

(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
2023 年 11 月 16 日 (16.11.2023)



(10) 国际公布号
WO 2023/217263 A1

(51) 国际专利分类号:
G06N 3/08 (2023.01)

(21) 国际申请号: PCT/CN2023/093805

(22) 国际申请日: 2023 年 5 月 12 日 (12.05.2023)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202210524932.0 2022年5月13日 (13.05.2022) CN

(71) 申请人: 北京字跳网络技术有限公司
(**BEIJING ZITIAO NETWORK TECHNOLOGY CO., LTD.**) [CN/CN]; 中国北京市海淀区紫金数码园
4号楼2层0207, Beijing 100190 (CN)。

(72) 发明人: 刘松伟 (**LIU, Songwei**); 中国北京市朝阳区七圣中街 12 号院融中心 B1 小邮局, Beijing 100028 (CN)。李明蹊 (**LI, Mingxi**); 中国北京市朝阳区七圣中街 12 号院融中心 B1 小邮局, Beijing 100028 (CN)。孔方圆 (**KONG, Fangyuan**); 中国北京市朝阳区七圣中街 12 号院融中心 B1 小邮局, Beijing 100028 (CN)。陈芳民 (**CHEN, Fangmin**); 中国北京市朝阳区七圣中街 12 号院融中心 B1 小邮局, Beijing 100028 (CN)。拜阳 (**BAI, Yang**); 中国北京市朝阳区七圣中街 12 号院融中心 B1 小邮局, Beijing 100028 (CN)。

(74) 代理人: 中国贸促会专利商标事务所有
限公司 (**CCPIT PATENT AND TRADEMARK LAW OFFICE**); 中国北京市西城区复兴门内大街 158
号远洋大厦 F10 层, Beijing 100031 (CN)。

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

(54) 发明名称: 数据处理方法、装置、设备及介质

对原始神经网络中的候选网络层, 按照预设的多个剪枝率分
别进行剪枝处理获取对应的多个子神经网络 S101

将测试数据集分别输入原始神经网络和多个子神经网络进行
处理, 基于原始神经网络和多个子神经网络的输出数据集,
获取原始神经网络对应的参考性能指标, 以及多个子神经网络
对应的多个测试性能指标 S102

根据多个测试性能指标相对于参考性能指标的性能损失, 分
析原始神经网络中的候选网络层参数在不同剪枝率下的参数
冗余度 S103

图 1

- S101 Respectively perform pruning processing on candidate network layers in an original neural network according to a plurality of preset pruning rates to obtain a plurality of corresponding sub-neural networks
- S102 Respectively input test data sets into the original neural network and the plurality of sub-neural networks for processing, and obtain, on the basis of output data sets of the original neural network and the plurality of sub-neural networks, a reference performance index corresponding to the original neural network and a plurality of test performance indexes corresponding to the plurality of sub-neural networks
- S103 Analyze, according to performance losses of the plurality of test performance indexes relative to a reference performance index, parameter redundancies of parameters of the candidate network layers in the original neural network under different pruning rates

(57) **Abstract:** Embodiments of the present disclosure relate to a data processing method and apparatus, a device, and a medium. The method comprises: respectively performing pruning processing on candidate network layers in an original neural network according to a plurality of preset pruning rates to obtain a plurality of corresponding sub-neural networks; respectively inputting test data sets into the original neural network and the plurality of sub-neural networks for processing, and obtaining, on the basis of output data sets of the original neural network and the plurality of sub-neural networks, a reference performance index corresponding to the original neural

(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, 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))。

network and a plurality of test performance indexes corresponding to the plurality of sub-neural networks; and analyzing, according to performance losses of the plurality of test performance indexes relative to a reference performance index, parameter redundancies of parameters of the candidate network layers in the original neural network under different pruning rates.

