

(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
2020 年 2 月 13 日 (13.02.2020)



(10) 国际公布号
WO 2020/029592 A1

(51) 国际专利分类号:
G06F 17/50 (2006.01)

(21) 国际申请号: PCT/CN2019/080510

(22) 国际申请日: 2019 年 3 月 29 日 (29.03.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810913895.6 2018年8月10日 (10.08.2018) CN

(71) 申请人: 中科寒武纪科技股份有限公司 (CAMBRICON TECHNOLOGIES CORPORATION LIMITED) [CN/CN]; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。

(72) 发明人: 刘少礼 (LIU, Shaoli); 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。 梁军 (LIANG, Jun); 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。 郭崎 (GUO, Qi); 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。

(74) 代理人: 北京华进京联知识产权代理有限公司 (ACIP LAW OFFICES); 中国北京市海淀区知春路 7 号致真大厦 A1403, Beijing 100191 (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, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) Title: CONVERSION METHOD, APPARATUS, COMPUTER DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 转换方法、装置、计算机设备及存储介质

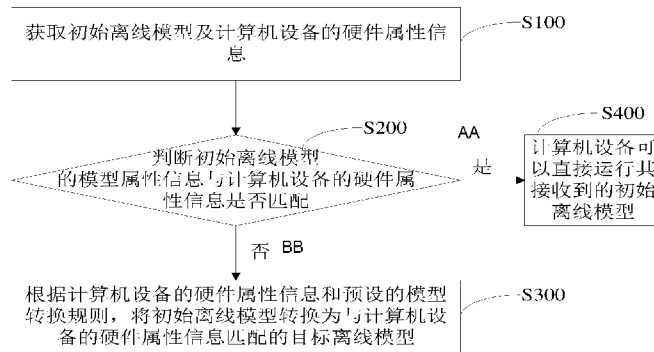


图 5

S100 Acquire an initial offline model and hardware attribute information of a computer device
S200 Determine whether model attribute information of the initial offline model matches the hardware attribute information of the computer device
S300 According to the hardware attribute information of the computer device and a preset model conversion rule, convert the initial offline model into a target offline model that matches the hardware attribute information of the computer device
S400 The computer device may directly run the initial offline model received thereby
AA Yes
BB No

(57) Abstract: A model conversion method, an apparatus, a computer device, and a storage medium, wherein the method may convert an initial offline model into a target offline model. The described model conversion method, apparatus, computer device, and storage medium greatly reduce the data input volume of a computer device, and the conversion process is simple; thus, the data processing volume of the computer device may be reduced, processing efficiency may thereby be increased, and power consumption is reduced.

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, 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: 一种模型转换方法、装置、计算机设备及存储介质, 该方法能够将初始离线模型转换为目标离线模型。上述的模型转换方法、装置、计算机设备和存储介质, 极大的降低了计算机设备的数据输入量, 转换过程简单, 从而可以降低该计算机设备的数据处理量, 进而可以提高处理效率, 降低功耗。

转换方法、装置、计算机设备及存储介质

相关申请

本申请要求 2018 年 08 月 10 日申请的, 申请号为 201810913895.6, 名称为“转换方法、装置、计算机设备及存储介质”的中国专利申请的优先权, 在此将其全文引入作为参考。

技术领域

本申请涉及计算机技术领域, 具体涉及一种转换方法、装置、计算机设备及存储介质。

背景技术

随着人工智能技术的发展, 如今深度学习已无处不在且必不可少, 并随之产生了许多可扩展的深度学习系统, 例如, TensorFlow、MXNet、Caffe 和 PyTorch 等等, 上述深度学习系统可以用于提供各种能够在 CPU 或 GPU 等处理器上运行的神经网络模型或其他机器学习模型。当神经网络模型等机器学习模型需在不同的处理器上运行时, 往往需要对神经网络模型等机器学习模型进行转换。

传统的转换方法是: 开发者可以为同一神经网络模型配置多个不同的转换模型, 多个转换模型可以适用于不同的处理器, 在实际使用中, 计算机设备需同时接收该神经网络模型及其对应的多个转换模型, 用户需要根据当前计算机设备的类型从上述的多种模型中选择一个模型, 从而使得该神经网络模型能够在当前计算机设备上运行。但上述转换方法的数据输入量及数据处理量均较大, 上述较大的数据输入量及数据处理量容易超出计算机设备的存储容量及处理极限, 导致计算机设备的处理速度低甚至无法正常工作。

发明内容

为至少在一定程度上克服相关技术中存在的问题, 本申请提供转换方法、装置、计算机设备和存储介质。

本申请提供了一种模型转换方法, 所述方法包括如下步骤:

获取初始离线模型及计算机设备的硬件属性信息;

根据所述初始离线模型和所述计算机设备的硬件属性信息, 判断所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息是否匹配;

若所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息不匹配, 则根据所述计算机设备的硬件属性信息和预设的模型转换规则, 将所述初始离线模型转换为与所述计算机设备的硬件属性信息匹配的目标离线模型。

在一个实施例中, 所述的根据所述计算机设备的硬件属性信息和预设的模型转换规则, 将所述初始离线模型转换为与所述计算机设备的硬件属性信息匹配的目标离线模型的步骤, 包括:

根据所述计算机设备的硬件属性信息确定所述目标离线模型的模型属性信息;

根据所述初始离线模型的模型属性信息和所述目标离线模型的模型属性信息, 从多个所述预设的模型转换规则中选择一个所述模型转换规则作为目标模型转换规则;

根据所述初始离线模型的模型属性信息和所述目标模型转换规则, 将所述初始离线模型转换为所述目标离线模型。

在一个实施例中，所述的根据所述初始离线模型的模型属性信息和所述目标离线模型的模型属性信息，从多个所述预设的模型转换规则中选择一个所述模型转换规则作为目标模型转换规则的步骤，包括：

根据所述初始离线模型的模型属性信息和所述目标离线模型的模型属性信息，从多个所述预设的模型转换规则中选择一个以上的可用模型转换规则；

对所述一个以上的可用模型转换规则进行优先级排序，并将优先级最高的可用模型转换规则作为所述目标转换规则。

在一个实施例中，所述的对一个以上的所述可用模型转换规则进行优先级排序的步骤，包括如下步骤：

分别获取采用各个所述可用模型转换规则，将所述初始离线模型转换为所述目标离线模型的过程参数，其中，所述过程参数包括转换速度、功耗、内存占用率及磁盘 I/O 占用率中的一种或多种；

根据各个所述可用模型转换规则的过程参数，对一个以上的可用模型转换规则进行优先级排序。

在一个实施例中，所述初始离线模型的模型属性信息、所述目标离线模型的模型属性信息及所述可用模型转换规则三者之间一一对应存储。

在一个实施例中，所述的根据所述初始离线模型和所述计算机设备的硬件属性信息，判断所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息是否匹配的步骤包括：

若根据所述初始离线模型的模型属性信息和所述计算机设备的硬件属性信息，确定所述计算机设备能够支持所述初始离线模型的运行时，则判定所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息匹配；

若根据所述初始离线模型的模型属性信息和所述计算机设备的硬件属性信息，确定所述计算机设备不支持所述初始离线模型的运行时，则判定所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息不匹配。

在一个实施例中，所述的获取初始离线模型的步骤，包括：

通过所述计算机设备上的应用软件获取所述初始离线模型和所述计算机设备的硬件属性信息。

在一个实施例中，所述方法还包括如下步骤：

将所述目标离线模型存储于所述计算机设备的第一存储器或第二存储器。

在一个实施例中，所述方法还包括如下步骤：

获取所述目标离线模型，并运行所述目标离线模型，其中，所述目标离线模型中包含原始网络中各个离线节点对应的网络权值、指令以及各个离线节点与所述原始网络中的其他计算节点之间的接口数据。

本申请提供了一种模型转换装置，所述装置包括：

获取模块，用于获取初始离线模型及计算机设备的硬件属性信息；

判断模块，用于根据所述初始离线模型和所述计算机设备的硬件属性信息，判断所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息是否匹配；

转换模块，用于在所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息不匹配时，根据所述计算机设备的硬件属性信息和预设的模型转换规则，将所述初始离

线模型转换为与所述计算机设备的硬件属性信息匹配的目标离线模型。

本申请提供了一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现上述转换方法的步骤。

在一个实施例中，所述处理器包括运算单元和控制器单元，所述运算单元包括主处理电路和多个从处理电路；

所述控制器单元用于获取数据、机器学习模型以及计算指令；

所述控制器单元还用于解析所述计算指令得到多个运算指令，并将所述多个运算指令以及所述数据发送给所述主处理电路；

所述主处理电路用于对所述数据以及在所述主处理电路与所述多个从处理电路之间进行传输的数据和运算指令执行前序处理；

所述多个从处理电路用于依据从所述主处理电路传输的数据以及运算指令并行执行中间运算得到多个中间结果，并将多个中间结果传输给所述主处理电路；

所述主处理电路还用于对所述多个中间结果执行后续处理得到所述计算指令的计算结果。

一种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上述转换方法的步骤。

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。

图 1 为一个实施例中计算机设备的结构框图；

图 2 为图 1 中处理器的一个实施例的结构框图；

图 3 为图 1 中处理器的一个实施例的结构框图；

图 4 为图 1 中处理器的一个实施例的结构框图；

图 5 为一个实施例中模型转换方法的流程示意图；

图 6 为一个实施例中模型转换方法的流程示意图；

图 7 为一个实施例中离线模型生成方法的流程示意图；

图 8 为一个实施例中离线模型生成方法的流程示意图；

图 9 为一实施例的网络模型的网络结构图；

图 10 为图 9 中网络模型的离线模型生成过程示意图。

具体实施方式

这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

图 1 为一实施例的计算机设备的框图，该计算机设备可以是手机或平板电脑等移动终端，或台式电脑、板卡或云端服务器等终端。该计算机设备可以应用于机器人、打印机、

扫描仪、行车记录仪、导航仪、相机、摄像机、投影仪、手表、移动存储、可穿戴设备、交通工具、家用电器、和/或医疗设备。其中，交通工具可以包括飞机、轮船和/或车辆；家用电器可以包括电视、空调、微波炉、冰箱、电饭煲、加湿器、洗衣机、电灯、燃气灶、油烟机；医疗设备可以包括核磁共振仪、B超仪和/或心电图仪等等。

