传播一层就会对有参数的层计算出该层的梯度值（有些层不会计算出梯度，因为该层并没有参数，比如 loss 层，drop 层，relu 层，只有 inner product 层才会计算出梯度）并存储在该层中。这一过程计算出来的梯度参数会非常非常多，但是整个计算过程却非常迅速，这是这一过程的特点。

在计算卷积层的第二梯度值时，该过程只是反向传播的后半部分，即从 pool5 传播到 convolution1 的过程。在卷积层相关层的反向传播的过程中，每反向传播一层就会对有参数的层会计算出该层的梯度（有些层不会计算出梯度，因为该层并没有参数，比如 relu 层，norm 层，pool 层，只有 convolution 层才会计算出梯度）并存储在该层中。这一过程计算出来的梯度参数会非常非常少，但是整个计算过程会非常慢，这是这一过程的特点。

步骤 302，接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值；

在本申请实施例中，所述第三梯度值具体可以通过如下步骤获得：

采用第二计算队列计算在所述全连接层下的第三梯度值。

在具体实现中，所述第二图形处理单元 GPU 即从卡上的第二计算队列计算所述全连接层下的第三梯度值的过程，与主卡上的第一计算队列计算第一梯度值同时并行进行。

在本申请实施例中，所述接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值的步骤具体可以包括如下子步骤：

子步骤 3021，采用第一通信队列接收第二通信队列发送的第三梯度值。

在本申请实施例中，为了将主卡与从卡的计算与通信过程区分开，可以采用计算队列执行相应的计算过程，采用通信队列执行相应的数据发送与接收，因此，可以采用第一通信队列来接收第二通信队列发送的第三梯度值。

在具体实现中，可以利用流并行的方式，将卷积层的计算和全连接层

的参数通信并行执行，即在主卡采用第一计算队列计算所述第二梯度值时，采用第一通信队列接收第二通信队列发送的第三梯度值，使计算和通信的过程获得时间上的重叠，二者互不干扰。

步骤 303，依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数；

所述第一梯度值与第三梯度值分别是主卡和从卡计算的所述类 Alexnet 网络的全连接层的梯度，因此，当从卡的数据汇总到主卡上后，需要根据主卡和从卡的数据对全连接层的数据进行更新。

在本申请的一种优选实施例中，所述依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数的步骤具体可以包括如下子步骤：

子步骤 3031，计算所述第一梯度值和第三梯度值的平均值，获得所述类 Alexnet 网络的第一模型参数。

所述第一模型参数即为更新后的全连接层梯度。

步骤 304，接收所述第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第四梯度值；

在本申请实施例中，所述第四梯度值具体可以通过如下步骤获得：

采用第二计算队列计算在所述卷积层下的第四梯度值。

在具体实现中，所述第二图形处理单元 GPU 即从卡上的第二计算队列计算所述卷积层下的第四梯度值的过程，与主卡上的第一计算队列计算第二梯度值同时并行进行。

在本申请实施例中，所述接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第四梯度值的步骤具体可以包括如下子步骤：

子步骤 3041，采用第一通信队列接收第二通信队列发送的第四梯度值。

在本申请实施例中，为了将主卡与从卡的计算与通信过程区分开，可以采用计算队列执行相应的计算过程，采用通信队列执行相应的数据发送与接收，因此，可以采用第一通信队列来接收第二通信队列发送的第四梯度值。

步骤 305，依据所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模型参数；

所述第二梯度值与第四梯度值分别是主卡和从卡计算的所述类 Alexnet 网络的卷积层的梯度，因此，当从卡的数据汇总到主卡上后，需要根据主卡和从卡的数据对卷积层的数据进行更新。

在本申请的一种优选实施例中，所述依据所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模型参数的步骤具体可以包括如下子步骤：

子步骤 3051，计算所述第二梯度值和第四梯度值的平均值，获得所述类 Alexnet 网络的第二模型参数。

所述第二模型参数即为更新后的卷积层梯度。

步骤 306，采用所述第一模型参数和第二模型参数训练所述类 Alexnet 网络的模型。

在本申请实施例中，当分别获得所述第一模型参数和第二模型参数后，主卡可以针对所述第一模型参数和第二模型参数对自身的模型参数进行更新，以获得新的训练模型。

在已有技术中，在 Alexnet 网络下进行模型训练时，每一轮的计算均为依次进行发送/接收模型、Forward 前向传播、Backward 反向传播，接收/发送梯度值、参数更新。假设 Forward 前向传播的过程时间为 a ，Backward 反向传播的过程中涉及到全连接层的计算时间为 b ，涉及到卷积层的计算时间为 c ，发送/接收全连接层梯度值的时间为 m ，发送/接收卷积层梯度值的时间为 n ，那么按照已有技术完成整个过程的总时间 $T1$ 为：

$$T1 = a + b + c + m + n, \text{ 其中 } c \gg b, m \gg n$$

而采用本申请实施例的方法，利用流并行方式将反向传播过程中，卷积层的计算和全连接参数通信并行起来后，总时间 $T2$ 为：

$$T2 = a + b + \max(c, m) + n$$

$$\text{由于 } T1 - T2 = c + m - \max(c, m) > 0, \text{ 所以: } T1 > T2$$

由上式可知，利用流并行方式将通信和计算并行来优化类 Alexnet 网络的方案可以有效减少整个过程耗费的时间。

在本申请实施例中，通过分别在第一图形单元 GPU（主卡）和第二图形单元 GPU（从卡）上构建出相应的计算队列和通信队列，采用计算队列执行计算过程，采用通信队列进行数据通信，使计算和通信两个过程分开进行，并进一步使类 Alexnet 网络的卷积层的计算和全连接参数通信并行，有效地减少了模型训练过程耗费的时间，提高了模型训练的运行效率。

参照图 5，示出了本申请的一种类 Alexnet 网络的模型训练方法实施例二的步骤流程图，具体可以包括如下步骤：

步骤 501，判断网络是否为类 Alexnet 网络；

通常，如果将网络一般化并分为两部分，分别称为 M 和 N ，且在反向

传播时先进行 M 部分计算，后进行 N 部分计算，那么如果 M 部分的计算量占据整个时间的很小一部分，而所有用的参数占据所有参数量的很大一部分，并且 N 部分具有和 M 相反的特性，那么可以将称此种网络看作为类 Alexnet 网络。

