

(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
2024 年 12 月 26 日 (26.12.2024)



(10) 国际公布号
WO 2024/259853 A1

(51) 国际专利分类号:
G06N 3/082 (2023.01) **G06F 18/241** (2023.01)
G06F 18/214 (2023.01) **G06F 18/25** (2023.01)
G06F 18/213 (2023.01)

(21) 国际申请号: PCT/CN2023/128556

(22) 国际申请日: 2023 年 10 月 31 日 (31.10.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202310731911.0 2023年6月19日 (19.06.2023) CN

(71) 申请人: 腾讯科技(深圳)有限公司 (TENCENT TECHNOLOGY (SHENZHEN) COMPANY LIMITED) [CN/CN]; 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。

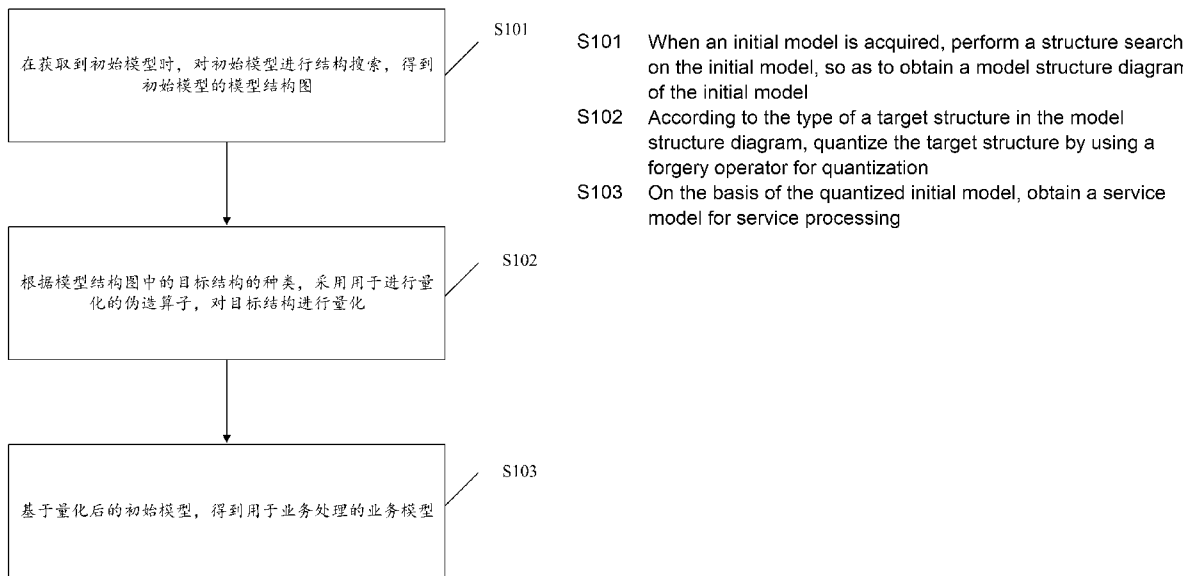
(72) 发明人: 程雅慧 (CHENG, Yahui); 中国广东省深圳市南山区高新区科技中一路腾讯大厦 35 层, Guangdong 518057 (CN)。

(74) 代理人: 北京励诚知识产权代理有限公司 (BEIJING LISENG INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市海淀区阜成路 73 号裕惠大厦 B 座 807, Beijing 100142 (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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD,

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

(54) 发明名称: 一种数据处理方法、装置、计算机设备及存储介质



(57) Abstract: Disclosed in the embodiments of the present application are a data processing method and apparatus, and a computer device and a storage medium, which can be applied to an artificial intelligence scenario. The method comprise: when an initial model is acquired, performing a structure search on the initial model, so as to obtain a model structure diagram of the initial model; according to the type of a target structure in the model structure diagram, quantizing the target structure by using a forgery operator for quantization, wherein the forgery operator comprises a quantization operator and a dequantization operator; and on the basis of the quantized initial model, obtaining a service model for service processing. By using the embodiments of the present application, the speed of model quantization can be increased, and universality of model quantization is achieved.



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, CV, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SC, 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, ME, 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: 本申请实施例公开了一种数据处理方法、装置、计算机设备及存储介质，可应用于人工智能场景，包括：在获取到初始模型时，对初始模型进行结构搜索，得到初始模型的模型结构图；根据模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对目标结构进行量化；伪造算子包括量化算子和反量化算子；基于量化后的初始模型，得到用于业务处理的业务模型。采用本申请实施例，可以提升模型量化的速度，实现模型量化的通用性。

一种数据处理方法、装置、计算机设备及存储介质

本申请要求于 2023 年 06 月 19 日提交中国专利局、申请号为 202310731911.0、发明名称为“一种数据处理的方法、装置、电子设备及存储介质”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。

技术领域

本申请涉及计算机技术领域，尤其涉及一种数据处理方法、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不断发展，人工智能模型（即 AI 模型）的结构也被设计的越来越复杂，AI 模型的权重参数也越来越庞大，那么在进行前向推理时对硬件的要求也越来越高。而在一些对实时性要求很高，同时算力资源有限的终端设备上，要想很好的部署 AI 模型，就需要对 AI 模型进行一些“压缩”处理，而量化就是一种模型压缩的方式。

传统量化方法往往需要人工手动去修改 AI 模型，以插入相应的伪造算子（即包括量化算子和反量化算子），进而导致代码侵入性很高，不利于代码的复用，即针对不同的模型结构需要设计不同的量化方案，不仅影响量化速度，还导致模型量化不具备通用性。

发明内容

本申请实施例提供一种数据处理方法、装置、计算机设备及存储介质，可以提升模型量化的速度，实现模型量化的通用性。

本申请实施例一方面提供一种数据处理方法，包括：

在获取到初始模型时，对初始模型进行结构搜索，得到初始模型的模型结构图；

根据模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对目标结构进行量化，伪造算子包括量化算子和反量化算子；

基于量化后的初始模型，得到用于业务处理的业务模型。

本申请实施例一方面提供一种数据处理装置，包括：

搜索模块，用于在获取到初始模型时，对初始模型进行结构搜索，得到初始模型的模型结构图；

量化模块，用于根据模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对目标结构进行量化，伪造算子包括量化算子和反量化算子；

业务模型确定模块，用于基于量化后的初始模型，得到用于业务处理的业务模型。

其中，该量化模块包括：

第一量化单元，用于在模型结构图中的目标结构包括具有权重的第一网络层的情况下，在具有权重的第一网络层的输入位置处，添加该伪造算子，且采用伪造算子对具有权重的第一网络层的权重进行量化。

其中，该量化模块还包括：

第二量化单元，用于在具有权重的第一网络层具有多个输出位置的情况下，在具有权重的第一网络层的多个输出位置处，分别添加该伪造算子。

其中，该量化模块还包括：

10 第三量化单元，用于在具有权重的第一网络层具有一个输出位置，且具有权重的第一网络层的输出位置处未连接激活层的情况下，在具有权重的第一网络层的输出位置处，添加该伪造算子。

其中，在具有权重的第一网络层具有一个输出位置，且具有权重的第一网络层的输出位置处连接激活层的情况下，不在具有权重的第一网络层的输出位置处，添加该伪造算子。

15 其中，该量化模块还包括：

第四量化单元，用于在模型结构图中的目标结构包括叠加层、叠加层的第一输入分支和叠加层的第二输入分支的情况下，在叠加层的输出位置处添加伪造算子；

其中，第一输入分支包括依次连接的具有权重的第一网络层和归一化层，归一化层与叠加层连接，第二输入分支的类型为非加权操作的输入类型，具有权重的第一网络层为卷积层。

20 其中，该量化模块还包括：

第五量化单元，用于在第二输入分支的输入位置处添加伪造算子。

其中，该量化模块还包括：

25 第六量化单元，用于在模型结构图中的目标结构包括第一叠加层，且第一叠加层的第一输入分支的类型与第一叠加层的第二输入分支的类型均为非加权操作的输入类型的情况下，在第一输入分支的输入位置处添加用于进行量化的伪造算子，在第二输入分支的输入位置处添加伪造算子，且在第一叠加层的输出位置处添加伪造算子。

其中，该量化模块还包括：

30 第七量化单元，用于在模型结构图中的目标结构包括第一拼接层，且第一拼接层的输入分支与第一拼接层的输出分支的数量相同的情况下，在第一拼接层的N个输出分支的每

个输出分支的输出位置处，分别添加用于进行量化的伪造算子；N 为大于 1 的正整数；每个输出分支处所添加的伪造算子的量化算子均相同。

其中，该业务模型确定模块包括：

第一训练单元，用于将量化后的初始模型确定为第一模型，对第一模型进行微调训练，

5 得到第二模型；

第一确定单元，用于基于第二模型，得到用于业务处理的业务模型；业务模型包括具有权重的第二网络层、和添加在具有权重的第二网络层的输入位置处的量化算子；在业务模型中，具有权重的第二网络层是根据具有权重的第一网络层得到的，且具有权重的第二网络层的权重是采用量化算子，对第二模型中的具有权重的第一网络层的权重进行处理后
10 所得到的。

其中，目标结构中的具有权重的第一网络层为卷积层，且目标结构还包括与卷积层连接的归一化层，以及与归一化层连接的激活层；

该业务模型确定模块还包括：

第一融合单元，用于对具有权重的第一网络层、归一化层和激活层进行融合，得到具
15 有权重的第二网络层。

其中，目标结构包括叠加层、叠加层的第一输入分支和叠加层的第二输入分支；第一输入分支包括依次连接的具有权重的第一网络层和归一化层；归一化层与叠加层连接，第二输入分支的类型为非加权操作的输入类型，具有权重的第一网络层为卷积层；

该业务模型确定模块还包括：

20 第二融合单元，用于对具有权重的第一网络层、归一化层和叠加层进行融合，得到具有权重的第二网络层。

其中，该业务模型确定模块还包括：

第二训练单元，用于将量化后的初始模型确定为第一模型，对第一模型进行微调训练，得到第二模型；

25 第二确定单元，用于基于第二模型，得到用于业务处理的业务模型；业务模型包括第二叠加层、添加在第二叠加层的第一输入分支处的量化算子、和添加在第二叠加层的第二输入分支处的量化算子；在业务模型中，第二叠加层是根据量化算子以及第二模型中的第一叠加层所得到的。

其中，该业务模型确定模块还包括：

30 第三训练单元，用于将量化后的初始模型确定为第一模型，对第一模型进行微调训练，

得到第二模型；

第三确定单元，用于基于第二模型，得到用于业务处理的业务模型；业务模型包括第二拼接层、和添加在第二拼接层的每个输入分支的输入位置处的量化算子；在业务模型中，第二拼接层的每个输入分支的输入位置处的量化算子是将添加在第二模型中的第一拼接层的每个输出分支的输出位置处的量化算子，分别等价向前传播至第二模型中的第一拼接层的每个输入分支的输入位置处后所得到的。

其中，该装置还包括：

特征获取模块，用于若业务模型的业务处理为图像识别，则在获取到待识别图像时，获取待识别图像的图像特征；

10 特征输入模块，用于将图像特征输入至业务模型；业务模型中的优化结构包括目标网络层、和添加在目标网络层的输入位置处的量化算子；目标网络层是根据量化算子，以及初始模型中的目标结构的原网络层所得到的；

精度转换模块，用于基于添加在目标网络层的输入位置处的量化算子，将优化结构的输入特征的精度由第一精度转换为第二精度；第一精度高于第二精度；

15 特征提取模块，用于通过目标网络层，对属于第二精度的输入特征进行特征提取处理，由目标网络层输出输入特征对应的输出特征；

类别确定模块，用于基于输出特征，得到待识别图像的预测类别。

本申请一方面提供了一种计算机设备，包括：处理器、存储器、网络接口；

20 处理器与存储器、网络接口相连，其中，网络接口用于提供数据通信功能，存储器用于存储计算机程序，处理器用于调用计算机程序，以使得计算机设备执行本申请实施例提供的方法。

本申请实施例一方面提供了一种计算机可读存储介质，计算机可读存储介质存储有计算机程序，该计算机程序适于由处理器加载并执行，以使得具有该处理器的计算机设备执行本申请实施例提供的方法。

25 本申请实施例一方面提供了一种计算机程序产品，该计算机程序产品包括计算机程序，该计算机程序存储在计算机可读存储介质中；计算机设备的处理器从计算机可读存储介质读取该计算机程序，处理器执行该计算机程序，使得该计算机设备执行本申请实施例中的方法。

30 在本申请实施例中，具有模型量化功能的计算机设备在获取到初始模型时，无需人工手动向初始模型添加伪造算子，而是在对初始模型的模型结构图进行查找后，直接根据初

始模型的模型结构图中的目标结构的种类，自动采用用于进行量化的伪造算子（这里的伪造算子可以包括量化算子和反量化算子），对目标结构进行量化，进而在量化结束后，可以基于量化后的初始模型，确定用于业务处理的业务模型。换言之，本申请实施例提供的数据处理方法来进行模型量化无需人工参与，对于任意一种初始模型而言，均可以根据模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，自动对初始模型中存在的可加速的待量化结构（即目标结构）进行量化，不仅可以有效降低人工成本，提升了模型量化的速度，还实现了模型量化的通用性。

附图说明

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

图 1 是本申请实施例提供的一种网络架构的结构示意图；

图 2 是本申请实施例提供的一种多个待量化结构的结构示意图；

图 3 是本申请实施例提供的一种数据处理方法的流程示意图；

图 4 是本申请实施例提供的一种对目标结构进行量化的示意图一；

图 5 是本申请实施例提供的一种对目标结构进行量化的示意图二；

图 6 是本申请实施例提供的一种对目标结构进行量化的示意图三；

图 7 是本申请实施例提供的一种对目标结构进行量化的示意图四；

图 8 是本申请实施例提供的一种对目标结构进行量化的示意图五；

图 9 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图一；

图 10 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图二；

图 11 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图三；

图 12 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图四；

图 13 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图五；

图 14 是本申请实施例提供的一种数据处理方法的流程示意图；

图 15 是本申请实施例提供的一种对业务模型进行模型优化的场景示意图；

图 16 是本申请实施例提供的一种数据处理装置的结构示意图；

图 17 是本申请实施例提供的一种计算机设备的示意图。

具体实施方式

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，都属于本申请保护的范围。

其中，本申请实施例针对复杂结构的人工智能模型，提出了一种通用型的量化方法，该量化方法涉及人工智能领域，其中，所谓人工智能（Artificial Intelligence，AI）是利用数字计算机或者数字计算机控制的计算模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，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。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习、自动驾驶、智慧交通等几大方向。

其中，机器学习（Machine Learning，ML）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

深度学习算法由一个个计算单元组成，本申请实施例可以将这些计算单元称为算子（Operator，OP）。在网络模型中，算子对应网络层中的计算逻辑，例如：卷积层（Convolution Layer）对应的算子（即卷积算子）可以用于表示做一次卷积计算；全连接层（Fully-connected Layer，FC layer）中的权值求和过程也是一个算子；激活层（activating layer）对应的算子（即激活算子）为在网络模型中被用做激活函数的算子（例如，tanh、ReLU等）。

计算机视觉技术（Computer Vision，CV）是一门研究如何使机器“看”的科学，更进一步的说，就是指用摄影机和电脑代替人眼对目标进行识别和测量等机器视觉，并进一步

做图形处理，使电脑处理成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科，计算机视觉研究相关的理论和技术，试图建立能够从图像或者多维数据中获取信息的人工智能系统。计算机视觉技术通常包括图像处理、图像识别、图像语义理解、图像检索、OCR、视频处理、视频语义理解、视频内容/行为识别、三维物体重建、3D 技术、虚拟现实、增强现实、同步定位与地图构建、自动驾驶、智慧交通等技术，还包括常见的人脸识别、指纹识别等生物特征识别技术。

自然语言处理（Nature Language processing, NLP）是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

语音技术（Speech Technology）的关键技术有自动语音识别技术和语音合成技术以及声纹识别技术。让计算机能听、能看、能说、能感觉，是未来人机交互的发展方向，其中语音成为未来最被看好的人机交互方式之一。

应当理解，随着深度学习技术在多个领域的快速应用，具体包括计算机视觉、自然语言处理、语言技术等，出现了大量的基于深度学习的网络模型。为了解决这些模型因模型结构越来越复杂而导致不便于部署在一些低成本的终端设备中这一问题，本申请实施例可以采用一种通用型的量化方法，对人工智能模型进行量化，进而使得量化后的模型存在以下优点：具有更少的存储开销和带宽需求、更快的计算速度、更低的能耗与占用面积、尚可接受的精度损失，支持低精度（例如，int8）等。

请参见图 1，图 1 是本申请实施例提供的一种网络架构的结构示意图。如图 1 所示，该网络架构可以包括服务器 10S 和终端设备集群。该终端设备集群可以包括一个或者多个终端设备，这里将不对终端设备的数量进行限制。如图 1 所示，该终端设备集群具体可以包括终端设备 100a、终端设备 100b、终端设备 100c、...、终端设备 100n。如图 1 所示，终端设备 100a、终端设备 100b、终端设备 100c、...、终端设备 100n 可以分别与上述服务器 10S 进行网络连接，以便于每个终端设备可以通过该网络连接与服务器 10S 进行数据交互。其中，这里的网络连接不限定连接方式，可以通过有线通信方式进行直接或间接地连接，也可以通过无线通信方式进行直接或间接地连接，还可以通过其他方式，本申请在此不做限制。

其中，该终端设备集群中的每个终端设备均可以包括：智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表、车载终端、智能电视等具有数据处理功能的智能终端。应当理解，如图 1 所示的终端设备集群中的每个终端设备均可以安装有业务应用（即应用客户端），当该应用客户端运行于各终端设备中时，可以分别与上述图 1 所示的服务器 10S 之间进行数据交互。其中，该应用客户端可以包括社交客户端、多媒体客户端（例如，视频客户端）、娱乐客户端（例如，游戏客户端）、信息流客户端、教育客户端、直播客户端等应用客户端。其中，该应用客户端可以为独立的客户端，也可以为集成在某客户端（例如，社交客户端、教育客户端以及多媒体客户端等）中的嵌入式子客户端，在此不做限定。

如图 1 所示，本申请实施例中的服务器 10S 可以为该应用客户端对应的服务器。该服务器 10S 可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云计算服务的云服务器。其中，本申请实施例将不对服务器的数量进行限制。

为便于理解，本申请实施例可以在图 1 所示的多个终端设备中选择一个终端设备作为业务对象所使用的业务终端设备。例如，本申请实施例可以将图 1 所示的终端设备 100a 作为业务终端设备，该业务终端设备中可以集成有业务应用（即应用客户端）。此时，该业务终端设备可以通过该应用客户端对应的业务数据平台与服务器 10S 之间实现数据交互。其中，在模型量化场景中，为了有效控制人工智能模型中的各个模块的计算精度，在对某一具有复杂结构的初始模型进行优化时，这里的业务对象（例如，用户）可以指定该初始模型中的某些网络层不参与量化，其中，本申请实施例可以将业务对象所指定的不参与量化的网络层称之为指定网络层（即属于黑名单中的网络层）。

在本申请实施例中，具有模型量化功能的计算机设备可以为服务器，也可以为图 1 所示的终端设备集群中的任意一个终端设备，例如，终端设备 100b，这里将不对计算机设备的具体形式进行限定。为便于理解，本申请实施例中的计算机设备可以以服务器（例如，图 1 所示的服务器 10S）为例，用以阐述该计算机设备通过通用型量化工具，对初始模型进行自动量化的具体实施方式。

其中，本申请实施例中的通用型量化工具利用了丰富的推理经验，设计各种适合低精度（例如，int8）加速的待量化结构，即针对不同种类的待量化结构，设定不同的量化规则，以得到初始规则集合，使得后续可以直接根据该初始规则集合所确定的量化规则集合，自动对获取到的初始模型进行量化。