该计算机设备可以包括处理器 100、与该处理器 100 连接的第一存储器 200 及第二存储器 300。可选地，处理器 100 可以是通用处理器，如 CPU（Central Processing Unit，中央处理器）、GPU（Graphics Processing Unit，图形处理器）或 DSP（Digital Signal Processing，数字信号处理），该处理器 100 还可以为 IPU（Intelligence Processing Unit，智能处理器）等网络模型处理器。当然，该处理器 100 还可以是指令集处理器、相关芯片组、专用微处理器（如，专用集成电路（ASIC））或用于缓存用途的板载存储器等等。

可选地，如图 2 所示，该处理器 100 可以包括控制器单元 110 和运算单元 120，其中，控制器单元 110 与运算单元 120 连接，该运算单元 120 可以包括一个主处理电路 121 和多个从处理电路 122。控制器单元 110 用于获取数据、机器学习模型以及计算指令。该机器学习模型具体可以包括网络模型，该网络模型可以为神经网络模型和/或非神经网络模型。控制器单元 110 还用于解析其获取的计算指令得到运算指令，并将多个运算指令以及数据发送给主处理器电路。主处理电路用于对数据以及该主处理电路与多个从处理电路之间传输的数据和运算指令执行前序处理。多个从处理电路用于依据从主处理电路传输的数据以及运算指令并行执行中间运算得到多个中间结果，并将多个中间结果传输给主处理电路；主处理电路还用于对多个中间结果执行后续处理得到计算指令的计算结果。

可选地，该控制器单元 110 可以包括指令存储单元 111、指令处理单元 112 和存储队列单元 114；指令存储单元 111 用于存储机器学习模型关联的计算指令；指令处理单元 112 用于对计算指令解析得到多个运算指令；存储队列单元 114 用于存储指令队列，该指令队列包括：按该队列的前后顺序待执行的多个运算指令或计算指令。可选地，该控制器单元 110 还可以包括依赖关系处理单元 113，用于在具有多个运算指令时，确定第一运算指令与第一运算指令之前的第零运算指令是否存在关联关系；如第一运算指令与第零运算指令存在关联关系，则将第一运算指令缓存在指令存储单元内，在第零运算指令执行完毕后，从指令存储单元提取第一运算指令传输至运算单元。具体地，若依赖关系处理单元 113 依据第一运算指令提取第一运算指令中所需数据（例如矩阵）的第一存储地址区间，依据第零运算指令提取第零运算指令中所需矩阵的第零存储地址区间；如第一存储地址区间与第零存储地址区间具有重叠的区域，则确定第一运算指令与第零运算指令具有关联关系；如第一存储地址区间与第零存储地址区间不具有重叠的区域，则确定第一运算指令与第零运算指令不具有关联关系。

在一个实施例中，如图 3 所示，运算单元 120 还可以包括分支处理电路 123，其中，主处理电路 121 与分支处理电路 123 连接，分支处理电路 123 与多个从处理电路 122 连接；分支处理电路 123，用于执行转发主处理电路 121 与从处理电路 122 之间的数据或指令。在此实施例中，主处理电路 121 具体用于将一个输入神经元分配成多个数据块，将多个数据块中的至少一个数据块、权值以及多个运算指令中的至少一个运算指令发送给分支处理电路；分支处理电路 123 用于转发主处理电路 121 与多个从处理电路 122 之间的数据块、权值以及运算指令；多个从处理电路 122 用于依据该运算指令对接收到的数据块以及权值执行运算得到中间结果，并将中间结果传输给分支处理电路 123；主处理电路 121 还

用于将分支处理电路发送的中间结果进行后续处理得到该计算指令的结果，将该计算指令的结果发送给所述控制器单元。

在另一种可选实施例中，如图4所示，运算单元120可以包括一个主处理电路121和多个从处理电路122。其中，多个从处理电路呈阵列分布；每个从处理电路与相邻的其他从处理电路连接，主处理电路连接多个从处理电路中的k个从处理电路，k个从处理电路为：第1行的n个从处理电路、第m行的n个从处理电路以及第1列的m个从处理电路，需要说明的是，如图1C所示的K个从处理电路仅包括第1行的n个从处理电路、第m行的n个从处理电路以及第1列的m个从处理电路，即该k个从处理电路为多个从处理电路中直接与主处理电路连接的从处理电路。K个从处理电路用于在主处理电路以及多个从处理电路之间的数据以及指令的转发。

可选地，上述的主处理电路121可以包括转换处理电路、激活处理电路、加法处理电路中的一种或任意组合；转换处理电路用于将主处理电路接收的数据块或中间结果执行第一数据结构与第二数据结构之间的互换（例如连续数据与离散数据的转换）；或将主处理电路接收的数据块或中间结果执行第一数据类型与第二数据类型之间的互换（例如定点类型与浮点类型的转换）；激活处理电路用于执行主处理电路内数据的激活运算；加法处理电路用于执行加法运算或累加运算。

该第一存储器200或第二存储器300可以存储有计算机程序，该计算机程序用于实现本申请实施例中提供的模型转换方法。具体地，该模型转换方法用于在计算机设备接收的初始离线模型的模型属性信息与该计算机设备的硬件属性信息不匹配时，将该初始离线模型转换为与该计算机设备的硬件属性信息相匹配的目标离线模型，以便该计算机设备能够运行该初始离线模型。上述的模型转换方法中，计算机设备的数据输入量少，且无需人为干预即可自动完成初始离线模型到目标离线模型的转换，转换过程简单，转换效率高。进一步地，第一存储器200或第二存储器300上存储的计算机程序还能够用于实现本申请实施例中提供的离线模型生成方法。

可选地，第一存储器200可以用于存储网络模型运行过程中的相关数据，如网络输入数据、网络输出数据、网络权值及指令等等。该第一存储器200可以是内存储器，如缓存等易失性存储器。该第二存储器300可以用于存储网络模型对应的离线模型，例如，第二存储器300可以是非易失性存储器。

上述计算机设备的工作原理与下文中的模型转换方法中各个步骤的执行过程一致，该计算机设备的处理器100在执行存储器200中的计算机程序时，实现模型转换方法中的各个步骤，具体可参见下文中的描述。

当然，在其他实施例中，该计算机设备还可以包含处理器和一个存储器，该计算机设备可以包含处理器与该处理器连接的存储器。该处理器可以采用图2-4所示的处理器，其具体结构可参见上文中关于处理器100的描述。该存储器可以包括多个存储单元，例如该存储器可以包括第一存储单元、第二存储单元和第三存储单元，其中，该第一存储单元可以用于存储计算机程序，该计算机程序用于实现本申请实施例中提供的模型转换方法。该第二存储单元可以用于存储原始网络运行过程中相关数据，该第三存储单元用于存储原始网络对应的离线模型及其该离线模型对应的目标离线模型及预设的模型转换规则等等。进一步地，该存储器包含的存储单元的数量还可以大于三个，此处不做具体限定。

如图5所示，本申请实施例提供了一种模型转换方法，可应用于上述的计算机设备中。

该模型转换方法用于在计算机设备接收的初始离线模型的模型属性信息与该计算机设备的硬件属性信息不匹配时,将该初始离线模型转换为与该计算机设备的硬件属性信息相匹配的目标离线模型,以便该计算机设备能够运行该初始离线模型。具体地,上述方法可以包括如下步骤:

S100: 获取初始离线模型及计算机设备的硬件属性信息;

具体地,上述的初始离线模型是指计算机设备直接获取的离线模型,该初始离线模型可以存储于第二存储器 300 中。其中,离线模型可以包含一原始网络中各个计算节点的网络权值以及指令等必要的网络结构信息,指令可以用于表明该计算节点用于执行何种计算功能,其具体可以包括该原始网络中各个计算节点的计算属性以及各个计算节点之间的连接关系等信息。该原始网络可以是网络模型,如神经网络模型等等,参见图 9 所示。因此,计算机设备可以通过运行该原始网络对应的离线模型,实现该原始网络的运算功能,无需重复对同一原始网络进行编译等操作,从而可以缩短处理器运行该网络时的运行时间,进而提高处理器的处理速度及效率。可选地,本申请实施例中的初始离线模型可以是根据原始网络直接生成的离线模型,也可以是具有其他模型属性信息的离线模型经过一次或多次转换后获得的离线模型,此处不做具体限定。

S200: 根据初始离线模型和计算机设备的硬件属性信息,判断初始离线模型的模型属性信息与计算机设备的硬件属性信息是否匹配。

具体地,该初始离线模型的模型属性信息可以包括该初始离线模型的数据结构及数据类型等等。例如,初始离线模型中各个网络权值及指令的摆放方式、各个网络权值的类型及各个指令的类型等等。上述的计算机设备的硬件属性信息包括该计算机设备的型号、该计算机设备能够处理的数据类型(如定点或浮点等等)及数据结构等等。计算机设备可以根据上述初始离线模型的模型属性信息和计算机设备的硬件属性信息,判断该计算机设备能否支持上述初始离线模型的运行(即该计算机设备能否运行上述的初始离线模型),从而确定该初始离线模型的模型属性信息与计算机设备的硬件属性信息是否匹配。

可选地,当该计算机设备能够支持上述初始离线模型的运行时,则判定该初始离线模型的模型属性信息与计算机设备的硬件属性信息匹配。当该计算机设备不支持上述初始离线模型的运行时,则判定该初始离线模型的模型属性信息与该计算机设备的硬件属性信息不匹配。

例如,该计算机设备可以根据其硬件属性信息确定该计算机设备能够处理的数据类型及数据结构,根据该初始离线模型的模型属性信息确定该初始离线模型的数据类型及数据结构。若该计算机设备根据其自身的硬件属性信息以及上述初始离线模型的模型属性信息,确定该计算机设备能够支持上述初始离线模型的数据类型及数据结构时,则可以判定该初始离线模型的模型属性信息与该计算机设备的硬件属性信息匹配。若该计算机设备根据其自身的硬件属性信息以及上述的初始离线模型的模型属性信息,确定该计算机设备不能够支持上述初始离线模型的数据类型及数据结构时,则可以判定该初始离线模型的模型属性信息与该计算机设备的硬件属性信息不匹配。

