(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 6 月 25 日 (25.06.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/125236 A1

(51) 国际专利分类号:
G06N 3/04 (2006.01)

(21) 国际申请号: PCT/CN2019/115082

(22) 国际申请日: 2019 年 11 月 1 日 (01.11.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201811545184.4 2018年12月17日 (17.12.2018) 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: 裴建国 (**PEI, Jianguo**); 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: 深圳市深佳知识产权代理事务所 (普通合伙) (**SHENPAT INTELLECTUAL PROPERTY AGENCY**); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦B座18C2、18D、18E、18E2, Guangdong 518001 (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, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,

WO 2020/125236 A1

(54) **Title:** DATA PROCESSING METHOD AND DEVICE, STORAGE MEDIUM, AND ELECTRONIC DEVICE

(54) 发明名称: 数据处理方法及装置、存储介质和电子装置

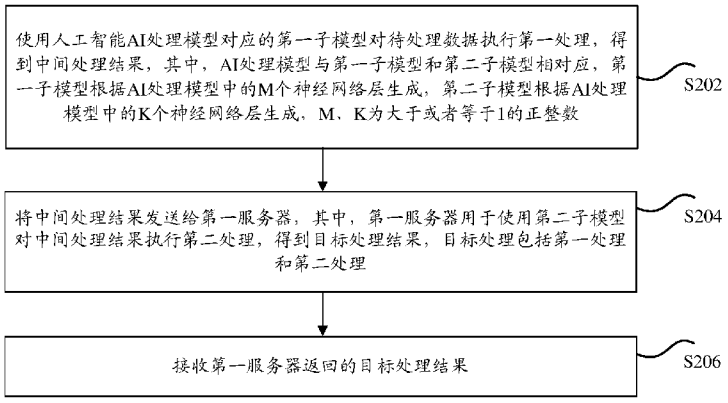

图 2

S202 Use a first sub-model corresponding to an artificial intelligence (AI) processing model to perform first processing on data to be processed, and obtain an intermediate processing result, wherein the AI processing model corresponds to the first sub-model and a second sub-model, the first sub-model is generated on the basis of M neural network layers in the AI processing model, the second sub-model is generated on the basis of K neural network layers in the AI processing model, and M and K are positive integers greater than or equal to 1

S204 Send the intermediate processing result to a first server, wherein the first server performs, by means of the second sub-model, second processing on the intermediate processing result and obtains a target processing result, and target processing comprises the first processing and the second processing

S206 Receive the target processing result returned by the first server

(57) **Abstract:** The present application discloses a data processing method and device, a storage medium, and an electronic device. The method comprises: using a first sub-model corresponding to an AI processing model to perform first processing on data to be processed, and obtaining an intermediate processing result, wherein the AI processing model corresponds to the first sub-model and a second sub-model, the first sub-model is generated on the basis of M neural network layers in the AI processing model, the second sub-model is generated on the basis of K neural network layers in the AI processing model, and M and K are integers greater than or equal to

[见续页]

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))。

1; sending the intermediate processing result to a first server, wherein the first server performs, by means of the second sub-model, second processing on the intermediate processing result and obtains a target processing result; and receiving the target processing result returned by the first server.

(57) 摘要：本申请公开了一种数据处理方法及装置、存储介质和电子装置。其中，该方法包括：使用AI处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI处理模型与所述第一子模型和第二子模型相对应，第一子模型根据所述AI处理模型中的M个神经网络层生成，第二子模型根据所述AI处理模型中的K个神经网络层，M、K为大于或者等于1的整数；将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果；接收第一服务器返回的目标处理结果。

数据处理方法及装置、存储介质和电子装置

本申请要求于 2018 年 12 月 17 日提交中国专利局、申请号为 CN201811545184.4、发明名称为“数据处理方法及装置、存储介质和电子装置”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

本申请涉及计算机领域，具体而言，涉及数据处理技术。

背景技术

目前，随着移动端设备（例如，手机，智能可穿戴设备）的流行，使得人工智能 AI（Artificial Intelligence，人工智能）模型（例如，深度学习模型）在资源（内存、CPU、能耗和带宽等）有限的便携式设备上部署变得越来越重要。

AI 处理模型（例如，人脸识别模型）通常采用大型神经网络进行数据处理，拥有高精度的识别效果，用户对处理效果体验满意。但是，人工智能模型包含的大型神经网络具有大量的层级与结点，内存开销和计算量巨大，受资源的限制，在移动端设备中部署 AI 处理模型非常困难。

目前，为了在移动端设备中能够部署 AI 处理模型，考虑如何减少它们所需要的内存与计算量就非常重要，通常采用模型压缩技术对 AI 处理模型进行模型压缩。然而，对于高精度的大神经网络，使用当前的模型压缩技术之后，其计算量和内存开销对于移动端设备依然很大，无法部署运行。也就是说，相关技术中存在由于模型压缩技术的模型压缩能力有限导致无法在移动端设备中部署 AI 处理模型的问题。

发明内容

本申请实施例中提供了一种数据处理方法及装置、存储介质和电子装置，以至少解决相关技术中存在由于模型压缩技术的模型压缩能力有限导致无法在移动端设备中部署 AI 处理模型的技术问题。

根据本申请实施例的一个方面，提供了一种数据处理方法，包括：使用 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI 处理模型用于对待处理数据执行目标处理，得到目标处理结果，AI 处理模型与所述第一子模型和第二子模型相对应，第一子模型根据所述 AI 处理模型中的 M 个神经网络层生成，第二子模型根据所述 AI 处理模型中的 K 个神经网络层生成，M、K 为大于或者等于 1 的正整数；将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果；接收第一服务器返回的目标处理结果。

根据本申请实施例的另一方面，还提供了一种数据处理装置，包括：处理单元，用于使用 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI 处理模型用于对待处理数据执行目标处理，得到目标处理结果，AI 处理模型与所述第一子模型和第二子模型相对应，所述

第一子模型根据所述 AI 处理模型中的 M 个神经网络层生成，所述第二子模型根据所述 AI 处理模型中的 K 个神经网络层生成，M、K 为大于或者等于 1 的正整数；发送单元，用于将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果，目标处理包括第一处理和第二处理；第一接收单元，用于接收第一服务器返回的目标处理结果。

根据本申请实施例的又一方面，还提供了一种存储介质，该存储介质中存储有计算机程序，其中，该计算机程序被设置为运行时执行上述方法。

根据本申请实施例的又一方面，还提供了一种电子装置，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其中，上述处理器通过计算机程序执行上述的方法。

在本申请实施例中，使用 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI 处理模型与第一子模型和第二子模型相对应，第一子模型根据 AI 处理模型中的 M 个神经网络层生成，第二子模型根据 AI 处理模型中的 K 个神经网络层生成，M、K 为大于或者等于 1 的正整数；将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果，目标处理包括第一处理和第二处理；接收第一服务器返回的目标处理结果，由于将 AI 处理模型中的包含 M 个神经网络层的第一子模型和包含 K 个神经网络层的第二子模型部署到不同的设备上，可以根据需要设定 M、K 的值，灵活设置第一子模型和第二子模型的规模，从而控制各子模型的计算量和内存开销，方便在移动端设备中进行AI处理模型的部署，达到了在移动端设备中部署AI处理模型的目的，从而实现了提高用户体验的技术效果，进而解决了相关技术中存在由于模型压缩技术的模型压缩能力有限导致无法在移动端设备中部署 AI 处理模型的问题。

附图说明

此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

图 1 是根据本申请实施例的一种数据处理方法的应用环境的示意图；
图 2 是根据本申请实施例的一种可选的数据处理方法的流程示意图；
图 3 是根据本申请实施例的一种可选的数据处理方法的示意图；
图 4 是根据本申请实施例的另一种可选的数据处理方法的示意图；
图 5 是根据本申请实施例的又一种可选的数据处理方法的示意图；
图 6 是根据本申请实施例的一种可选的图片对象识别结果的示意图；
图 7 是根据本申请实施例的又一种可选的数据处理方法的示意图；
图 8 是根据本申请实施例的一种可选的数据处理装置的结构示意图；以及
图 9 是根据本申请实施例的一种可选的电子装置的结构示意图。

具体实施方式

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

需要说明的是，本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

根据本申请实施例的一个方面，提供了一种数据处理方法，可选地，上述数据处理方法可以但不限于应用于如图 1 所示的应用环境中。移动端设备 102 通过网络 104 与服务器 106 相连。在移动端设备 102 中部署有根据 AI 处理模型生成的第一子模型，在服务器 106 中部署有根据 AI 处理模型生成的第二子模型，其中，第一子模型对应于 AI 处理模型的 M 个神经网络层，第二子模型对应于 AI 处理模型的 K 个神经网络层，M、K 均为大于或者等于 1 的正整数。