可以理解的是，若业务对象未配置不参与量化的指定网络层，则这里的量化规则集合为通用型量化工具所获取到的初始规则集合；若业务对象配置有不参与量化的指定网络层，则这里的量化规则集合是基于业务对象所配置的指定网络层，对初始规则集合进行过滤后所得到的规则集合。这意味着在本申请实施例中，业务对象配置的指定网络层的量化

5 优先级高于通用的初始规则集合，具有最高量化优先级。

其中，这里的量化规则集合可以包括N个量化规则，N为正整数，一个量化规则对应一种待量化结构。该计算机设备在获取到待优化的初始模型时，需要对初始模型进行结构搜索，以得到初始模型的模型结构图，进而可以在该模型结构图中查询是否存在目标结构，

10 这里的目标结构属于量化规则集合所对应的N种待量化结构。若模型结构图中存在目标结构，则该计算机设备可以自动从N个量化规则中确定与目标结构相匹配的目标量化规则，进而可以基于目标量化规则以及伪造算子，对目标结构进行量化，且在量化结束后，基于量化后的初始模型，确定用于业务处理的业务模型。

其中，这里的伪造算子（例如，fake op）主要用于将输入到该伪造算子的参数和输入先量化后再反量化，即等同于QDQ算子，其结构组成是量化算子（quant op）以及反量化

15 算子（dequant op）。该伪造算子可以用于存储当前层的量化信息，具体可以包括缩放因子（scale）和偏移量（zero-point），本申请实施例还可以基于伪造算子中的量化算子和反量化算子的位置来决定每一层算子的精度。

可以理解的是，量化算子主要用于将属于第一精度的特征转换为属于第二精度的特征，反量化算子主要用于将属于第二精度的特征转换回属于第一精度的特征，这里的第一

20 精度高于第二精度。例如，量化可以将浮点型实数（例如，FP32）转化为整型数（例如，INT8），而反量化可以将整型数（例如，INT8）转化回浮点型实数（例如，FP32）。

其中，这里的初始模型可以包括各种人工智能场景下的模型，例如，用于进行内容审核的检测模型、用于进行关键部位识别的检测模型（例如，人脸检测模型）以及图像识别模型等，这里将不对初始模型的应用场景进行限定。换言之，对于任意一种复杂结构的初

25 始模型而言，该计算机设备均可以根据预先设定好的量化规则集合，自动对初始模型中存在的可加速的待量化结构进行量化，以基于量化后的初始模型，确定用于业务处理的业务模型，这种量化方法无需人工针对不同模型设计不同的量化方案，而是能够自动对初始模型中存在的可加速的待量化结构进行量化，不仅可以有效降低人工成本，还能够提升模型量化速度，此外，本申请实施例这种量化方法因不需要人工插入伪造算子，使得代码侵入

30 性降低，从而实现了模型量化通用性。

为便于理解，进一步地，请参见图 2，图 2 是本申请实施例提供的一种多个待量化结构的结构示意图。其中，本申请实施例中的计算机设备可以为具有模型量化功能的计算机设备，该计算机设备可以为上述图 1 所示的终端设备集群中的任意一个终端设备，例如，终端设备 100b，该计算机设备也可以为上述图 1 所示的服务器 10S，这里将不对计算机设备进行限定。

应当理解，本申请实施例可以基于某一深度学习推理框架中的丰富的推理经验，设计出各种适合低精度加速的待量化结构（Patterns），进而可以基于不同的待量化结构设计出不同的量化规则。其中，这里的待量化结构属于本申请搜索的最小粒度类型，该待量化结构可以为由多个网络层所构成的模块，也可以为单个网络层所构成的模块。

如图 2 所示，规则集合 20R 是本申请实施例中的计算机设备获取到的初始规则集合，该规则集合 20R 可以包括 M 个通用型的量化规则，M 为正整数，一个量化规则对应一种待量化结构。为便于阐述，该规则集合 20R 中的量化规则的数量可以以 7 个为例，具体可以包括量化规则 R1、量化规则 R2、量化规则 R3、量化规则 R4、量化规则 R5、量化规则 R6 以及量化规则 R7。

其中，量化规则 R1、量化规则 R2 以及量化规则 R3 均为针对具有权重的网络层这一待量化结构所设计的量化规则，这里的具有权重的网络层即为具有加权操作的网络层，例如，卷积层（conv）、转置卷积层（例如，transpose conv）以及通用矩阵乘法（例如，gemm）；而在具有权重的网络层为卷积层时，可以进一步设计出量化规则 R4 和量化规则 R5。

比如，如图 2 所示，在结构 G1（例如，注意力结构（Attention 结构））中具有权重的网络层具有多个输出位置的情况下，该结构 G1 对应的量化规则 R1 可以指示在具有权重的网络层的输入位置处、以及具有权重的网络层的多个输出位置处，分别添加伪造算子，且采用伪造算子对具有权重的网络层的权重进行量化。例如，在结构 G1 为注意力结构（Attention 结构）时，该结构可以包括卷积算子（conv 算子）、矩阵变维算子（reshape 算子）、排列算子（permute 算子）、bmm 矩阵乘算子。

在结构 G2 包括具有权重的网络层时，该量化规则 R2 可以指示在保留高精度计算时，在具有权重的网络层的输入位置处添加伪造算子，且采用伪造算子对具有权重的网络层的权重进行量化；在保留低精度计算时，具有权重的网络层的输入位置处、具有权重的网络层的输出位置处，分别添加伪造算子，且采用伪造算子对具有权重的网络层的权重进行量化。

其中，根据推理经验可知，具有权重的网络层一般后面会跟激活层，若在该网络层的

输出位置处添加伪造算子，那么在保留高精度计算时，将会导致不可接受的精度损失，基于此，在结构 G2 中包括的具有权重的网络层具有一个输出位置，且该具有权重的第一网络层的输出位置未连接激活层的情况下，该结构 G1 对应的量化规则 R2 可以在具有权重的网络层的输入位置处、以及具有权重的网络层的输出位置处，分别添加伪造算子，且采用
5 伪造算子对具有权重的网络层的权重进行量化。而在结构 G3 中包括的具有权重的网络层具有一个输出位置，且该具有权重的第一网络层的输出位置连接激活层的情况下，该结构 G3 对应的量化规则 R3 可以在具有权重的网络层的输入位置处添加伪造算子，采用伪造算子对具有权重的网络层的权重进行量化，即可以不在该具有权重的网络层的输出位置处添加伪造算子。

10 另外，根据推理经验可知，由卷积层、归一化层（例如，BN）以及激活层（例如，ReLU）构成的结构（例如，图 2 所示的结构 G4）在模型推理阶段能够自动对原网络层的算子进行融合，因此，该结构 G4 对应的量化规则 R4 可以指示在卷积层的输入位置处添加伪造算子，且采用伪造算子对卷积层的权重进行量化。

根据推理经验可知，叠加层的非加权操作的输入的精度决定了结构融合后的算子的输出精度，换言之，若叠加层的非加权操作的输入不参与量化，则叠加层的输出后的量化算子将不会被融合，这样将导致融合后的算子存在精度转换的过程，进而影响速度，基于此，在结构 G5 包括叠加层、叠加层的第一输入分支（包括依次连接的卷积层和归一化层）、和叠加层的第二输入分支（即类型为非加权操作的输入类型）在叠加层包括第一输入分支（具有加权操作的分支）和第二输入分支的情况下，该结构 G5 对应的量化规则 R5 可以指
15 示在需要保持较高的速度时，在卷积层的输入位置处、叠加层的第二输入分支的输入位置处以及叠加层的输出位置处，分别添加伪造算子，且采用伪造算子对卷积层的权重进行量化；而在对速度无要求时，可以在卷积层的输入位置处、以及叠加层的输出位置处，分别添加伪造算子，且采用伪造算子对卷积层的权重进行量化，即可以不在叠加层的第二输入分支的输入位置处添加伪造算子。例如，图 2 所示的结构 G5 可以为与叠加层（例如，Add）
20 相关联的跳跃链接（skip connections）结构，例如，残差网络（resnet）或轻量级网络（Efficient Net）中的跳跃链接结构。

根据推理经验可知，叠加层的两个输入分支中存在无加权操作的低精度输入时，那么叠加层的输出分支必须要量化，基于此，在图 2 所示的结构 G6 可以包括叠加层，且该叠加层的两个输入分支的类型均为非加权操作的输入类型的情况下，该结构 G6 对应的量化
30 规则 R6 可以指示在叠加层的两个输入位置处、和叠加层的输出位置处，分别添加伪造算

子。

其中，根据推理经验可知，在一定情况下，伪造算子中的量化算子（即 Q 算子）可以等价向前传播，伪造算子中的反量化算子（即 DQ 算子）可以等价向后传播，因此如果在拼接层（例如，Concat）的每个输出位置处伪造算子，这样将使得拼接层后的计算属于第二精度，但是要想拼接层对应的拼接算子也被量化，则需要令两个输出分支处所添加的伪造算子中的量化算子相同，基于此，在图 2 所示的结构 G7 包括拼接层，且拼接层的输入分支与拼接层的输出分支的数量（例如，N 个）相同的情况下，结构 G7 对应的量化规则 R7 可以指示在拼接层的每个输出分支的输出位置处，分别添加具有相同量化算子的伪造算子，其中，这里的 N 为大于 1 的正整数，本申请实施例可以以 2 为例。

可以理解的是，后续在对初始模型进行量化时，业务对象的指定网络层的量化优先级具有最高量化优先级。比如，若业务对象针对初始模型的业务配置包括不参与量化的指定网络层（例如，激活层），此时，计算机设备可以先过滤掉包括指定网络层的待量化结构所对应的量化规则（例如，量化规则 R3、量化规则 R4），得到过滤后的第一规则集合，再基于过滤后的第一规则集合，检测初始模型的模型结构图中是否存在可加速的待量化结构（即与过滤后的 5 种结构相同的目标结构）。

其中，具有模型量化功能的计算机设备根据通用型量化工具，自动对初始模型中存在的可加速的待量化结构进行量化的具体实现方式可以参见下述图 3-图 15 所对应的实施例。

进一步地，请参见图 3，图 3 是本申请实施例提供的一种数据处理方法的流程示意图。如图 3 所示，该方法可以由具有模型量化功能的计算机设备执行，该计算机设备可以为终端设备（例如，上述图 1 所示的终端设备集群中的任意一个终端设备，例如，终端设备 100b，也可以为服务器（例如，上述图 1 所示的服务器 10S），在此不做限定。为便于理解，本申请实施例以该方法由具有模型量化功能的计算机设备为例进行说明，该方法至少可以包括以下步骤 S101-步骤 S103：

步骤 S101，计算机设备在获取到初始模型时，对初始模型进行结构搜索，得到初始模型的模型结构图。

具体地，该计算机设备在获取到待优化的初始模型时，可以调用通用型量化工具，基于该通用型量化工具中的搜索图工具，对初始模型进行结构搜索，进而可以得到该初始模型的模型结构图。

步骤 S102，计算机设备根据模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对目标结构进行量化。

其中，这里的伪造算子可以包括量化算子和反量化算子。可以理解的是，这里的目标结构的种类可以为第一结构、第二结构或第三结构。这里的第一结构是指包括具有权重的网络层的结构，本申请实施例可以将初始模型的目标结构中的具有权重的网络层称之为具有权重的第一网络层。其中，上述图 2 所示的结构 G1、结构 G2、结构 G3、结构 G4 以及结构 G5 均属于第一结构。这里的第二结构（例如，上述图 2 所示的结构 G6）包括输入分支的类型为非加权操作的输入类型的叠加层，本申请实施例可以将初始模型的目标结构中所包括的输入分支的类型均为非加权操作的输入类型的叠加层，称之为第一叠加层。这里的第三结构（例如，上述图 2 所示的结构 G7）包括输入分支与输出分支的数量相同的拼接层，本申请实施例可以将初始模型的目标结构中所包括的输入分支与输出分支的数量相同的拼接层称之为第一拼接层。

应当理解，该计算机设备可以获取预先根据推理经验所设计的包括 M 个通用型的量化规则，进而可以将这 M 个量化规则所构成的规则集合确定为初始规则集合（例如，上述图 2 所示的规则集合 20R）。其中，这里的 M 为正整数。与此同时，该计算机设备还需要获取业务对象针对初始模型的业务配置。其中，这里的业务配置是指业务对象（例如，用户）针对初始模型所设定的配置信息，例如，该初始模型后续所要部署的业务平台、业务对象所指定的不参与量化的指定网络层等，这里将不对其进行限定。可以理解的是，由于初始规则集合包括 M 个量化规则，且一个量化规则对应一种待量化结构，因此，该初始规则集合与 M 个待量化结构相关。进一步地，该计算机设备可以基于该业务配置以及初始规则集合，确定最终用于进行模型量化的量化规则集合。

其中，当业务对象指定某一网络层不参与量化时，这意味着业务配置包括不参与量化的指定网络层，此时，该计算机设备需要在 M 个待量化结构中查询指定网络层，得到查询结果。若查询结果指示 M 个待量化结构中存在包括指定网络层的待量化结构，则该计算机设备可以将查询到的待量化结构确定为过滤结构，进而可以在初始规则集合中删除过滤结构对应的量化规则，将删除后的初始规则集合确定为量化规则集合。

如图 2 所示，若业务配置包括不参与量化的指定网络层为激活层，则该计算机设备需要在图 2 所示的 7 个待量化结构中查询激活层，以得到查询结果。由于图 2 所示的结构 G3、以及结构 G4 均包括激活层，因此该查询结果指示 7 个待量化结构中存在包括指定网络层的待量化结构，此时，该计算机设备可以将结构 G3 和结构 G4 分别确定为过滤结构，然后在规则集合 20R 中删除该过滤结构对应的量化规则（即量化规则 R3 和量化规则 R4），进而将删除后的规则集合 20R 确定为量化规则集合。这时，这里的量化规则集合可以包括图

2 所示的量化规则 R1、量化规则 R2、量化规则 R5、量化规则 R6 以及量化规则 R7。

可选的，当业务对象未指定不参与量化的网络层时，这意味着业务配置中不存在指定网络层，此时，该计算机设备可以直接将初始规则集合确定为量化规则集合。如图 2 所示，若业务配置不包括指定网络层，则该计算机设备可以直接将图 2 所示的规则集合 20R 确定为量化规则集合。此时，这里的量化规则集合可以包括图 2 所示的量化规则 R1、量化规则 R2、量化规则 R3、量化规则 R4、量化规则 R5、量化规则 R6 以及量化规则 R7。

其中，由于量化规则集合可以包括 N 个量化规则，且一个量化规则对应一种待量化结构，因此，该计算机设备需要在模型结构图中使用深度优先搜索 (Depth First Search, DFS)，来找出所有符合上述量化规则集合中的 N 个待量化结构的目标结构，进而可以根据模型结构图中的目标结构的种类，从 N 个量化规则中，确定与目标结构相匹配的目标量化规则，确定伪造算子在初始模型中的添加位置，进而可以在确定的添加位置中添加伪造算子，以实现

10 对目标结构的量化。

比如，在模型结构图中的目标结构包括具有权重的第一网络层的情况下，该计算机设备可以在具有权重的第一网络层的输入位置处，添加用于进行量化的伪造算子，且采用伪造算子对具有权重的第一网络层的权重进行量化。

15 其中，在具有权重的第一网络层具有多个输出位置的情况下，该计算机设备还需要在具有权重的第一网络层的多个输出位置处，分别添加用于进行量化的伪造算子。

换言之，在目标结构包括具有权重的第一网络层，且具有权重的第一网络层具有多个输出位置的情况下，该计算机设备需要从量化规则集合中，获取与目标结构相匹配的量化规则，进而可以将该匹配到的量化规则（例如，上述图 2 所示的量化规则 R1）确定为目标量化规则。然后，该计算机设备可以基于该量化规则 R1，将具有权重的第一网络层的输入位置、具有权重的第一网络层的输出位置、以及具有权重的第一网络层的权重位置分别确定为添加位置，进而可以在添加位置处添加伪造算子。其中，每个添加位置处添加的伪造算子可以相同，也可以不同，这里将不对其进行限定。

25 可选的，在具有权重的第一网络层具有一个输出位置，且该具有权重的第一网络的输出位置处未连接激活层的情况下，该计算机设备还需要在具有权重的第一网络层的输出位置处，添加用于进行量化的伪造算子。

可选的，在具有权重的第一网络层具有一个输出位置，且具有权重的第一网络层的输出位置处连接激活层的情况下，为了保留高精度的计算，减少量化输出所导致的精度丢失，该计算机设备可以不在具有权重的第一网络层的输出位置处，添加用于进行量化的伪造算

子。

为便于理解，进一步地，请参见图 4，图 4 是本申请实施例提供的一种对目标结构进行量化的示意图一。如图 4 所示，该结构 40G 可以为该计算机设备在初始模型的模型结构图中所查找到的目标结构，该结构 40G 包括具有权重的第一网络层（例如，卷积层），该具有权重的第一网络层的具有一个输出位置，且该输出位置处并未连接激活层。

如图 4 所示，本申请实施例中的输入（例如，X）在传输至该结构 40G 时，可以通过该结构 40G 中的具有权重的第一网络层，对 X 进行特征提取处理，从而得到该具有权重的第一网络层的输出（即结构 40G 的输出，例如，Y）。若这里的 X 属于第一精度，那么该结构 40G 的输出也为第一精度。进一步地，在对结构 40G 进行量化时，该计算机设备可以在量化规则集合中，查找与结构 40G 相匹配的量化规则，进而可以将查找到的量化规则确定为目标量化规则（例如，上述图 2 所示的量化规则 R2）。

基于此，该计算机设备可以基于该目标量化规则，将具有权重的第一网络层的输入位置确定为位置 A1，将具有权重的第一网络层中的权重位置确定为位置 A2，将具有权重的第一网络层的输出位置确定为位置 A3，进而可以将该位置 A1、位置 A2 和位置 A3 分别确定为伪造算子在结构 40G 中的添加位置。

进一步地，该计算机设备可以在位置 A1 处添加伪造算子（例如，伪造算子 4F1），使得 X 在经过伪造算子 4F1 中的量化算子后转换为第二精度，再经过伪造算子 4F1 中的反量化算子后，还原为第一精度。同理，该计算机设备还可以在位置 A2 处添加伪造算子，对具有权重的第一网络层的权重进行量化，使得该具有权重的第一网络层的权重在经过伪造算子 4F2 中的量化算子后，转换为第二精度，再经过伪造算子 4F2 中的反量化算子后，还原为第一精度。该计算机设备还可以在位置 A3 处添加伪造算子（例如，伪造算子 4F3），使得具有权重的第一网络层的输出在经过伪造算子 4F3 中的量化算子后，转换为第二精度，再经过伪造算子 4F3 中的反量化算子后，还原为第一精度。

由此可见，本申请实施例在检测到结构 40G 这样的目标结构时，可以通过与结构 40G 相匹配的量化规则以及伪造算子（包括伪造算子 4F1、伪造算子 4F2 和伪造算子 4F3），自动对结构 40G 进行量化，以得到量化后的结构 40G（例如，图 4 所示的量化结构 41G）。

又比如，由于推理经验可知，在模型结构图中的目标结构包括依次连接的具有权重的第一网络层，归一化层以及激活层，且具有权重的第一网络层为卷积层的情况下，后续可以对这样的目标结构进行融合，因此本申请实施例还可以对这样的目标结构进行量化，即计算机设备可以从量化规则集合中，获取与目标结构相匹配的量化规则，进而可以将该匹

配到的量化规则（例如，上述图 2 所示的量化规则 R4）确定为目标量化规则。然后，该计算机设备可以基于该量化规则 R4，在具有权重的第一网络层的输入位置添加伪造算子，且采用量化算子，对具有权重的第一网络层的权重进行量化。