若初始离线模型的模型属性信息与计算机设备的硬件属性信息不匹配,则执行步骤 S300: 根据计算机设备的硬件属性信息和预设的模型转换规则,将初始离线模型转换为与计算机设备的硬件属性信息匹配的目标离线模型。

具体地,若该初始离线模型的模型属性信息与该计算机设备的硬件属性信息不匹配,

则说明该计算机设备不能支持该初始离线模型的运行，此时需对该初始离线模型进行转换处理，以便该计算机设备能够使用该初始离线模型实现相应的运算。即当该初始离线模型的模型属性信息与计算机设备的硬件属性信息不匹配时，则可以根据计算机设备的硬件属性信息和预设的模型转换规则，将上述的初始离线模型转换为与该计算机设备的硬件属性信息匹配的目标离线模型。其中，该目标离线模型的模型属性信息与该计算机设备的硬件属性信息匹配，该计算机设备能够支持该目标离线模型的运行。同上，该目标离线模型也包括一原始网络中各个计算节点的网络权值以及指令等必要的网络结构信息。

可选地，上述的预设的模型转换规则可以预先存储于计算机设备的第一存储器或第二存储器中。若初始离线模型的模型属性信息与计算机设备的硬件属性信息不匹配时，计算机设备可以从第一存储器或第二存储器中获得相应的模型转换规则，以将初始离线模型转换为目标离线模型。可选地，上述预设的模型转换规则可以是一个以上，一个以上的模型转换规则可以与上述的初始离线模型及目标离线模型一一对应存储。例如，上述的初始离线模型、目标离线模型以及预设的模型转换规则可以通过映射表的方式进行对应存储。可选地，初始离线模型的模型属性信息、目标离线模型的模型属性信息及可用模型转换规则三者之间一一对应存储。

若初始离线模型的模型属性信息与计算机设备的硬件属性信息匹配，则可以执行步骤S400：计算机设备可以直接运行其接收到的初始离线模型，即计算机设备可以根据该初始离线模型包含的网络权值及指令进行运算，以实现原始网络的运算功能。具体地，若该初始离线模型的模型属性信息与该计算机设备的硬件属性信息匹配，则说明该计算机设备能够支持该初始离线模型的运行，此时无需对该初始离线模型进行转换处理，计算机设备能够运行该初始离线模型。应当清楚的是，本实施例中，直接运行该初始离线模型是指，使用该初始离线模型运行该原始网络对应的机器学习算法（如神经网络算法），通过执行前向运算实现算法的目标应用（如语音识别等人工智能应用）。

本申请实施例中的模型转换方法，计算机设备只需接收一个初始离线模型，即可根据预设的模型转换规则和该计算机设备的硬件属性信息，将该初始离线模型转换为目标离线模型，无需获得多个不同的模型数据，极大的降低了计算机设备的数据输入量，避免了数据输入量过大导致的超出计算机设备的存储容量等问题，保证计算机设备的正常运行。同时，上述模型转换方法可以降低该计算机设备的数据处理量，进而可以提高处理效率，降低功耗。同时，上述的模型转换过程中，无需人工干预，自动化程度高，使用便捷。

可选地，如图6所示，上述的步骤S300可以包括：

S310：根据计算机设备的硬件属性信息确定目标离线模型的模型属性信息。

具体地，该计算机设备可以根据其自身的硬件属性信息，确定该计算机设备能够支持的数据类型及数据结构等等，从而可以确定目标离线模型的数据类型及数据结构等模型属性信息。即该计算机设备能够根据其硬件属性信息确定目标离线模型的模型属性信息，该目标离线模型的模型属性信息可以包括目标离线模型的数据结构及数据类型等信息。

S320：根据初始离线模型的模型属性信息和目标离线模型的模型属性信息，从多个预设的模型转换规则中选择一个模型转换规则作为目标模型转换规则。

具体地，由于初始离线模型、目标离线模型和预设的模型转换规则存在一一对应的映射关系，因此，可以根据该初始离线模型的模型属性信息和目标离线模型的模型属性信息确定出目标模型转换规则。其中，该模型转换规则可以包括数据类型的转换方法及数据结

构的转换方法等等。

S330: 根据初始离线模型的模型属性信息和目标模型转换规则, 将初始离线模型转换为目标离线模型。

具体地, 计算机设备可以根据该目标模型转换规则提供的转换方法, 将初始离线模型转换为目标离线模型, 从而使得该计算机设备能够运行该目标离线模型进行运算。

可选地, 若初始离线模型和目标离线模型之间存在多个可用的模型转换规则时, 计算机设备可以通过根据预设算法自动选定目标离线模型。具体地, 上述的步骤 S320 进一步包括如下步骤:

根据初始离线模型的模型属性信息和目标离线模型的模型属性信息, 从多个预设的模型转换规则中选择一个以上的可用模型转换规则。

对一个以上的可用模型转换规则进行优先级排序, 并将优先级最高的可用模型转换规则作为目标转换规则。

具体地, 上述的可用模型转换规则是指能够将初始离线模型转换为目标离线模型的转换规则。当存在一个以上的可用模型转换规则时, 说明将初始离线模型转换为目标离线模型的方式有多种不同的方式。上述的优先级排序方法可以是预设的, 也可以是用户自定义的。

可选地, 计算机设备可以获取各个可用模型转换规则对应的过程参数, 并根据各个可用模型转换规则对应的过程参数, 对一个以上的可用模型转换规则进行优先级排序。其中, 该过程参数可以是采用该可用模型转换规则将初始离线模型转换为目标离线模型时所涉及的计算机设备的性能参数。可选地, 该过程参数可以包括转换速度、功耗、内存占用率及磁盘 I/O 占用率中的一种或多种。

例如, 计算机设备可以根据将初始离线模型转换为目标离线模型过程中的转换速度、转换功耗、内存占用率及磁盘 I/O 占用率等一种或多种过程参数的组合, 对上述的各个可用模型转换规则进行评分(如对各参考因素进行加权计算以获得评分值), 并将评分最高的可用模型转换规则作为优先级最高的可用模型转换规则。即可以将评分最高的可用模型转换规则作为该目标转换规则。之后, 可以执行上述步骤 S330, 根据初始离线模型的模型属性信息和目标模型转换规则, 将初始离线模型转换为目标离线模型。这样, 可以通过结合模型转换过程中的设备性能因素选择一种较优的模型转换规则, 以提高计算机设备的处理速度及效率。

可选地, 如图 3 所示, 上述方法还包括如下步骤:

S500: 将目标离线模型存储于计算机设备的第一存储器或第二存储器。

具体地, 计算机设备可以将其获得的目标离线模型存储于该计算机设备的本地存储器上(如第一存储器)。可选地, 计算机设备还可以将其获得的目标离线模型存储于与该计算机设备连接的外部存储器(如第二存储器), 该外部存储器可以是云端存储器或其他存储器等等。这样, 当该计算机设备需要重复使用该目标离线模型时, 无需重复执行上述的转换操作, 只需直接读取该目标离线模型即可实现相应的运算。

可选地, 计算机设备执行上述的目标离线模型的过程可以包括如下步骤:

获取目标离线模型, 并运行目标离线模型, 以执行原始网络。其中, 目标离线模型中包含原始网络中各个离线节点对应的网络权值、指令以及各个离线节点与原始网络中的其他计算节点之间的接口数据等必要的网络结构信息。

具体地，当计算机设备需要利用该目标离线模型实现运算时，其可以直接从第一存储器或第二存储器中获取该目标离线模型，并根据目标离线模型中的网络权值及指令等进行运算，从而实现原始网络的运算功能。这样，当该计算机设备需要重复使用该目标离线模型时，无需重复执行上述的转换操作，只需直接读取该目标离线模型即可实现相应的运算。

应当清楚的是，本实施例中，直接运行该目标离线模型是指，使用该目标离线模型运行该原始网络对应的机器学习算法（如神经网络算法），通过执行前向运算实现算法的目标应用（如语音识别等人工智能应用）。

可选地，该计算机设备上安装有应用软件（Application），可以通过计算机设备上的应用软件获取初始离线模型，该应用软件（Application）能够提供从存储器 200 或外部存储器中读取初始离线模型的接口，从而可以通过该应用软件获取该初始离线模型。可选地，该应用软件还可以提供读取计算机设备的硬件属性信息的接口，从而可以通过该应用软件获取该计算机设备的硬件属性信息。可选地，该应用软件还可以提供能够读取预设的模型规则的接口，从而可以通过该应用软件获取预设的模型转换规则。

当然，在其他实施例中，该计算机设备还可以提供输入/输出接口（如 I/O 接口）等，该输入/输出接口用于获取初始离线模型，或输出目标离线模型。该计算机设备的硬件属性信息可以是预设存储于该计算机设备中的。

在一个实施例中，如图 7 所示，本申请实施例还提供了一种根据原始网络模型生成离线模型的方法，用于根据获取的原始网络的相关数据生成并存储该原始网络的离线模型，从而在处理器再次运行该原始网络时，可以直接运行该原始网络对应的离线模型，无需再次对同一原始网络进行编译等操作，从而缩短处理器运行该网络时的运行时间，进而提高处理器的处理速度及效率。其中，原始网络可以是神经网络或非神经网络等网络模型，例如，原始网络可以是图 9 所示的网络。具体地，上述方法包括如下步骤：