移动端设备 102 获取到待处理数据之后，使用第一子模型对待处理数据执行第一处理，得到中间处理结果，并将中间处理结果通过网络 104 发送给服务器 106。

服务器 106 接收到中间处理结果之后，使用第二子模型对中间处理结果进行第二处理，得到待处理数据的目标处理结果，并将得到的目标处理结果通过网络 104 发送给移动端设备 102。

移动端设备 102 接收服务器 106 返回的目标处理结果。

可选地，在本实施例中，上述移动端设备 102 可以包括但不限于以下至少之一：手机、平板电脑、笔记本电脑、智能可穿戴设备等。上述网络 104 可以包括但不限于有线网络和无线网络，其中，上述有线网络可以包括：局域网、城域网和广域网，上述无线网络可以包括：蓝牙、WIFI（Wireless Fidelity，无线保真）、移动网络及其他实现无线通信的网络。上述服务器 106 可以包括但不限于以下至少之一：PC 机及其他用于提供服务的设备。上述只是一种示例，本实施例对此不做任何限定。

可选地，作为一种可选的实施方式，如图 2 所示，上述数据处理方法可以包括：

步骤 S202，使用 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI 处理模型与所述第一子模型和第二子模型

相对应，第一子模型根据 AI 处理模型中的 M 个神经网络层生成，第二子模型根据 AI 处理模型中的 K 个神经网络层生成，M、K 为大于或者等于 1 的正整数。

步骤 S204，将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果，目标处理包括第一处理和第二处理；

步骤 S206，接收第一服务器返回的目标处理结果。

上述数据处理方法可以应用于移动端设备通过传感器（例如，移动终端的摄像头）获取待处理数据的过程中。下面以通过移动终端的摄像头获取待处理数据的应用场景为例进行说明。

移动终端中安装的应用（例如，网络直播应用）调用移动终端的摄像头进行网络直播，调用摄像头的同时触发使用移动终端中部署的第一子模型对通过摄像头获取到的图像数据进行处理（例如，对图像数据中的特定目标进行处理，可以包括但不限于滤镜处理、瘦脸处理、大眼处理等），并将处理后得到的中间数据通过无线网络（例如，移动网络、WIFI 网络）传输到云端服务器，由云端服务器使用其部署的第二子模型对中间数据进行后续处理，并将得到的图像数据返回给移动终端，在移动终端上进行显示，或者传输给观看直播的其他移动终端进行显示。

获取待处理数据的操作以及对待处理数据的操作可以是实时的，也可以是非实时的。具体的获取方式，本实施例中对此不作限定。

目前，高效的深度学习方法可以显著地影响分布式系统、嵌入式设备和用于人工智能的 FPGA（FPGA Field Programmable Gate Array，现场可编程门阵列）等。例如，ResNet-101（Residual Neural Network，残差神经网络）具有 101 层卷积网络、超过 200MB 的储存需求和计算每一张图片所需要的浮点数乘法时间。对于只有兆字节资源的手机和 FPGA 等设备，部署这种大模型就非常困难。

常用的高精度大模型包括: R-CNN（Region-Convolutional Neural Network，区域卷积神经网络），SSD（Single Shot Multibox Detector，单发多盒探测器）和 YOLO（You Only Love One，一种基于卷积神经网络的目标检测模型）。R-CNN 的识别精度随这基础网络模型的减小，而下降。如表 1 所示，表 1 示出了 R-FCN（Region-base Fully Convolutional Network，基于区域的全卷积网络）的不同基础模型的精度。

表 1

	training data	test data	ResNet-50	ResNet-101	ResNet-152
R-FCN	07+12	07	77.0	79.5	79.6
R-FCN multi-sc train	07+12	07	78.7	80.5	80.4

然而，这些高精度的大模型（即使最小的基础网络模型 ResNet50），在移动端设备中根本无法部署。

移动端设备的数据采集 sensor（传感器）获取的数据越来越大，比如手机摄像头，越来越高清，带来图片数据很大，传送到云端计算，速度慢，流量消耗大。因此，用户不愿意使用高精度的 DNN（Deep Neural Network，深度神经网络）大模型。

如图 3 所示，为了能够在移动端设备中部署 AI 处理模型，相关的处理方式为：采用模型压缩算法对高精度的 AI 处理模型进行整体压缩，将大模型整体压缩为一个小模型，以损失 AI 处理模型的识别精度为代价，最终降低神经网络的计算量。上述处理方法一般分两个步骤：

步骤 1，在超强计算能力的 GPGPU（General Purpose Computer on GPU，在图形处理器上进行通用运算）集群上，训练出高识别精度和高计算量神经网络模型。

步骤 2，在高精度的大模型基础上，神经网络压缩技术（模型压缩技术），训练一个低识别精度和计算量的小模型，最终满足嵌入式设备（移动端设备）的计算能力。

目前的模型压缩技术可以大致分为四类：参数修剪和共享（parameter pruning and sharing）、低秩分解（low-rank factorization）、迁移/压缩卷积滤波器（transfered/compact convolutional filter）和知识精炼（knowledge distillation）。

然而，由于模型压缩技术的压缩能力有限，很多高精度大模型压缩后，依然无法部署到移动端。表 2 和表 3 为相关模型压缩技术的性能数据，其中，表 2 示出了低秩分解技术在 ILSVRC-2012 数据集上的性能对比结果，表 3 示出了迁移/压缩卷积滤波器在 CIFAR-10 和 CIFAR-100 数据集上的性能对比结果。

表 2

Model	TOP-5 Accuracy	Speed-up	Compression Rate
Alex Net	80.3%	1.	1.
BN Low-rank	80.56%	1.09	4.94
CP Low-rank	79.66%	1.82	5.
VGG-16	90.60%	1.	1.
BN Low-rank	90.47%	1.53	2.72
CP Low-rank	90.31%	2.05	2.75
GoogleNet	92.21%	1.0	1.
BN Low-rank	91.88%	1.08	2.79
CP Low-rank	91.79%	1.20	2.84

表 3

Model	CIFAR-100	CIFAR-10	Compression Rate
VGG-16	34.26%	9.85%	1.
MBA [45]	33.66%	9.76%	2.
CRELU [44]	34.57	9.92%	2.
CIRC [42]	35.15%	10.23%	4.
DCNN [43]	33.57%	9.65%	1.62

表 2 和表 3 示出的当前模型压缩技术的性能数据，压缩比（Compression

Rate）都是5倍以下，对于高精度的大神经网络，即使压缩5倍，其计算量和内存开销对于移动端设备依然很大，无法部署运行。

同时，相关的AI处理模型在移动端的部署方案，均会损失最终的AI处理模型识别精度。有些AI处理模型，最终压缩后的精度，可能完全无法被用户接受，体验极差。

在本申请中，根据AI处理模型（AI处理模型）生成两个子模型：第一子模型和第二子模型，并将第一子模型部署在移动端设备，将第二子模型部署在服务器中，通过第一子模型结合第二子模型配合使用的方式实现使用AI处理模型处理的计算任务，以此减少AI处理模型部署在移动端设备中的模型的计算量和内存开销，达到可以在移动端设备中部署AI处理模型的目的，同时由于向服务器传输的仅为第一子模型的处理结果，在较少传输的数据量的同时，保护了用户隐私，提高了用户体验。

下面结合图2所示的步骤详述本申请的实施例。

在步骤S202提供的技术方案中，移动端设备使用AI处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果。

上述AI处理模型用于对待处理数据执行目标处理，得到目标处理结果。待处理数据可以是待处理多媒体数据，可以包括但不限于：图像数据、视频数据、语音数据。上述目标处理包括但不限于：图像处理（图像去模糊、图像对象识别、图像美化、图像渲染、文字翻译）、语音处理（语音翻译、语音去噪）。得到目标处理结果可以包括但不限于：图像处理结果（图像去模糊的结果、对象识别的结果、图像美化的而结果、图像渲染的结果、文字翻译的结果），语音处理结果（语音去噪的结果、语音翻译的结果）。

根据AI处理模型可以生成对应的第一子模型和第二子模型，具体的，第一子模型根据AI处理模型中的M个神经网络层生成，第二子模型根据AI处理模型中的K个神经网络层生成，其中，M、K均为大于或者等于1的正整数。第一子模型可以部署在移动端设备，第二子模型可以部署在第一服务器中（云端服务器）。其中，所述M个神经网络层和所述K个神经网络层为所述AI处理模型中不同层级的网络层。

上述AI处理模型与第一子模型和第二子模型相对应，所谓相对应可以理解是为第一子模型和第二子模型相结合能够实现AI处理模型的功能以达到相同或者近似的处理效果。

在部署AI处理模型之前，可以首先对初始AI处理模型进行训练，得到包括N个神经网络层的AI处理模型。

可选地，可以使用训练数据（可以是多媒体数据，例如，图像数据、视频数据、语音数据等）对初始AI处理模型进行训练，得到AI处理模型（高精度的AI处理模型）。具体的训练过程可以结合相关技术，本实施例中对此不作赘述。

在得到 AI 处理模型之后，可以从 AI 处理模型中拆分出两个计算段，其中，拆出的 M 个神经网络层为第一计算段，拆出的 K 个神经网络层为第二计算段，其中，N 大于或者等于 M 与 K 的和。