5 在本申请实施例中，对于包括有 m 个结构层的某一网络，可以首先判断该网络是否为类 Alexnet 网络。

在本申请的一种优选实施例中，所述判断网络是否为类 Alexnet 网络的步骤具体可以包括如下子步骤：

10 子步骤 5011，对所述网络进行预训练，获得每个结构层的计算时间和参数量；

子步骤 5012，根据所述计算时间和参数量，获得所述网络的计算总时间和总参数量；

子步骤 5013，按照预设传输顺序，逐层累加所述 m 个结构层的计算时间，分别获得截至第 p 层的计算时间之和；

15 子步骤 5014，当所述截至第 p 层的计算时间之和与所述计算总时间的比值满足第一预设条件时，累加剩余 m-p 层的参数量，获得所述剩余 m-p 层的参数量之和；

子步骤 5015，判断所述剩余 m-p 层的参数量之和与所述总参数量的比值是否满足第二预设条件；

20 在本申请实施例中，可以利用计算机去判断当前网络是否属于类 Alexnet 网络，即可以通过预训练获取到的时间参数进行分析。

首先，对所述网络进行预训练，可以获得每个结构层的计算时间和参数量；然后根据所述计算时间和参数量，能够获得所述网络的计算总时间和总参数量；按照预设传输顺序（一般地，所述预设传输顺序可以是反向传输方向，即从所述网络的最后一层传输至第一层的过程），逐层累加所述 m 个结构层的计算时间，分别获得截至第 p 层的计算时间之和；当所述截至第 p 层的计算时间之和与所述计算总时间的比值满足第一预设条件时，累加剩余 m-p 层的参数量，获得所述剩余 m-p 层的参数量之和；最后判断所述剩余 m-n 层的参数量之和与所述总参数量的比值是否满足第二预设条件。

通常，由于 Alexnet 网络的特点在于计算量大的部分参数量小，而计算量小的部分参数量却很大，因此，本领域技术人员可以据此设置第一预设条件和第二预设条件的具体数值，本申请对所述数值不作具体限定。

具体地，可以把预训练的网络分为两个部分，即 M 部分和 N 部分，

那么问题就可以划归为如何选取分割 M 和 N 的分界点。进一步地, 所述分界点的选取过程可以按照如下方式进行: 将该网络在训练前进行若干次预训练过程, 并计算每次运行时每个层进行反向传播时的计算时间和拥有的参数量。然后, 将所有层的反向传播时间累加和记为 T, 所有层所拥有的参数量记为 V, 以反向传播的起始层为起点, 不断累加下一层的计算时间, 并记为 t。当 $t/T > 0.1$ 时停止累加, 并将当前层记为第 p 层。将从第 p 层到反向传播的最后一层所拥有的参数量总和记为 v, 若此时 $v/V < 0.1$, 那么可以认为此网络类型为类 Alexnet 网络, 可以继续执行子步骤 5016, 若 $v/V > 0.1$, 则可以认为此网络并不是类 Alexnet 网络。

子步骤 5016, 将所述网络划分为类 Alexnet 网络。

在本申请的一种优选实施例中, 所述将所述网络划分为类 Alexnet 网络的子步骤可以进一步包括:

将所述网络的前 p 层划分为类 Alexnet 网络的全连接层;

将所述剩余 m-p 层划分为类 Alexnet 网络的卷积层。

当所述网络被确认为属于类 Alexnet 网络时, 可以根据获得的分解点, 即子步骤 5011-5015 中获得的第 p 层, 将所述网络具体划分为全连接层部分和卷积层部分。

参照图 6, 是本申请的一种的数据反向传播计算时间和参数量统计图, 图中横坐标为层数, 左部分为反向传播每一层的计算时间统计, 纵坐标单位为 ms, 右部分为每层所拥有的的参数量, 单位为字节。从反向传播的起始层开始到上文计算出来的第 p 层之间的所有层计算时间之和为整个反向传播总时间的 10%, 而此时如果从第 p 层到反向传播的最后一层之间的所有层所拥有的参数量之和占据总参数量的 10% 左右, 从而可以说明从反向传播的第一层到第 p 层具有 90% 的参数量。进而可以认定, 从反向传播的最后一层到第 p 层的部分为 M, 即全连接层, 而从第 p 层到反向传播的第一层的部分为 N, 即卷积层部分。上述判断过程可以通过如图 7 所示的算法流程图表示。

步骤 502, 采用第一计算队列计算在所述全连接层下的第一梯度值和
在所述卷积层下的第二梯度值;

在本申请实施例中, 可以分别在第一图形处理单元 GPU 即主卡, 和第二图形处理单元 GPU 即从卡上分别构建出计算队列和通信队列。具体地, 可以在主卡上构建第一计算队列和第一通信队列, 在从卡上构建第二计算队列和第二通信队列, 由第一计算队列和第二计算队列执行相应的计算过程, 由第一通信队列和第二通信队列执行相应的通信过程, 以获得时

间上的叠加。

因此，在本申请实施例中，可以采用第一计算队列计算在所述全连接层下的第一梯度值和在所述卷积层下的第二梯度值。

在具体实现中，在计算全连接层下的第一梯度值时，该过程只是反向传播的前半部分。反向传播的完整过程是从 loss 层开始（最后一层），按照箭头相反的方向逐层传播。例如从 loss 层传播到 inner produce8，再到 drop7...，一直到 convolution1。而计算第一梯度值的过程只是包含图中从 loss 层传播到 inner product6 的过程。在全连接层相关层的反向传播的过程中，每反向传播一层就会对有参数的层计算出该层的梯度值（有些层不会计算出梯度，因为该层并没有参数，比如 loss 层，drop 层，relu 层，只有 inner product 层才会计算出梯度）并存储在该层中。这一过程计算出来的梯度参数会非常非常多，但是整个计算过程却非常迅速，这是这一过程的特点。

在计算卷积层的第二梯度值时，该过程只是反向传播的后半部分，即从 pool5 传播到 convolution1 的过程。在卷积层相关层的反向传播的过程中，每反向传播一层就会对有参数的层会计算出该层的梯度（有些层不会计算出梯度，因为该层并没有参数，比如 relu 层，norm 层，pool 层，只有 convolution 层才会计算出梯度）并存储在该层中。这一过程计算出来的梯度参数会非常非常少，但是整个计算过程会非常慢，这是这一过程的特点。

步骤 503，采用第一通信队列接收第二通信队列发送的第三梯度值；

在本申请实施例中，所述第三梯度值可以通过第二图形处理单元 GPU 即从卡上的第二计算队列计算获得的，计算第三梯度值的过程，与主卡上的第一计算队列计算第一梯度值同时并行进行。