S010: 获取原始网络的模型数据集及模型结构参数。

具体地，计算机设备的处理器可以获取原始网络的模型数据集及模型结构参数，通过该原始网络的模型数据集及模型结构参数可以获得该原始网络的网络结构图。其中，模型数据集包括原始网络中各个计算节点对应的网络权值等数据，图 9 所示的神经网络中的 W1~W6 即用于表示计算节点的网络权值。模型结构参数包括原始网络中多个计算节点的连接关系及各个计算节点的计算属性，其中，计算节点之间的连接关系用于表示计算节点之间是否有数据传递，例如，当多个计算节点之间具有数据流的传递时，则可以说明多个计算节点之间具有连接关系。进一步地，计算节点的连接关系可以包括输入关系和输出关系等等。如图 9 所示，计算节点 F1 输出作为计算节点 F4 和 F5 的输入，则可以说明计算节点 F1 和计算节点 F4 之间具有连接关系，计算节点 F1 和计算节点 F4 之间具有连接关系。再如，计算节点和计算节点 F2 之间没有数据传递，则可以说明计算节点 F1 和计算节点 F2 之间不存在连接关系。

各个计算节点的计算属性可以包括相应计算节点的计算类型及计算参数，其中计算节点的计算类型是指该计算节点用于完成何种计算，如计算节点的计算类型可以包括加法运算、减法运算及卷积运算等等，相应的，该计算节点可以是用于实现加法运算的计算节点、用于实现减法运算的计算节点或用于实现卷积运算的计算节点等等。计算节点的计算参数可以是完成该计算节点对应的计算类型所需的必要参数。例如，计算节点的计算类型可以是用于实现加法运算的计算节点，相应的，该计算节点的计算参数可以为加法运算中的加

数, 该加法运算中的被加数可以作为输入数据通过获取模块获取, 或者, 该加法运算中的被加数可以是该计算节点的上一计算节点的输出数据等等。

S020: 根据原始网络的模型数据集和模型结构参数运行原始网络, 获得原始网络中各个计算节点对应的指令。

具体地, 计算机设备的处理器可以根据原始网络的模型数据集和模型结构参数运行该原始网络, 并获得原始网络中各个计算节点对应的指令。进一步地, 处理器还可以获取该原始网络的输入数据, 根据原始网络的输入数据、网络模型数据集和模型结构参数运行原始网络, 获得该原始网络中各个计算节点对应的指令。更进一步地, 上述运行该原始网络获得各个计算节点的指令的过程实质上是编译的过程, 该编译过程可以通过计算机设备的处理器或虚拟设备实现。即计算机设备的处理器或虚拟设备根据原始网络的模型数据集和模型结构参数运行原始网络。其中, 虚拟设备指的是在存储器的内存空间中虚拟出一段处理器运行空间。

应当清楚的是, 本实施例中的运行原始网络是指, 处理器使用人工神经网络模型数据运行某种机器学习算法(如神经网络算法), 通过执行前向运算实现算法的目标应用(如语音识别等人工智能应用)。

S030: 根据原始网络的各个计算节点对应的网络权值及指令, 生成原始网络对应的离线模型, 并将原始网络对应的离线模型存储至非易失性存储器(数据库)中。

具体地, 该处理器的控制模块可以根据原始网络的各个计算节点对应的网络权值和指令, 生成该原始网络对应的离线模型, 例如, 该处理器的控制模块可以将原始网络的各个计算节点对应的网络权值和指令存储至非易失性存储器中, 以实现离线模型的生成及存储。其中, 针对原始网络的每个计算节点, 该计算节点的网络权值及指令二者之间一一对应进行存储。这样, 当再次运行该原始网络时, 可以直接从非易失性存储器中获取该原始网络对应的离线模型, 并根据与其对应的离线模型运行原始网络, 无需在线对该原始网络的各个计算节点进行编译获得指令, 提高了系统的运行速度及效率。

应当清楚的是, 本实施例中, 直接运行该原始网络对应的离线模型是指, 使用离线模型运行该原始网络对应的机器学习算法(如神经网络算法), 通过执行前向运算实现算法的目标应用(如语音识别等人工智能应用)。

可选地, 如图 8 所示, 上述步骤 S200 可以包括:

S021: 根据原始网络的模型结构参数, 获得原始网络中各个计算节点的执行顺序。

具体地, 处理器可以根据原始网络的模型结构参数, 获得原始网络中各个计算节点的执行顺序, 根据原始网络中各个计算节点的连接关系, 获得原始网络中各个计算节点的执行顺序。例如, 如图 9 所示, 计算节点 F4 的输入数据为计算节点 F1 的输出数据以及计算节点 F2 的输出数据, 计算节点 F6 的输入数据为计算节点 F4 的输出数据以及计算节点 F5 的输出数据。因此, 图 9 所示的神经网络中各个计算节点的执行顺序可以为 F1-F2-F3-F4-F5-F6 或 F1-F3-F2-F5-F4-F6 等等。当然, 计算节点 F1、F2 和 F3 可以并行执行, 计算节点 F4 和 F5 也可以并行执行, 此处仅举例说明, 并不具体限定其执行顺序。

S022: 按照原始网络中各个计算节点的执行顺序运行原始网络, 分别获得原始网络中各个计算节点对应的指令。

具体地, 处理器可以根据原始网络中各个计算节点的执行顺序运行该原始网络, 以获得原始网络中各个计算节点对应的指令, 即处理器可以将原始网络的模型数据集等数据进

行编译获得各个计算节点对应的指令，通过各个计算节点对应的指令可以获知该计算节点用于实现何种计算功能，即可以获得该计算节点的计算类型及计算参数等计算属性。

进一步地，如图 8 所示，上述步骤 S300 还包括：

S031：根据原始网络的模型数据集和模型结构参数，获得原始网络的内存分配方式。

具体地，处理器可以根据原始网络的模型数据集和模型结构参数，获得原始网络的内存分配方式；根据原始网络的模型结构参数，获得原始网络中各个计算节点的执行顺序；并根据原始网络中各个计算节点的执行顺序确定当前网络的内存分配方式。例如，按各个计算节点的执行顺序将各个计算节点在运行过程中的相关数据保存至一个栈内。其中，内存分配方式是指确定原始网络中各个计算节点相关的数据（包括输入数据、输出数据、网络权值数据及中间结果数据等等）在内存空间（如第一存储器）上的存储位置。例如，可以采用数据表存储各个计算节点相关的数据（输入数据、输出数据、网络权值数据及中间结果数据等等）和内存空间的映射关系。

S032：根据原始网络的内存分配方式，将原始网络运行过程中的相关数据存储至第一存储器中，其中，原始网络运行过程中的相关数据包括原始网络的各个计算节点对应的网络权值、指令、输入数据、中间计算结果及输出数据等等。例如，如图 9 所示，X1 和 X2 表示该神经网络的输入数据，Y 表示该神经网络的输出数据，处理器可以将该神经网络的输出数据转换为控制机器人或不同数字接口的控制命令。W1~W6 用于表示计算节点 F1、F2 和 F3 对应的网络权值，计算节点 F1~F5 的输出数据可以作为中间计算结果。处理器可以根据已确定的内存分配方式，将原始网络运行过程中的相关数据存储至第一存储器，如内存储器或缓存等易失性存储器，其具体的存储方式可参见图 10 中左半部分存储空间。

S033：从第一存储器中获取原始网络的各个计算节点对应的网络权值及指令，并将原始网络的各个计算节点对应的网络权值及指令存储于第二存储器中，生成离线模型。其中，第二存储器可以为外部存储器等非易失性存储器。该离线模型的生成过程具体可参见图 10 所示，图 10 中右半部分的存储空间内存储的即为原始网络的对应的离线模型。

如图 9 和图 10 所示，下面结合附图说明上述的离线模型生成过程：

首先，处理器可以获得该原始网络的模型数据集、模型结构参数以及输入数据，从而根据该原始网络的模型数据集和模型结构参数可以获得该原始网络的网络结构图，如图 9 所示。

其次，处理器可以根据原始网络的模型结构参数，获得原始网络各个计算节点的连接关系，并根据各个计算节点的连接关系获得原始网络中各个计算节点的执行顺序，以及原始网络在运行过程中的内存分配方式，从而可以获得原始网络在运行过程中相关数据的存储位置。如图 10 中左半部分存储空间所示，原始网络在运行过程中的相关数据可以按照各个计算节点执行顺序存储在一个栈中。

最后，处理器可以将原始网络的各个计算节点对应的网络权值及指令存储于非易失性的第二存储器中，生成离线模型，该离线模型的存储方式可参见图 8 中右半部分存储空间所示。并且，该离线模型仅仅包含运行该原始网络所必需的网络权值及指令等数据，而不需对原始网络运行过程中的输入数据、输出数据或中间计算结果等进行存储，从而可以减小第二存储器中的存储空间的消耗。

作为进一步地改进，离线模型中还包括节点接口数据，节点接口数据用于表示原始网络的各个计算节点的连接关系。具体地，节点接口数据可以包括各个计算节点的输入数据

来源和输出数据来源。例如，如图 9 所示，节点接口数据可以包括计算节点 F1、F2 和 F3 为起始计算节点，其输入分别为预设的输入数据，计算节点 F1 的输出数据作为计算节点 F4 和计算节点 F5 的输入数据等等。这样，在再次运行该原始网络时，只需获得原始网络的起始计算节点和输入数据，之后，便可以根据该原始网络对应的离线模型执行该原始网络。

应该理解的是，虽然图 5-8 的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图 5-8 中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一非易失性计算机可读存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程 ROM(PROM)、电可编程 ROM(EPROM)、电可擦除可编程 ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM 以多种形式可得，诸如静态 RAM(SRAM)、动态 RAM(DRAM)、同步 DRAM(SDRAM)、双数据率 SDRAM(DDRSDRAM)、增强型 SDRAM(ESDRAM)、同步链路(Synchlink) DRAM(SLDRAM)、存储器总线(Rambus)直接 RAM(RDRAM)、直接存储器总线动态 RAM(DRDRAM)、以及存储器总线动态 RAM(RDRAM)等。

本申请实施例还提供了一种模型转换装置，上述模型转换装置包括获取模块、判断模块和转换模块。其中，

获取模块用于获取初始离线模型及计算机设备的硬件属性信息；

判断模块用于根据初始离线模型和计算机设备的硬件属性信息，判断初始离线模型的模型属性信息与计算机设备的硬件属性信息是否匹配；

转换模块用于在初始离线模型的模型属性信息与计算机设备的硬件属性信息不匹配时，根据计算机设备的硬件属性信息和预设的模型转换规则，将初始离线模型转换为与计算机设备的硬件属性信息匹配的目标离线模型。