由于第一计算段仅包含了与高精度 AI 处理模型对应的部分神经网络层，第一计算段的计算量和内存开销可以控制在移动端设备可承受的范围内。因而拆分出的第一计算段可作为第一子模型直接部署到移动端设备中。

为了进一步控制第一计算段的计算量和内存开销，可以采用模型压缩算法（例如，蒸馏法）对第一计算段中包含的神经网络层进行压缩，得到第一子模型。

可选地，在使用与 AI 处理模型对应的第一子模型对待处理数据进行处理之前，移动端设备接收第二服务器发送的第一子模型，其中，第二服务器用于使用目标压缩算法对上述 M 个神经网络层进行压缩，得到第一子模型，其中，目标压缩算法用于对神经网络进行压缩。

第二服务器可以使用用于对神经网络进行压缩的目标压缩算法（模型压缩算法）对第一计算段中包含的 M 个神经网络层进行压缩，得到第一子模型。对应的，将第二计算段中包含的 K 个神经网络层作为第二子模型。

通过对 M 个神经网络层进行模型压缩，可以进一步减少第一计算段的计算量和内存开销，同时可以减少使用过程中向第一服务器传输的数据量，降低对移动端设备的资源消耗，提高用户体验。

可选地，第二子模型用于补偿相对于 AI 处理模型，由于对包含 M 个神经网络层的第一子模型进行压缩所损失的处理精度。

在对第一计算段进行模型压缩之后，为了保证整个 AI 处理模型的精度，可以对第二计算段中包含的神经网络层进行训练，调整第二计算段中包含的神经网络层的参数信息，以补偿由于对第一计算段进行模型压缩而造成的精度损失。

可选地，在使用与 AI 处理模型对应的第一子模型对待处理数据进行处理之前，第二服务器可以使用目标压缩算法，对 M 个神经网络层进行压缩，得到第一子模型；获取第一子模型对训练对象执行第一处理得到的第一处理结果；对 K 个神经网络层进行训练，得到第二子模型，其中，K 个神经网络层的输入为第一处理结果，训练的约束条件为：第二子模型的输出结果与第二处理结果的处理精度差值小于或等于目标阈值，其中，第二处理结果为使用 AI 处理模型（未进行过模型压缩时的 AI 处理模型）对训练对象进行处理得到的处理结果。

在对第二计算段中包含的神经网络层进行训练时，可以以第一子模型的输出作为第二计算段的输入，结合 AI 处理模型（未进行模型压缩）对训练数据的处理结果，使用模型训练算法对第二计算段中包含的神经网络层，训练出一个不损失最终精度的网络识别层（第二子模型）。

在第二服务器模型训练完成，得到第一子模型和第二子模型之后，可以将

第一子模型部署到移动端设备，将第二子模型部署到第一服务器（云端处理器）。

例如，为了在移动端设别中直接部署高精度的 R-FCN 大模型（其基础模型采用 ResNet-101），如图 4 所示，将其拆分为 2 段，云段计算部分（第二计算段）和移动端计算部分（第一计算段），移动端设备只部署压缩的前 3 层网络，除了前 3 层以外的网络部署在云端（这种情况下，N 等于 M 与 K 的和）。

在对网络进行压缩时，可以采用蒸馏法训练前 3 层网络，得到压缩网络（第一子模型），并将其部署移动端设备。图 5 展示了训练网络压缩层一种方法，基于这个损失函数可以蒸馏出，最接近原始 ResNet-101 前 3 层的输出结果。

对于云端计算部分，可以利用压缩网络的输出作为输入，利用迁移学习的方法，训练出一个不损失最终精度的网络识别层（第二子模型），部署到云端（例如，第一服务器）。

在将第一子模型部署到移动端设备之后，移动端设备可以通过获取待处理数据，上述待处理数据可以是实时获取的数据，也可以是非实时获取的数据，获取的待处理数据的方式可以有多种。

作为一种可选的实施方式，待处理数据可以是移动端设备从其他设备（其他移动端设备、终端设备、服务器设备等）中接收到的。

作为另一种可选的实施方式，待处理数据可以是移动端设备通过自身的数据采集部件获取到的。上述数据采集部件可以但不限于：传感器（例如，摄像头），麦克风等。

可选地，在使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理之前，通过移动端设备的目标传感器获取待处理数据。

在获取到待处理数据之后，移动端设备可以使用第一子模型对待处理数据执行第一处理，得到中间处理结果。

对于不同类型的待处理数据，可以采用不同的 AI 处理模型进行处理。对于同一待处理数据，可以采用一种或多个 AI 处理模型进行处理。

可选地，在待处理数据为待处理图像的情况下，AI 处理模型为 AI 识别模型，用于识别待处理图像中包含的目标对象，第一子模型执行的第一处理为第一识别处理。

使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果可以包括：使用第一子模型对待处理图像执行第一识别处理，得到中间识别结果。

可选地，在待处理数据为待处理图像的情况下，AI 处理模型为 AI 去模糊模型，用于对待处理图像进行去模糊处理，得到去模糊结果的处理结果，第一子模型执行的第一处理为第一去模糊处理。

使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果可以包括：使用第一子模型对待处理图像执行第一去模糊处理，得到中间去模糊结果。

可选地，在待处理数据为待翻译数据（可以是语音，也可以是图片或其他需要进行翻译的数据）的情况下，AI处理模型为AI翻译模型，用于将待翻译数据中包含的使用第一语言的第一语言数据，翻译为使用第二语言的第二语言数据，第一子模型执行的第一处理为第一翻译处理。

使用与AI处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果可以包括：使用第一子模型对待翻译数据执行第一翻译处理，得到中间翻译结果。

在步骤S204提供的技术方案中，将中间处理结果发送给第一服务器。

上述发送可以是通过移动端设备的移动网络进行的，也可以是移动端设备通过其他无线网络进行的。由于发送的是第一子网络的处理结果（计算结果），数据大小一般远小于原始待处理数据的大小，同时用户隐私方面也得到保证。

第一服务器接收到中间处理结果之后，可以使用部署的第二子模型对中间处理结果执行第二处理，得到目标处理结果。

对于不同类型的待处理数据，可以采用与不同的AI处理模型对应的第二子模型对于中间处理结果进行处理。对于同一待处理数据，可以采用一种或多个与AI处理模型对应的第二子模型对于中间处理结果进行处理。

可选地，在待处理数据为待处理图像的情况下，第二处理为第二识别处理，目标识别结果用于指示待处理图像中的目标对象。

第一服务器使用第二子模型对中间识别结果执行第二识别处理得到，得到待处理图像中的目标对象的识别结果。

可选地，在待处理数据为待处理图像的情况下，第二处理为第二去模糊处理，目标识别结果用于指示待处理图像的去模糊结果。

第一服务器使用第二子模型对中间去模糊结果执行第二去模糊处理，得到待处理图像中的目标去模糊结果。

可选地，在待处理数据为待翻译数据的情况下，第二处理为第二翻译处理，目标识别结果为包含第二语言数据的目标翻译结果。

第一服务器使用第二子模型对中间翻译结果执行第二翻译处理，得到包含第二语言数据的目标翻译结果。

在得到目标处理结果之后，第一服务器将目标翻译结果发送给移动端设备。

在步骤S206提供的技术方案中，移动端设备接收第一服务器返回的目标处理结果。

在接收到第一服务器返回的目标处理结果之后，移动端设备可以在移动端设备上显示目标识别结果（待处理图像），或者，播放目标识别结果（待处理语音数据，例如，待翻译数据）。

下面结合具体示例进行说明。高精度AI处理模型可以应用到视频产品中，用于对视频产品的AI处理，可以包括但不限于：AI自动去模糊，AI图片自动识别等，例如，AI自动图片识别技术。常用的高精度大模型包括：R-CNN、

SSD 和 YOLO 等，识别效果如图 6 所示。

本示例中的对象识别方法采用对 AI 处理模型分段计算的方式，根据高精度 AI 处理模型，生成与之对应的两个计算部分：移动端计算部分和云端计算部分。移动端计算部分对应于 AI 处理模型的前 m 层神经网络层（例如，前 3 层），部署在移动端（移动端设备，嵌入式设备，如，手机等）中。云端计算部分对应于 AI 处理模型的剩余部分（除前 m 层神经网络层以外的其他神经网络层），部署在云端（云集群，云计算中心，云服务器）。

在进行传感器数据处理时，在移动端设备上，通过移动端计算部分执行计算，获得中间结果，并通过连接移动端和云端的网络（例如，移动网络，无线网络等），将获得的中间结果发送给云端。云端接收移动端发送的中间结果，使用 AI 处理模型的剩余部分接力计算，并将最终处理结果发送给移动端。

通过上述分段计算的方式，可以减少移动端的计算量和内存开销，保证在移动端部署高精度 AI 处理模型的可行性。并且，将移动端计算部分的中间结果作为云端计算部分的输入，可以减少移动端的数据通讯量，同时保护用户的隐私，从而提高用户体验。