为便于理解，进一步地，请参见图 5，图 5 是本申请实施例提供的一种对目标结构进行量化的示意图二。如图 5 所示，该结构 50G 可以为该计算机设备在初始模型的模型结构图中所查找到的目标结构，该结构 50G 可以包括依次连接的卷积层、归一化层以及激活层。

如图 5 所示，在结构 50G 中的卷积层、归一化层以及激活层的精度均为第一精度（例如，FP32）时，这意味着该卷积层中的输入（例如，X）的精度和权重的精度也为第一精度，换言之，该结构 50G 对应的输入在经过卷积计算、归一化计算以及激活计算后所得到的输出（例如，Y）仍为第一精度。可以理解的是，在对结构 50G 进行量化时，计算机设备可以根据结构 50G 的种类，在量化规则集合中查找与结构 50G 相匹配的量化规则，进而可以将查找到的量化规则确定为目标量化规则（例如，上述图 2 所示的量化规则 R4）。

基于此，该计算机设备可以基于该目标量化规则，将结构 50G 中的卷积层的输入位置确定为位置 B1，且将结构 50G 中的卷积层的权重位置确定为位置 B2，进而可以将该位置 B1 和位置 B2 分别确定为伪造算子在结构 50G 中的添加位置。

进一步地，该计算机设备可以在位置 B1 处添加伪造算子（例如，伪造算子 5F1），使得属于第一精度的输入在经过伪造算子 5F1 中的量化算子后，转换为第二精度，再经过伪造算子 5F1 中的反量化算子后，还原为第一精度。与此同时，该计算机设备还可以采用伪造算子 5F2，对卷积层的权重进行量化，即在位置 B2 处添加伪造算子 5F2，使得属于第一精度的权重在经过伪造算子 5F2 中的量化算子后，转换为第二精度，再经过伪造算子 5F2 中的反量化算子后，还原为第一精度。其中，这里的伪造算子 5F1 与伪造算子 5F2 可以相同，也可以不同，这里将不对其进行限定。

由此可见，本申请实施例在检测到结构 50G 这样的目标结构时，可以通过与结构 50G 相匹配的量化规则以及伪造算子（包括伪造算子 5F1 和伪造算子 5F2），自动对结构 50G 进行量化，以得到量化后的结构 50G（例如，图 5 所示的量化结构 51G）。

又比如，在模型结构图中的目标结构包括叠加层、叠加层的第一输入分支和叠加层的第二输入分支的情况下，该计算机设备可以在叠加层的输出位置处添加伪造算子，也可以在第二输入分支的输入位置处添加伪造算子。其中，这里的第一输入分支包括依次连接的具有权重的第一网络层和归一化层，该归一化层与叠加层连接，第二输入分支的类型为非加权操作的输入类型，具有权重的第一网络层为卷积层。

换言之，在模型结构图中的目标结构包括的具有权重的第一网络层为卷积层，且该卷积层的输出位置与归一化层连接，该归一化层的输出位置与叠加层连接的情况下，该计算机设备不仅需要在具有权重的第一网络层的输入位置处，添加用于进行量化的伪造算子，采用伪造算子对具有权重的第一网络层的权重进行量化，还需要在叠加层的输出位置处添加伪造算子。此外，由于第二输入分支的精度决定了融合后的具有权重的网络层的输出精度，为了提高速度，该计算机设备还可以在第二输入分支的输入位置处添加伪造算子。

为便于理解，进一步地，请参见图 6，图 6 是本申请实施例提供的一种对目标结构进行量化的示意图三。如图 6 所示，该结构 60G 可以为该计算机设备在初始模型的模型结构图中所查找到的目标结构，该结构 60G 是一种与叠加层相关联的跳跃链接结构。其中，该叠加层的第一输入分支可以包括依次连接的卷积层和归一化层，第二输入分支的类型为非加权操作的输入类型，即该第二输入分支对应的输入 X2 可以为一个非加权操作的输入。

如图 6 所示，该结构 60G 中的叠加层可以对经过卷积计算、归一化计算后的第一输入（例如，X1）和第二输入（例如，X2）进行叠加，得到结构 60G 的输出（例如，Y）。由于该第一输入和第二输入均属于第一精度，因此该结构 60G 的输出仍为第一精度。可以理解的是，在对结构 60G 进行量化时，该计算机设备可以在量化规则集合中，查找与结构 60G 相匹配的量化规则，进而可以将查找到的量化规则确定为目标量化规则（例如，上述图 2 所示的量化规则 R5）。

基于此，为了提高后续模型处理速度，该计算机设备可以基于该目标量化规则，将结构 60G 中的卷积层的输入位置确定为位置 C1，且将卷积层的权重位置确定为位置 C2，将第二输入分支的输入位置确定为位置 C3，将叠加层的输出位置确定为位置 C4，进而可以将该位置 C1、位置 C2、位置 C3 和位置 C4 分别确定为伪造算子在初始模型中的添加位置。

进一步地，该计算机设备可以在位置 C1 处添加伪造算子（例如，伪造算子 6F1），使得属于第一精度的第一输入（例如，X1）在经过伪造算子 6F1 中的量化算子后，转换为第二精度，再经过伪造算子 6F1 中的反量化算子后，还原为第一精度。同理，该计算机设备还可以在位置 C2 处添加伪造算子（例如，伪造算子 6F2），采用伪造算子 6F2，对卷积层的权重进行量化，使得属于第一精度的权重在经过伪造算子 6F2 中的量化算子后，转换为第二精度，再经过伪造算子 6F2 中的反量化算子后，还原为第一精度。该计算机设备还可以在位置 C3 处添加伪造算子（例如，伪造算子 6F3），使得属于第一精度的第二输入（例如，X2）在经过伪造算子 6F3 中的量化算子后，转换为第二精度，再经过伪造算子 6F2 中的反量化算子后，还原为第一精度。该计算机设备还可以在位置 C4 处添加伪造算子（例

如，伪造算子 6F4），使得叠加层的输出在经过伪造算子 6F4 中的量化算子后，转换为第二精度，再经过伪造算子 6F4 中的反量化算子后，还原为第一精度。其中，这里的各添加位置处所添加的伪造算子可以相同，也可以不同，这里将不对其进行限定。

由此可见，本申请实施例在检测到结构 60G 这样的目标结构时，可以通过与结构 60G 相匹配的量化规则以及伪造算子（包括伪造算子 6F1、伪造算子 6F2、伪造算子 6F3、伪造算子 6F4），自动对结构 60G 进行量化，以得到量化后的结构 60G（例如，图 6 所示的量化结构 61G）。

又比如，在模型结构图中的目标结构包括第一叠加层，且该第一叠加层的第一输入分支的类型与该第一叠加层的第二输入分支的类型均为非加权操作的输入类型的情况下，该计算机设备需要分别在第一输入分支的输入位置处添加用于进行量化的伪造算子、在第二输入分支的输入位置处添加伪造算子，且在第一叠加层的输出位置处添加伪造算子。

为便于理解，进一步地，请参见图 7，图 7 是本申请实施例提供的一种对目标结构进行量化的示意图四。如图 7 所示，该结构 70G 可以为该计算机设备在初始模型的模型结构图中所查找到的目标结构，该结构 70G 可以为包括两个输入分支和一个输出分支的叠加层。

如图 7 所示，本申请实施例中的两个输入（例如，X1 和 X2）在传输至该结构 70G 时，可以通过该结构 70G 中的叠加层，对 X1 和 X2 进行叠加处理，从而得到该叠加层的输出（即结构 70G 的输出，例如，Y）。若这里的 X1 和 X2 均属于第一精度，那么该结构 70G 的输出也为第一精度。可以理解的是，在对结构 70G 进行量化时，该计算机设备可以在量化规则集合中，查找与结构 70G 相匹配的量化规则，进而可以将查找到的量化规则确定为用于对结构 70G 进行量化的目标量化规则（例如，上述图 2 所示的量化规则 R6）。

基于此，该计算机设备可以基于该目标量化规则，将叠加层的第一输入分支的输入位置（即 X1 的输入位置）确定为位置 D1，将叠加层的第二输入分支的输入位置（即 X2 的输入位置）确定为位置 D2，将叠加层的输出位置确定为位置 D3，进而可以将该位置 D1、位置 D2 和位置 D3 分别确定为伪造算子在结构 70G 中的添加位置。

进一步地，该计算机设备可以在位置 D1 处添加伪造算子（例如，伪造算子 7F1），使得输入至结构 71G 的第一输入（例如，X1）在经过伪造算子 7F1 中的量化算子后转换为第二精度，再经过伪造算子 7F1 中的反量化算子后，还原为第一精度。同理，该计算机设备还可以在位置 D2 处添加伪造算子（例如，伪造算子 7F2），使得输入至结构 71G 的第二输入（例如，X2）在经过伪造算子 7F2 中的量化算子后转换为第二精度，再经过伪造算子 7F2 中的反量化算子后，还原为第一精度。该计算机设备还可以在位置 D3 处添加伪

造算子（例如，伪造算子 7F3），使得叠加层的输出在经过伪造算子 7F3 中的量化算子后，转换为第二精度，再经过伪造算子 7F3 中的反量化算子（例如，反量化算子 2）后，还原为第一精度。

由此可见，本申请实施例在检测到结构 70G 这样的目标结构时，可以通过与结构 70G 相匹配的量化规则以及伪造算子（包括伪造算子 7F1、伪造算子 7F2 和伪造算子 7F3），自动对结构 70G 进行量化，以得到量化后的结构 70G（例如，图 7 所示的量化结构 71G）。

又比如，在模型结构图中的目标结构包括第一拼接层，且该第一拼接层的输入分支与第一拼接层的输出分支的数量相同的情况下，该计算机设备可以于第一拼接层的 N 个输出分支的每个输出分支的输出位置处，分别添加用于进行量化的伪造算子；其中，这里的 N 可以为大于 1 的正整数，且为了后续能够优化结构，这里的每个输出分支处所添加的伪造算子的量化算子均相同。

为便于理解，进一步地，请参见图 8，图 8 是本申请实施例提供的一种对目标结构进行量化的示意图五。如图 8 所示，该结构 80G 可以为该计算机设备在初始模型的模型结构图中所查找到的目标结构，该结构 80G 可以包括拼接层，该拼接层的输入分支的数量与输出分支的数量相同，为便于阐述，本申请实施例可以以两个输入分支和两个输出分支为例。

如图 8 所示，该结构 80G 中的拼接层可以对两个输入分支的输入（例如，X1 和 X2）进行拼接处理，从而得到拼接层的输出（例如，Y1 和 Y2）。若这里的 X1 和 X2 均属于第一精度，那么该结构 80G 的两个输出均为第一精度。可以理解的是，在对结构 80G 进行量化时，该计算机设备可以在量化规则集合中，查找与结构 80G 相匹配的量化规则，进而可以将查找到的量化规则确定为目标量化规则（例如，上述图 2 所示的量化规则 R7）。

基于此，该计算机设备可以基于该目标量化规则，将拼接层的第一输出分支的输出位置确定为位置 E1，且将拼接层的第二输出分支的输出位置确定为位置 E2，进而可以将该位置 E1 和位置 E2 分别确定为伪造算子在结构 80G 中的添加位置。

进一步地，该计算机设备可以在位置 E1 处添加伪造算子（例如，伪造算子 8F1），使得拼接层的第一输出在经过伪造算子 8F1 中的量化算子后，转换为第二精度，再经过伪造算子 8F1 中的反量化算子（例如，反量化算子 1）后，还原为第一精度。同理，该计算机设备还可以在位置 E2 处添加伪造算子（例如，伪造算子 8F2），使得拼接层的第二输出在经过伪造算子 8F2 中的量化算子后，转换为第二精度，再经过伪造算子 8F2 中的反量化算子（例如，反量化算子 2）后，还原为第一精度。其中，该伪造算子 8F1 与伪造算子 8F2 中的量化算子相同，而伪造算子 8F1 与伪造算子 8F2 中的反量化算子可以相同，也可以不

同，这里将不对其进行限定。

由此可见，本申请实施例在检测到结构 80G 这样的目标结构时，可以通过与结构 80G 相匹配的量化规则以及伪造算子（包括伪造算子 8F1 和伪造算子 8F2），自动对结构 80G 进行量化，以得到量化后的结构 80G（例如，图 8 所示的量化结构 81G）。

5 步骤 S103，计算机设备基于量化后的初始模型，得到用于业务处理的业务模型。

具体地，该计算机设备可以对模型结构图中存在的各种可加速的待量化结构分别进行量化，在量化结束后，该计算机设备可以将量化后的初始模型确定为第一模型。为了使得量化后的模型精度满足业务需求，该计算机设备还可以对第一模型进行微调训练，得到第二模型，进而可以基于该第二模型，得到用于业务处理的业务模型。

10 其中，在对第一模型进行微调训练时，该计算机设备可以获取样本业务数据以及样本业务数据对应的样本业务标签，进而可以将样本业务数据的样本特征输入至第一模型，通过第一模型中添加的伪造算子以及第一模型中的各网络层，对样本业务数据进行预测，得到样本业务数据的预测业务标签。此时，该计算机设备可以基于样本业务标签以及预测业务标签，对第一模型进行微调训练，得到第二模型，再对第二模型进行优化处理，得到用
15 于业务处理的业务模型。

其中，这里的样本业务数据是基于初始模型所进行的业务处理所确定的。比如，若初始模型的业务处理为关键部位检测（例如，人脸检测）处理，则这里的样本业务数据可以为对获取到的某一人脸数据库（例如人脸数据集（Annotated Facial Landmarks in the Wild））中的人脸图像，采用局部切块、翻转、高斯模糊等方式，所生成人脸/非人脸训练集。此时，
20 这里的样本业务标签可以用于指示样本业务数据的真实类别（例如，人脸或非人脸）。

其中，在本申请实施例中，涉及到的人脸（或其他生物特征）识别技术，在本申请以上实施例运用到具体产品或技术中时，相关数据收集、使用和处理过程应该遵守国家法律法规要求，收集人脸信息前应该告知信息处理规则并征求目标对象的单独同意，并严格遵照法律法规要求和个人信息处理规则处理人脸信息，采取技术措施保障相关数据安全。

25 又比如，若初始模型的业务处理为内容审核处理，则这里的样本业务数据可以为获取到的多媒体数据（例如，视频、音频、文本、图片等），此时，这里的样本业务标签可以用于指示样本业务数据的内容质量的真实影响度，影响度越高，意味着该样本业务数据的内容质量越高。

又比如，若初始模型的业务处理为图像识别处理，则这里的样本业务数据可以为获取
30 到的图片数据，此时，这里的样本业务标签可以用于指示样本业务数据的真实类别（例如，

人、猫、狗等）。当然，该初始模型的业务处理还可以包括其他处理，这里将不对其进行一一举例。

在本申请中，涉及到的数据抓取技术方案实施（例如，内容审核处理时所抓取到的多媒体数据或图像识别处理时所抓取到的图片数据），在本申请以上实施例运用到具体产品或技术中时，相关数据收集、使用和处理过程应该遵守国家法律法规要求，符合合法、正当、必要的原则，不涉及获取法律法规禁止或限制的数据类型，不会妨碍目标网站的正常运行。

可以理解的是，当第二模型的网络结构比较复杂时，本申请实施例需要对第二模型中的待优化结构进行优化处理，使得第二模型中的尽可能多的算子进行融合，以得到业务模型。其中，由于第二模型是经过微调训练的已量化的初始模型，因此，该第二模型的模型结构图、第一模型的模型结构图均与初始模型的模型结构图相同，本申请实施例可以将第二模型中具有权重的网络层也称之为具有权重的第一网络层，依次类推，将第二模型中的输入分支的类型均为非加权操作的输入类型的叠加层称之为第一叠加层；将第二模型中的输入分支与输出分支的数量相同的拼接层称之为第一拼接层。

比如，在初始模型的目标结构包括具有权重的第一网络层的情况下，第二模型中的待优化结构也包括具有权重的第一网络层，那么，基于该第二模型所最终得到的业务模型可以包括具有权重的网络层（即第二网络层）、添加在具有权重的第二网络层的输入位置处的量化算子；在该业务模型中，具有权重的第二网络层是根据具有权重的第一网络层得到的，且该具有权重的第二网络层的权重是采用量化算子，对第二模型中的具有权重的第一网络层的权重进行处理后所得到的。

为便于理解，进一步地，请参见图 9，图 9 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图一。如图 9 所示，这里的结构 91G 可以为第二模型中的待优化结构，该结构 91G 是该计算机设备对初始模型中的目标结构进行量化，并微调训练后所得到的，该结构 91G 与上述图 4 所示的结构 41G 的结构相同。

其中，该结构 91G 中可以包括伪造算子 F1（即由量化算子 1 和反量化算子 1 所构成的伪造算子）、伪造算子 F2（即由量化算子 2 和反量化算子 2 所构成的伪造算子）以及伪造算子 F3（即由量化算子 3 和反量化算子 3 所构成的伪造算子）。

应当理解，该计算机设备在对结构 91G 进行优化处理时，可以根据结构 91G 中的具有权重的第一网络层，得到具有权重的第二网络层，进而可以采用量化算子，对结构 91G 中的具有权重的第一网络层的权重进行处理，得到具有权重的第二网络层的权重，比如，该

计算机设备可以过滤掉反量化算子 1 和反量化算子 2，并采用量化算子 2 以及量化算子 3，对结构 91G 中的具有权重的第一网络层的权重进行处理，得到具有权重的第二网络层的权重。

这意味着该计算机设备对第二模型中的结构 91G 进行优化处理后，可以得到业务模型中的优化结构（即图 9 所示的结构 92G）。如图 9 所示，该结构 92G 可以包括具有权重的第二网络层，添加在该具有权重的第二网络层的输入位置处的量化算子（例如，量化算子 1）。

又比如，在初始模型的目标结构中的具有权重的第一网络层为卷积层，且该目标结构还包括与卷积层连接的归一化层以及与归一化层连接的激活层的情况下，该计算机设备还需要对具有权重的第一网络层、归一化层和激活层进行融合，得到具有权重的第二网络层。

为便于理解，进一步地，请参见图 10，图 10 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图二。如图 10 所示，这里的结构 101G 可以为第二模型中的待优化结构，该结构 101G 是该计算机设备对初始模型中的目标结构进行量化，并微调训练后所得到的，这里的目标结构中具有权重的第一网络层为卷积层，且该目标结构还包括与该卷积层连接的归一化层以及与归一化层连接的激活层。该结构 101G 与上述图 5 所示的结构 51G 的结构相同。

其中，该结构 101G 中可以包括伪造算子 F1（即由量化算子 1 和反量化算子 1 所构成的伪造算子）和伪造算子 F2（即由量化算子 2 和反量化算子 2 所构成的伪造算子）。应当理解，该计算机设备在对结构 101G 进行优化处理时，可以根据结构 101G 中的卷积层，得到结构 102G 中的融合后的卷积层。比如，该计算机设备可以对结构 101G 中的卷积层、归一化层和激活层进行融合，得到融合后的卷积层（即具有权重的第二网络层）。

与此同时，该计算机设备还可以采用量化算子，对结构 101G 中的卷积层的权重进行处理，得到融合后的卷积层的权重，比如，该计算机设备可以过滤掉反量化算子 1 和反量化算子 2，并采用量化算子 2，对结构 101G 中的卷积层的权重进行处理，得到融合后的卷积层的权重。

这意味着该计算机设备对第二模型中的结构 101G 进行优化处理后，可以得到业务模型中的优化结构（即图 10 所示的结构 102G）。如图 10 所示，该结构 102G 可以包括具有权重的第二网络层，添加在该具有权重的第二网络层的输入位置处的量化算子（例如，量化算子 1）。