可选地，该模型转换装置可以是安装于该计算机设备上的应用软件(Application)。该应用软件(Application)能够提供从存储器 200 或外部存储器中读取初始离线模型的接口，从而可以通过该应用软件获取该初始离线模型。可选地，该应用软件还可以提供读取计算机设备的硬件属性信息的接口，从而可以通过该应用软件获取该计算机设备的硬件属性信息。可选地，该应用软件还可以提供能够读取预设的模型规则的接口，从而可以通过该应用软件获取预设的模型转换规则。进一步地，该应用软件可以根据初始离线模型和计算机设备的硬件属性信息以及预设的模型转换规则，在该初始离线模型的模型属性信息与计算机设备的硬件属性信息不匹配时，将该初始离线模型转换为与该计算机设备的硬件属性信息匹配的目标离线模型，并将该目标离线模型存储至第一存储器或第二存储器中。

关于模型转换装置的具体限定可以参见上文中对于模型转换方法的限定,在此不再赘述。上述模型转换装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中,也可以以软件形式存储于计算机设备中的存储器中,以便于处理器调用执行以上各个模块对应的操作。

此外,本申请实施例还提供了一种计算机可读存储介质,其上存储有计算机程序,上述计算机程序被处理器执行时实现权利要求上述任一实施例的方法的步骤。可选地,该计算机可读存储介质可以包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限,RAM以多种形式可得,诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

具体地,上述计算机程序被处理器执行时实现如下步骤:

S100: 获取初始离线模型及计算机设备的硬件属性信息;

具体地,上述的初始离线模型是指计算机设备直接获取的离线模型,该初始离线模型可以存储于第二存储器300中。其中,离线模型可以包含一原始网络中各个计算节点的网络权值以及指令等必要的网络结构信息,其中,指令可以用于表明该计算节点用于执行何种计算功能,其具体可以包括该原始网络中各个计算节点的计算属性以及各个计算节点之间的连接关系等信息。可选地,本申请实施例中的初始离线模型可以是根据原始网络直接生成的离线模型,也可以是具有其他模型属性信息的离线模型经过一次或多次转换后获得离线模型,此处不做具体限定。

S200: 根据初始离线模型和计算机设备的硬件属性信息,判断初始离线模型的模型属性信息与计算机设备的硬件属性信息是否匹配。

具体地,该初始离线模型的模型属性信息可以包括该初始离线模型的数据结构及数据类型等等。例如,初始离线模型中各个网络权值及指令的摆放方式、各个网络权值的类型及各个指令的类型等等。上述的计算机设备的硬件属性信息包括该计算机设备的型号、该计算机设备能够处理的数据类型(如定点或浮点等等)及数据结构等等。

例如,该计算机设备可以根据其硬件属性信息确定该计算机设备能够处理的数据类型及数据结构,根据该初始离线模型的模型属性信息确定该初始离线模型的数据类型及数据结构。若该计算机设备根据其自身的硬件属性信息以及上述初始离线模型的模型属性信息,确定该计算机设备能够支持上述初始离线模型的数据类型及数据结构时,则可以判定该初始离线模型的模型属性信息与该计算机设备的硬件属性信息匹配。若该计算机设备根据其自身的硬件属性信息以及上述的初始离线模型的模型属性信息,确定该计算机设备不能够支持上述初始离线模型的数据类型及数据结构时,则可以判定该初始离线模型的模型属性信息与该计算机设备的硬件属性信息不匹配。

若初始离线模型的模型属性信息与计算机设备的硬件属性信息不匹配,则执行步骤S300: 根据计算机设备的硬件属性信息和预设的模型转换规则,将初始离线模型转换为与计算机设备的硬件属性信息匹配的目标离线模型。

具体地，若该初始离线模型的模型属性信息与该计算机设备的硬件属性信息不匹配，则说明该计算机设备不能支持该初始离线模型的运行，此时需对该初始离线模型进行转换处理，以便该计算机设备能够使用该初始离线模型实现相应的运算。即当该初始离线模型的模型属性信息与计算机设备的硬件属性信息不匹配时，则可以根据计算机设备的硬件属性信息和预设的模型转换规则，将上述的初始离线模型转换为与该计算机设备的硬件属性信息匹配的目标离线模型。其中，该目标离线模型的模型属性信息与该计算机设备的硬件属性信息匹配，该计算机设备能够支持该目标离线模型的运行。同上，该目标离线模型也包括一原始网络中各个计算节点的网络权值以及指令等必要的网络结构信息。

若初始离线模型的模型属性信息与计算机设备的硬件属性信息匹配，则可以执行步骤S400：计算机设备可以直接运行其接收到的初始离线模型，即计算机设备可以根据该初始离线模型包含的网络权值及指令进行运算，以实现原始网络的运算功能。具体地，若该初始离线模型的模型属性信息与该计算机设备的硬件属性信息匹配，则说明该计算机设备能够支持该初始离线模型的运行，此时无需对该初始离线模型进行转换处理，计算机设备能够运行该初始离线模型。

应当清楚的是，处理器执行上述计算机程序的过程，与上述实施例的模型转换方法的执行过程一致，具体可参见上文中的描述，此处不再赘述。

可以理解的是，上述各实施例中相同或相似部分可以相互参考，在一些实施例中未详细说明的内容可以参见其他实施例中相同或相似的内容。

需要说明的是，在本申请的描述中，术语“第一”、“第二”等仅用于描述目的，而不能理解为指示或暗示相对重要性。此外，在本申请的描述中，除非另有说明，“多个”的含义是指至少两个。

流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现特定逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本申请的可选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本申请的实施例所属技术领域的技术人员所理解。

应当理解，本申请的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或它们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本申请各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用，也可以存储在一个计算机可读存储介质中。

上述提到的存储介质可以是只读存储器，磁盘或光盘等。

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一定指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何的一个或多个实施例或示例中以合适的方式结合。

尽管上面已经示出和描述了本申请的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本申请的限制，本领域的普通技术人员在本申请的范围内可以对上述实施例进行变化、修改、替换和变型。

权利要求

1、一种模型转换方法，其特征在于，所述方法包括如下步骤：

获取初始离线模型及计算机设备的硬件属性信息；

根据所述初始离线模型和所述计算机设备的硬件属性信息，判断所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息是否匹配；

若所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息不匹配，则根据所述计算机设备的硬件属性信息和预设的模型转换规则，将所述初始离线模型转换为与所述计算机设备的硬件属性信息匹配的目标离线模型。

2、根据权利要求1所述的模型转换方法，其特征在于，所述的根据所述计算机设备的硬件属性信息和预设的模型转换规则，将所述初始离线模型转换为与所述计算机设备的硬件属性信息匹配的目标离线模型的步骤，包括：

根据所述计算机设备的硬件属性信息确定所述目标离线模型的模型属性信息；

根据所述初始离线模型的模型属性信息和所述目标离线模型的模型属性信息，从多个所述预设的模型转换规则中选择一个所述模型转换规则作为目标模型转换规则；

根据所述初始离线模型的模型属性信息和所述目标模型转换规则，将所述初始离线模型转换为所述目标离线模型。

3、根据权利要求2所述的模型转换方法，其特征在于，所述的根据所述初始离线模型的模型属性信息和所述目标离线模型的模型属性信息，从多个所述预设的模型转换规则中选择一个所述模型转换规则作为目标模型转换规则的步骤，包括：

根据所述初始离线模型的模型属性信息和所述目标离线模型的模型属性信息，从多个所述预设的模型转换规则中选择出一个以上的可用模型转换规则；

对所述一个以上的可用模型转换规则进行优先级排序，并将优先级最高的可用模型转换规则作为所述目标转换规则。

4、根据权利要求3所述的模型转换方法，其特征在于，所述的对一个以上的可用模型转换规则进行优先级排序的步骤，包括如下步骤：

分别获取采用各个所述可用模型转换规则，将所述初始离线模型转换为所述目标离线模型的过程参数，其中，所述过程参数包括转换速度、功耗、内存占用率及磁盘 I/O 占用率中的一种或多种；

根据各个所述可用模型转换规则的过程参数，对一个以上的所述可用模型转换规则进行优先级排序。

5、根据权利要求3所述的模型转换方法，其特征在于，所述初始离线模型的模型属性信息、所述目标离线模型的模型属性信息及所述可用模型转换规则三者之间一一对应存储。

6、根据权利要求1-5任一项所述的模型转换方法，其特征在于，所述的根据所述初始离线模型和所述计算机设备的硬件属性信息，判断所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息是否匹配的步骤包括：

若根据所述初始离线模型的模型属性信息和所述计算机设备的硬件属性信息，确定所述计算机设备能够支持所述初始离线模型的运行时，则判定所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息匹配；

若根据所述初始离线模型的模型属性信息和所述计算机设备的硬件属性信息，确定所

述计算机设备不支持所述初始离线模型的运行时，则判定所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息不匹配。

7、根据权利要求 1-5 任一项所述的模型转换方法，其特征在于，所述的获取初始离线模型的步骤，包括：

通过所述计算机设备上的应用软件获取所述初始离线模型和所述计算机设备的硬件属性信息。

8、根据权利要求 1-5 任一项所述的模型转换方法，其特征在于，所述方法还包括如下步骤：

将所述目标离线模型存储于所述计算机设备的第一存储器或第二存储器。

9、根据权利要求 8 所述的模型转换方法，其特征在于，所述方法还包括如下步骤：

获取所述目标离线模型，并运行所述目标离线模型，其中，所述目标离线模型中包含原始网络中各个离线节点对应的网络权值、指令以及各个离线节点与所述原始网络中的其他计算节点之间的接口数据。

10、一种模型转换装置，其特征在于，所述装置包括：

获取模块，用于获取初始离线模型及计算机设备的硬件属性信息；

判断模块，用于根据所述初始离线模型和所述计算机设备的硬件属性信息，判断所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息是否匹配；