在本申请实施例中，由于主卡与从卡的计算与通信过程分开进行，计算队列执行相应的计算过程，通信队列执行相应的数据发送与接收，因此，可以采用第一通信队列来接收第二通信队列发送的第三梯度值。

在具体实现中，可以利用流并行的方式，将卷积层的计算和全连接层的参数通信并行执行，即在主卡采用第一计算队列计算所述第二梯度值时，采用第一通信队列接收第二通信队列发送的第三梯度值，使计算和通信的过程获得时间上的重叠，二者互不干扰。

步骤 504，计算所述第一梯度值和第三梯度值的平均值，获得所述类 Alexnet 网络的第一模型参数；

在本申请实施例中，所述第一梯度值与第三梯度值分别是主卡和从卡计算的所述类 Alexnet 网络的全连接层的梯度，因此，当从卡的数据汇总

到主卡上后, 需要根据主卡和从卡的数据对全连接层的数据进行更新。在具体实现中, 可以将所述第一梯度值和第三梯度值相加求平均值, 以获得第一模型参数, 所述第一模型参数即为更新后的全连接层梯度。

步骤 505, 采用第一通信队列接收第二通信队列发送的第四梯度值;

在本申请实施例中, 所述第四梯度值可以采用第二计算队列计算获得, 然后可以采用第二通信队列将第四梯度值发送至主卡。

在具体实现中, 所述第二图形处理单元 GPU 即从卡上的第二计算队列计算所述卷积层下的第四梯度值的过程, 与主卡上的第一计算队列计算第二梯度值同时并行进行。

步骤 506, 计算所述第二梯度值和第四梯度值的平均值, 获得所述类 Alexnet 网络的第二模型参数;

在具体实现中, 可以将所述第二梯度值和第四梯度值相加求平均值, 以获得第二模型参数, 所述第二模型参数即为更新后的卷积层梯度。

步骤 507, 采用所述第一模型参数和第二模型参数训练所述类 Alexnet 网络的模型。

在本申请实施例中, 在采用某一网络进行模型训练前, 可以对所述网络进行预训练, 通过对获得时间参数进行分析, 以判断所述网络是否属于类 Alexnet 网络。

为了便于理解, 请参照图 4, 下面以一个完整的示例对本申请实施例的主卡与从卡的计算和通信过程作一说明:

1、主卡的 BroadCast Model 和从卡的 Receive Model: 主卡的 BroadCast Model 是将主卡的模型发送到从卡上, 而从卡的 Receive Model 是负责接收主卡的模型。这一过程是为了让两张卡都保留有相同的 Alexnet 网络结构。之所以要将梯度发送到主卡上再更新, 是因为在模型训练时只会更新主卡的模型, 而不管从卡。因为在第二轮 Forward 开始前, 需要将主卡的模型广播到从卡上。因此, 可以始终以主卡模型为基础。

2、主卡的 Forward 过程和从卡的 Forward 过程: 该过程两张卡的行为相同, 属于前向传播, 是按照图 1 中 Alexnet 网络中每层和每层之间的箭头所指方向进行。例如从 Data 传播到 Convolution1, 再到 relu1..., 一直到最后的 loss 层。这时 loss 层会得出一个 loss 值, 该值被称为损失值, 而后一过程 Backward (反向传播) 能够进行的先决条件是需要得出 loss 值, 因此必须先进行 Forward, 后进行 Backward。对于 Forward 过程来说, 先经过的是卷积层相关层, 后经过的是全连接层相关层。每一层的计算方式均

有些区别，这是由于各层的计算公式不同所致。

3、主卡和从卡的 Backward For Inner Product 过程：该过程两张卡的行为为相同，该过程只是 Backward 完整的前半部分。Backward 完整过程是从 loss 层开始（即最后一层），按照图 1 中箭头相反的方向逐层传播。例如从 loss 传播到 inner produce8，再到 drop7...，一直到 convolution1。而 Backward For Inner Product 过程只是包含图 1 中全连接层部分的反向传播。即从 loss 传播到 inner product6 的过程。所以，Backward For Inner Product 这一过程是全连接层相关层的反向传播过程。在全连接层相关层的反向传播的过程中，每反向传播一层就会对有参数的层会计算出该层的梯度（有些层不会计算出梯度，因为该层并没有参数，比如 loss 层，drop 层，relu 层，只有 inner product 层才会计算出梯度）并存储在该层中。这一过程计算出来的梯度参数会非常非常多，但是整个计算过程却非常迅速，这是这一过程的特点。

4、主卡和从卡的 Backward For Convolution 过程：该过程两张卡的行为为相同，该过程只是 Backward 完整的后半部分。而 Backward For Convolution 过程只是包含图 1 中卷积层部分的反向传播。即从 pool5 传播到 convolution1 的过程。所以，Backward For convolution 这一过程是卷积层相关层的反向传播过程。在卷积层相关层的反向传播的过程中，每反向传播一层就会对有参数的层会计算出该层的梯度（有些层不会计算出梯度，因为该层并没有参数，比如 relu 层，norm 层，pool 层，只有 convolution 层才会计算出梯度）并存储在该层中。这一过程计算出来的梯度参数会非常非常少，但是整个计算过程会非常慢，这是这一过程的特点。

5、主卡和从卡部分的第一个过程 Receive inner product Gradients 和 Send inner product Gradients：这两个过程是梯度的发送与接收过程。该过程在主卡上是 receive 过程。即接收从卡计算出来的梯度，在从卡上是 send 过程，即发送计算的梯度的过程。这两个过程在各自的卡上都是跟在 Backward For Innerproduct 过程之后，代表该过程必须等待 Backward For Innerproduct 之后才能进行，但该过程是处在通信队列中的过程，而计算过程处在计算队列中，所以其与 Backward For Convolution 同时并行执行。

6、主卡和从卡部分的第二个过程 Update Inner product Gradients 过程：这两个过程是更新全连接层梯度的过程。但是两张卡的行为不一样，主卡是对全连接层梯度取平均的过程，而从卡这个过程是个空过程，即不执行任何行为。但该过程是处在通信队列中的过程，与 Backward For Convolution 同时并行执行

7、主卡和从卡部分的第三个过程 Receive Convolution Gradients 和 Send Convolution Gradients 过程：即主卡接收从卡发送的卷积层梯度参数，而从卡向主卡发送卷积层梯度参数的过程。虽然该过程处在通信队列中，但是它与计算队列的 Backward For Convolution 有依赖关系，因此这一过程必须
5 要等待 Backward For Convolution 和 Update Inner product Gradients 完成之后才能执行。