又比如，在初始模型的目标结构包括叠加层、叠加层的第一输入分支和叠加层的第二

输入分支，且第一输入分支包括依次连接的具有权重的第一网络层和归一化层的情况下，该计算机设备还可以对具有权重的第一网络层、归一化层和叠加层进行融合，得到具有权重的第二网络层。其中，这里的归一化层与叠加层连接，第二输入分支的类型为非加权操作的输入类型，具有权重的第一网络层为卷积层。

5 为便于理解，进一步地，请参见图 11，图 11 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图三。如图 11 所示，这里的结构 111G 可以为第二模型中的待优化结构，该结构 111G 是该计算机设备对初始模型中的目标结构进行量化，并微调训练后所得到的，这里的目标结构包括叠加层，叠加层的第一输入分支、叠加层的第二输入分支。其中，该第一输入分支包括依次连接的具有权重的第一网络层和归一化层，且该具有权重的第一网络层为卷积层，归一化层与叠加层连接。该第二输入分支的类型为非加
10 权操作的输入类型。该结构 111G 与上述图 6 所示的结构 61G 的结构相同。

其中，该结构 111G 中可以包括伪造算子 F1（即由量化算子 1 和反量化算子 1 所构成的伪造算子）、伪造算子 F2（即由量化算子 2 和反量化算子 2 所构成的伪造算子）、伪造算子 F3（即由量化算子 3 和反量化算子 3 所构成的伪造算子）、伪造算子 F4（即由量化
15 算子 4 和反量化算子 4 所构成的伪造算子）。

应当理解，该计算机设备在对结构 111G 进行优化处理时，可以根据结构 111G 中的卷积层，得到结构 112G 中的融合后的卷积层。比如，该计算机设备可以对结构 111G 中的卷积层、归一化层和激活层进行融合，得到融合后的卷积层（即具有权重的第二网络层）。

与此同时，该计算机设备还可以采用量化算子，对结构 111G 中的卷积层的权重进行处理，得到融合后的卷积层的权重，比如，该计算机设备可以过滤掉反量化算子 1、反量化算子 2 和反量化算子 3，并采用量化算子 2 以及量化算子 4，对结构 111G 中的卷积层的
20 权重进行处理，得到融合后的卷积层的权重。

这意味着该计算机设备对第二模型中的结构 111G 进行优化处理后，可以得到业务模型中的优化结构（即图 11 所示的结构 112G）。如图 11 所示，该结构 112G 可以包括具有
25 权重的第二网络层，添加在该具有权重的第二网络层的输入位置处的量化算子（例如，量化算子 1 和量化算子 3）。

又比如，在初始模型的目标结构包括第一叠加层，且第一叠加层的第一输入分支的类型与第一叠加层的第二输入分支的类型均为非加权操作的输入类型的情况下，那么，基于该第二模型所最终得到的业务模型可以包括第二叠加层、添加在第二叠加层的第一输入分
30 支处的量化算子、和添加在第二叠加层的第二输入分支处的量化算子。在该业务模型中，

第二叠加层是根据量化算子以及第二模型中的第一叠加层所得到的。

为便于理解，进一步地，请参见图 12，图 12 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图四。如图 12 所示，这里的结构 121G 可以为第二模型中的待优化结构，该结构 121G 是该计算机设备对初始模型中的目标结构进行量化，并微调训练后所得到的，这里的目标结构可以包括第一叠加层，且第一叠加层的第一输入分支的类型与第一叠加层的第二输入分支的类型均为非加权操作的输入类型。该结构 121G 可以为上述图 7 所示的结构 71G 的结构相同。

其中，该结构 121G 中可以包括伪造算子 F1（即由量化算子 1 和反量化算子 1 所构成的伪造算子）、伪造算子 F2（即由量化算子 2 和反量化算子 2 所构成的伪造算子）以及伪造算子 F3（即由量化算子 3 和反量化算子 3 所构成的伪造算子）。应当理解，该计算机设备在对结构 121G 进行优化处理时，可以根据量化算子以及第二模型中的第一叠加层，得到结构 122G 中的第二叠加层。比如，该计算机设备可以过滤掉反量化算子 1 和反量化算子 2，并采用量化算子 3 对第一叠加层进行处理后，得到第二叠加层。

这意味着该计算机设备对第二模型中的结构 121G 进行优化处理后，可以得到业务模型中的优化结构（即图 12 所示的结构 122G）。如图 12 所示，该结构 122G 可以包括第二叠加层、添加在第二叠加层的第一输入分支处的量化算子（即量化算子 1）、和添加在第二叠加层的第二输入分支处的量化算子（即量化算子 2）。

可以理解的是，本申请实施例可以根据等价传播规则适当移动量化算子和反量化算子的位置，以便于尽可能多的将待优化模型的多个算子进行融合，简单来说，需要尽可能的将反量化算子推迟，以推迟反量化操作，尽可能的将量化算子提前，以提前量化操作，这样将使得参数以低精度的形式在模型中的尽量多的网络层中进行传播，从而提升模型速度。

基于此，在初始模型的目标结构包括第一拼接层，且第一拼接层的输入分支与第一拼接层的输出分支的数量相同的情况下，那么，基于该第二模型所最终得到的业务模型可以包括第二拼接层、添加在第二拼接层的每个输入分支的输入位置处的量化算子。在该业务模型中，第二拼接层的每个输入分支的输入位置处的量化算子是将添加在第二模型中的第一拼接层的每个输出分支的输出位置处的量化算子，分别等价向前传播至第二模型中的第一拼接层的每个输入分支的输入位置处后所得到的。

为便于理解，进一步地，请参见图 13，图 13 是本申请实施例提供的一种对第二模型中的待优化结构进行处理的示意图五。如图 13 所示，这里的结构 131G 可以为第二模型中

的待优化结构，该结构 131G 是该计算机设备对初始模型中的目标结构进行量化后所得到的，该目标结构可以包括第一拼接层，且第一拼接层的输入分支与第一拼接层的输出分支的数量相同，这里的数量以 2 为例。其中，该结构 131G 与上述图 8 所示的结构 81G 的结构相同。

- 5 其中，该结构 131G 中可以包括伪造算子 F1（即由量化算子 1 和反量化算子 1 所构成的伪造算子）和伪造算子 F2（即由量化算子 2 和反量化算子 2 所构成的伪造算子），且这里的量化算子 1 等于量化算子 2。

基于此，该计算机设备在对结构 131G 进行优化处理时，可以将量化算子 1 的位置由第一拼接层的第一输出位置等价向前传播至第一拼接层的第一输入位置（即 X1 对应的输入位置），同理，还需要将量化算子 2 的位置由第一拼接层的第二输出位置等价向前传播至第一拼接层的第二输入位置（即 X2 对应的输入位置），进而可以得到结构 132G。此时，该计算机设备可以将结构 132G 中的拼接层称之为第二拼接层。

这意味着该计算机设备对第二模型中的结构 131G 进行优化处理后，可以得到业务模型中的优化结构（即图 13 所示的结构 132G）。如图 13 所示，该结构 132G 可以包括第二
15 拼接层、添加在第二拼接层的每个输入分支的输入位置处的量化算子。在结构 132G 中，该计算机设备可以通过量化算子 1 将 X1 的精度由第一精度转换为第二精度，通过量化算子 2 将 X2 的精度由第一精度转换为第二精度，进而可以通过结构 132G 中的第二拼接层，将属于第二精度的 X1 与属于第二精度的 X2 进行拼接处理，从而使得后续拼接处理后所得到的参数仍为第二精度。

20 在本申请实施例中，初始模型对应的业务对象无需手动在初始模型中的相应位置处添加伪造算子，而是直接将任意一种初始模型提交至计算机设备，以使计算机设备调用通用型量化工具，来获取预先设定好的量化规则集合，进而基于该通用型量化工具以及量化规则集合，对初始模型中存在的可加速的待量化结构进行量化，整个量化过程可在通用型量化工具内部完成，从而可以有效降低人工成本，提升了模型量化速度，实现了量化通用性。
25 此外，由于该通用型量化工具还支持对量化后的初始模型进行微调训练，使得后续得到的业务模型在满足业务指标的同时还能够获取最高的加速比。又因对第二模型中的待优化结构进行优化处理，这样将使得后续得到的业务模型能够提升模型的每秒查询率（query per second，QPS），进而还可减少模型部署内存以及显存占用等。

进一步地，请参见图 14，图 14 是本申请实施例提供的一种数据处理方法的流程示意图。该方法可以由具有模型量化功能的终端设备（例如，上述图 1 所示的终端设备集群中
30

的任意一个终端设备，例如，终端设备 100a) 执行，也可以由具有模型量化功能的服务器（例如，上述图 1 所示的服务器 10S）执行，还可以由具备模型应用功能的终端设备和具备模型量化功能的服务器交互执行，在此不做限定。该方法至少可以包括步骤 S201-步骤 S209。

5 步骤 S201，在获取到初始模型时，对初始模型进行结构搜索，得到初始模型的模型结构图。

步骤 S202，根据模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对目标结构进行量化。

其中，伪造算子包括量化算子和反量化算子。

10 步骤 S203，基于量化后的初始模型，确定用于业务处理的业务模型。

其中，本申请实施例中的数据处理方法可以包括模型量化过程以及模型应用过程。可以理解的是，该步骤 S201-步骤 S203 阐述了模型量化过程，该模型训练过程的具体实施方式可参见上述图 3 所对应实施例中对步骤 S101-步骤 S103 的描述，这里将不再继续进行赘述。

15 为便于理解，进一步地，请参见图 15，图 15 是本申请实施例提供的一种对业务模型进行模型优化的场景示意图。其中，本申请实施例中的计算机设备可以为具有模型量化功能的计算机设备，该计算机设备可以为上述图 1 所示的终端设备集群中的任意一个终端设备，例如，终端设备 100b，该计算机设备也可以为上述图 1 所示的服务器 10S，这里将不对计算机设备进行限定。

20 图 15 所示的模型 151W 为该计算机设备所获取到的初始模型。该计算机设备在获取到图 15 所示的模型 151W 时，可以调用通用型量化工具，对模型 151W 进行结构搜索，以得到该模型 151W 的模型结构图。进一步地，该计算机设备可以基于预先设定好的量化规则集合对应的 N 个待量化结构，在模型 151W 的模型结构图中查询是否存在属于 N 个待量化结构的目标结构。

25 由于该模型 151W 的模型结构图中的结构 15G1 属于可加速的待量化结构，因此，该计算机设备可以从量化规则集合中的 N 个量化规则中确定与结构 15G1 相匹配的目标量化规则（例如，上述图 2 所示的量化规则 R4），进一步地，该计算机设备可以基于目标量化规则以及伪造算子，对结构 15G1 进行量化，以得到量化后的目标结构（例如，图 15 所示的结构 15G2）。其中，该计算机设备对结构 15G1 进行量化的具体实施方式可以参见上述
30 图 5 所示的对结构 50G 进行量化的具体实施方式，这里将不再对其进行赘述。

在量化结束后，该计算机设备可以得到图 15 所示的模型 152W（即第一模型），进一步地，该计算机设备可以对模型 152W 进行微调训练，进而可以将微调训练后的模型 152W 称之为第二模型，然后该计算机设备可以在第二模型中遍历待优化结构（即已经量化并训练的目标结构），进而可以根据待优化结构中的量化算子以及待优化结构中的原网络层，
5 得到优化结构。比如，该计算机设备在遍历到第二模型中的结构 15G2 时，可以参见上述图 10 所示的优化处理方式，对结构 15G2 进行优化处理，以得到结构 15G2 对应的优化结构（例如，图 15 所示的结构 15G3）。

在遍历完成时，该计算机设备可以将优化处理后的第二模型确定为第三模型，进而基于第三模型所部署的业务平台，对第三模型进行模型推理，得到用于业务处理的业务模型
10 （例如，图 15 所示的模型 153W）。

由此可见，本申请实施例在获取到初始模型时，通过通用型量化工具，能够自动对初始模型进行模型量化、微调训练以及模型推理等优化过程，从而使得后续得到的业务模型能够提升模型的每秒查询率，进而还可减少模型部署内存以及显存占用等。

可以理解的是，这里的业务模型可以适用于多种应用场景，因此，这里的业务模型的业务处理可以为关键部位检测处理，也可以为内容审核处理，还可以为图像识别处理，这里将不对其进行限定。为便于理解，本申请实施例可以以图像识别为例，用以阐述业务模型的应用过程，其中，该模型应用过程具体可以参见下述步骤 S204-步骤 S208 的描述。

步骤 S204，若业务模型的业务处理为图像识别，则在获取到待识别图像时，获取待识别图像的图像特征。

具体地，若业务模型的业务处理为图像识别，则该计算机设备在获取到待识别图像时，
20 可以对该待识别图像进行编码处理，以得到待识别图像的图像特征。

步骤 S205，将图像特征输入至业务模型。

具体地，该计算机设备可以将待识别图像的图像特征输入至业务模型，由于这里的业务模型是对高精度的初始模型进行模型量化、微调训练以及推理处理后所得到的模型，这
25 意味着该业务模型中存在优化结构（例如，图 15 所示的结构 15G3）。其中，这里的优化结构包括目标网络层和添加在目标网络层的输入位置处的量化算子（例如，图 15 所示的结构 15G3 的输入位置之前量化算子 1）；该目标网络层是根据量化算子，以及初始模型中的目标结构的原网络层所得到的。

步骤 S206，基于添加在目标网络层的输入位置处的量化算子，将优化结构的输入特征
30 的精度由第一精度转换为第二精度。

具体地，该计算机设备可以基于添加在目标网络层的输入位置处的量化算子，对优化结构的输入特征进行量化操作，以将输入特征的精度由第一精度转换为第二精度。其中，第一精度高于第二精度，这意味着本申请实施例可以通过量化算子，将输入至该优化结构的参数的精度由高精度转换低精度，从而达到加速的目的。

5 步骤 S207，通过目标网络层，对属于第二精度的输入特征进行特征提取处理，由目标网络层输出输入特征对应的输出特征；

其中，目标网络层的输出特征仍为第一精度。

步骤 S208，基于输出特征，得到待识别图像的预测类别。

具体地，该计算机设备可以将优化结构的输出特征（即目标网络层的输出特征）传输
10 至该业务模型中的优化结构的下一网络层，直到由该业务模型预测出该待识别图像针对多个分类类别的影响度，进而可以在这多个分类类别的影响度中选择最大影响度对应的分类类别作为该待识别图像的预测类别。

比如，若业务模型能够识别的分类类别的数量为 3 个，具体可以包括类别 1（例如，人）、类别 2（例如，猫），类型 3（例如，狗），则该计算机设备在通过该业务模型可
15 以预测出待识别图像针对类别 1 的影响度为 15%，针对类别 2 的影响度为 80%，针对类别 3 的影响度为 5%，然后，该计算机设备可以将最大影响度对应的分类类别（即类别 2），作为待识别图像的预测类别。

在本申请实施例中，初始模型对应的业务对象无需手动在初始模型中的相应位置处添加伪造算子，而是直接将任意一种初始模型提交至计算机设备，以使计算机设备调用通用
20 型量化工具，来获取预先设定好的量化规则集合，进而基于该通用型量化工具以及量化规则集合，对初始模型中存在的可加速的待量化结构进行量化，整个量化过程可在通用型量化工具内部完成，从而可以有效降低人工成本，提升了模型量化速度，实现了量化通用性。此外，由于该通用型量化工具还支持对量化后的初始模型进行微调训练，使得后续得到的业务模型在满足业务指标的同时还能够获取最高的加速比，即该计算机设备可以通过业务
25 模型，快速且准确的进行业务处理（例如，图像识别处理等）。

进一步地，请参见图 16，图 16 是本申请实施例提供的一种数据处理装置的结构示意图。如图 16 所示，该数据处理装置 1 可以包括：搜索模块 10，量化模块 20，业务模型确定模块 30，特征获取模块 40，特征输入模块 50，精度转换模块 60，特征提取模块 70 以及类别确定模块 80。

30 该搜索模块 10，用于在获取到初始模型时，对初始模型进行结构搜索，得到初始模型

的模型结构图。

该量化模块 20，用于根据模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对目标结构进行量化；伪造算子包括量化算子和反量化算子。

其中，该量化模块 20 包括：第一量化单元 201，第二量化单元 202，第三量化单元 203，
5 第四量化单元 204，第五量化单元 205，第六量化单元 206 以及第七量化单元 207。

该第一量化单元 201，用于在模型结构图中的目标结构包括具有权重的第一网络层的情况下，在具有权重的第一网络层的输入位置处，添加用于进行量化的伪造算子，且采用伪造算子对具有权重的第一网络层的权重进行量化。

该第二量化单元 202，用于在具有权重的第一网络层具有多个输出位置的情况下，在
10 具有权重的第一网络层的多个输出位置处，分别添加用于进行量化的伪造算子。

该第三量化单元 203，用于在具有权重的第一网络层具有一个输出位置，且具有权重的第一网络层的输出位置处未连接激活层的情况下，在具有权重的第一网络层的输出位置处，添加用于进行量化的伪造算子。

其中，在具有权重的第一网络层具有一个输出位置，且具有权重的第一网络层的输出
15 位置处连接激活层的情况下，不在具有权重的第一网络层的输出位置处，添加用于进行量化的伪造算子。

该第四量化单元 204，用于在模型结构图中的目标结构包括叠加层、叠加层的第一输入分支和叠加层的第二输入分支的情况下，在叠加层的输出位置处添加伪造算子；

其中，第一输入分支包括依次连接的具有权重的第一网络层和归一化层，归一化层与
20 叠加层连接，第二输入分支的类型为非加权操作的输入类型，具有权重的第一网络层为卷积层。

该第五量化单元 205，用于在第二输入分支的输入位置处添加伪造算子。

该第六量化单元 206，用于在模型结构图中的目标结构包括第一叠加层，且第一叠加层的第一输入分支的类型与第一叠加层的第二输入分支的类型均为非加权操作的输入类
25 型的情况下，在第一输入分支的输入位置处添加用于进行量化的伪造算子，在第二输入分支的输入位置处添加伪造算子，且在第一叠加层的输出位置处添加伪造算子。

该第七量化单元 207，用于在模型结构图中的目标结构包括第一拼接层，且第一拼接层的输入分支与第一拼接层的输出分支的数量相同的情况下，在第一拼接层的 N 个输出分支的每个输出分支的输出位置处，分别添加用于进行量化的伪造算子；N 为大于 1 的正整
30 数；每个输出分支处所添加的伪造算子的量化算子均相同。

其中，该第一量化单元 201，第二量化单元 202，第三量化单元 203，第四量化单元 204，第五量化单元 205，第六量化单元 206 以及第七量化单元 207 的具体实现方式可以参见上述图 3 所对应实施例中对步骤 S102 的描述，这里将不再继续进行赘述。

该业务模型确定模块 30，用于基于量化后的初始模型，得到用于业务处理的业务模型。

5 其中，该业务模型确定模块 30 包括：第一训练单元 301，第一确定单元 302，第一融合单元 303，第二融合单元 304，第二训练单元 305，第二确定单元 306，第三训练单元 307 以及第三确定单元 308。

该第一训练单元 301，用于将量化后的初始模型确定为第一模型，对第一模型进行微调训练，得到第二模型；

10 该第一确定单元 302，用于基于第二模型，得到用于业务处理的业务模型；业务模型包括具有权重的第二网络层、和添加在具有权重的第二网络层的输入位置处的量化算子；在业务模型中，具有权重的第二网络层是根据具有权重的第一网络层得到的，且具有权重的第二网络层的权重是采用量化算子，对第二模型中的具有权重的第一网络层的权重进行处理后所得到的。

15 其中，目标结构中的具有权重的第一网络层为卷积层，且目标结构还包括与卷积层连接的归一化层，以及与归一化层连接的激活层；

该第一融合单元 303，用于对具有权重的第一网络层、归一化层和激活层进行融合，得到具有权重的第二网络层。

20 其中，目标结构包括叠加层、叠加层的第一输入分支和叠加层的第二输入分支；第一输入分支包括依次连接的具有权重的第一网络层和归一化层；归一化层与叠加层连接，第二输入分支的类型为非加权操作的输入类型，具有权重的第一网络层为卷积层；