转换模块，用于在所述初始离线模型的模型属性信息与所述计算机设备的硬件属性信息不匹配时，根据所述计算机设备的硬件属性信息和预设的模型转换规则，将所述初始离线模型转换为与所述计算机设备的硬件属性信息匹配的目标离线模型。

11、一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求 1 至 7 中任一项所述方法的步骤。

12、根据权利要求 11 所述的计算机设备，其特征在于，所述处理器包括运算单元和控制器单元，所述运算单元包括主处理电路和多个从处理电路；

所述控制器单元用于获取数据、机器学习模型以及计算指令；

所述控制器单元还用于解析所述计算指令得到多个运算指令，并将所述多个运算指令以及所述数据发送给所述主处理电路；

所述主处理电路用于对所述数据以及在所述主处理电路与所述多个从处理电路之间进行传输的数据和运算指令执行前序处理；

所述多个从处理电路用于依据从所述主处理电路传输的数据以及运算指令并行执行中间运算得到多个中间结果，并将多个中间结果传输给所述主处理电路；

所述主处理电路还用于对所述多个中间结果执行后续处理得到所述计算指令的计算结果。

13、一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求 1 至 7 中任一项所述的方法的步骤。

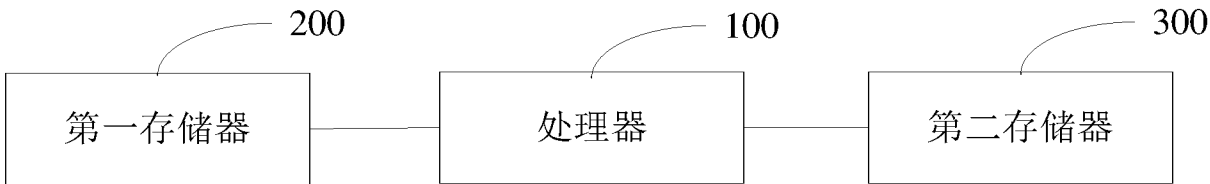


图 1

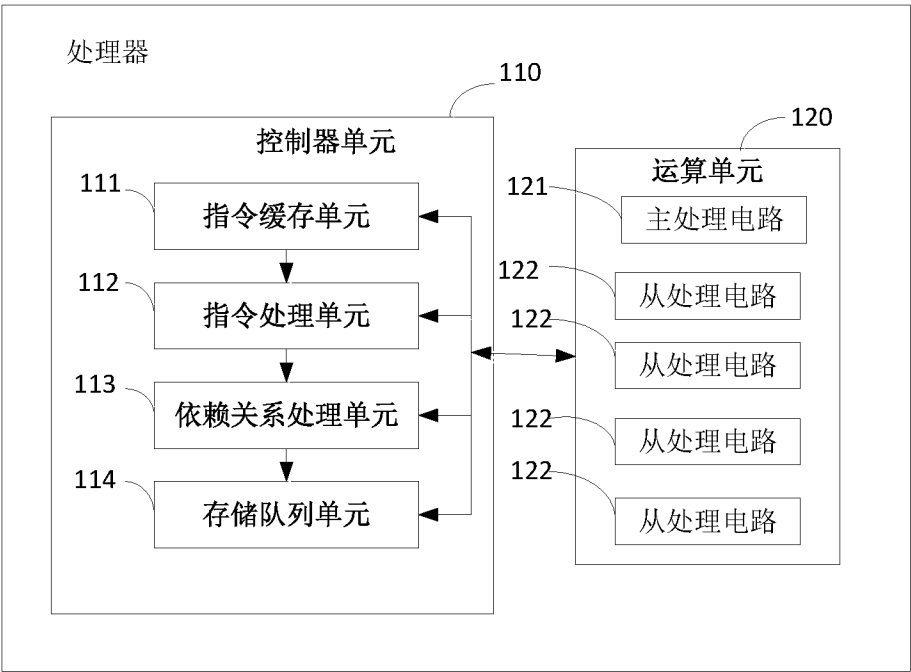


图 2

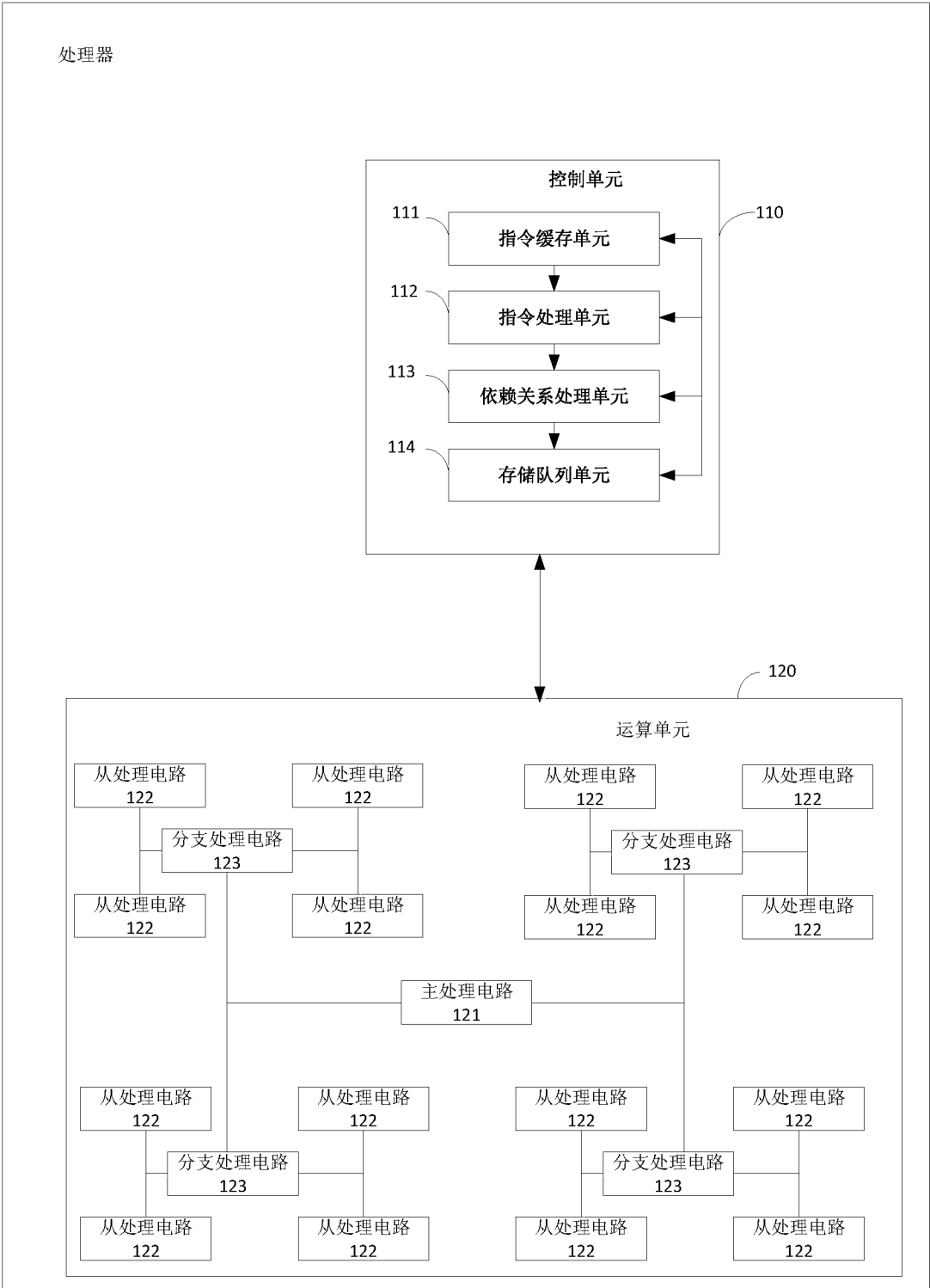


图 3

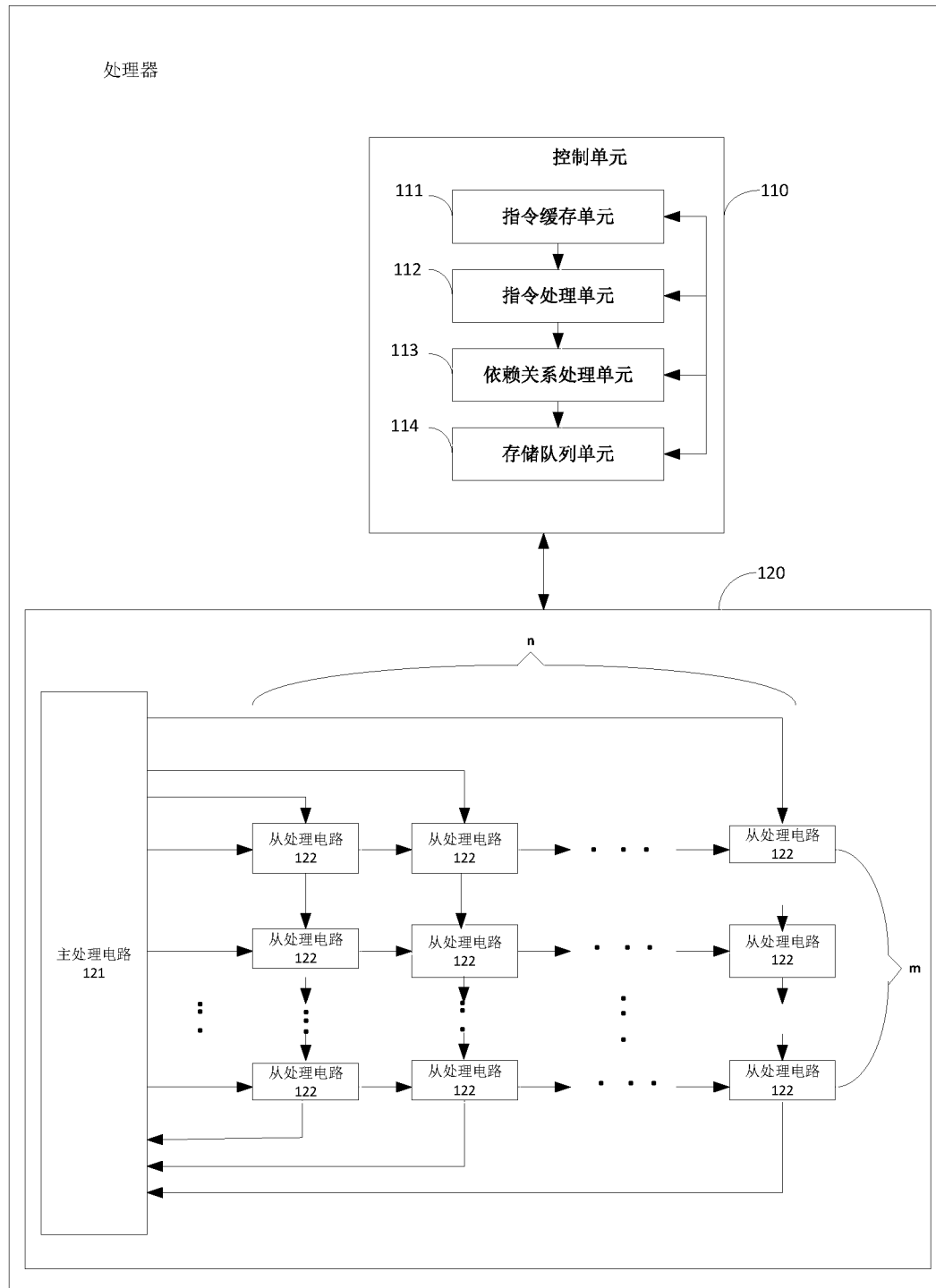


图 4

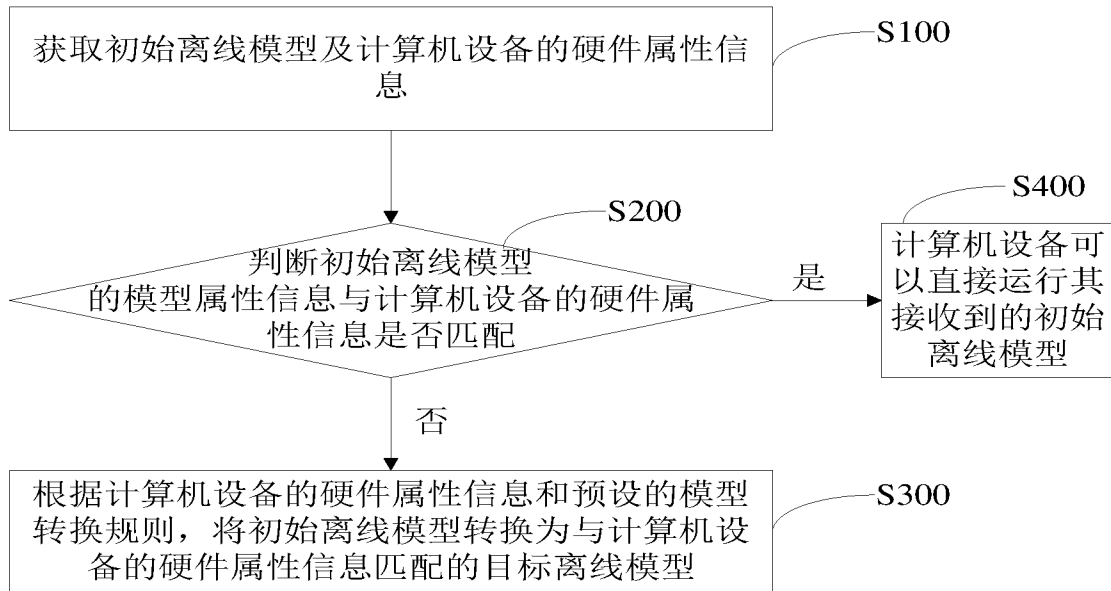


图 5

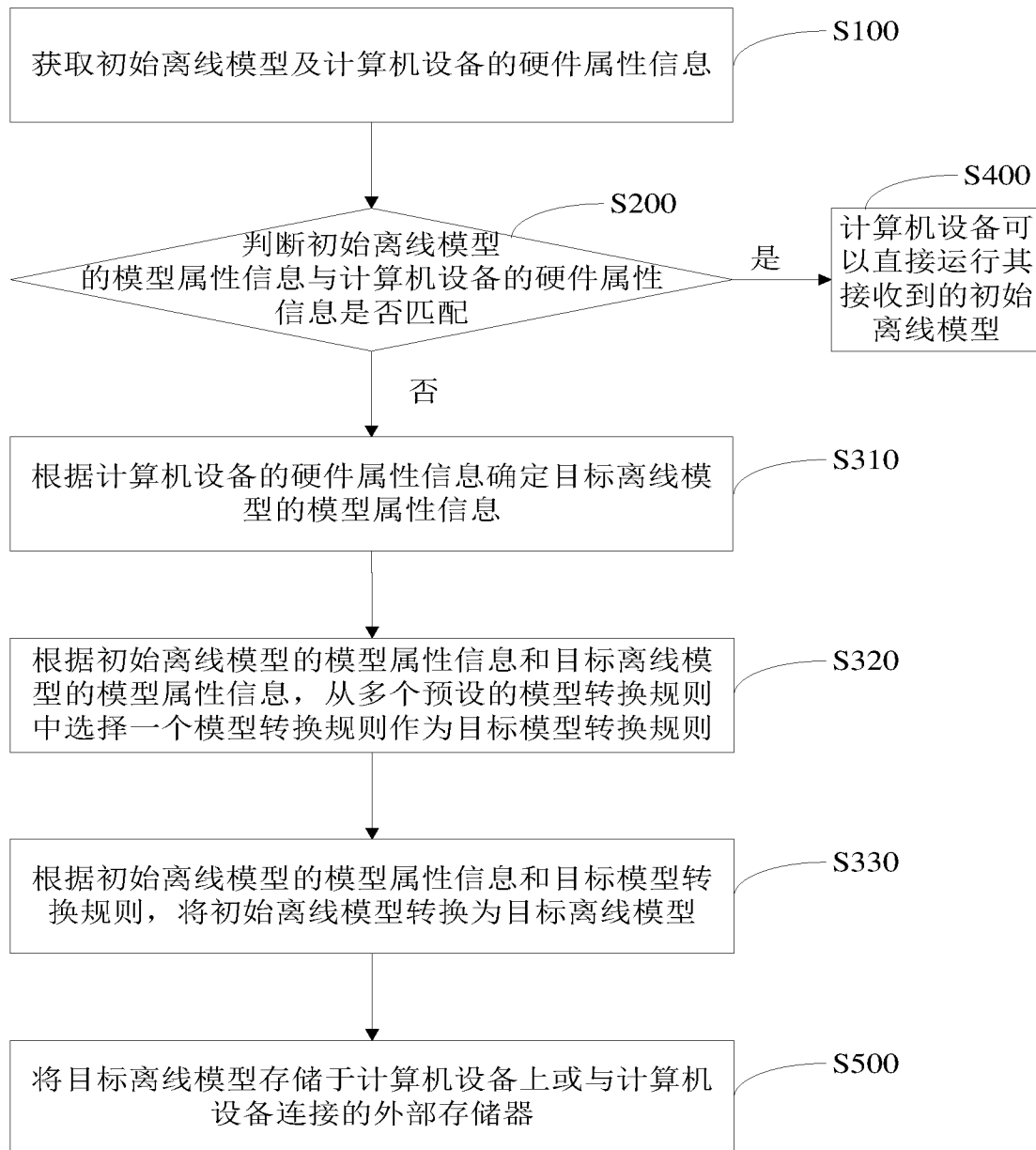


图 6

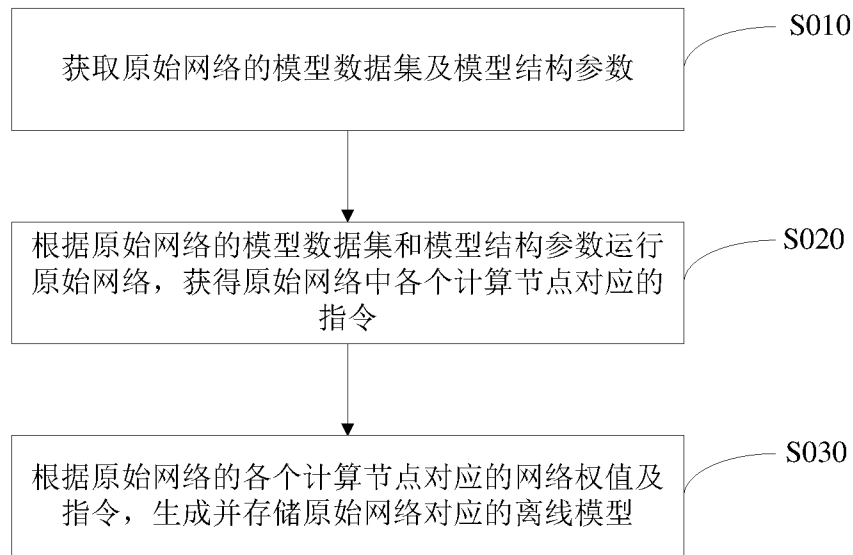


图 7

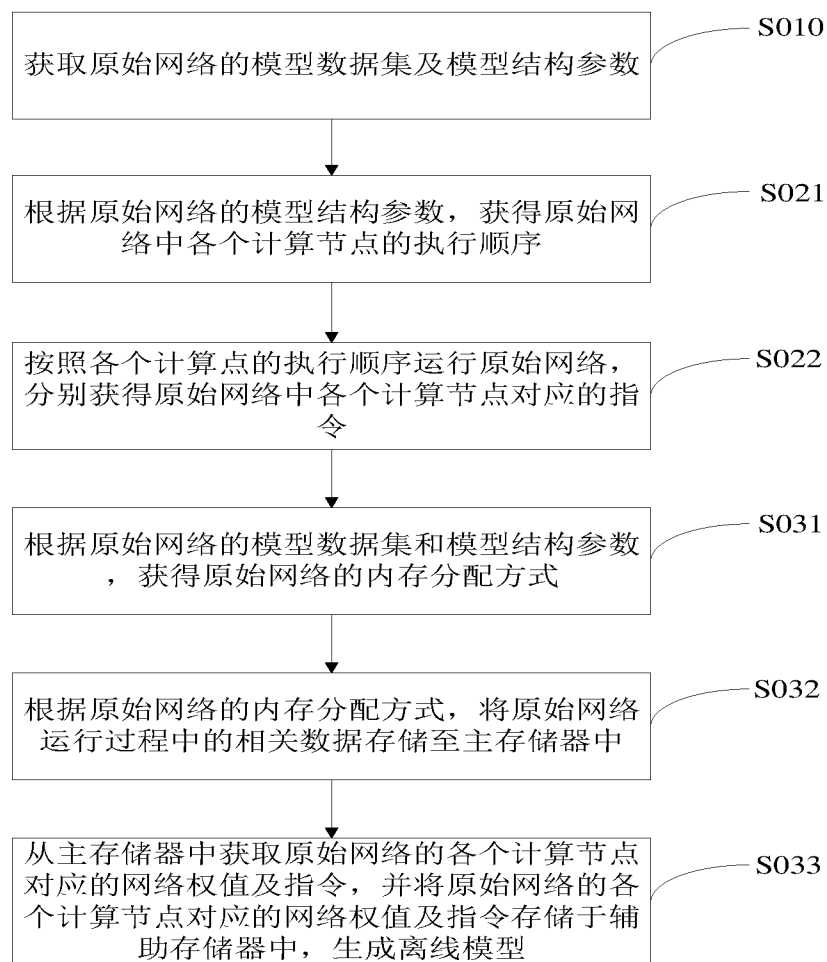


图 8

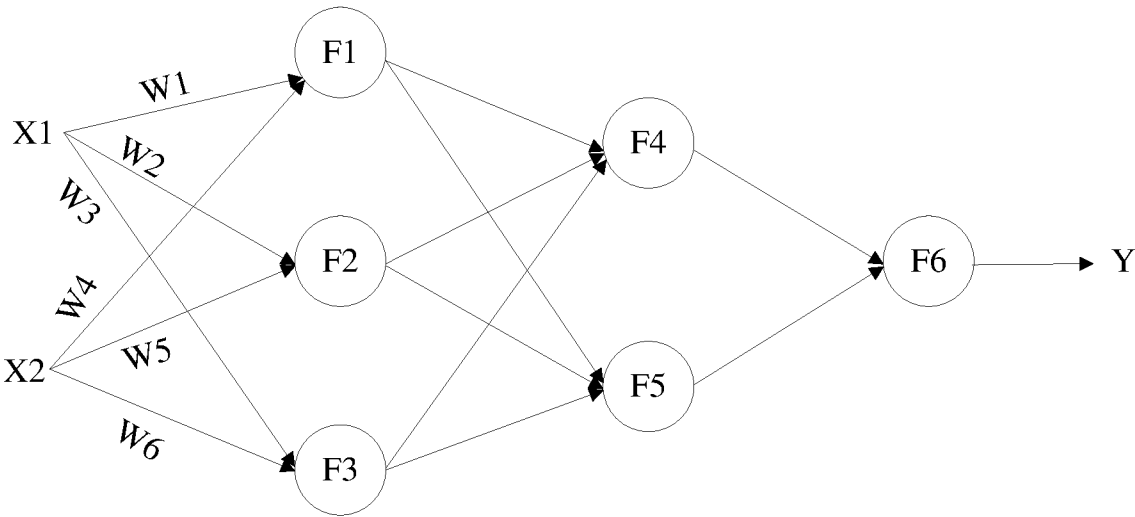


图 9

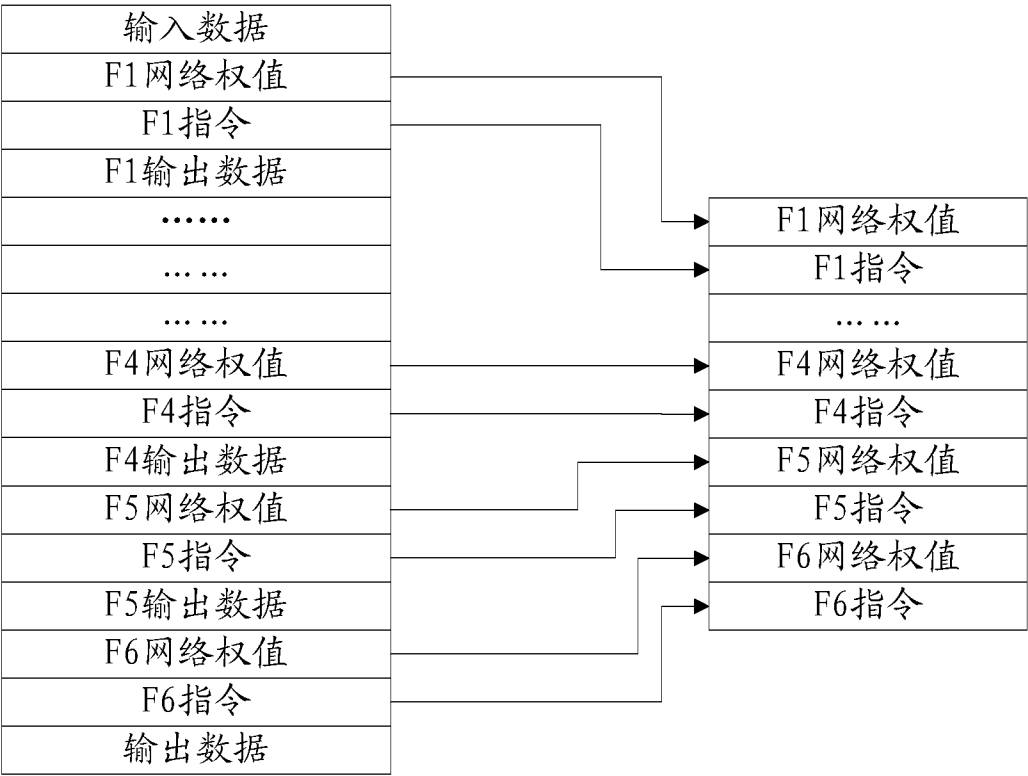


图 10

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/080510

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/50(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)

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)

CNPAT, WPI, EPODOC, CNKI, GOOGLE: 模型, 转换, 离线, 硬件, 属性, 匹配, 规则, offline, model, conversion, hardware, attribute, match, rule