(57) 摘要: 本公开实施例涉及一种数据处理方法、装置、设备及介质, 其中该方法包括: 对原始神经网络中的候选网络层, 按照预设的多个剪枝率分别进行剪枝处理, 以获取对应的多个子神经网络, 将测试数据集分别输入原始神经网络和多个子神经网络进行处理, 基于原始神经网络和多个子神经网络的输出数据集, 获取原始神经网络对应的参考性能指标, 以及多个子神经网络对应的多个测试性能指标, 根据多个测试性能指标相对于参考性能指标的性能损失, 分析原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。

数据处理方法、装置、设备及介质

本申请以申请号为 202210524932.0，申请日为 2022 年 5 月 13 日的中国申请为基础，并主张其优先权，该中国申请的公开内容在此作为整体引入本申请中。

5 技术领域

本公开涉及计算机技术领域，尤其涉及一种数据处理方法、装置、设备及介质。

背景技术

10 基于神经网络的人工智能技术在移动终端上的应用，智能移动终端飞速发展满足人们的各种应用需求。其中，其主要实现技术包括基于训练好的神经网络模型数据在视频处理语言识别、图像识别与理解、游戏视觉等应用领域的数据处理。基于移动终端有限的计算资源，考虑到绝大多数的卷积神经网络均存在一定程度上的参数冗余，因此通过剪枝去除神经网络中各层冗余的卷积内核或卷积内核上的神经元，
15 在移动终端上得到计算资源和存储资源更小的神经网络。

发明内容

为了解决上述技术问题或者至少部分地解决上述技术问题，本公开提供了一种数据处理方法、装置、设备及介质。

20 本公开实施例提供了一种数据处理方法，所述方法包括：
对原始神经网络中的候选网络层，按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络；

将测试数据集分别输入所述原始神经网络和多个所述子神经网络进行处理，基于所述原始神经网络和多个所述子神经网络的输出数据集，获取所述原始神经网络对应的参考性能指标，以及多个所述子神经网络对应的多个测试性能指标；
25

根据所述多个测试性能指标相对于所述参考性能指标的性能损

失，分析所述原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。

一种可选的实施方式中，所述的方法，还包括：

获取网络压缩需求；

5

根据所述网络压缩需求设置所述多个剪枝率，其中，所述多个剪枝率之间的差异与网络压缩度正相关。

一种可选的实施方式中，所述按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络，包括：

对所述候选网络层中的权重分布进行范数计算；

10

如果根据计算结果确定所述权重分布属于预设的第一区域分布的候选网络层，则使用预设的第一剪枝器进行剪枝处理，其中，所述第一区域分布的范数区间大于预设的区间阈值，且所述第一区域分布的范数最小值为零；

15

如果根据计算结果确定所述权重分布属于预设的第二区域分布的候选网络层，则使用预设的第二剪枝器进行剪枝处理，其中，所述第二区域分布的范数方差大于预设的方差阈值，且所述第二区域分布的范数最小值不为零。

20

一种可选的实施方式中，所述测试数据集包括：多媒体数据，其中，所述多媒体数据为音频数据、视频数据、图像数据中的一种或者多种组合。

25

一种可选的实施方式中，所述将测试数据集分别输入所述原始神经网络和多个所述子神经网络进行处理，基于所述原始神经网络和多个所述子神经网络的输出数据集，获取所述原始神经网络对应的参考性能指标，以及多个所述子神经网络对应的多个测试性能指标，包括：

将测试图像数据集分别输入所述原始神经网络和每个所述子神经网络进行处理，基于所述原始神经网络和多个所述子神经网络的输出图像数据集与所述测试图像数据集之间的像素处理结果，获取与所述原始神经网络对应的峰值信噪比作为所述参考性能指标，以及与每个

所述子神经网络对应的峰值信噪比作为所述测试性能指标；

或者，

将测试音频数据集分别输入所述原始神经网络和每个所述子神经网络进行处理，基于所述原始神经网络和多个所述子神经网络的输出的识别文字数据集与所述测试音频数据集的标注文字之间的对比结果，获取与所述原始神经网络对应的准确率作为所述参考性能指标，以及与所述子神经网络对应的准确率作为所述测试性能指标。