该第二融合单元 304，用于对具有权重的第一网络层、归一化层和叠加层进行融合，得到具有权重的第二网络层。

25 该第二训练单元 305，用于将量化后的初始模型确定为第一模型，对第一模型进行微调训练，得到第二模型；

该第二确定单元 306，用于基于第二模型，得到用于业务处理的业务模型；业务模型包括第二叠加层、添加在第二叠加层的第一输入分支处的量化算子、和添加在第二叠加层的第二输入分支处的量化算子；在业务模型中，第二叠加层是根据量化算子以及第二模型中的第一叠加层所得到的。

30 该第三训练单元 307，用于将量化后的初始模型确定为第一模型，对第一模型进行微

调训练，得到第二模型；

该第三确定单元 308，用于基于第二模型，得到用于业务处理的业务模型；业务模型包括第二拼接层、和添加在第二拼接层的每个输入分支的输入位置处的量化算子；在业务模型中，第二拼接层的每个输入分支的输入位置处的量化算子是将添加在第二模型中的第一拼接层的每个输出分支的输出位置处的量化算子，分别等价向前传播至第二模型中的第一拼接层的每个输入分支的输入位置处后所得到的。

其中，该第一训练单元 301，第一确定单元 302，第一融合单元 303，第二融合单元 304，第二训练单元 305，第二确定单元 306，第三训练单元 307 以及第三确定单元 308 的具体实现方式可以参见上述图 3 所对应实施例中对步骤 S103 的描述，这里将不再继续进行赘述。

该特征获取模块 40，用于若业务模型的业务处理为图像识别，则在获取到待识别图像时，获取待识别图像的图像特征。

该特征输入模块 50，用于将图像特征输入至业务模型；业务模型中的优化结构包括目标网络层、和添加在目标网络层的输入位置处的量化算子；目标网络层是根据量化算子，以及初始模型中的目标结构的原网络层所得到的。

该精度转换模块 60，用于基于添加在目标网络层的输入位置处的量化算子，将优化结构的输入特征的精度由第一精度转换为第二精度；第一精度高于第二精度。

该特征提取模块 70，用于通过目标网络层，对属于第二精度的输入特征进行特征提取处理，由目标网络层输出输入特征对应的输出特征。

该类别确定模块 80，用于基于输出特征，得到待识别图像的预测类别。

其中，该搜索模块 10，量化模块 20，业务模型确定模块 30，特征获取模块 40，特征输入模块 50，精度转换模块 60，特征提取模块 70 以及类别确定模块 80 的具体实现方式可以参见上述图 14 所对应实施例中对步骤 S201-步骤 S208 的描述，这里将不再继续进行赘述。另外，对采用相同方法的有益效果描述，也不再赘述。

进一步地，请参见图 17，图 17 是本申请实施例提供的一种计算机设备的示意图。如图 17 所示，该计算机设备 1000 可以包括：至少一个处理器 1001，例如，CPU，至少一个网络接口 1004，存储器 1005，至少一个通信总线 1002。其中，通信总线 1002 用于实现这些组件之间的连接通信。网络接口 1004 可选地可以包括标准的有线接口、无线接口（如 WI-FI 接口）。存储器 1005 可以是高速 RAM 存储器，也可以是非不稳定的存储器（non-volatile memory），例如至少一个磁盘存储器。存储器 1005 可选地还可以是至少一

个位于远离前述处理器 1001 的存储装置。如图 17 所示，作为一种计算机存储介质的存储器 1005 可以包括操作系统、网络通信模块、用户接口模块以及设备控制应用程序。其中，在一些实施例中，该计算机设备还可以包括图 17 所示的用户接口 1003，比如，若该计算机设备为图 1 所示的具有模型量化功能的终端设备（例如，终端设备 100b），则该计算机设备还可以包括该用户接口 1003，其中，该用户接口 1003 可以包括显示屏（Display）、键盘（Keyboard）等。

在图 17 所示的计算机设备 1000 中，网络接口 1004 主要用于进行网络通信；而用户接口 1003 主要用于为用户提供输入的接口；而处理器 1001 可以用于调用存储器 1005 中存储的设备控制应用程序，以实现：

在获取到初始模型时，对初始模型进行结构搜索，得到初始模型的模型结构图；

根据模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对目标结构进行量化；伪造算子包括量化算子和反量化算子；

基于量化后的初始模型，得到用于业务处理的业务模型。

应当理解，本申请实施例中所描述的计算机设备 1000 可执行前文图 3 和图 14 所对应实施例中对该数据处理方法的描述，也可执行前文图 16 所对应实施例中对该数据处理装置 1 的描述，在此不再赘述。另外，对采用相同方法的有益效果描述，也不再进行赘述。

本申请实施例还提供一种计算机可读存储介质，该计算机可读存储介质存储有计算机程序，该计算机程序包括程序指令，该程序指令被处理器执行时实现图 3 和图 14 中各个步骤所提供的数据处理方法，具体可参见图 3 以及图 14 各个步骤所提供的实现方式，在此不再赘述。

计算机可读存储介质可以是前述任一实施例提供的数据传输装置或者计算机设备的内部存储单元，例如计算机设备的硬盘或内存。该计算机可读存储介质也可以是该计算机设备的外部存储设备，例如该计算机设备上配备的插接式硬盘，智能存储卡（smart media card, SMC），安全数字（secure digital, SD）卡，闪存卡（flash card）等。进一步地，该计算机可读存储介质还可以既包括该计算机设备的内部存储单元也包括外部存储设备。该计算机可读存储介质用于存储该计算机程序以及该计算机设备所需的其他程序和数据。该计算机可读存储介质还可以用于暂时地存储已经输出或者将要输出的数据。

本申请实施例还提供了一种计算机程序产品，该计算机程序产品包括计算机程序，该计算机程序存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机程序，处理器执行该计算机程序，使得该计算机设备可执行前文各实施例中对

数据处理方法或装置的描述，在此不再赘述。另外，对采用相同方法的有益效果描述，也不再赘述。

本申请实施例的说明书和权利要求书及附图中的术语“第一”、“第二”等是用于区别不同对象，而非用于描述特定顺序。此外，术语“包括”以及它们任何变形，意图在于覆盖不排除的包括。例如包括了一系列步骤或单元的过程、方法、装置、产品或设备没有5 限定于已列出的步骤或模块，而是可选地还包括没有列出的步骤或模块，或可选地还包括对于这些过程、方法、装置、产品或设备固有的其他步骤单元。

本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和10 软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

以上所揭露的仅为本申请较佳实施例而已，当然不能以此来限定本申请之权利范围，15 因此依本申请权利要求所作的等同变化，仍属本申请所涵盖的范围。

权 利 要 求 书

1.一种数据处理方法，其特征在于，包括：

在获取到初始模型时，对所述初始模型进行结构搜索，得到所述初始模型的模型结构图；

5 根据所述模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对所述目标结构进行量化，所述伪造算子包括量化算子和反量化算子；

基于量化后的初始模型，得到用于业务处理的业务模型。

2.根据权利要求1所述的数据处理方法，其特征在于，所述根据所述模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对所述目标结构进行量化，包括：

10 在所述模型结构图中的目标结构包括具有权重的第一网络层的情况下，在所述具有权重的第一网络层的输入位置处，添加所述伪造算子，且采用所述伪造算子对所述具有权重的第一网络层的权重进行量化。

15 3.根据权利要求2所述的数据处理方法，其特征在于，所述数据处理方法还包括：

在所述具有权重的第一网络层具有多个输出位置的情况下，在所述具有权重的第一网络层的多个输出位置处，分别添加所述伪造算子。

4.根据权利要求2所述的数据处理方法，其特征在于，所述数据处理方法还包括：

20 在所述具有权重的第一网络层具有一个输出位置，且所述具有权重的第一网络层的输出位置处未连接激活层的情况下，在所述具有权重的第一网络层的输出位置处，添加所述伪造算子。

5.根据权利要求2所述的数据处理方法，其特征在于，所述数据处理方法还包括：

25 在所述具有权重的第一网络层具有一个输出位置，且所述具有权重的第一网络层的输出位置处连接激活层的情况下，不在所述具有权重的第一网络层的输出位置处，添加所述伪造算子。

6.根据权利要求2所述的数据处理方法，其特征在于，所述数据处理方法还包括：

30 在所述模型结构图中的目标结构包括叠加层、所述叠加层的第一输入分支和所述叠加

层的第二输入分支的情况下，在所述叠加层的输出位置处添加所述伪造算子；

其中，所述第一输入分支包括依次连接的所述具有权重的第一网络层和归一化层，所述归一化层与所述叠加层连接，所述第二输入分支的类型为非加权操作的输入类型，所述具有权重的第一网络层为卷积层。

5

7.根据权利要求6所述的数据处理方法，其特征在于，所述数据处理方法还包括：
在所述第二输入分支的输入位置处添加所述伪造算子。

8.根据权利要求1所述的数据处理方法，其特征在于，所述根据所述模型结构图中的
10 目标结构的种类，采用用于进行量化的伪造算子，对所述目标结构进行量化，包括：

在所述模型结构图中的目标结构包括第一叠加层，且所述第一叠加层的第一输入分支的类型与所述第一叠加层的第二输入分支的类型均为非加权操作的输入类型的情况下，在所述第一输入分支的输入位置处添加所述伪造算子，在所述第二输入分支的输入位置处添加所述伪造算子，且在所述第一叠加层的输出位置处添加所述伪造算子。

15

9.根据权利要求1所述的数据处理方法，其特征在于，所述根据所述模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对所述目标结构进行量化，包括：

在所述模型结构图中的目标结构包括第一拼接层，且所述第一拼接层的输入分支与所述第一拼接层的输出分支的数量相同的情况下，在所述第一拼接层的N个输出分支的每个
20 输出分支的输出位置处，分别添加所述伪造算子，N为大于1的正整数，每个输出分支处所添加的所述伪造算子的量化算子均相同。

10.根据权利要求2所述的数据处理方法，其特征在于，所述基于量化后的初始模型，得到用于业务处理的业务模型，包括：

25 将量化后的初始模型确定为第一模型，对所述第一模型进行微调训练，得到第二模型；
基于所述第二模型，得到用于业务处理的业务模型，

其中，所述业务模型包括具有权重的第二网络层、和添加在所述具有权重的第二网络层的输入位置处的量化算子，在所述业务模型中，所述具有权重的第二网络层是根据具有权重的第一网络层得到的，且所述具有权重的第二网络层的权重是采用量化算子，对所述
30 第二模型中的具有权重的第一网络层的权重进行处理后所得到的。

11.根据权利要求 10 所述的数据处理方法，其特征在于，所述目标结构中的具有权重的第一网络层为卷积层，且所述目标结构还包括与所述卷积层连接的归一化层，以及与所述归一化层连接的激活层；

5 所述数据处理方法还包括：

对所述具有权重的第一网络层、所述归一化层和所述激活层进行融合，得到所述具有权重的第二网络层。

12.根据权利要求 10 所述的数据处理方法，其特征在于，所述目标结构包括叠加层、
10 所述叠加层的第一输入分支和所述叠加层的第二输入分支；所述第一输入分支包括依次连接的所述具有权重的第一网络层和归一化层；所述归一化层与所述叠加层连接，所述第二输入分支的类型为非加权操作的输入类型，所述具有权重的第一网络层为卷积层；

所述数据处理方法还包括：

15 对所述具有权重的第一网络层、所述归一化层和所述叠加层进行融合，得到所述具有权重的第二网络层。

13.根据权利要求 8 所述的数据处理方法，其特征在于，所述基于量化后的初始模型，得到用于业务处理的业务模型，包括：

将量化后的初始模型确定为第一模型，对所述第一模型进行微调训练，得到第二模型；

20 基于所述第二模型，得到用于业务处理的业务模型，

其中，所述业务模型包括第二叠加层、添加在所述第二叠加层的第一输入分支处的量化算子、和添加在所述第二叠加层的第二输入分支处的量化算子，在所述业务模型中，所述第二叠加层是根据量化算子以及所述第二模型中的第一叠加层所得到的。

25 14.根据权利要求 9 所述的数据处理方法，其特征在于，所述基于量化后的初始模型，得到用于业务处理的业务模型，包括：

将量化后的初始模型确定为第一模型，对所述第一模型进行微调训练，得到第二模型；

基于所述第二模型，得到用于业务处理的业务模型，

30 其中，所述业务模型包括第二拼接层、和添加在所述第二拼接层的每个输入分支的输入位置处的量化算子，在所述业务模型中，所述第二拼接层的每个输入分支的输入位置处

的量化算子是将添加在所述第二模型中的第一拼接层的每个输出分支的输出位置处的量化算子，分别等价向前传播至所述第二模型中的第一拼接层的每个输入分支的输入位置处后所得到的。

5 15.根据权利要求1所述的数据处理方法，其特征在于，所述数据处理方法还包括：

若所述业务模型的业务处理为图像识别，则在获取到待识别图像时，获取所述待识别图像的图像特征；

10 将所述图像特征输入至所述业务模型；所述业务模型中的优化结构包括目标网络层、和添加在所述目标网络层的输入位置处的量化算子；所述目标网络层是根据量化算子，以及所述初始模型中的所述目标结构的原网络层所得到的；

基于添加在所述目标网络层的输入位置处的量化算子，将所述优化结构的输入特征的精度由第一精度转换为第二精度，所述第一精度高于所述第二精度；

通过所述目标网络层，对属于所述第二精度的输入特征进行特征提取处理，由所述目标网络层输出所述输入特征对应的输出特征；

15 基于所述输出特征，得到所述待识别图像的预测类别。

16.一种数据处理装置，其特征在于，包括：

搜索模块，用于在获取到初始模型时，对所述初始模型进行结构搜索，得到所述初始模型的模型结构图；

20 量化模块，用于根据所述模型结构图中的目标结构的种类，采用用于进行量化的伪造算子，对所述目标结构进行量化，所述伪造算子包括量化算子和反量化算子；

业务模型确定模块，用于基于量化后的初始模型，得到用于业务处理的业务模型。

17.一种计算机设备，其特征在于，包括：处理器和存储器以及网络接口；

25 所述处理器与所述存储器、所述网络接口相连，其中，所述网络接口用于提供数据通信功能，所述存储器用于存储计算机程序，所述处理器用于调用所述计算机程序，以使得所述计算机设备执行权利要求1至15任一项所述的数据处理方法。

18.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有计算机程序，所述计算机程序适于由处理器加载并执行，以使得具有所述处理器的计算机设备

30

执行权利要求 1 至 15 任一项所述的数据处理方法。

- 19.一种计算机程序产品，其特征在于，所述计算机程序产品包括计算机程序，所述计算机程序存储在计算机可读存储介质中，所述计算机程序适于由处理器读取并执行，以使
- 5 得具有所述处理器的计算机设备执行权利要求 1 至 15 任一项所述的数据处理方法。