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 109492241 A (CAMBRICON TECHNOLOGIES CO., LTD.) 19 March 2019 (2019-03-19) claims 1-13	1-13
X	CN 101430647 A (BEIJING VIMICRO CORPORATION) 13 May 2009 (2009-05-13) description, p. 2, paragraph 2 to p. 5, paragraph 9	1-13
A	CN 101937351 A (SHENZHEN CITY RENXING NETWORK TECHNOLOGY CO., LTD.) 05 January 2011 (2011-01-05) entire document	1-13
A	CN 106331858 A (BEIJING QIHOO TECHNOLOGY CO., LTD. ET AL.) 11 January 2017 (2017-01-11) entire document	1-13

☐ 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

18 June 2019

Date of mailing of the international search report

08 July 2019

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/CN2019/080510

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	109492241	A	19 March 2019	None	
CN	101430647	A	13 May 2009	None	
CN	101937351	A	05 January 2011	None	
CN	106331858	A	11 January 2017	None	

国际检索报告

国际申请号

PCT/CN2019/080510

A. 主题的分类 G06F 17/50 (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) 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)) CNPAT, WPI, EPDOC, CNKI, GOOGLE: 模型, 转换, 离线, 硬件, 属性, 匹配, 规则, offline, model, conversion, hardware, attribute, match, rule																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 109492241 A (北京中科寒武纪科技有限公司) 2019年 3月 19日 (2019 - 03 - 19) 权利要求1-13</td> <td>1-13</td> </tr> <tr> <td>X</td> <td>CN 101430647 A (北京中星微电子有限公司) 2009年 5月 13日 (2009 - 05 - 13) 