可选地，可以对与 AI 处理模型对应的计算部分进行局部压缩：对移动端计算部分进行模型压缩，作为局部高压缩比部分（高压缩比计算段，即高压缩比神经网络层），而云端计算部分不进行压缩，作为其他无压缩部分（高精度计算段，即高精度神经网络层）。

在进行传感器数据处理时，在移动端设备上，通过高压缩比神经网络层的计算，获得一个近似原始网络的第一层或者前几层的网络中间结果（高压缩比神经网络层的输出结果），并通过连接移动端和云端的网络，将中间结果发送给云端。

云端接收移动端的中间结果，使用剩余计算部分接力计算，并将最终识别结果发送给移动端。

由于高压缩比神经网络层所描述的网络层次简单，可以大幅度压缩计算量，所损失的信息比全网络压缩要少很多。并且，通过高压缩比神经网络层的计算，可以获得一个近似原始网络的第一层或者前几层的输出结果，而上述输出结果的数据大小一般远小于原始传感器数据的大小（也小于无压缩时移动端计算部分的输出结果的数据大小），可以降低最终的传输数据量。同时，通过其他无压缩部分的计算，可以保持或者尽量保持使用 AI 处理模型处理时的原有识别精度，从而提高用户体验。

对于对象识别的流程，如图 7 所示，本示例中的对象识别方法包括以下步骤：

步骤 1，在移动端和云端部署 AI 处理模型。

在使用 AI 处理模型进行处理之前，可以对 AI 处理模型进行训练，得到移动端计算部分和云端计算部分，并将移动端计算部分部署到移动端，将云端计算部分部署到云端。

在对 AI 处理模型进行训练时，对于无压缩的 AI 处理模型，可以对原始模型进行训练，得到高精度 A 处理 I 模型。对于局部压缩的 AI 处理模型，在上述训练的基础上，可以用得到的高精度 AI 处理模型的中间数据作为标签，采用用于对模型进行压缩的第一算法（例如，蒸馏法）对 AI 处理模型的前 n 层（例如，前 3 层）神经网络层进行训练压缩，得到移动端计算部分；可以用移动端计算部分的输出结果作为输入，用高精度 AI 处理模型的中间数据和/最终数据作为标签，采用第二算法（例如，迁移学习）训练 AI 处理模型中除前 n 层（例如，前 3 层）以外的其他神经网络层，得到云端计算部分。

上述AI处理模型的局部压缩方式可以应用在所有移动端AI处理模型的部署中。

以目前精度较好的 R-FCN 大模型为例进行说明，该模型的基础模型采用 ResNet-101。如图 4 所示，根据 ResNet-101 生成两个计算部分：移动端计算部分和云端计算部分。

对于移动端计算部分，该部分可以为压缩网络。可以用高精度 AI 处理模型的中间数据作为标签，采用蒸馏法指导训练压缩 AI 处理模型的前 3 层神经网络层，得到一个高压缩比的神经网络层（移动端计算部分），最终部署到移动端。

图 5 示出了训练网络压缩层的一种方式。基于图 5 中示出的损失函数可以蒸馏出最接近原始 ResNet-101 的前 3 层的输出结果。

对于云端计算部分，该部分可以为高精度网络。可以采用迁移学习的方法（transfer learning）对 AI 处理模型中除了前 3 层神经网络层以外的神经网络层进行训练，得到云端计算部分。该高精度网络的输入为压缩网络的输出，训练过程可以参考高精度 AI 处理模型的输出结果（中间数据和/或最终数据），以得到一个不损失或者尽量不损失最终精度的网络识别层，最终部署到移动端。

在训练时，将高压缩比神经网络层的输出结果，作为高精度模型的剩余子模块的输入，再次训练该子模型，确保模型适应压缩后的输入变化，以达到原始 AI 处理模型的精度。

在进行 AI 处理模型部署时，上述训练过程可以是由特定的设备（目标设备）或者特定的服务器执行的，也可以由云端执行的。在压缩网络和高精度网络训练完成之后，可以通过得到的压缩网络和高精度网络发送至移动端和云端，分别进行网络部署。

步骤 2，通过移动端设备的传感器获取传感器数据。

在 AI 处理模型部署完成之后，移动端设备可以通过移动端设备的传感器获取传感器数据。上述传感器数据可以是任意通过传感器获取到的数据，可以包括但不限于：语音数据、图像数据或者其他数据。

步骤 3，通过移动端设备的高压缩比的神经网络计算得到中间结果，并将中间结果发送给云计算中心。

在移动端设备中，使用 AI 处理模型的移动端计算部分（高压缩比的神经网络）对传感器数据进行计算，得到中间结果，并将中间结果通过网络发送给云计算中心。

步骤 4，通过云计算中心的高精度的神经网络计算得到最终识别结果，并将最终识别结果发送给移动端设备。

通过本示例的上述技术方案，对高精度 AI 处理模型进行分段计算，根据 AI 处理模型生成与之对应的高压缩比计算段和高精度计算段，分别部署移动端和云端，接力计算完成一个与高精度 AI 处理模型具有同等性能的计算任务，进一步地，采用对 AI 处理模型进行局部压缩的方式，利用高压缩比的神经网络替换高精度 AI 处理模型最开始的一层或者几层的子网络（而不是整个模型网络），完成第一层或者前几层的计算量，而不是同其他全网络压缩方法一样，以损失精度为代价，降低计算量，通过上述方式，可以使得用户在损失少量流量的情况下，可以借助云端，获取高精度的神经网络模型的体验，为移动端部署高精度 AI 处理模型提供了便利，扩大了对云端的计算需求。

通过上述步骤 S202 至步骤 S206，终端侧设备使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI 处理模型用于对待处理数据执行目标处理，得到目标处理结果，AI 处理模型与第一子模型和第二子模型相对应，第一子模型对应于 AI 处理模型中的 M 个神经网络层，第二子模型对应于 AI 处理模型中的 K 个神经网络层，M、K 为大于或者等于 1 的正整数；将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果，目标处理包括第一处理和第二处理；接收第一服务器返回的目标处理结果，解决了相关技术中存在由于模型压缩技术的模型压缩能力有限导致无法在移动端设备中部署 AI 处理模型的问题，达到了提高用户体验的效果。

作为一种可选的技术方案，在使用与 AI 处理模型对应的第一子模型对待处理数据进行处理之前，上述方法还包括：

S1，接收第二服务器发送的第一子模型，其中，第二服务器用于使用目标压缩算法对 M 个神经网络层进行压缩，得到第一子模型，其中，目标压缩算法用于对神经网络进行压缩。

可选地，第二子模型用于补偿 AI 处理模型由于对 M 个神经网络层进行压缩所损失的处理精度。

可选地，在使用所述 AI 处理模型的所述第一子模型对待处理数据进行处理之前，上述方法还包括：

S1，第二服务器使用目标压缩算法，对 M 个神经网络层进行压缩，得到第一子模型；

S2，第二服务器获取第一子模型对训练对象执行第一处理得到的第一处理结果；

S3，第二服务器对 K 个神经网络层进行训练，得到第二子模型，其中，K

个神经网络层的输入为第一处理结果，训练的约束条件为：第二子模型的输出结果与第二处理结果的处理精度差值小于或等于目标阈值，其中，第二处理结果为使用 AI 处理模型对训练对象进行处理得到的处理结果。

通过本实施例，使用目标压缩算法对 M 个神经网络层进行压缩，得到第一子模型，可以减少第一子模型的计算量和内存开销，以及移动端设备向第一服务器发送的中间处理结果所包含的数据量，提高用户体验。进一步地，通过对第一子模型和第二子模型进行训练，以通过第二子模型补偿由于第一子模型压缩所造成的精度损失，提高待处理数据的处理精度，进而提高用户体验。

作为一种可选的技术方案，

使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果包括：使用第一子模型对待处理图像执行第一识别处理，得到中间处理结果，其中，待处理数据为待处理图像，第一处理为第一识别处理，待处理图像中包含目标对象，AI 处理模型用于识别待处理图像中的目标对象；

接收第一服务器返回的目标处理结果包括：接收第一服务器返回的目标识别结果，其中，目标处理结果为目标识别结果，目标识别结果由第一服务器使用第二子模型对中间识别结果执行第二识别处理得到，第二处理为第二识别处理，目标识别结果用于指示待处理图像中的目标对象。

作为另一种可选的技术方案，

使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果包括：使用第一子模型对待处理图像执行第一去模糊处理，得到中间处理结果，其中，待处理数据为待处理图像，第一处理为第一去模糊处理，AI 处理模型用于对待处理图像执行去模糊操作，得到目标去模糊结果；

接收第一服务器返回的目标处理结果包括：接收第一服务器返回的目标去模糊结果，其中，目标处理结果为目标去模糊结果，目标去模糊结果由第一服务器使用第二子模型对中间去模糊结果执行第二去模糊处理得到，第二处理为第二去模糊处理。