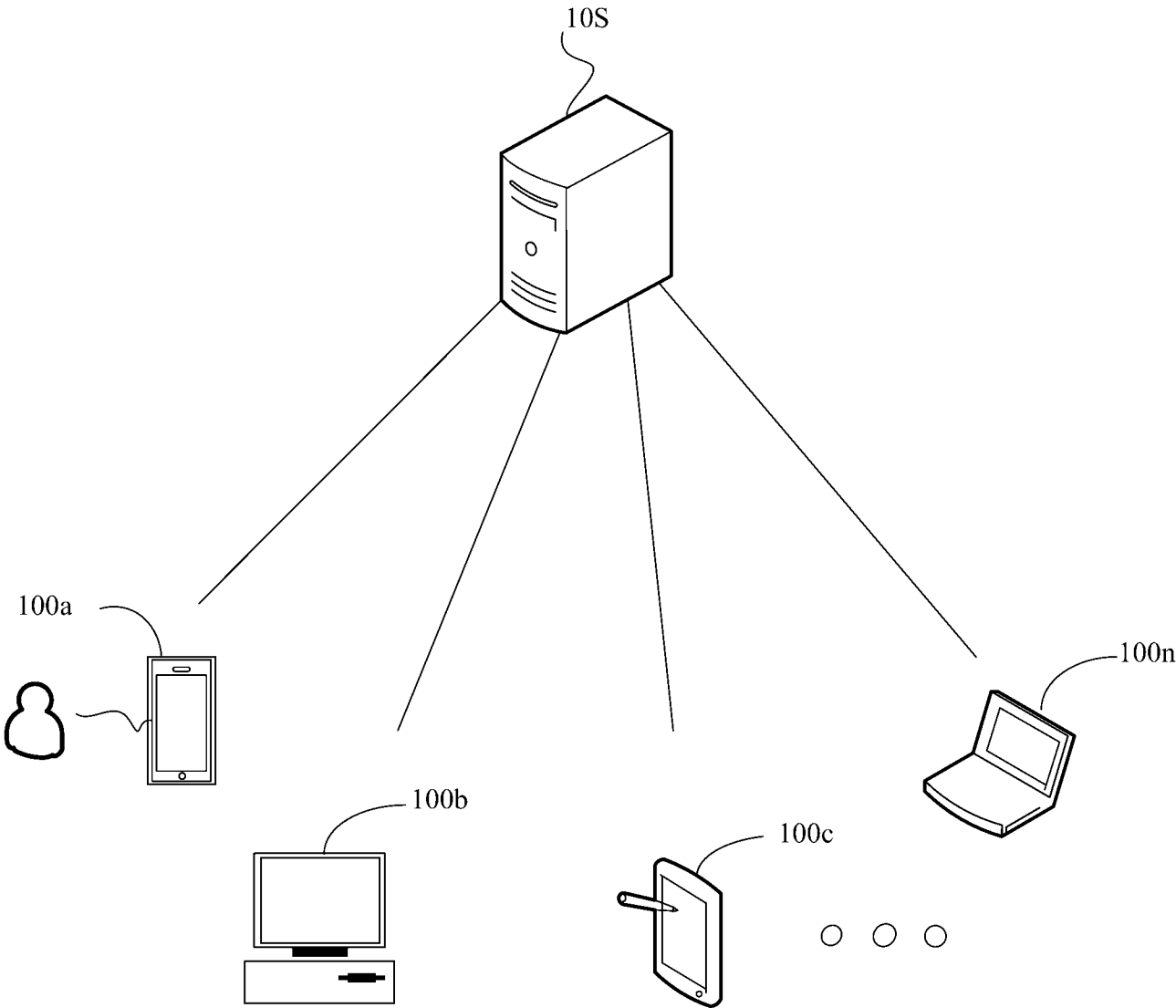


图 1

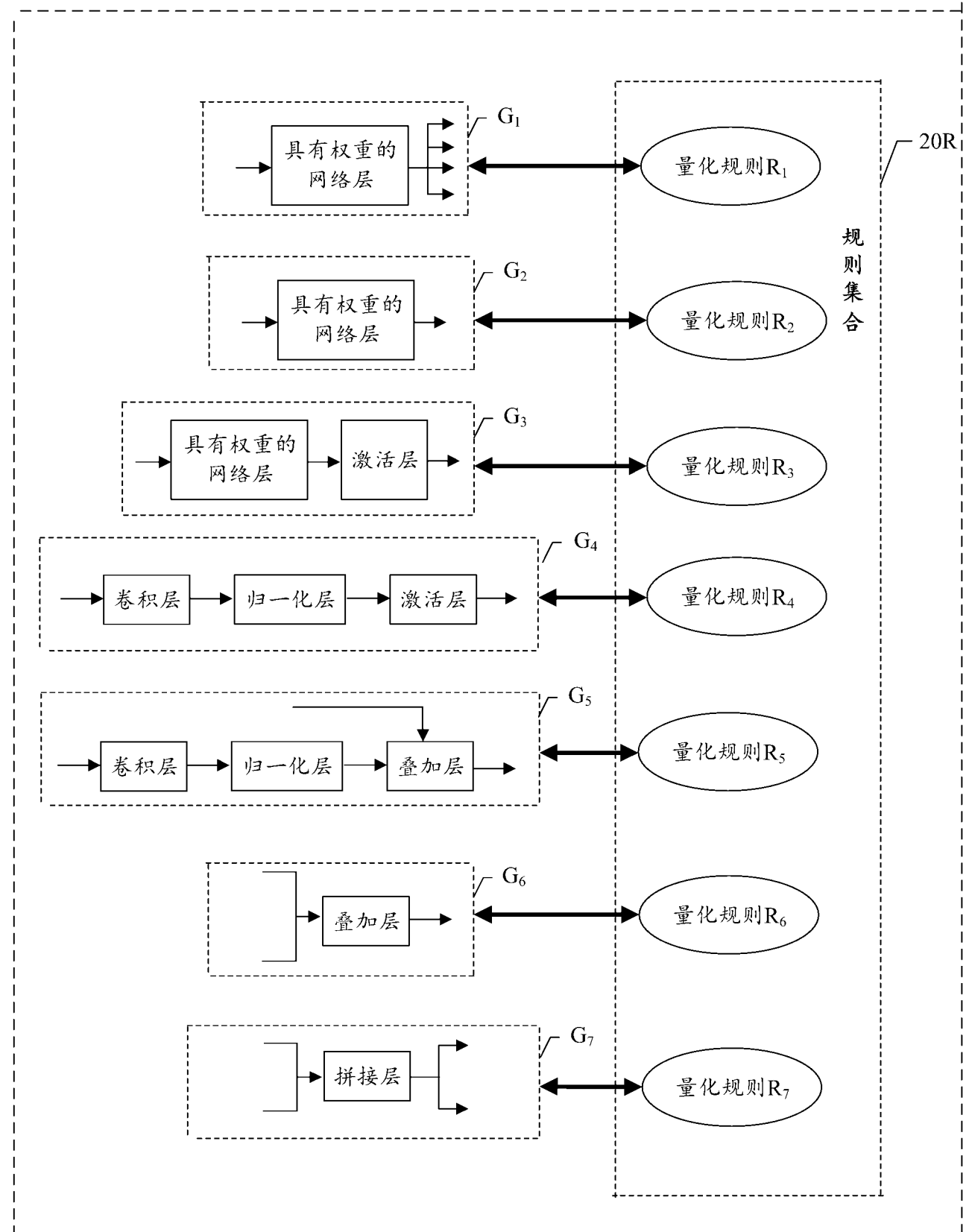


图 2

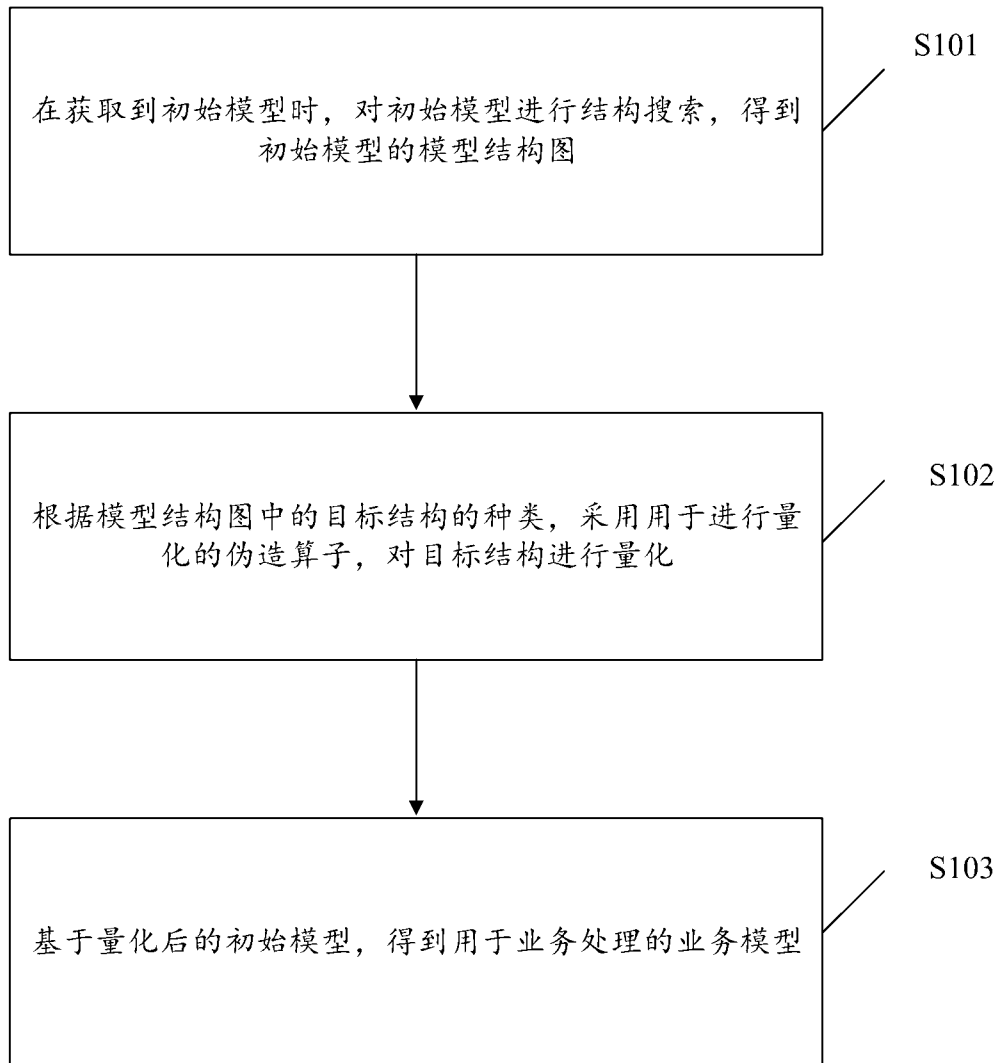


图 3

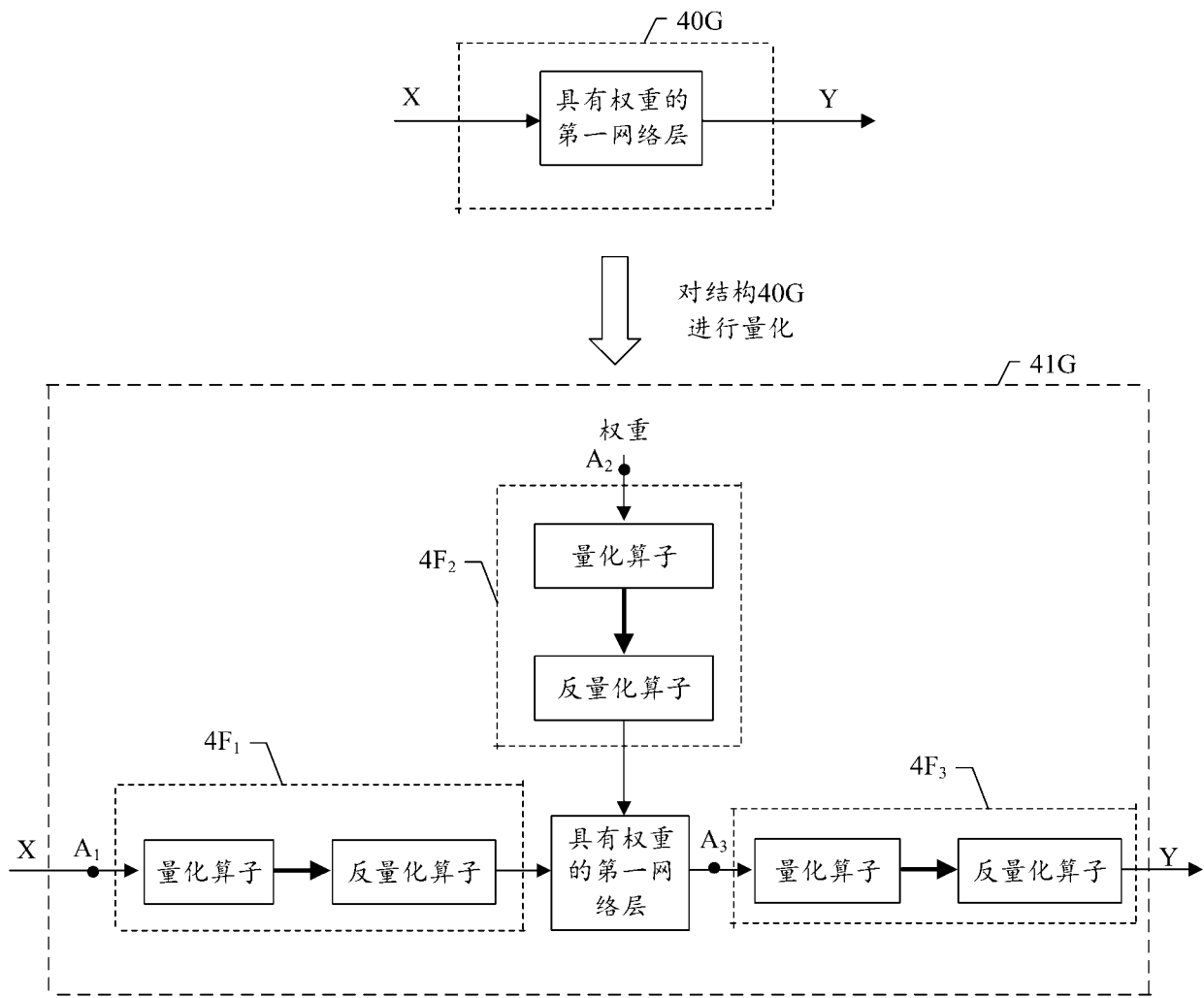


图 4

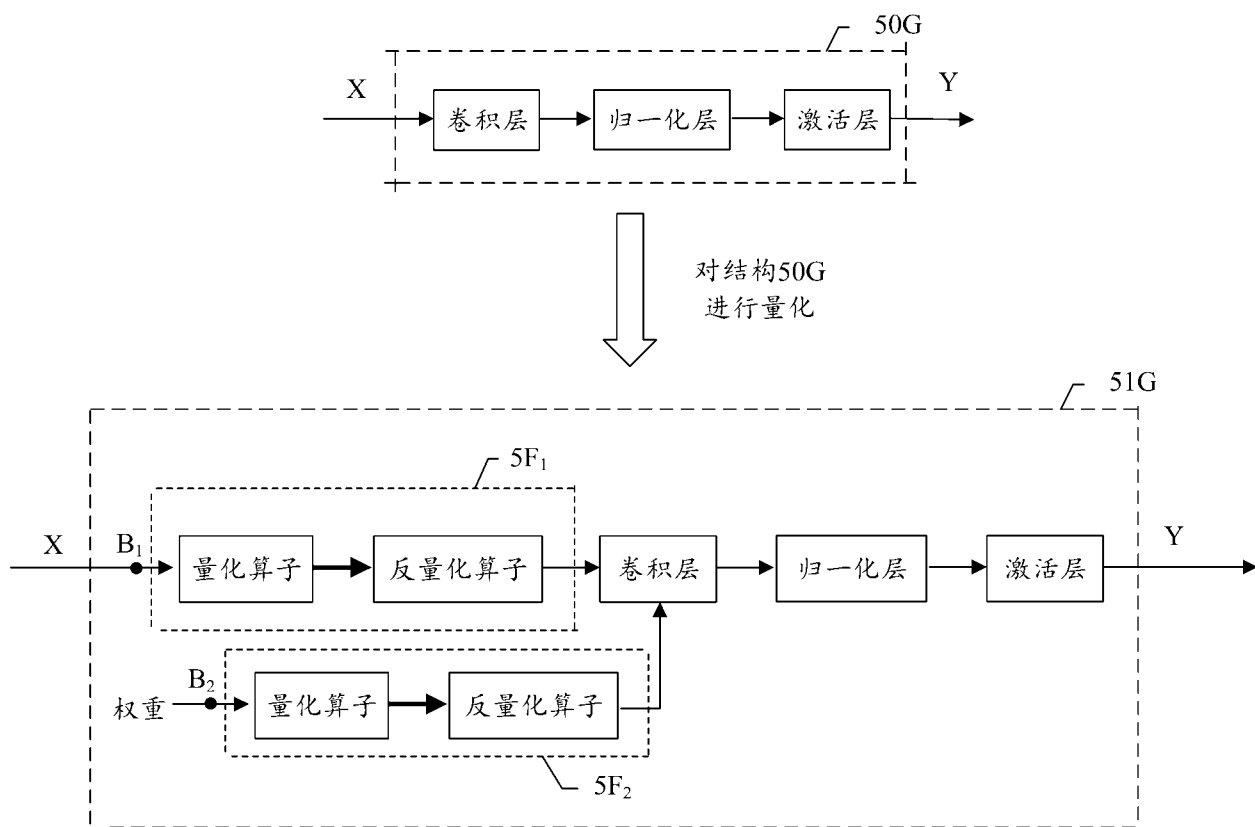


图 5

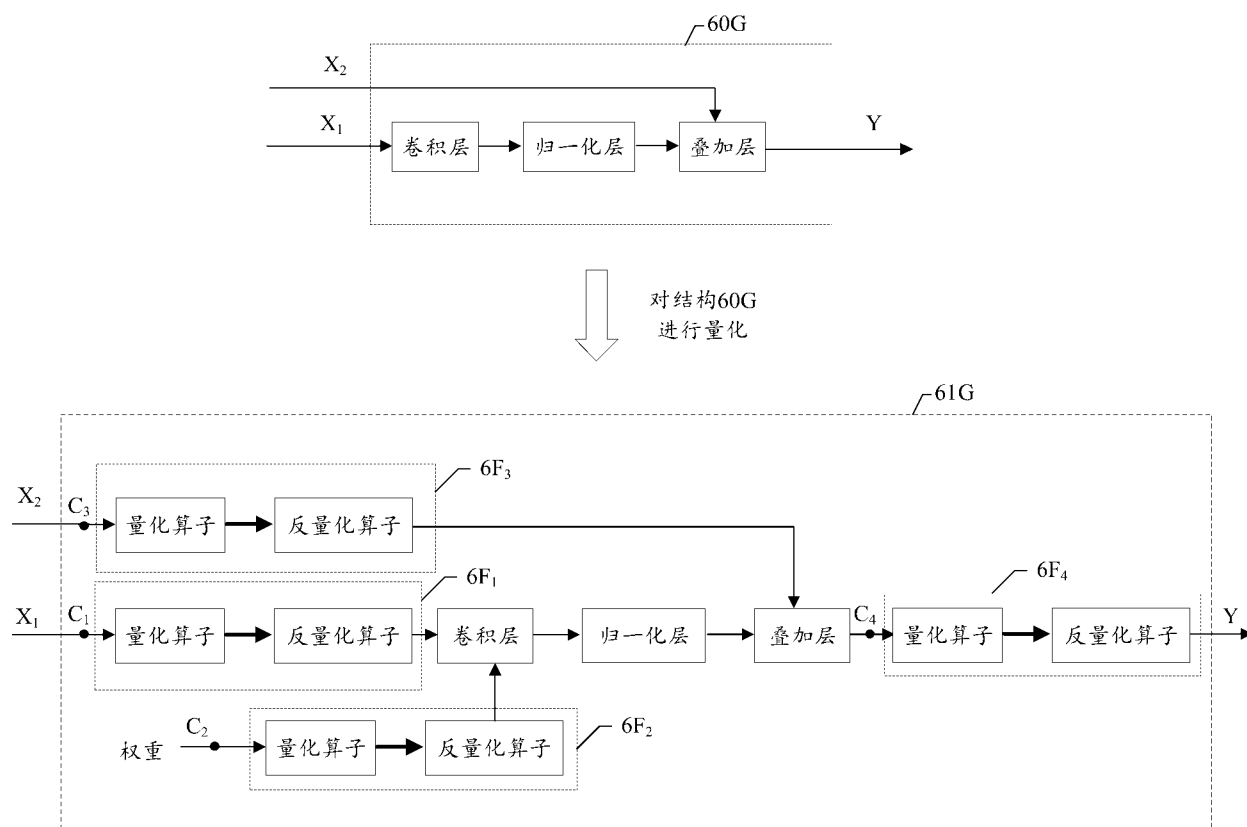


图 6

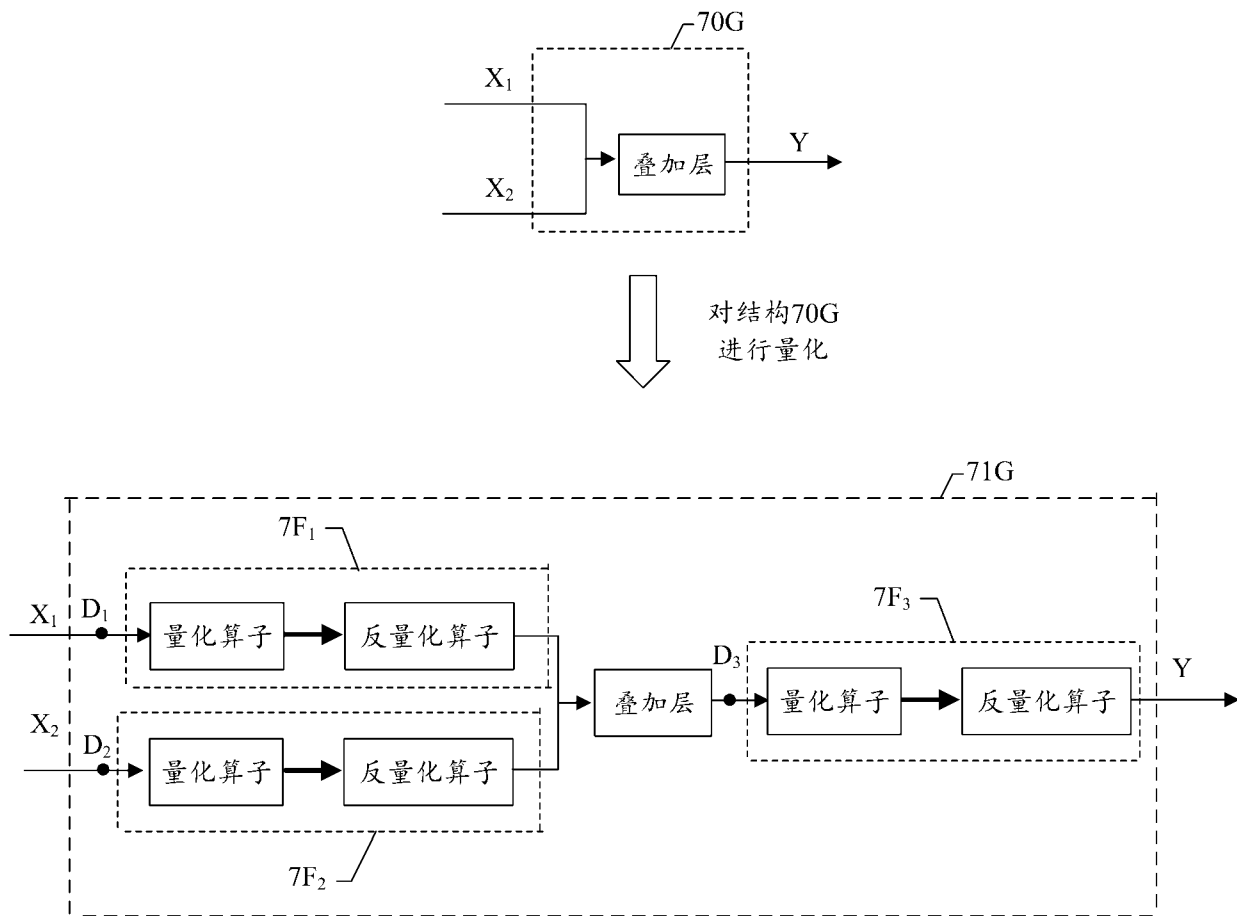