8、主卡和从卡的 Update Convolution Gradients 过程：这两个过程是对卷积层梯度取平均的过程。但是两张卡的行为不一样，主卡是对卷积层梯度取平均的过程，而从卡这个过程是个空过程，即不执行任何行为。

10 需要说明的是，对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请实施例并不受所描述的动作顺序的限制，因为依据本申请实施例，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的
15 的实施例均属于优选实施例，所涉及的动作并不一定是本申请实施例所必须的。

参照图 8，示出了本申请的一种类 Alexnet 网络的模型训练装置实施例的结构框图，具体可以包括如下模块：

20 第一计算模块 801，用于采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值；

第一接收模块 802，用于接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值；

25 第二计算模块 803，用于依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数；

第二接收模块 804，用于接收所述第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第四梯度值；

第三计算模块 805，用于依据所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模型参数；

30 训练模块 806，用于采用所述第一模型参数和第二模型参数训练所述类 Alexnet 网络的模型。

在本申请实施例中，所述类 Alexnet 网络可以由全连接层和卷积层组成，所述第一计算模块 801 具体可以包括如下子模块：

第一计算子模块 8011，用于采用第一图形处理单元 GPU 计算在所述

全连接层下的第一梯度值和在所述卷积层下的第二梯度值。

在本申请实施例中，所述第一图形处理单元 GPU 可以包括有第一计算队列，所述第一计算子模块 8011 具体可以包括如下单元：

5 第一计算单元 8011A，用于采用第一计算队列计算在所述全连接层下的第一梯度值和在所述卷积层下的第二梯度值。

在本申请实施例中，所述第一图形处理单元 GPU 还可以包括有第一通信队列，所述第二图形处理单元 GPU 可以包括有第二通信队列，所述第一接收模块 802 具体可以包括如下子模块：

10 第一接收子模块 8021，用于采用第一通信队列接收第二通信队列发送的第三梯度值；

所述第二接收模块 804 具体可以包括如下子模块：

第二接收子模块 8041，用于采用第一通信队列接收第二通信队列发送的第四梯度值。

15 在本申请实施例中，所述第二图形处理单元还可以包括有第二计算队列，所述第三梯度值和所述第四梯度值可以分别通过如下模块获得：

第四计算模块 807，用于采用第二计算队列计算在所述全连接层下的第三梯度值；以及，

第五计算模块 808，用于采用第二计算队列计算在所述卷积层下的第四梯度值。

20 在本申请实施例中，所述第二计算模块 803 具体可以包括如下子模块：

第一模型参数计算子模块 8031，用于计算所述第一梯度值和第三梯度值的平均值，获得所述类 Alexnet 网络的第一模型参数。

在本申请实施例中，所述第三计算模块 805 具体可以包括如下子模块：

25 第二模型参数计算子模块 8051，用于计算所述第二梯度值和第四梯度值的平均值，获得第二模型参数。

在本申请实施例中，所述装置还可以包括如下模块：

判断模块 808，用于判断网络是否为类 Alexnet 网络。

在本申请实施例中，所述网络可以包括有 m 个结构层，所述判断模块 808 具体可以包括如下子模块：

30 计算时间和参数量获得子模块 8081，用于对所述网络进行预训练，获得每个结构层的计算时间和参数量；

计算总时间和总参数量获得子模块 8082，用于根据所述计算时间和参数量，获得所述网络的计算总时间和总参数量；

计算时间之和获得子模块 8083，用于按照预设传输顺序，逐层累加所

述 m 个结构层的计算时间，分别获得截至第 p 层的计算时间之和；

参数量之和获得子模块 8084，用于在所述截至第 p 层的计算时间之和与所述计算总时间的比值满足第一预设条件时，累加剩余 m-p 层的参数量，获得所述剩余 m-p 层的参数量之和；

5 判断子模块 8085，用于判断所述剩余 m-p 层的参数量之和与所述总参数量的比值是否满足第二预设条件；

划分子模块 8086，用于在满足第二预设条件时，将所述网络划分为类 Alexnet 网络。

在本申请实施例中，所述划分子模块 8086 具体可以包括如下单元：

10 全连接层划分单元 8086A，用于将所述网络的前 p 层划分为类 Alexnet 网络的全连接层；

卷积层划分单元 8086B，用于将所述剩余 m-p 层划分为类 Alexnet 网络的卷积层。

15 对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相
20 参见即可。

本领域内的技术人员应明白，本申请实施例的实施例可提供为方法、装置、或计算机程序产品。因此，本申请实施例可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请
25 实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

在一个典型的配置中，所述计算机设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM) 和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的
30 示例。计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机

存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带, 磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质, 可用于存储可以被计算设备访问的信息。按照本文中的界定, 计算机可读介质不包括非持续性的电脑可读媒体 (transitory media), 如调制的数据信号和载波。

本申请实施例是参照根据本申请实施例的方法、终端设备(系统)、和计算机程序产品的流程图和 / 或方框图来描述的。应理解可由计算机程序指令实现流程图和 / 或方框图中的每一流程 and / 或方框、以及流程图和 / 或方框图中的流程 and / 或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理终端设备的处理器以产生一个机器, 使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以特定方式工作的计算机可读存储器中, 使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品, 该指令装置实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上, 使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理, 从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和 / 或方框图一个方框或多个方框中指定的功能的步骤。

尽管已描述了本申请实施例的优选实施例, 但本领域内的技术人员一旦得知了基本创造性概念, 则可对这些实施例做出另外的变更和修改。所以, 所附权利要求意欲解释为包括优选实施例以及落入本申请实施例范围的所有变更和修改。

最后, 还需要说明的是, 在本文中, 诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来, 而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且, 术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含, 从而使得包括一系列要素的过程、方法、物品或者终端设备不仅包括那些要素, 而且还包括没有明确列出的其他要素, 或者是还包括为这种过程、

方法、物品或者终端设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者终端设备中还存在另外的相同要素。

- 5 以上对本申请所提供的一种类 Alexnet 网络的模型训练方法和一种类 Alexnet 网络的模型训练装置，进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，
- 10 本说明书内容不应理解为对本申请的限制。

权 利 要 求 书

1、一种类 Alexnet 网络的模型训练方法，其特征在于，包括：

采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值；

5 接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值；

依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数；

10 接收所述第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第四梯度值；

依据所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模型参数；

采用所述第一模型参数和第二模型参数训练所述类 Alexnet 网络的模型。

15 2、根据权利要求 1 所述的方法，其特征在于，所述类 Alexnet 网络由全连接层和卷积层组成，所述采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值的步骤包括：