作为又一种可选的技术方案，

使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果包括：使用第一子模型对待翻译数据执行第一翻译处理，得到中间翻译结果，其中，待处理数据为待翻译数据，第一处理为第一翻译处理，AI 处理模型用于将待翻译数据中包含的使用第一语言的第一语言数据，翻译为使用第二语言的第二语言数据；

接收第一服务器返回的目标处理结果包括：接收第一服务器返回的目标翻译结果，其中，目标处理结果为目标翻译结果，目标翻译结果由第一服务器使用第二子模型对中间翻译结果执行第二翻译处理得到，第二处理为第二翻译处理，目标翻译结果中包含第二语言数据。

通过本实施例，对于不同类型的待处理数据执行一种或多种不同的处理（例如，对象识别、图像去模糊、数据翻译），满足用户不同的需求，提高终

端业务处理的能力，提高用户体验。

作为一种可选的技术方案，在使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理之前，上述方法还包括:

S1，通过移动端设备的目标传感器获取待处理数据。

通过本实施例，通过移动端设备的目标传感器获取待处理数据，可以实现对待处理数据的实时处理，适应不同的应用需求，提高用户体验。

需要说明的是，对于前述的各方法实施例，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，本申请并不受所描述的动作顺序的限制，因为依据本申请，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，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本申请所必须的。

根据本申请实施例，还提供了一种用于实施上述数据处理方法的数据处理装置。图 8 是根据本申请实施例的一种可选的数据处理装置的示意图，如图 8 所示，该装置可以包括:

（1）处理单元 82，用于使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI 处理模型用于对待处理数据执行目标处理，得到目标处理结果，AI 处理模型与第一子模型和第二子模型相对应，第一子模型根据 AI 处理模型中的 M 个神经网络层生成，第二子模型根据 AI 处理模型中的 K 各神经网络层生成，M、K 为大于或者等于 1 的正整数;

（2）发送单元 84，用于将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果，目标处理包括第一处理和第二处理;

（3）第一接收单元 86，用于接收第一服务器返回的目标处理结果。

上述数据处理装置可以但不限于应用于移动端设备通过传感器（例如，移动终端的摄像头）获取待处理数据的过程中。

需要说明的是，该实施例中的处理单元 82 可以用于执行本申请实施例中的步骤 S202，该实施例中的发送单元 84 可以用于执行本申请实施例中的步骤 S204，该实施例中的第一接收单元 86 可以用于执行本申请实施例中的步骤 S206。

通过本申请提供的实施例，终端侧设备使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI 处理模型用于对待处理数据执行目标处理，得到目标处理结果，AI 处理模型与第一子模型和第二子模型相对应，第一子模型根据 AI 处理模型中的 M 个神经网络层生成，第二子模型根据 AI 处理模型中的 K 个神经网络层生成，M、K 为大于或者等于 1 的正整数；将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果，目标处

理包括第一处理和第二处理；接收第一服务器返回的目标处理结果，解决了相关技术中存在由于模型压缩技术的模型压缩能力有限导致无法在移动端设备中部署 AI 处理模型的问题，达到了提高用户体验的效果。

作为一种可选的技术方案，上述装置还包括：

第二接收单元，用于在使用与 AI 处理模型对应的第一子模型对待处理数据进行处理之前，接收第二服务器发送的第一子模型，其中，第二服务器用于使用目标压缩算法对 M 个神经网络层进行压缩，得到第一子模型，其中，目标压缩算法用于对神经网络进行压缩。

通过本实施例，使用目标压缩算法对 M 个神经网络层进行压缩，得到第一子模型，可以减少第一子模型的计算量和内存开销，以及移动端设备向第一服务器发送的中间处理结果所包含的数据量，提高用户体验。

作为一种可选的技术方案，上述处理单元 82 包括：处理模块，第一接收单元 86 包括，接收模块，其中，

（1）处理模块，用于使用第一子模型对待处理图像执行第一识别处理，得到中间识别结果，其中，待处理数据为待处理图像，第一处理为第一识别处理，待处理图像中包含目标对象，AI 处理模型用于识别待处理图像中的目标对象；

（2）接收模块，用于接收第一服务器返回的目标识别结果，其中，目标处理结果为目标识别结果，目标识别结果由第一服务器使用第二子模型对中间识别结果执行第二识别处理得到，第二处理为第二识别处理，目标识别结果用于指示待处理图像中的目标对象。

作为另一种可选的技术方案，上述处理单元 82 包括：第一处理模块，第一接收单元 86 包括，第一接收模块，其中，

（1）第一处理模块，用于使用第一子模型对待处理图像执行第一去模糊处理，得到中间去模糊结果，其中，待处理数据为待处理图像，第一处理为第一去模糊处理，AI 处理模型用于对待处理图像执行去模糊操作，得到目标去模糊结果；

（2）第一接收模块，用于接收第一服务器返回的目标去模糊结果，其中，目标处理结果为目标去模糊结果，目标去模糊结果由第一服务器使用第二子模型对中间去模糊结果执行第二去模糊处理得到，第二处理为第二去模糊处理。

作为又一种可选的技术方案，上述处理单元 82 包括：第二处理模块，第一接收单元 86 包括，第二接收模块，其中，

（1）第二处理模块，用于使用第一子模型对待翻译数据执行第一翻译处理，得到中间翻译结果，其中，待处理数据为待翻译数据，第一处理为第一翻译处理，AI 处理模型用于将待翻译数据中包含的使用第一语言的第一语言数据，翻译为使用第二语言的第二语言数据；

（2）第二接收模块，用于接收第一服务器返回的目标翻译结果，其中，目标处理结果为目标翻译结果，目标翻译结果由第一服务器使用第二子模型对

中间翻译结果执行第二翻译处理得到，第二处理为第二翻译处理，目标翻译结果中包含第二语言数据。

通过本实施例，对于不同类型的待处理数据执行一种或多种不同的处理（例如，对象识别、图像去模糊、数据翻译），满足用户不同的需求，提高终端业务处理的能力，提高用户体验。

作为一种可选的技术方案，上述装置还包括:

获取单元，用于在使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理之前，通过移动端设备的目标传感器获取待处理数据。

通过本实施例，通过移动端设备的目标传感器获取待处理数据，可以实现对待处理数据的实时处理，适应不同的应用需求，提高用户体验。

根据本申请的实施例的又一方面，还提供了一种存储介质，该存储介质中存储有计算机程序，其中，该计算机程序被设置为运行时执行上述任一项方法实施例中的步骤。

可选地，在本实施例中，上述存储介质可以被设置为存储用于执行以下步骤的计算机程序:

S1，使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI 处理模型用于对待处理数据执行目标处理，得到目标处理结果，AI 处理模型与第一子模型和第二子模型相对应，第一子模型根据 AI 处理模型中的 M 个神经网络层生成，第二子模型根据 AI 处理模型中的 K 个神经网络层生成，M、K 为大于或者等于 1 的正整数;

S2，将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果，目标处理包括第一处理和第二处理;

S3，接收第一服务器返回的目标处理结果。

可选地，在本实施例中，本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令终端设备相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括: 闪存盘、只读存储器（Read-Only Memory，简称为 ROM）、随机存取器（Random Access Memory，简称为 RAM）、磁盘或光盘等。

根据本申请实施例的又一个方面，还提供了一种用于实施上述数据处理方法的电子装置，如图 9 所示，该电子装置包括: 处理器 902、存储器 904、数据总线 906 和传输装置 908 等。上述各部件可以通过数据总线 906 或者其他用于数据传输的线进行连接。该存储器中存储有计算机程序，该处理器被设置为通过计算机程序执行上述任一项方法实施例中的步骤。

可选地，在本实施例中，上述电子装置可以位于计算机网络的多个网络设备中的至少一个网络设备。

可选地，在本实施例中，上述处理器可以被设置为通过计算机程序执行以下步骤：

S1，使用与 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI 处理模型用于对待处理数据执行目标处理，得到目标处理结果，AI 处理模型与第一子模型和第二子模型相对应，第一子模型根据 AI 处理模型中的 M 个神经网络层生成，第二子模型根据 AI 处理模型中的 K 个神经网络层生成，M、K 为大于或者等于 1 的正整数；

S2，将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果，目标处理包括第一处理和第二处理；

S3，接收第一服务器返回的目标处理结果。

可选地，本领域普通技术人员可以理解，图 9 所示的结构仅为示意，电子装置也可以是智能设备、智能手机（如 Android 手机、ios 手机等）、平板电脑、掌上电脑以及移动互联网设备（Mobile Internet Devices，简称为 MID）、PAD 等终端设备。图 9 其并不对上述电子装置的结构造成限定。例如，电子装置还可包括比图 9 中所示更多或者更少的组件（如网络接口等），或者具有与图 9 所示不同的配置。