图 7

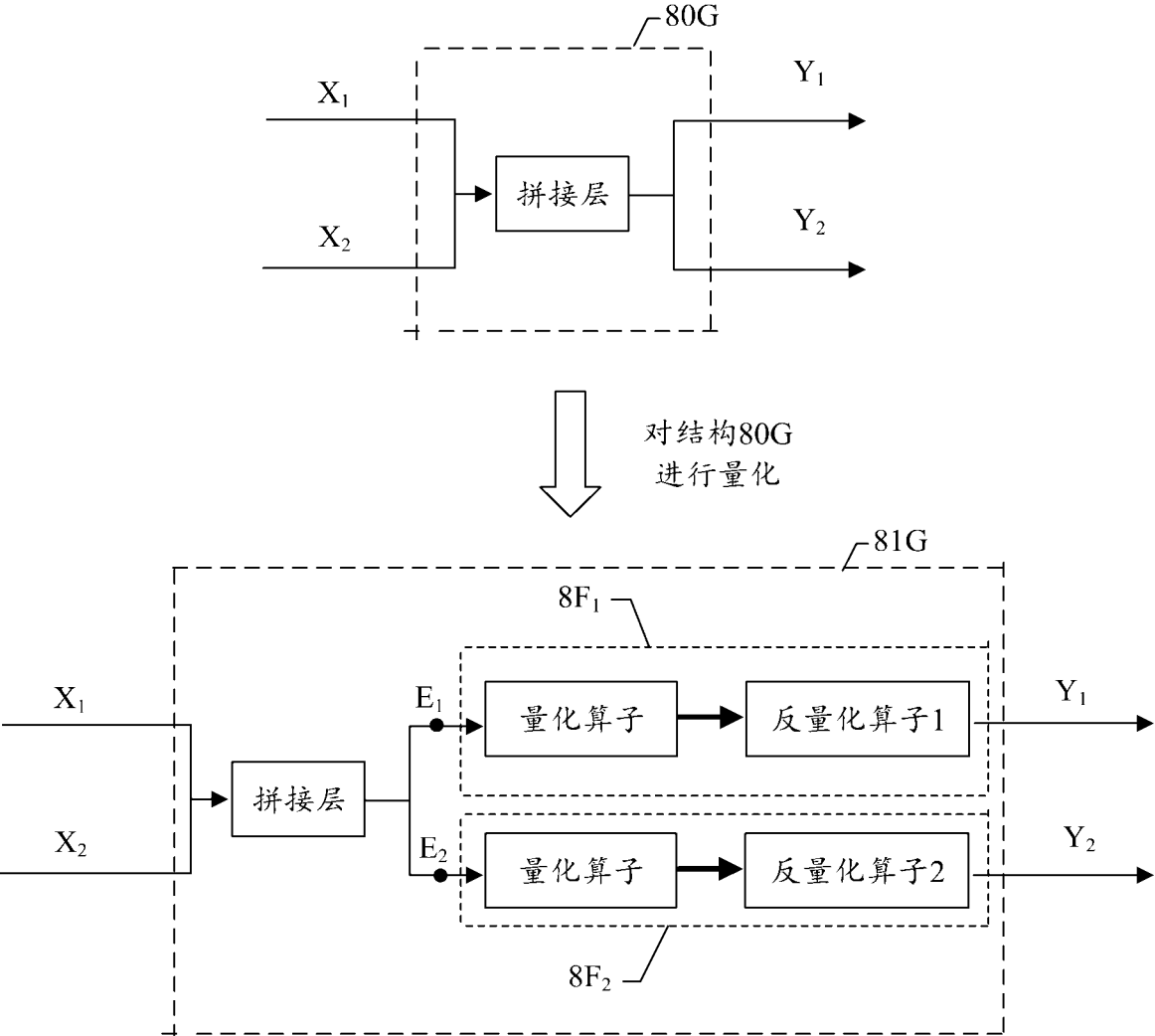


图 8

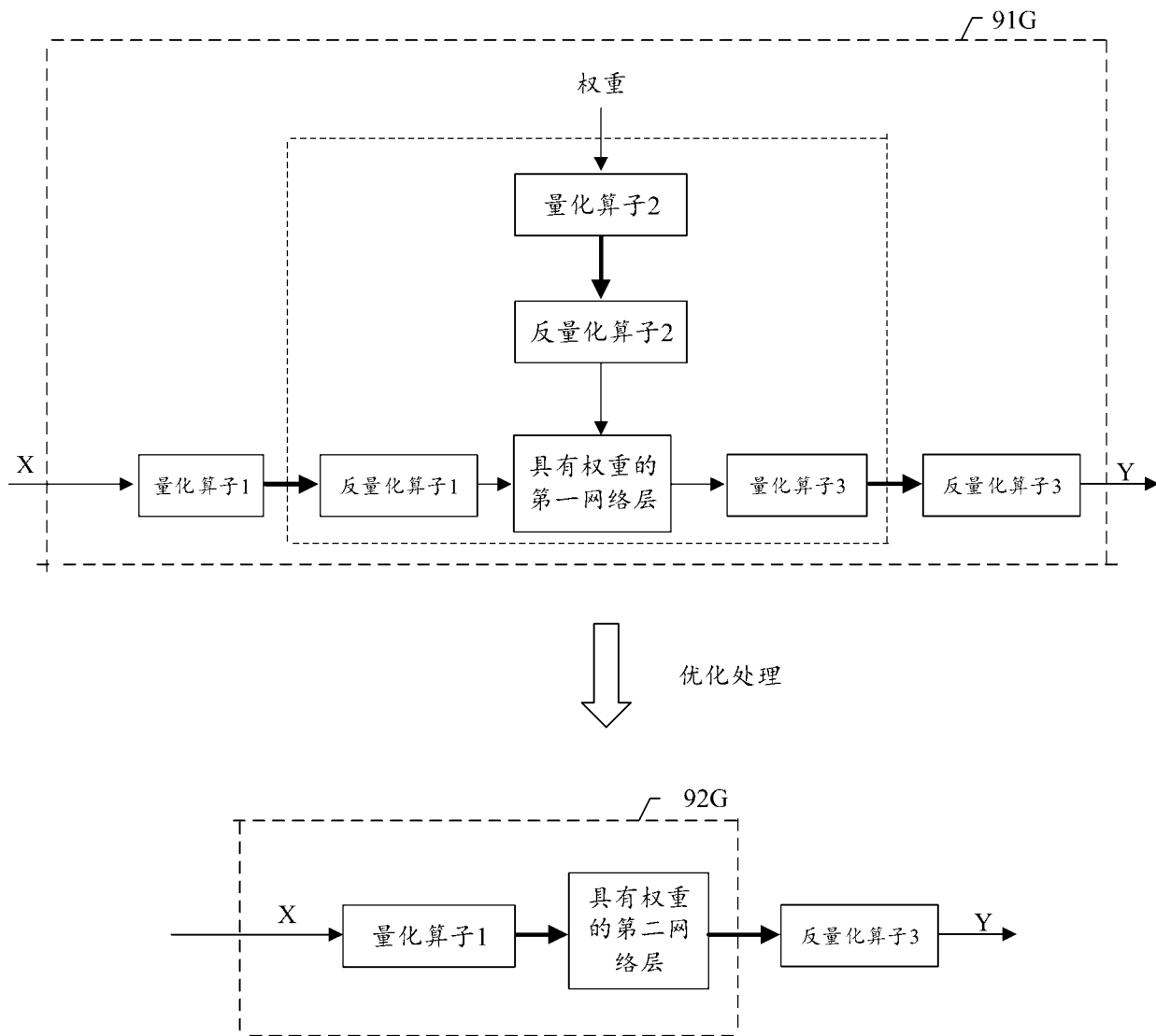


图 9

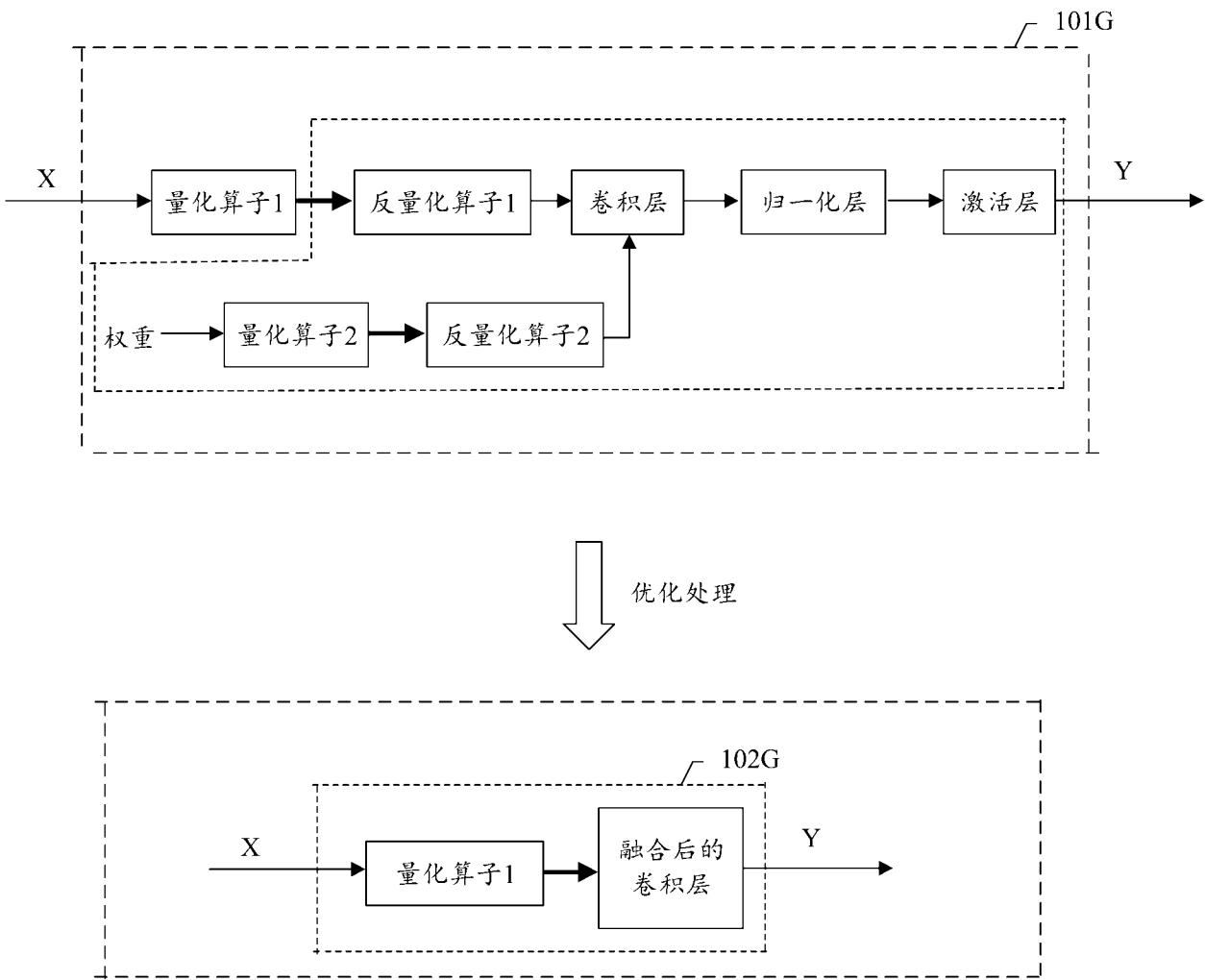


图 10

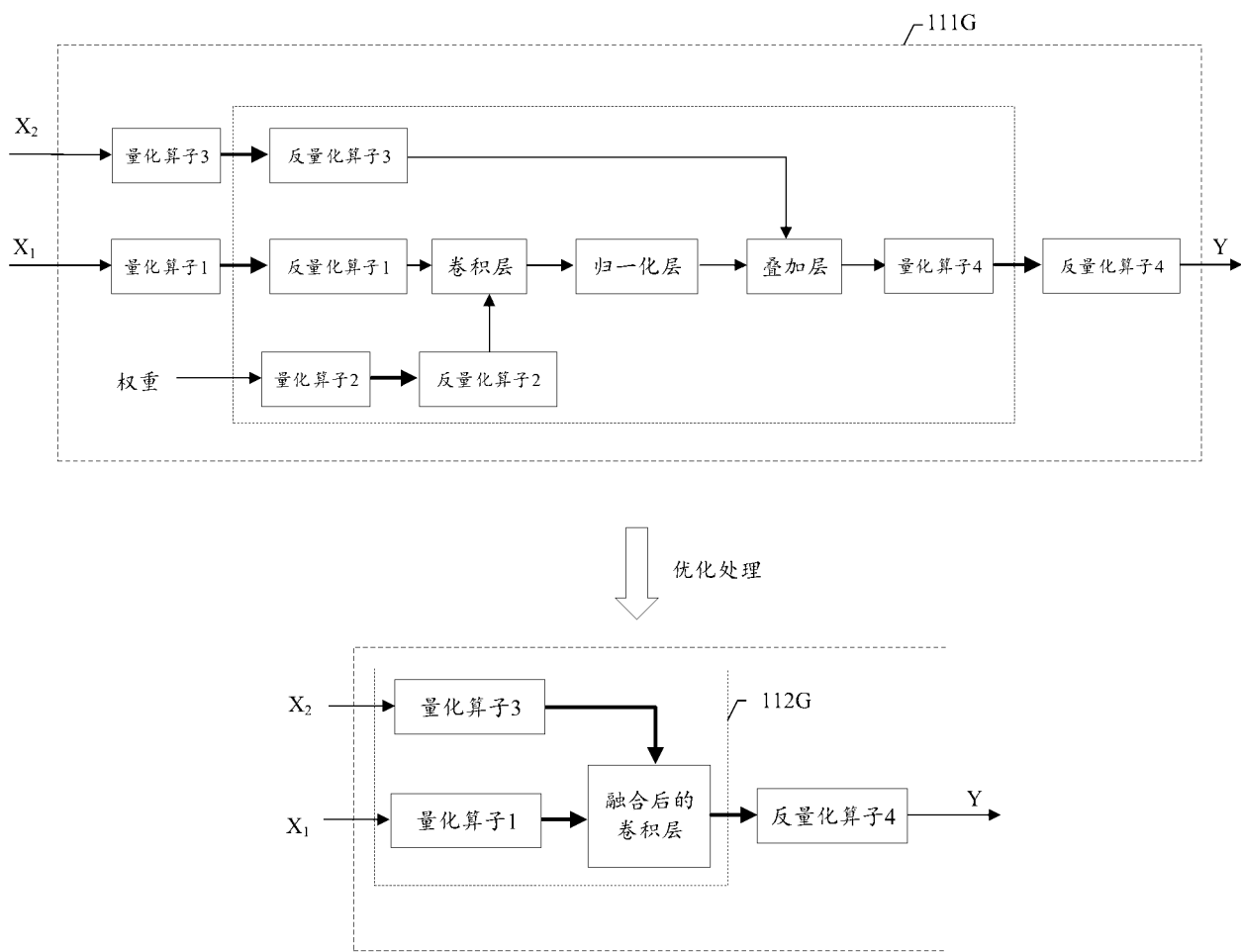


图 11

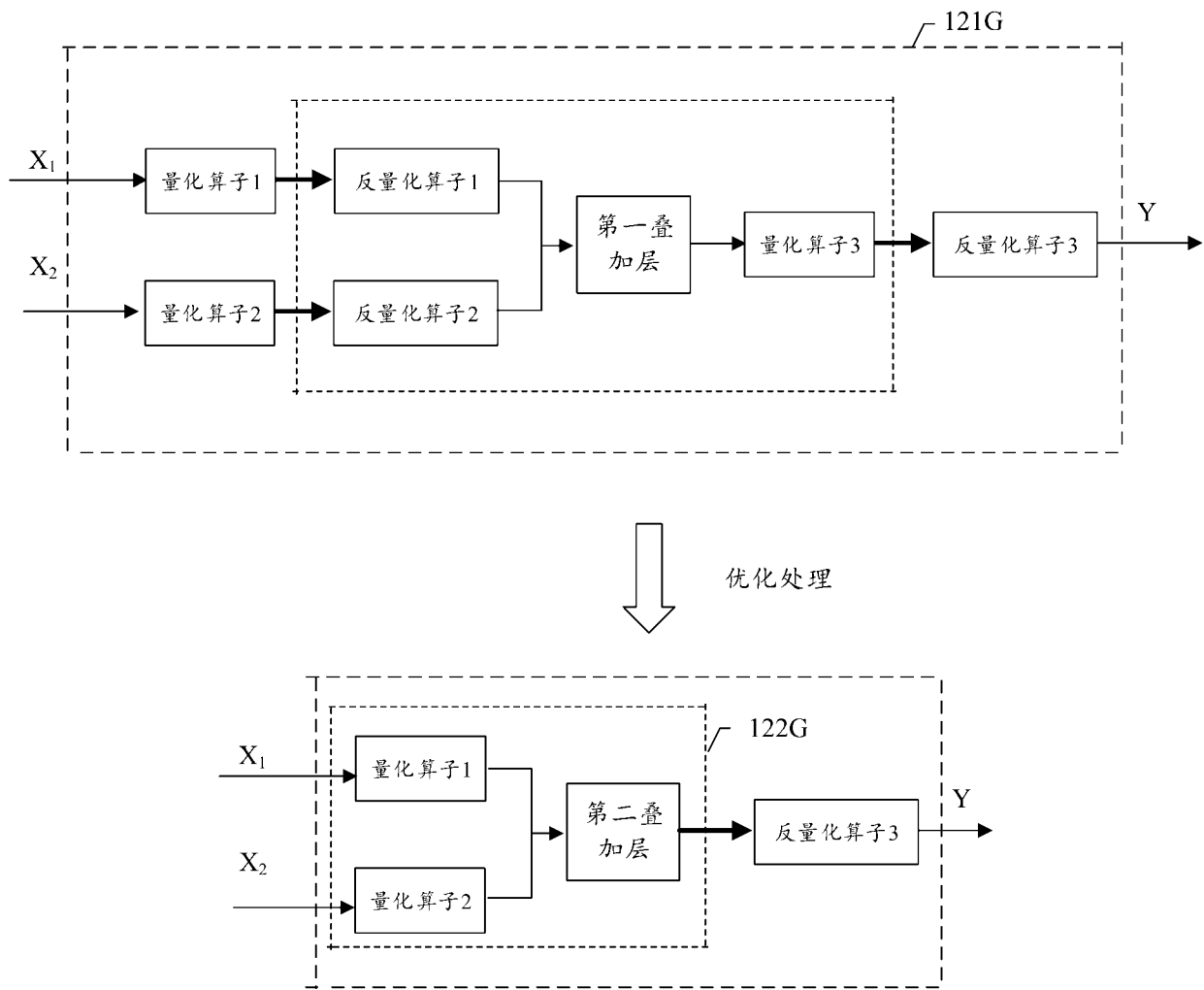
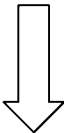
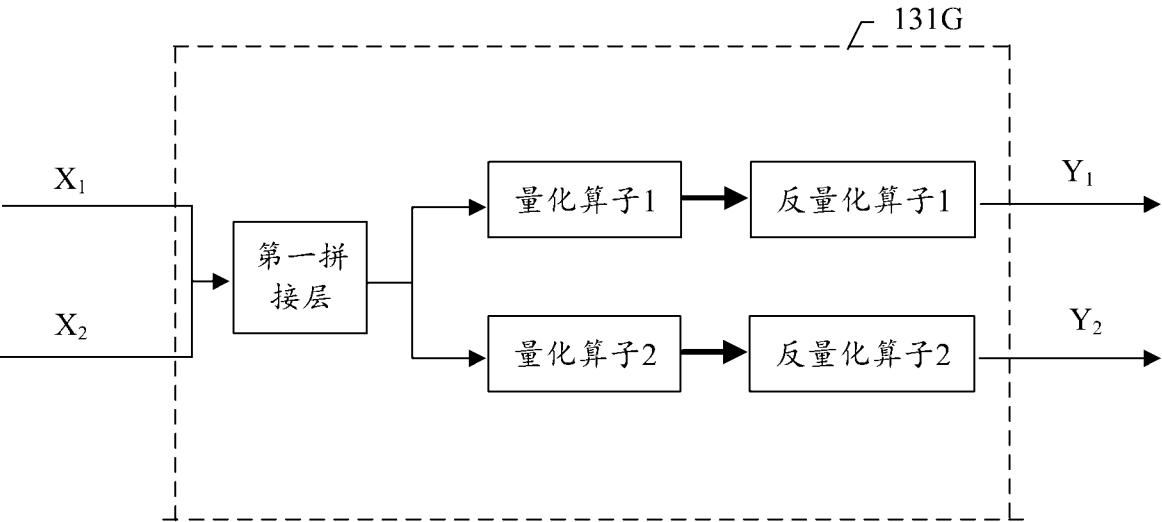