采用第一图形处理单元 GPU 计算在所述全连接层下的第一梯度值和
在所述卷积层下的第二梯度值。

20 3、根据权利要求 2 所述的方法，其特征在于，所述第一图形处理单元 GPU 包括第一计算队列，所述采用第一图形处理单元 GPU 计算在所述全连接层下的第一梯度值和
在所述卷积层下的第二梯度值的步骤包括：

采用第一计算队列计算在所述全连接层下的第一梯度值和
在所述卷积层下的第二梯度值。

25 4、根据权利要求 3 所述的方法，其特征在于，所述第一图形处理单

元 GPU 还包括第一通信队列，所述第二图形处理单元 GPU 包括第二通信队列，所述接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值的步骤包括：

采用第一通信队列接收第二通信队列发送的第三梯度值；

5 所述接收所述第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第四梯度值的步骤包括：

采用第一通信队列接收第二通信队列发送的第四梯度值。

5、根据权利要求 4 所述的方法，其特征在于，所述第二图形处理单元还包括第二计算队列，所述第三梯度值和所述第四梯度值分别通过如下
10 步骤获得：

采用第二计算队列计算在所述全连接层下的第三梯度值；以及，

采用第二计算队列计算在所述卷积层下的第四梯度值。

6、根据权利要求 1-5 任一所述的方法，其特征在于，所述依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数的步骤包
15 括：

计算所述第一梯度值和第三梯度值的平均值，获得所述类 Alexnet 网络的第一模型参数。

7、根据权利要求 6 所述的方法，其特征在于，所述采用所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模型参数的步骤包括：

20 计算所述第二梯度值和第四梯度值的平均值，获得所述类 Alexnet 网络的第二模型参数。

8、根据权利要求 1 或 2 或 3 或 4 或 5 或 7 所述的方法，其特征在于，在采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值的步骤前，还包括：

25 判断网络是否为类 Alexnet 网络。

9、根据权利要求 8 所述的方法，其特征在于，所述网络包括 m 个结构层，所述判断网络是否为类 Alexnet 网络的步骤包括：

对所述网络进行预训练，获得每个结构层的计算时间和参数量；

根据所述计算时间和参数量，获得所述网络的计算总时间和总参数量；

按照预设传输顺序，逐层累加所述 m 个结构层的计算时间，分别获得截至第 n 层的计算时间之和；

当所述截至第 p 层的计算时间之和与所述计算总时间的比值满足第一预设条件时，累加剩余 $m-p$ 层的参数量，获得所述剩余 $m-p$ 层的参数量之和；

判断所述剩余 $m-p$ 层的参数量之和与所述总参数量的比值是否满足第二预设条件；

若是，则将所述网络划分为类 Alexnet 网络。

10、根据权利要求 9 所述的方法，其特征在于，所述将所述网络划分为类 Alexnet 网络的步骤包括：

将所述网络的前 p 层划分为类 Alexnet 网络的全连接层；

将所述剩余 $m-p$ 层划分为类 Alexnet 网络的卷积层。

11、一种类 Alexnet 网络的模型训练装置，其特征在于，包括：

第一计算模块，用于采用第一图形处理单元 GPU 计算在类 Alexnet 网络下的第一梯度值和第二梯度值；

第一接收模块，用于接收第二图形处理单元 GPU 发送的在所述类 Alexnet 网络下的第三梯度值；

第二计算模块，用于依据所述第一梯度值和第三梯度值计算所述类 Alexnet 网络的第一模型参数；

第二接收模块，用于接收所述第二图形处理单元 GPU 发送的在所述

类 Alexnet 网络下的第四梯度值;

第三计算模块, 用于依据所述第二梯度值和第四梯度值计算所述类 Alexnet 网络的第二模型参数;

训练模块, 用于采用所述第一模型参数和第二模型参数训练所述类 Alexnet 网络的模型。

12、根据权利要求 11 所述的装置, 其特征在于, 所述类 Alexnet 网络由全连接层和卷积层组成, 所述第一计算模块包括:

第一计算子模块, 用于采用第一图形处理单元 GPU 计算在所述全连接层下的第一梯度值和在所述卷积层下的第二梯度值。

10 13、根据权利要求 12 所述的装置, 其特征在于, 所述第一图形处理单元 GPU 包括第一计算队列, 所述第一计算子模块包括:

第一计算单元, 用于采用第一计算队列计算在所述全连接层下的第一梯度值和在所述卷积层下的第二梯度值。

15 14、根据权利要求 13 所述的装置, 其特征在于, 所述第一图形处理单元 GPU 还包括第一通信队列, 所述第二图形处理单元 GPU 包括第二通信队列, 所述第一接收模块包括:

第一接收子模块, 用于采用第一通信队列接收第二通信队列发送的第三梯度值;

所述第二接收模块包括:

20 第二接收子模块, 用于采用第一通信队列接收第二通信队列发送的第四梯度值。

15 15、根据权利要求 14 所述的装置, 其特征在于, 所述第二图形处理单元还包括第二计算队列, 所述第三梯度值和所述第四梯度值分别通过如下模块获得:

25 第四计算模块, 用于采用第二计算队列计算在所述全连接层下的第三

梯度值；以及，

第五计算模块，用于采用第二计算队列计算在所述卷积层下的第四梯度值。

16、根据权利要求 11-15 任一所述的装置，其特征在于，所述第二计算模块包括：

第一模型参数计算子模块，用于计算所述第一梯度值和第三梯度值的平均值，获得所述类 Alexnet 网络的第一模型参数。

17、根据权利要求 16 所述的装置，其特征在于，所述第三计算模块包括：

第二模型参数计算子模块，用于计算所述第二梯度值和第四梯度值的平均值，获得第二模型参数。

18、根据权利要求 11 或 12 或 13 或 14 或 15 或 17 所述的装置，其特征在于，所述装置还包括：

判断模块，用于判断网络是否为类 Alexnet 网络。

19、根据权利要求 18 所述的装置，其特征在于，所述网络包括 m 个结构层，所述判断模块包括：

计算时间和参数量获得子模块，用于对所述网络进行预训练，获得每个结构层的计算时间和参数量；

计算总时间和总参数量获得子模块，用于根据所述计算时间和参数量，获得所述网络的计算总时间和总参数量；

计算时间之和获得子模块，用于按照预设传输顺序，逐层累加所述 m 个结构层的计算时间，分别获得截至第 p 层的计算时间之和；

参数量之和获得子模块，用于在所述截至第 p 层的计算时间之和与所述计算总时间的比值满足第一预设条件时，累加剩余 $m-p$ 层的参数量，获得所述剩余 $m-p$ 层的参数量之和；