其中，存储器 904 可用于存储软件程序以及模块，如本申请实施例中的数据处理方法和装置对应的程序指令/模块，处理器 902 通过运行存储在存储器 904 内的软件程序以及模块，从而执行各种功能应用以及数据处理，即实现上述签名信息的传输方法。存储器 904 可包括高速随机存储器，还可以包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器 904 可进一步包括相对于处理器 902 远程设置的存储器，这些远程存储器可以通过网络连接至终端。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

上述的传输装置 908 用于经由一个网络接收或者发送数据。上述的网络具体实例可包括有线网络及无线网络。在一个实例中，传输装置 908 包括一个网络适配器（Network Interface Controller，简称为 NIC），其可通过网线与其他网络设备与路由器相连从而可与互联网或局域网进行通讯。在一个实例中，传输装置 908 为射频（RadioFrequency，简称为 RF）模块或蓝牙，其用于通过无线方式与互联网进行通讯。

上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。

上述实施例中的集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在上述计算机可读取的存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在存储介质中，包括若干指令用以使得一台或多台计算机设备（可为个人计算机、服务器或者网络设备等）执行本申请各个实施例所述方法的全部或部分步

骤。

在本申请的上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

在本申请所提供的几个实施例中，应该理解到，所揭露的客户端，可通过其它的方式实现。其中，以上所描述的装置实施例仅仅是示意性的，例如所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，单元或模块的间接耦合或通信连接，可以是电性或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，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单元中，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单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

以上所述仅是本申请的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本申请原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本申请的保护范围。

权 利 要 求

1.一种数据处理方法，应用于移动端设备，包括:

使用人工智能 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，所述 AI 处理模型与所述第一子模型和第二子模型相对应，所述第一子模型根据所述 AI 处理模型中的 M 个神经网络层生成，所述第二子模型根据所述 AI 处理模型中的 K 个神经网络层生成，所述 M、K 为大于或者等于 1 的正整数;

将所述中间处理结果发送给第一服务器，其中，所述第一服务器用于使用所述第二子模型对所述中间处理结果执行第二处理，得到目标处理结果;

接收所述第一服务器返回的所述目标处理结果。

2.根据权利要求 1 所述的方法，在使用所述 AI 处理模型的所述第一子模型对所述待处理数据进行处理之前，所述方法还包括:

接收第二服务器发送的所述第一子模型，其中，所述第二服务器用于使用目标压缩算法对所述 AI 处理模型中拆分出的 M 个神经网络层进行压缩，得到所述第一子模型; 或，

接收第二服务器发送的所述第一子模型，其中，所述第二服务器用于从所述 AI 处理模型中拆分出的 M 个神经网络层作为所述第一子模型。

3.根据权利要求 2 所述的方法，所述第二子模型用于补偿所述 AI 处理模型由于对所述 M 个神经网络层进行压缩所损失的处理精度。

4.根据权利要求 2 所述的方法，在使用所述 AI 处理模型的所述第一子模型对待处理数据进行处理之前，所述方法还包括:

获取所述第一子模型对训练对象执行所述第一处理得到的第一处理结果;

对从所述 AI 处理模型中拆分的 K 个神经网络层进行训练，得到满足训练约束条件的所述第二子模型，其中，所述 K 个神经网络层的输入为所述第一处理结果，所述训练约束条件为: 所述第二子模型的输出结果与第二处理结果的处理精度差值小于或等于目标阈值，其中，所述第二处理结果为使用所述 AI 处理模型对所述训练对象进行处理得到的处理结果。

5.根据权利要求 1 所述的方法，所述使用所述 AI 处理模型对应的所述第一子模型对所述待处理数据执行第一处理，得到中间处理结果包括:

当所述待处理数据为待处理图像时，使用所述第一子模型对所述待处理图像执行第一识别处理，得到中间识别结果，其中，所述 AI 处理模型用于识别所述待处理图像中的目标对象;

接收所述第一服务器返回的所述目标处理结果包括: 接收所述第一服务器返回的目标识别结果，其中，所述目标识别结果由所述第一服务器使用所述第二子模型对所述中间识别结果执行第二识别处理得到，所述目标识别结果用于指示所述待处理图像中的所述目标对象。

6.根据权利要求 1 所述的方法，所述使用所述 AI 处理模型对应的所述第一子模型对所述待处理数据执行第一处理，得到中间处理结果包括:

当所述待处理数据为待处理图像时，使用所述第一子模型对所述待处理图像执行第一去模糊处理，得到中间去模糊结果，其中，所述 AI 处理模型用于对所述待处理图像执行去模糊操作，得到目标去模糊结果；

接收所述第一服务器返回的所述目标处理结果包括：接收所述第一服务器返回的所述目标去模糊结果，其中，所述目标去模糊结果由所述第一服务器使用所述第二子模型对所述中间去模糊结果执行第二去模糊处理得到。

7.根据权利要求 1 所述的方法，所述使用所述 AI 处理模型对应的所述第一子模型对所述待处理数据执行第一处理，得到中间处理结果包括：

当所述待处理数据为待翻译数据时，使用所述第一子模型对待翻译数据执行第一翻译处理，得到中间翻译结果，其中，所述 AI 处理模型用于将所述待翻译数据中包含的使用第一语言的第一语言数据，翻译为使用第二语言的第二语言数据；

接收所述第一服务器返回的所述目标处理结果包括：接收所述第一服务器返回的目标翻译结果，其中，所述目标翻译结果由所述第一服务器使用所述第二子模型对所述中间翻译结果执行第二翻译处理得到，所述目标翻译结果中包含所述第二语言数据。

8.根据权利要求 1 所述的方法，在使用所述 AI 处理模型中的所述第一子模型对所述待处理数据执行所述第一处理之前，所述方法还包括：

通过所述移动端设备自身的数据采集部件获取所述待处理数据，所述数据采集部件包括传感器和/或麦克风。

9.根据权利要求 1 至 8 中任一项所述的方法，所述第一子模型部署在移动端设备上，所述第一服务器为云端服务器。

10.一种数据处理装置，包括：

处理单元，用于使用人工智能 AI 处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，所述 AI 处理模型与所述第一子模型和第二子模型相对应，所述第一子模型根据所述 AI 处理模型中的 M 个神经网络层生成，所述第二子模型根据所述 AI 处理模型中的 K 个神经网络层生成，所述 M、K 为大于或者等于 1 的正整数；

发送单元，用于将所述中间处理结果发送给第一服务器，其中，所述第一服务器用于使用所述第二子模型对所述中间处理结果执行第二处理，得到所述目标处理结果，所述目标处理包括所述第一处理和所述第二处理；

第一接收单元，用于接收所述第一服务器返回的所述目标处理结果。

11.根据权利要求 10 所述的装置，所述装置还包括：

第二接收单元，用于接收第二服务器发送的所述第一子模型，其中，所述第二服务器用于使用目标压缩算法对所述 M 个神经网络层进行压缩，得到所述第一子模型；或，接收第二服务器发送的所述第一子模型，其中，所述第二服务器用于从所述 AI 处理模型中拆分出的 M 个神经网络层作为所述第一子模型。

12.根据权利要求 10 所述的装置，

所述处理单元包括：处理模块，用于当所述待处理数据为待处理图像时，使用所述第一子模型对所述待处理图像执行第一识别处理，得到中间识别结果，其中，所述 AI 处理模型用于识别所述待处理图像中的目标对象；

所述第一接收单元包括：接收模块，用于接收所述第一服务器返回的目标识别结果，其中，所述目标识别结果由所述第一服务器使用所述第二子模型对所述中间识别结果执行第二识别处理得到，所述目标识别结果用于指示所述待处理图像中的所述目标对象。

13.根据权利要求 10 至 12 中任一项所述的装置，所述装置还包括：

获取单元，用于在使用所述 AI 处理模型中的所述第一子模型对所述待处理数据执行所述第一处理之前，通过移动端设备自身的数据采集部件获取所述待处理数据，所述数据采集部件包括传感器和/或麦克风。