一种可选的实施方式中，所述的方法，还包括：

检测所述原始神经网络中是否存在具有通道依赖特征的关联网络层，其中，所述通道依赖特征包括：相邻的网络层具有相加的数据操作和相乘的数据操作中的至少一种；

如果存在所述关联网络层，将所述具有通道依赖特征的所有关联网络层设置为一个所述候选网络层。

一种可选的实施方式中，所述方法，还包括：

根据所述候选网络层参数在不同剪枝率下的参数冗余度，确定所述原始神经网络中被剪枝的目标网络层，以生成目标神经网络对目标数据集进行处理。

一种可选的实施方式中，所述根据所述候选网络层参数在不同剪枝率下的参数冗余度，确定所述原始神经网络中被剪枝的目标网络层，包括：

根据所述多个测试性能指标相对于所述参考性能指标的性能损失，绘制所述候选网络层与所述多个剪枝率对应的性能指标曲线；

计算所述性能指标曲线中每个剪枝率的斜率，根据斜率变化确定所述候选网络层的最大剪枝率，其中，所述最大剪枝率对应的性能指标表示所述候选网络层参数的最大参数冗余度；

根据目标剪枝率、每个所述候选网络层的所述最大参数冗余度对应的所述最大剪枝率，确定所述原始神经网络中被剪枝的目标网络层。

本公开实施例还提供了一种数据处理装置，所述装置包括：

剪枝处理模块，用于对原始神经网络中的候选网络层，按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络；

处理获取模块，用于将测试数据集分别输入所述原始神经网络和多个所述子神经网络进行处理，基于所述原始神经网络和多个所述子神经网络的输出数据集，获取所述原始神经网络对应的参考性能指标，以及多个所述子神经网络对应的多个测试性能指标；

确定模块，用于根据所述多个测试性能指标相对于所述参考性能指标的性能损失，分析所述原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。

本公开实施例还提供了一种电子设备，所述电子设备包括：处理器；用于存储可执行指令的存储器；其中所述可执行指令能够由所述处理器从所述存储器中读取，并执行以实现如本公开实施例提供的数据处理方法。

本公开实施例还提供了一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序用于执行如本公开实施例提供的数据处理方法。

本公开实施例还提供了一种计算机程序产品，所述计算机程序产品包括计算机程序/指令，所述计算机程序/指令被处理器执行时实现上述的方法。

本公开实施例还提供了一种计算机程序，包括：指令，所述指令当由处理器执行时使所述处理器执行如本公开实施例提供的数据处理方法。

附图说明

结合附图并参考以下具体实施方式，本公开各实施例的上述和其他特征、优点及方面将变得更加明显。贯穿附图中，相同或相似的附图标记表示相同或相似的元素。应当理解附图是示意性的，原件和元素不一定按照比例绘制。

图 1 为本公开实施例提供的一种数据处理方法的流程示意图；

图 2 为本公开实施例提供的另一种数据处理方法的流程示意图；

图 3 为本公开实施例提供的一种剪枝率和性能指标关系的示意图；

图 4 为本公开实施例提供的另一种剪枝率和性能指标关系的示意图；

图 5 为本公开实施例提供的一种数据处理装置的结构示意图；

图 6 为本公开实施例提供的一种电子设备的结构示意图。

具体实施方式

下面将参照附图更详细地描述本公开的实施例。虽然附图中显示了本公开的某些实施例，然而应当理解的是，本公开可以通过各种形式来实现，而且不应该被解释为限于这里阐述的实施例，相反提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是，本公开的附图及实施例仅用于示例性作用，并非用于限制本公开的保护范围。

应当理解，本公开的方法实施方式中记载的各个步骤可以按照不同的顺序执行，和/或并行执行。此外，方法实施方式可以包括附加的步骤和/或省略执行示出的步骤。本公开的范围在此方面不受限制。

本文使用的术语“包括”及其变形是开放性包括，即“包括但不限于”。术语“基于”是“至少部分地基于”。术语“一个实施例”表示“至少一个实施例”；术语“另一实施例”表示“至少一个另外的实施例”；术语“一些实施例”表示“至少一些实施例”。其他术语的相关定义将在下文描述中给出。