判断子模块，用于判断所述剩余 $m-p$ 层的参数量之和与所述总参数量的比值是否满足第二预设条件；

划分子模块，用于在满足第二预设条件时，将所述网络划分为类 Alexnet 网络。

- 5 20、根据权利要求 19 所述的装置，其特征在于，所述划分子模块包括：

全连接层划分单元，用于将所述网络的前 p 层划分为类 Alexnet 网络的全连接层；

- 10 卷积层划分单元，用于将所述剩余 $m-p$ 层划分为类 Alexnet 网络的卷积层。

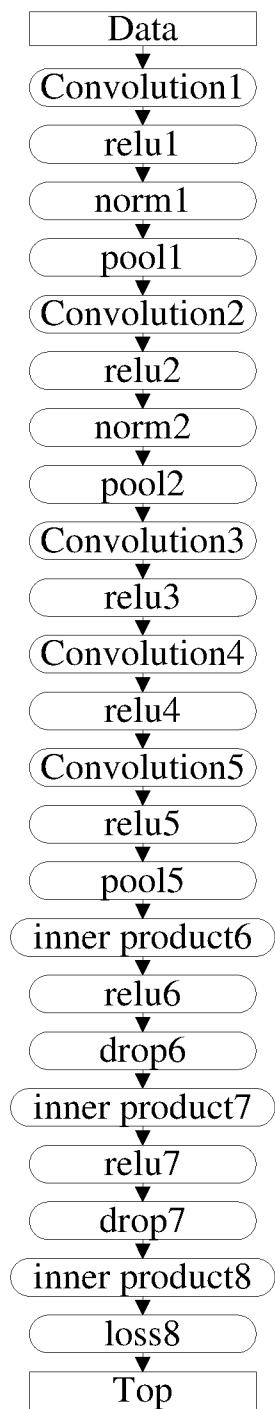


图 1

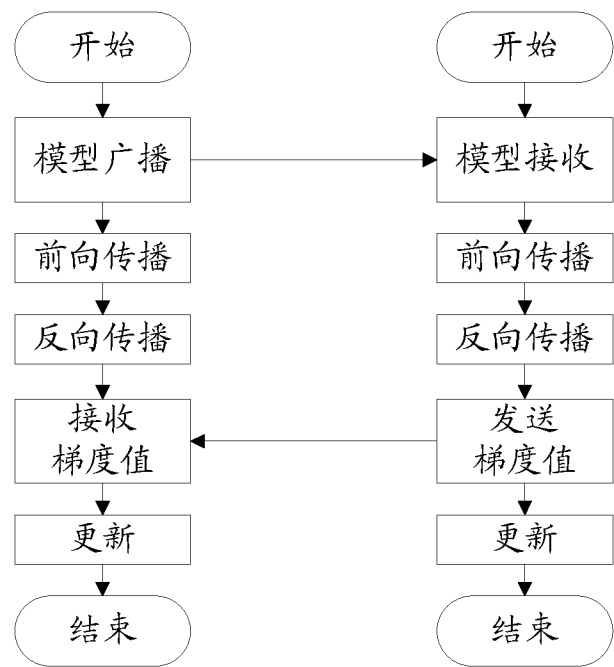


图 2



图 3

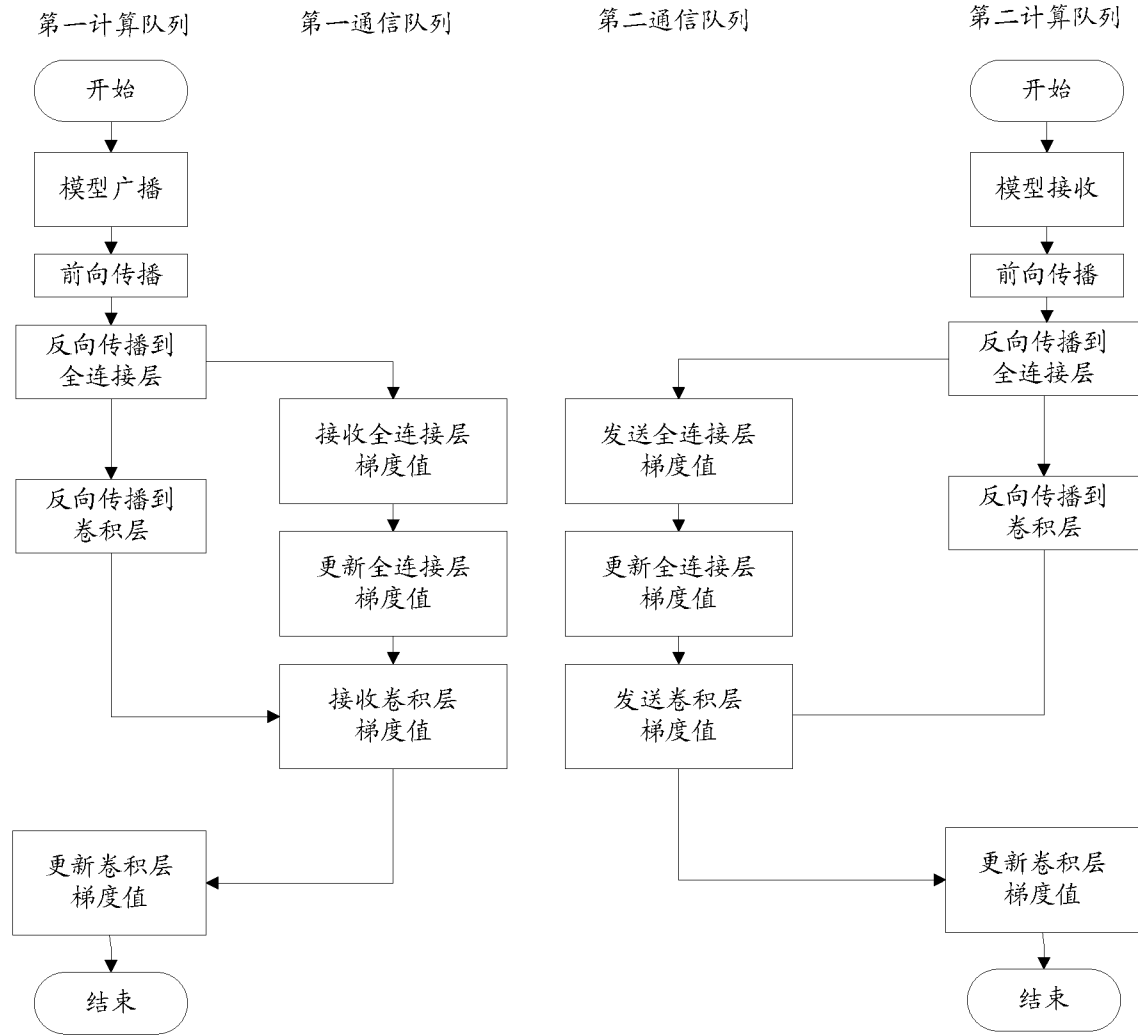


图 4

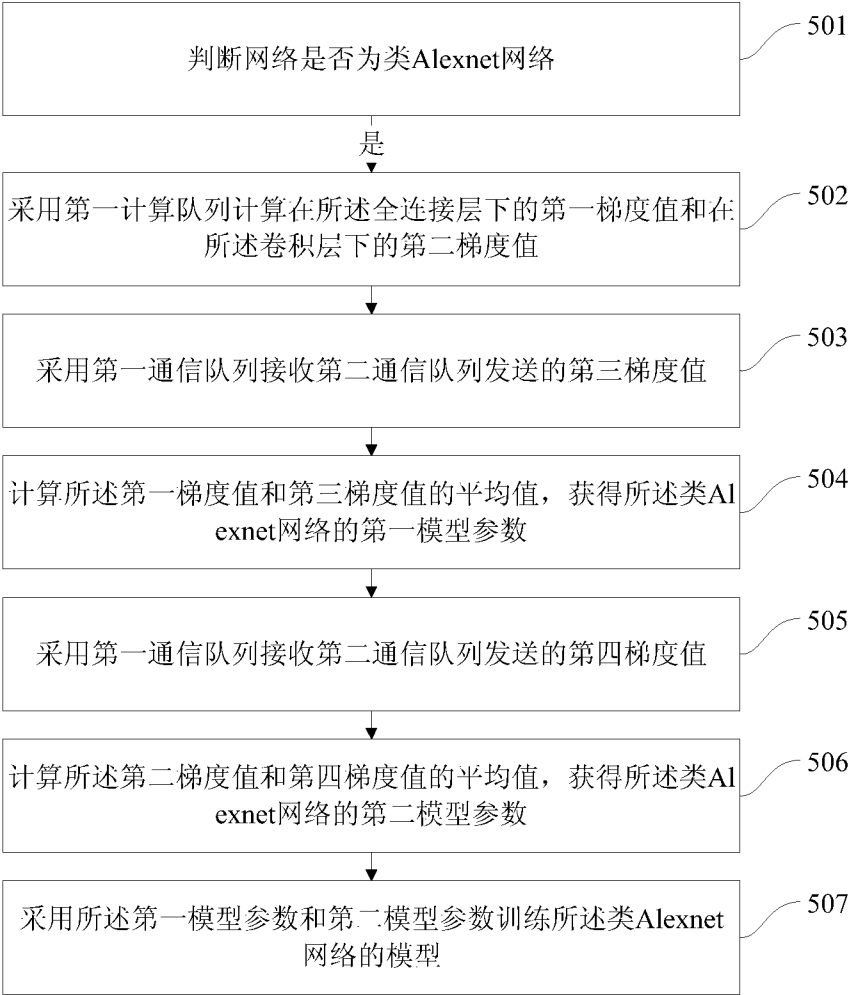


图 5

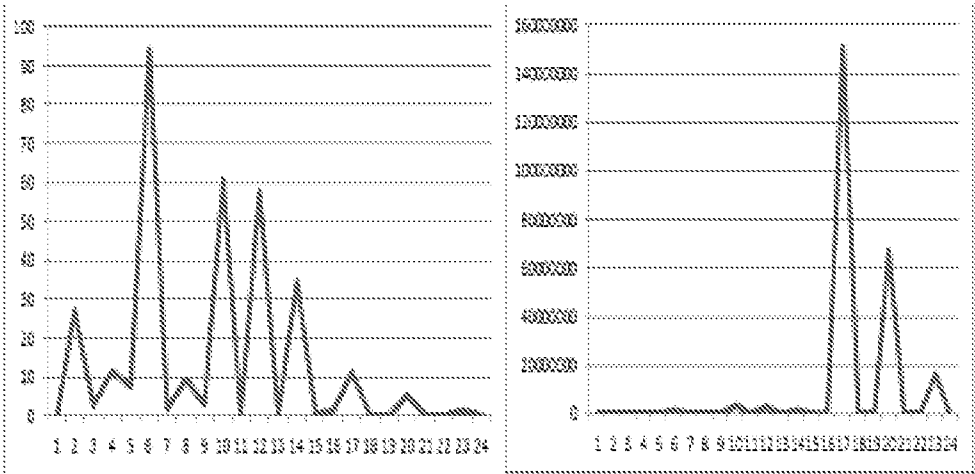


图 6

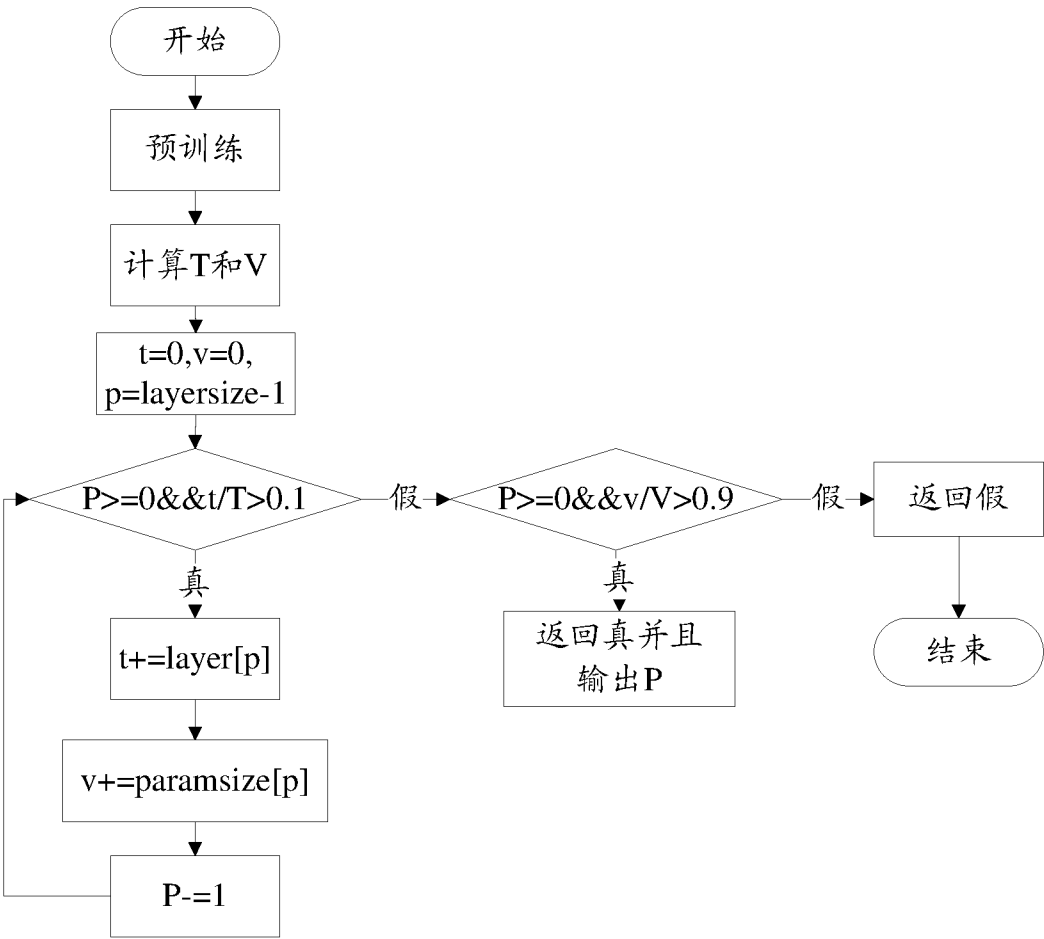


图 7

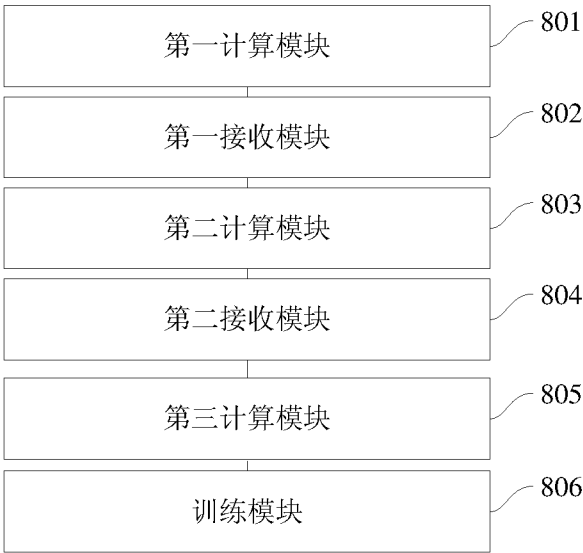


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/077897

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08 (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; G06T; 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)

CNKI, IEEE, VEN, CPRSABS, CNTXT, USTXT: multi-core multi-thread convolution back propagation ALEXNET BP GPU CNN GRAPHICS PROCESS+, BACK, PROPAGATION, MULTI+, PLURALITY, CORE?, THREAD?, GRADIENT? DISTRIBUT+ CONVOLUTION+ NEURAL NETWORK

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104036451 A (TENCENT INC.), 10 September 2014 (10.09.2014), description, paragraphs [0097]-[0145], and figures 9-16	1-20
A	CN 104035751 A (TENCENT INC.), 10 September 2014 (10.09.2014), the whole document	1-20