14.一种存储介质，所述存储介质中存储有计算机程序，其中，所述计算机程序被设置为运行时执行所述权利要求 1 至 8 任一项中所述的方法。

15.一种电子装置，包括存储器和处理器，所述存储器中存储有计算机程序，所述处理器被设置为通过所述计算机程序执行所述权利要求 1 至 8 任一项中所述的方法。

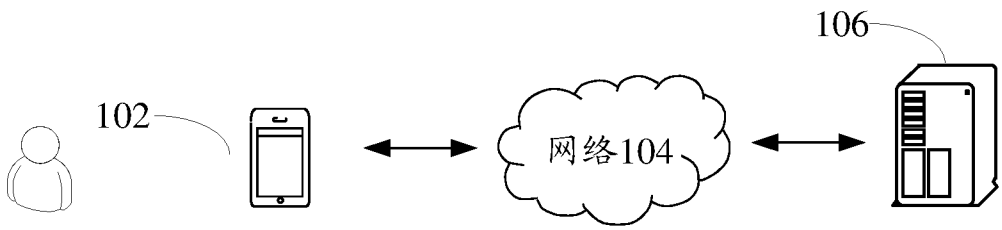

图 1

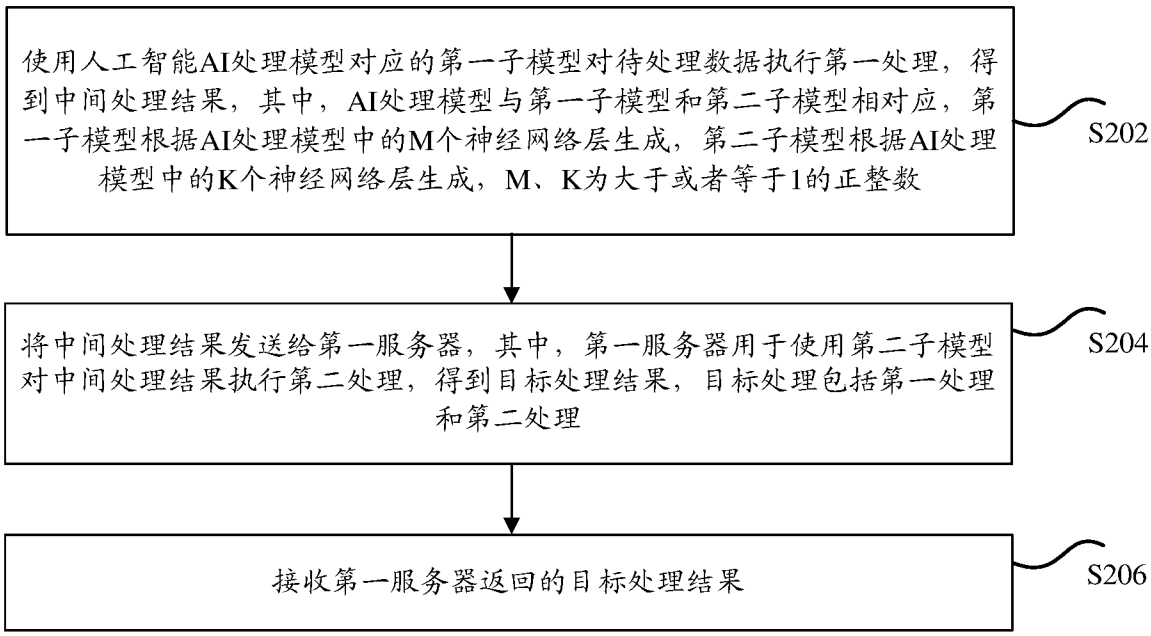
使用人工智能AI处理模型对应的第一子模型对待处理数据执行第一处理，得到中间处理结果，其中，AI处理模型与第一子模型和第二子模型相对应，第一子模型根据AI处理模型中的M个神经网络层生成，第二子模型根据AI处理模型中的K个神经网络层生成，M、K为大于或者等于1的正整数
S202
将中间处理结果发送给第一服务器，其中，第一服务器用于使用第二子模型对中间处理结果执行第二处理，得到目标处理结果，目标处理包括第一处理和第二处理
S204
接收第一服务器返回的目标处理结果
S206

图 2

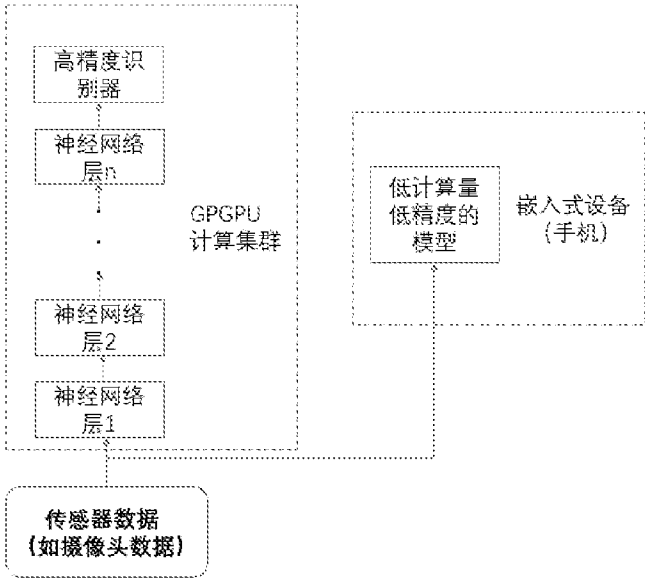
高精度识别器
神经网络层n
神经网络层2
神经网络层1
GPGPU计算集群
低计算量低精度的模型
嵌入式设备（手机）
传感器数据（如摄像头数据）

图 3

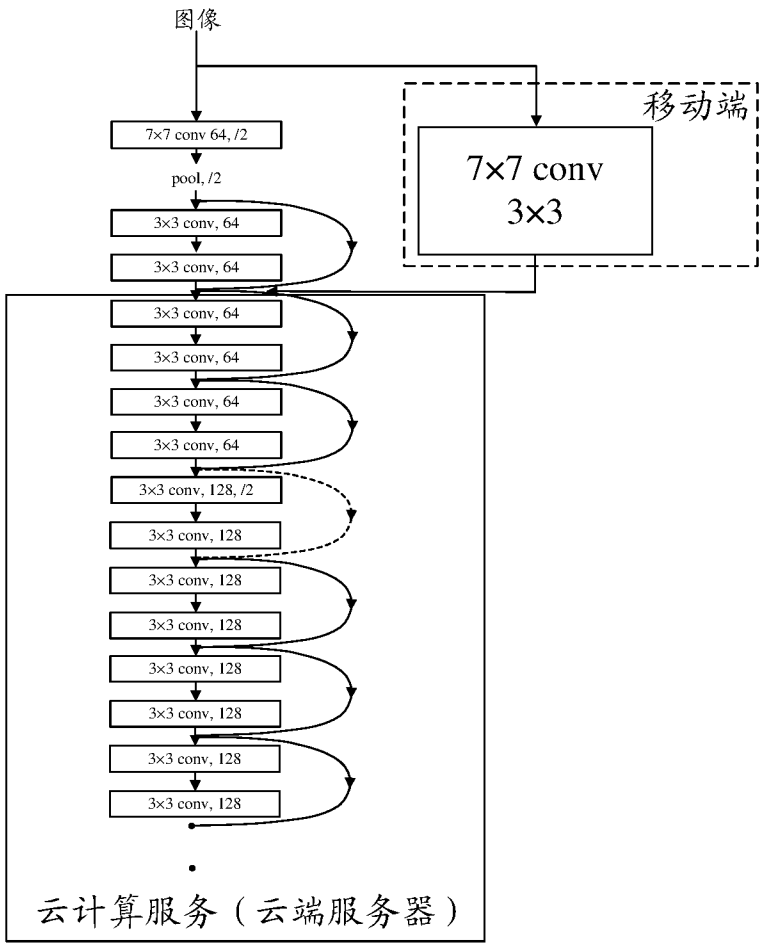
图像
移动端
7×7 conv 64, /2
pool, /2
3×3 conv, 64
3×3 conv, 64
3×3 conv, 64
3×3 conv, 64
3×3 conv, 64
3×3 conv, 64
3×3 conv, 128, /2
3×3 conv, 128
3×3 conv, 128
3×3 conv, 128
3×3 conv, 128
3×3 conv, 128
3×3 conv, 128
3×3 conv, 128
7×7 conv
3×3
云计算服务（云端服务器）

图 4

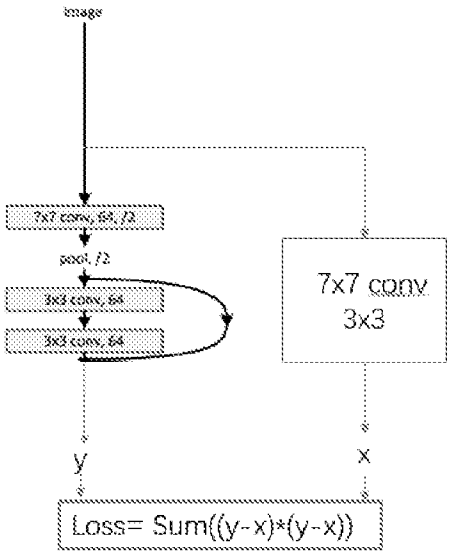
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
7x7 conv
3x3
y
x
Loss= Sum((y-x)*(y-x))