需要注意，本公开中提及的“第一”、“第二”等概念仅用于对不同的装置、模块或单元进行区分，并非用于限定这些装置、模块或单元所执行的功能的顺序或者相互依存关系。

需要注意，本公开中提及的“一个”、“多个”的修饰是示意性而非限制性的，本领域技术人员应当理解，除非在上下文另有明确指出，

否则应该理解为“一个或多个”。

本公开实施方式中的多个装置之间所交互的消息或者信息的名称仅用于说明性的目的，而并不是用于对这些消息或信息的范围进行限制。

5 相关技术中，剪枝方案使得剪枝后的网络处理性能不同，有的剪枝方案导致的网络处理性能比较差，从而导致后续数据处理的结果不可靠。

10 图 1 为本公开实施例提供的一种数据处理方法的流程示意图，该方法可以由数据处理装置执行，其中该装置可以采用软件和/或硬件实现，一般可集成在电子设备中。如图 1 所示，该方法包括：

步骤 101、对原始神经网络中的候选网络层，按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络。

15 为了保证模型效果，在神经网络中可能存在一定数量的参数冗余，通过剪枝处理，能够在保证神经网络精度的前提下，剪枝掉神经网络中各卷积层冗余的卷积内核（即，结构化剪枝）或卷积内核上的神经元（即，非结构化剪枝），从而得到计算资源和存储资源占用更小的“瘦身模型”，加速神经网络的推理过程，助力神经网络的边缘部署。

20 然而，不同的剪枝方案使得剪枝后的网络处理性能不同，有的剪枝方案所导致的网络处理性能比较差，从而导致数据处理的结果不可靠。

在本实施例中，原始神经网络为需要进行剪枝处理的神经网络模型，该神经网络模型可以是经过训练获得的，并且该神经网络模型可以根据应用场景和/或用户需求等进行设定，本实施例不作限制。

25 在本公开实施例中，通过特定的评估准则对原始神经网络中的所有神经元的相对重要性进行排序，然后根据预设的剪枝率剪掉网络中相对不重要的神经元，从而压缩网络模型。

在本公开实施例中，对原始神经网络中的候选网络层，按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络；其

中，原始神经网络包括多个候选网络层，比如原始神经网络中包括四个卷积层，分别为卷积层一 Conv1，卷积层二 Conv2、卷积层三 Conv3 和卷积层四 Conv4，则可以将卷积层一 Conv1，卷积层二 Conv2、卷积层三 Conv3 和卷积层四 Conv4 都作为原始神经网络的候选网络层，或者将卷积层一 Conv1 和卷积层二 Conv2 作为候选网络层，具体根据应用场景需要选择设置。

其中，预先根据每个候选网络层的重要性设置对应的剪枝率；剪枝率指的是候选网络层剪掉的卷积核的百分之比，比如候选网络层 A 具有 N 个卷积核，剪枝率为 p%，因此，候选网络层 A 需要将 N 乘以 p%个卷积核剪掉。

在本公开实施例中，每个候选网络层预先设置多个不同的剪枝率，从而按照预先设置多个不同的剪枝率分别对每个候选网络层进行剪枝处理后，可以得到每个候选网络层对应的多个子神经网络，比如预先设置十个剪枝率，每个剪枝率相差百分之十，分别为百分之十、百分之二十、百分之三十直到百分之百，从而基于十个不同的剪枝率分别对候选网络层比如卷积层一 Conv1 进行处理，从而可以得到卷积层一 Conv1 对应的十个子神经网络。