说明书第2页第2段-第5页第9段</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>CN 101937351 A (深圳市任子行网络技术股份有限公司) 2011年 1月 5日 (2011 - 01 - 05) 全文</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>CN 106331858 A (北京奇虎科技有限公司 等) 2017年 1月 11日 (2017 - 01 - 11) 全文</td> <td>1-13</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 109492241 A (北京中科寒武纪科技有限公司) 2019年 3月 19日 (2019 - 03 - 19) 权利要求1-13	1-13	X	CN 101430647 A (北京中星微电子有限公司) 2009年 5月 13日 (2009 - 05 - 13) 说明书第2页第2段-第5页第9段	1-13	A	CN 101937351 A (深圳市任子行网络技术股份有限公司) 2011年 1月 5日 (2011 - 01 - 05) 全文	1-13	A	CN 106331858 A (北京奇虎科技有限公司 等) 2017年 1月 11日 (2017 - 01 - 11) 全文	1-13
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
PX	CN 109492241 A (北京中科寒武纪科技有限公司) 2019年 3月 19日 (2019 - 03 - 19) 权利要求1-13	1-13															
X	CN 101430647 A (北京中星微电子有限公司) 2009年 5月 13日 (2009 - 05 - 13) 说明书第2页第2段-第5页第9段	1-13															
A	CN 101937351 A (深圳市任子行网络技术股份有限公司) 2011年 1月 5日 (2011 - 01 - 05) 全文	1-13															
A	CN 106331858 A (北京奇虎科技有限公司 等) 2017年 1月 11日 (2017 - 01 - 11) 全文	1-13															
☐ 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 2019年 6月 18日		国际检索报告邮寄日期 2019年 7月 8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		授权官员 赵天奇 电话号码 86-(10)-53961320															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/080510

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	109492241	A	2019年 3月 19日	无	
CN	101430647	A	2009年 5月 13日	无	
CN	101937351	A	2011年 1月 5日	无	
CN	106331858	A	2017年 1月 11日	无	