A	CN 104463324 A (CHANGSHA MASHA ELECTRONIC TECHNOLOGY CO., LTD.), 25 March 2015 (25.03.2015), the whole document	1-20
A	US 2015161522 A1 (INT BUSINESS MACHINES CORP.), 11 June 2015 (11.06.2015), the whole document	1-20

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 22 June 2017 (22.06.2017)	Date of mailing of the international search report 28 June 2017 (28.06.2017)
Name and mailing address of the ISA/CN: State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No.: (86-10) 62019451	Authorized officer HU, Yan Telephone No.: (86-10) 62089101

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2017/077897

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104036451 A	10 September 2014	WO 2015192806 A1	23 December 2015
		US 9607355 B2	28 March 2017
		EP 3158529 A1	26 April 2017
		EP 3158529 A4	31 May 2017
		US 2016321776 A1	03 November 2016
CN 104035751 A	10 September 2014	CN 104035751 B	12 October 2016
		WO 2015192812 A1	23 December 2015
		US 2016321777 A1	03 November 2016
		EP 3129870 A1	15 February 2017
CN 104463324 A	25 March 2015	None	
US 2015161522 A1	11 June 2015	US 9665823 B2	30 May 2017

A. 主题的分类 G06N 3/08 (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; G06T; 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)) CNKI, IEEE, VEN, CPRSABS, CNTXT, USTXT: 图形处理 多核 多线程 分布式 梯度 卷积 反向传播 卷积神经网络 ALEXNET BP GPU CNN GRAPHICS PROCESS+, BACK, PROPAGATION, MULTI+, PLURALITY, CORE?, THREAD?, GRADIENT? DISTRIBUT+ CONVOLUTION+ NEURAL NETWORK		