图 12



其中，量化算子1
等于量化算子2

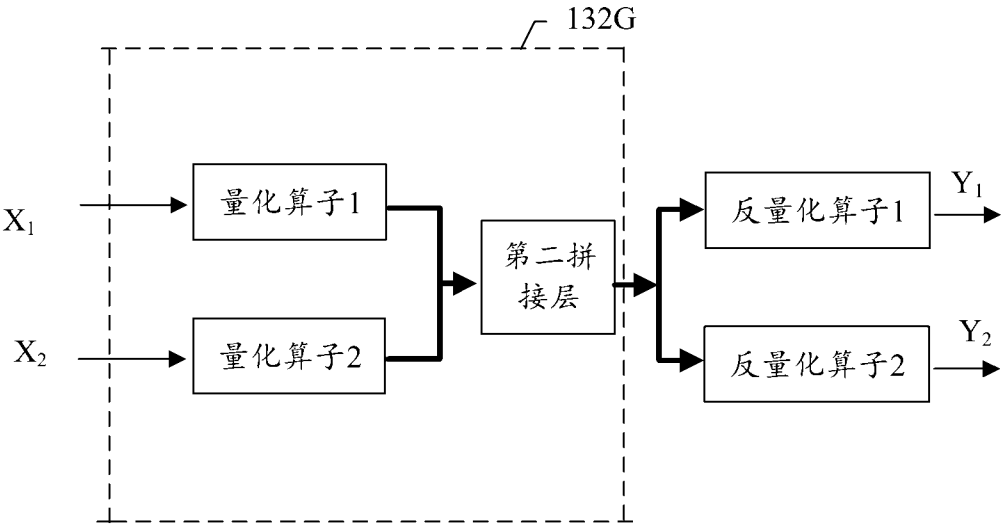


图 13

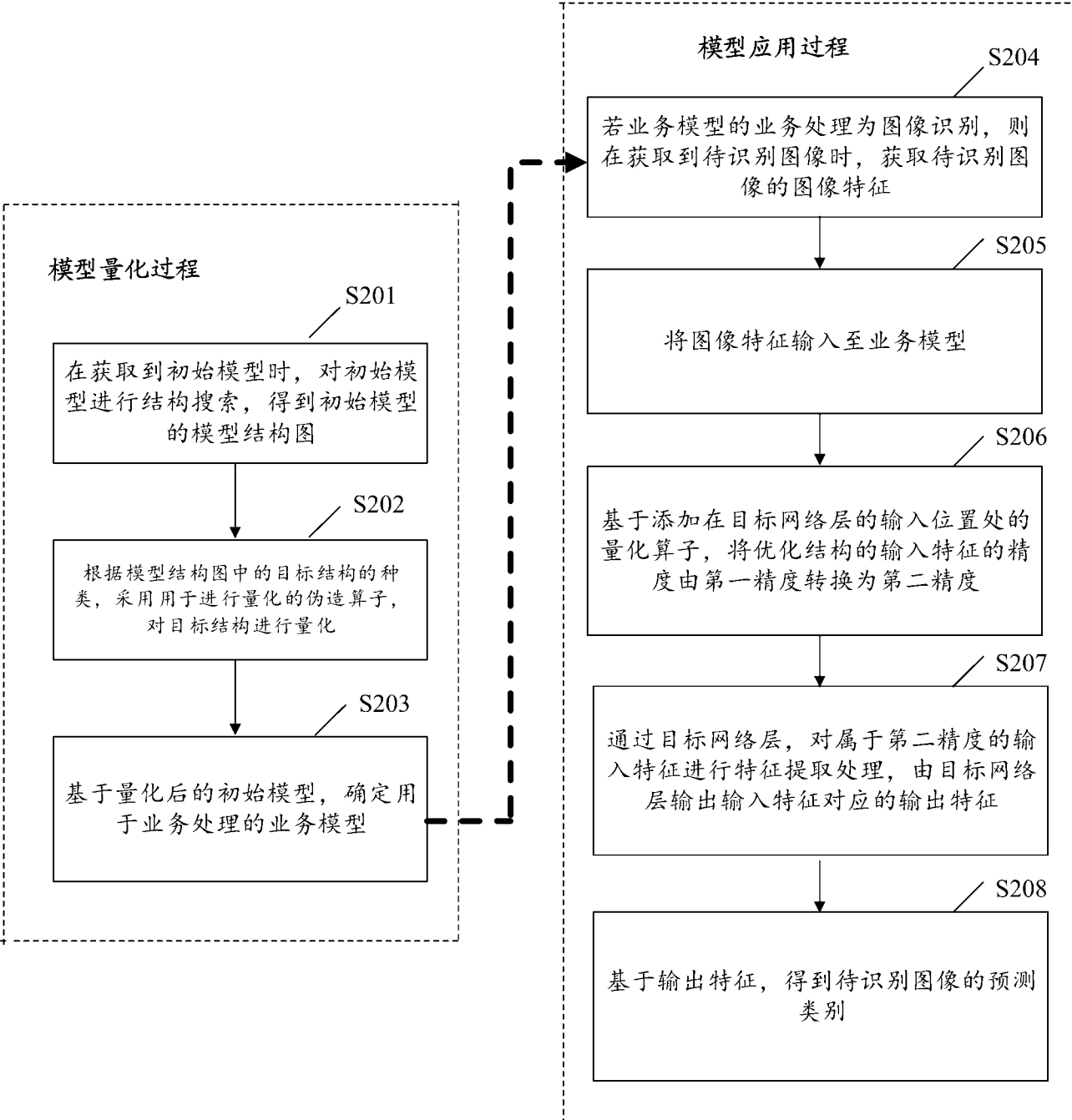


图 14

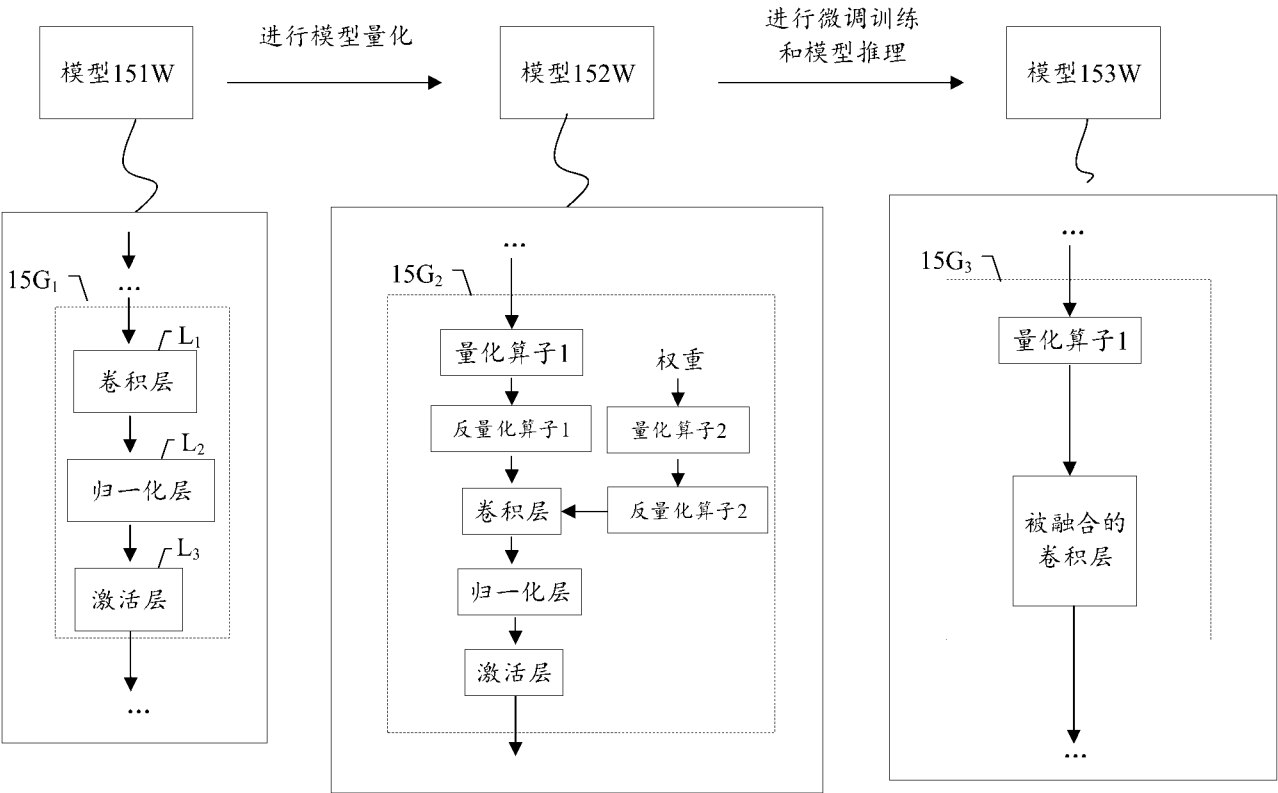


图 15

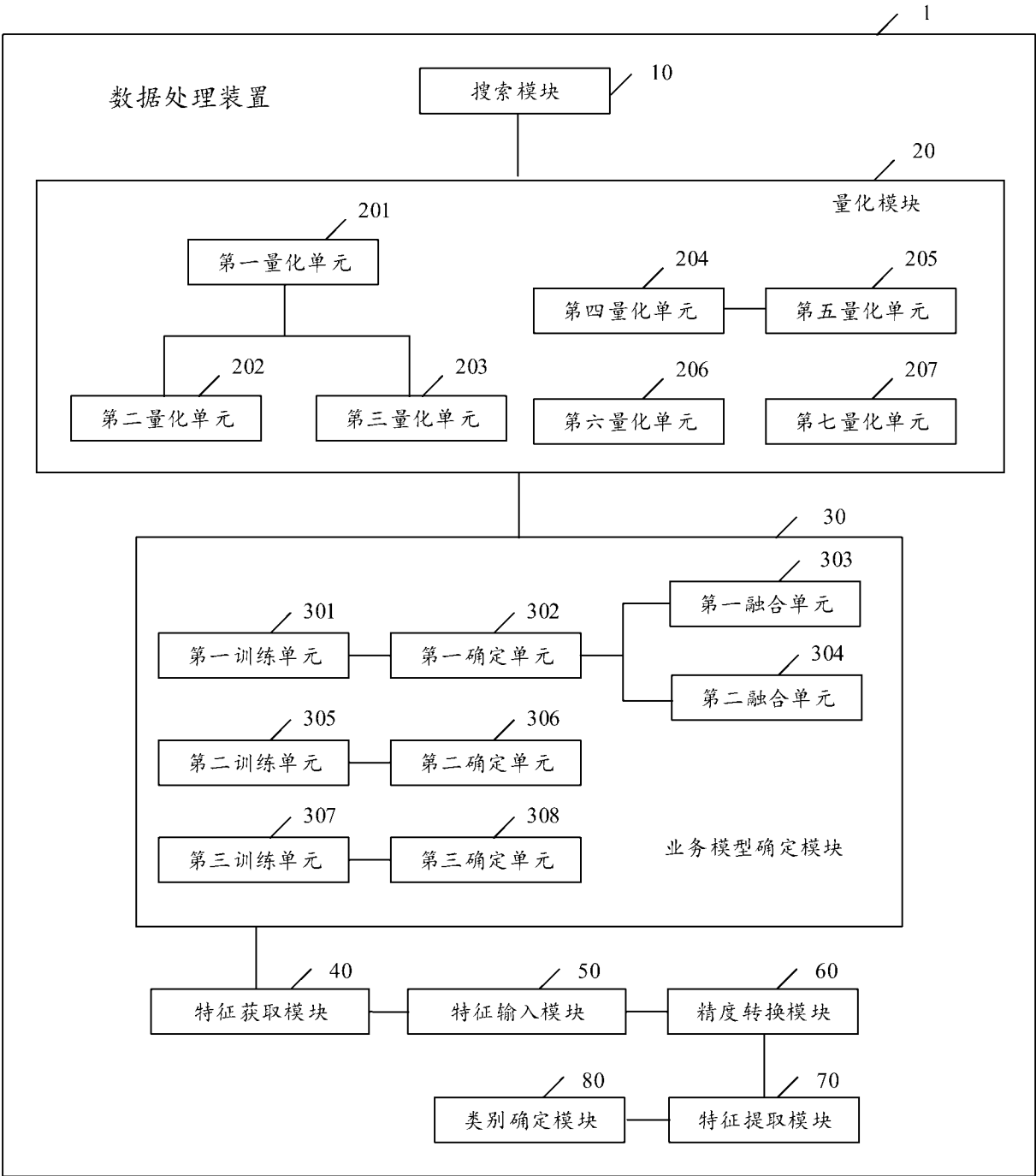


图 16

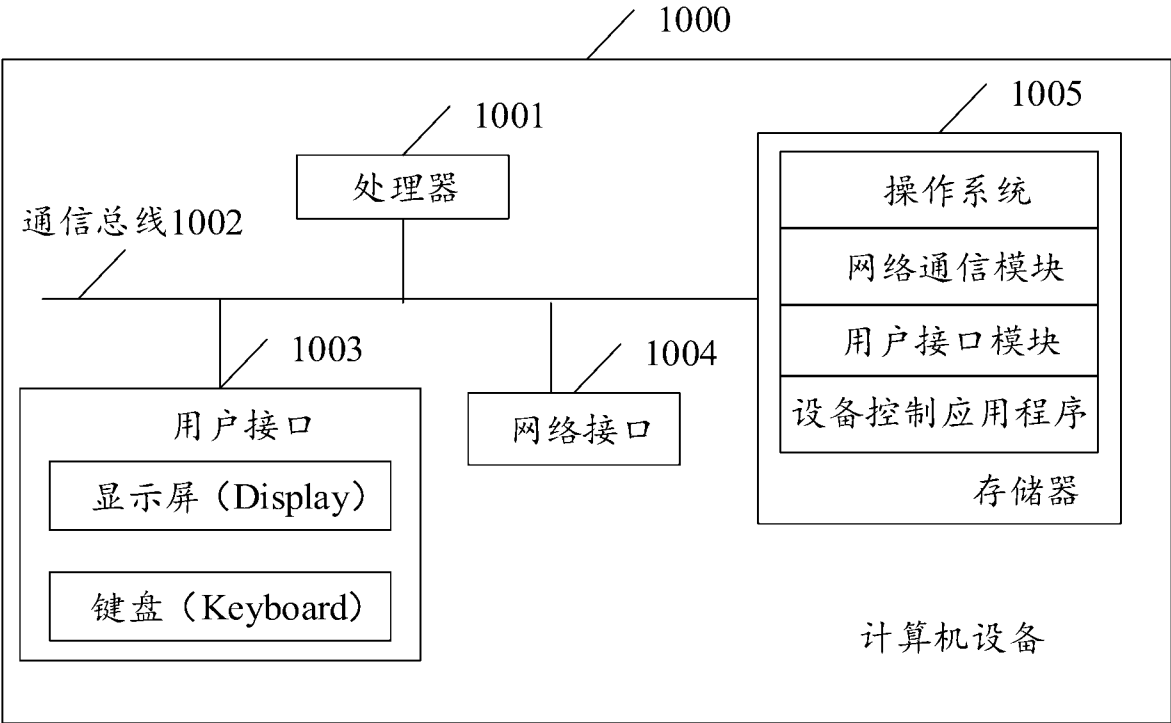


图 17

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/128556

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/082(2023.01)i; G06F 18/214(2023.01)i; G06F 18/213(2023.01)i; G06F 18/241(2023.01)i; G06F 18/25(2023.01)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)

IPC: 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)

CNABS; CNTXT; CNKI; VEN; DWPI; USTXT; WOTXT; EPTXT: 模型, 结构, 搜索, 神经网络, 量化, 权值, 网络层, 业务, 处理, model, frame, search, neural, network, quantization, weight, network layer, business, processing

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 113780551 A (BEIJING SENSETIME SCIENCE TECHNOLOGY DEVELOPMENT CO., LTD.) 10 December 2021 (2021-12-10) description, paragraphs [0004]-[0237], and figure 2B	1-19
A	CN 113610232 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO., LTD.) 05 November 2021 (2021-11-05) entire document	1-19
A	CN 115983365 A (GUANGDONG OPPO MOBILE COMMUNICATIONS CO., LTD.) 18 April 2023 (2023-04-18) entire document	1-19



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

“D” document cited by the applicant in the international application

“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

26 November 2023

Date of mailing of the international search report

05 December 2023

Name and mailing address of the ISA/CN

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

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/128556

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113780551	A	10 December 2021	HK	40065242	A0	22 July 2022
				WO	2023029349	A1	09 March 2023
				CN	113780551	B	24 March 2023
CN	113610232	A	05 November 2021	CN	113610232	B	22 February 2022
				WO	2023050707	A1	06 April 2023
CN	115983365	A	18 April 2023	None			

A. 主题的分类

G06N 3/082(2023.01)i; G06F 18/214(2023.01)i; G06F 18/213(2023.01)i; G06F 18/241(2023.01)i; G06F 18/25(2023.01)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)

IPC: 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))

CNABS;CNTXT;CNKI;VEN;DWPI;USTXT;WOTXT;EPTXT: 模型, 结构, 搜索, 神经网络, 量化, 权值, 网络层, 业务, 处理, model, frame, search, neural, network, quantization, weight, network layer, business, processing

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 113780551 A (北京市商汤科技开发有限公司) 2021年12月10日 (2021 - 12 - 10) 说明书第[0004]-[0237]段、图2B	1-19
A	CN 113610232 A (苏州浪潮智能科技有限公司) 2021年11月5日 (2021 - 11 - 05) 全文	1-19
A	CN 115983365 A (OPPO广东移动通信有限公司) 2023年4月18日 (2023 - 04 - 18) 全文	1-19

☐ 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

“D” 申请人在国际申请中引证的文件

“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

2023年11月26日

国际检索报告邮寄日期

2023年12月5日

ISA/CN的名称和邮寄地址

中国国家知识产权局
中国北京市海淀区蓟门桥西土城路6号 100088

授权官员

朱雪梅

电话号码 (+86) 0512-88995919

PCT/ISA/210 表(第2页) (2022年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2023/128556

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	113780551	A	2021年12月10日	HK	40065242	A0	2022年7月22日
				WO	2023029349	A1	2023年3月9日
				CN	113780551	B	2023年3月24日
CN	113610232	A	2021年11月5日	CN	113610232	B	2022年2月22日
				WO	2023050707	A1	2023年4月6日
CN	115983365	A	2023年4月18日	无			