图 5

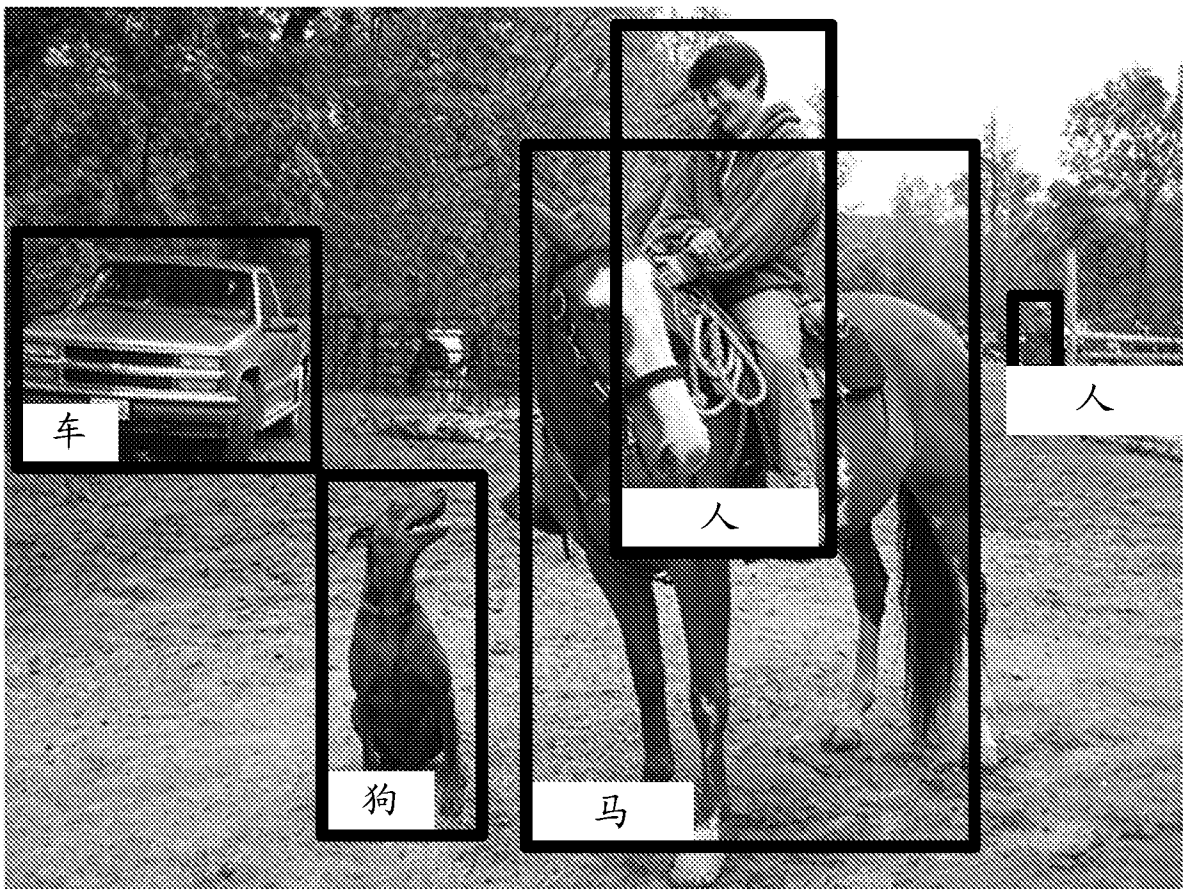
车
狗
人
马
人

图 6

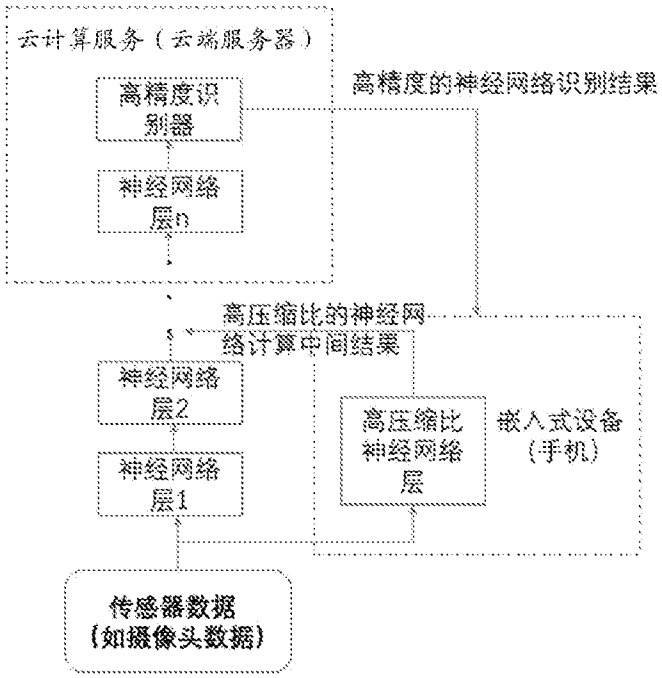
云计算服务（云端服务器）
高精度识别器
神经网络层n
高精度的神经网络识别结果
高压缩比的神经网络计算中间结果
神经网络层2
神经网络层1
高压缩比神经网络层
嵌入式设备（手机）
传感器数据（如摄像头数据）

图 7

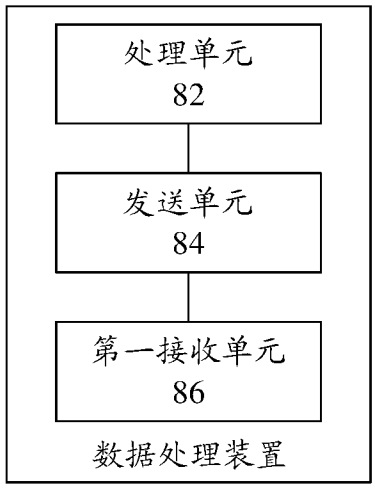
处理单元
82
发送单元
84
第一接收单元
86
数据处理装置

图 8

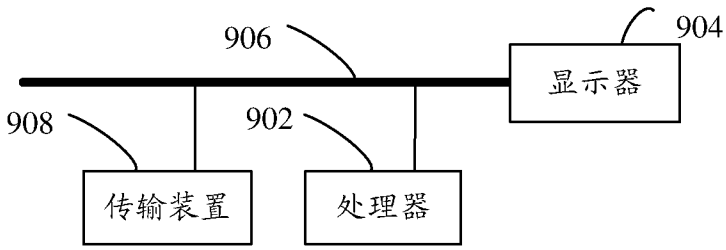
906
904
显示器
908
902
传输装置
处理器

图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/115082

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/04(2006.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N

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)

CNTXT, CNABS, SIPOABS, DWPI, CNKI: 神经网络, 模型, 层, 分段, 训练, 识别, 终端, 服务器, 人工智能; neural net, neural networks, model, layer, fragment, segmentation, training, sensing, recognition, identification, terminal, server, AI, artificial intelligence

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 108875899 A (BEIJING KUANGSHI TECHNOLOGY CO., LTD.) 23 November 2018 (2018-11-23) description, paragraphs 24-190	1, 2, 5-15
Y	CN 108875899 A (BEIJING KUANGSHI TECHNOLOGY CO., LTD.) 23 November 2018 (2018-11-23) description, paragraphs 31-65	3, 4
Y	CN 104933463 A (HANGZHOU LANGHE TECHNOLOGY LIMITED) 23 September 2015 (2015-09-23) description, paragraphs 36-55	3, 4
A	CN 107909147 A (SHENZHEN HARZONE TECHNOLOGY CO., LTD.) 13 April 2018 (2018-04-13) entire document	1-15
PX	CN 109685202 A (TENCENT TECHNOLOGY SHENZHEN CO., LTD.) 26 April 2019 (2019-04-26) entire document	1-15

☐ 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	Date of mailing of the international search report
08 January 2020	**16 January 2020**
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/CN2019/115082

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 108875899 A	23 November 2018	None	
CN 104933463 A	23 September 2015	CN 104933463 B	23 January 2018
CN 107909147 A	13 April 2018	None	
CN 109685202 A	26 April 2019	None	

国际检索报告

国际申请号
PCT/CN2019/115082

A. 主题的分类

G06N 3/04(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06N

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））

CNTXT, CNABS, SIPOABS, DWPI, CNKI: 神经网络，模型，层，分段，训练，识别，终端，服务器，人工智能; neural net, neural networks, model, layer, fragment, segmentation, training, sensing, recognition, identification, terminal, server, AI, artificial intelligence

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 108875899 A (北京旷视科技有限公司) 2018年 11月 23日 (2018-11-23) 说明书第24-190段	1,2,5-15
Y	CN 108875899 A (北京旷视科技有限公司) 2018年 11月 23日 (2018-11-23) 说明书第31-65段	3,4
Y	CN 104933463 A (杭州朗和科技有限公司) 2015年 9月 23日 (2015-09-23) 说明书第36-55段	3,4
A	CN 107909147 A (深圳市华尊科技股份有限公司) 2018年 4月 13日 (2018-04-13) 全文	1-15
PX	CN 109685202 A (腾讯科技深圳有限公司) 2019年 4月 26日 (2019-04-26) 全文	1-15

☐ 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	国际检索报告邮寄日期
2020年 1月 8日	2020年 1月 16日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 程琼 电话号码 86-(10)-62411647

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/115082

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 108875899 A	2018年 11月 23日	无	
CN 104933463 A	2015年 9月 23日	CN 104933463 B	2018年 1月 23日
CN 107909147 A	2018年 4月 13日	无	
CN 109685202 A	2019年 4月 26日	无	