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
X	CN 104036451 A (深圳市腾讯计算机系统有限公司) 2014年 9月 10日 (2014 - 09 - 10) 说明书第 [0097]-[0145]段, 及图9-16	1-20
A	CN 104035751 A (深圳市腾讯计算机系统有限公司) 2014年 9月 10日 (2014 - 09 - 10) 全文	1-20
A	CN 104463324 A (长沙马沙电子科技有限公司) 2015年 3月 25日 (2015 - 03 - 25) 全文	1-20
A	US 2015161522 A1 (INT BUSINESS MACHINES CORP) 2015年 6月 11日 (2015 - 06 - 11) 全文	1-20
☐ 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 2017年 6月 22日		国际检索报告邮寄日期 2017年 6月 28日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局 (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 胡燕 电话号码 (86-10)62089101

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/077897

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	104036451	A	2014年 9月 10日	WO	2015192806	A1	2015年 12月 23日
				US	9607355	B2	2017年 3月 28日
				EP	3158529	A1	2017年 4月 26日
				EP	3158529	A4	2017年 5月 31日
				US	2016321776	A1	2016年 11月 3日
CN	104035751	A	2014年 9月 10日	CN	104035751	B	2016年 10月 12日
				WO	2015192812	A1	2015年 12月 23日
				US	2016321777	A1	2016年 11月 3日
				EP	3129870	A1	2017年 2月 15日
CN	104463324	A	2015年 3月 25日	无			
US	2015161522	A1	2015年 6月 11日	US	9665823	B2	2017年 5月 30日