在本公开实施例中，按照预设的多个剪枝率分别进行剪枝处理以获取对应的多个子神经网络的方式有多种，可以根据应用场景等进行选择，本实施例不做限制，示例说明如下：

一种可选的实施方式中，对候选网络层中的权重分布进行范数计算，如果根据计算结果确定权重分布属于预设的第一区域分布的候选网络层，则使用预设的第一剪枝器进行剪枝处理，其中，第一区域分布的范数区间大于预设的区间阈值，且第一区域分布的范数最小值为零，如果根据计算结果确定权重分布属于预设的第二区域分布的候选网络层，则使用预设的第二剪枝器进行剪枝处理，其中，第二区域分布的范数方差大于预设的方差阈值，且第二区域分布的范数最小值不为零。

另一种可选的实施方式中，根据预设的多个剪枝率调用相关剪枝器直接对原始神经网络的候选网络层进行剪枝处理，得到多个子神经网络。

需要说明的是，每次针对一个候选网络层按照预设的一个剪枝率进行处理，其他候选网络层不变，得到一个子神经网络。

步骤 102、将测试数据集分别输入原始神经网络和多个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出数据集，获取原始神经网络对应的参考性能指标，以及多个子神经网络对应的多个测试性能指标。

在本公开实施例中，测试数据集可以根据应用场景选择设置，比如为多媒体数据，多媒体数据为音频数据、视频数据、图像数据中的一种或者多种组合。

其中，参考性能指标指的是原始神经网络对测试数据集进行处理后得到的输出数据集进行分析得到的性能数值，测试性能指标指的是经过剪枝率处理后的子神经网络对测试数据集进行处理后得到的输出数据集进行分析得到的性能数值。

具体地，在不同剪枝率剪枝处理后的子神经网络对测试数据集进行处理损失的精度不同，即性能损失不同，比如剪枝率为百分之三十对候选网络层剪枝处理后的子神经网络对测试数据集进行处理损失的精度越大，即性能损失越大，表示候选网络层在剪枝率百分之三十时的参数冗余度比较小。

在本公开实施例中，不同场景的测试数据集，得到的参考性能指标和测试性能指标不同，因此，将测试数据集分别输入原始神经网络和多个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出数据集，获取原始神经网络对应的参考性能指标，以及多个子神经网络对应的多个测试性能指标的方式有多种，可以根据应用场景等进行选择，本实施例不做限制，示例说明如下：

一种可选的实施方式中，比如画面质量增强场景，将测试图像数

数据集分别输入原始神经网络和每个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出图像数据集与测试图像数据集之间的像素处理结果，获取与原始神经网络对应的峰值信噪比作为参考性能指标，以及与每个子神经网络对应的峰值信噪比作为测试性能指标。

5 另一种可选的实施方式中，比如语音识别场景，将测试音频数据集分别输入原始神经网络和每个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出的识别文字数据集与测试音频数据集的标注文字之间的对比结果，获取与原始神经网络对应的准确率作为参考性能指标，以及与每个子神经网络对应的准确率作为测试性能指标。

10 步骤 103，根据多个测试性能指标相对于参考性能指标的性能损失，分析原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。

15 在本公开实施例中，针对每个候选网络层，设置多少不同的剪枝率得到多少个子神经网络，多少个子神经网络对应多少个测试性能指标，因此，计算多个测试性能指标相对于参考性能指标的性能损失，可以得到原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。

20 具体地，根据多个测试性能指标相对于参考性能指标的性能损失，绘制候选网络层与所述多个剪枝率对应的性能指标曲线，基于性能指标曲线分析候选网络层参数在不同剪枝率下的参数冗余度。

25 本公开实施例提供的数据处理方案，对原始神经网络中的候选网络层，按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络，将测试数据集分别输入原始神经网络和多个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出数据集，获取原始神经网络对应的参考性能指标，以及多个子神经网络对应的多个测试性能指标，根据多个测试性能指标相对于参考性能指标的性能损失，分析原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。

余度。采用上述技术方案，基于实际的数据集获取参数冗余度，提高后续剪枝的可靠性，从而提高剪枝之后神经网络的高精度，提高数据处理的效率和精确性。

5 在一些实施例中，根据候选网络层参数在不同剪枝率下的参数冗余度，确定原始神经网络中被剪枝的目标网络层，以生成目标神经网络对目标数据集进行处理。

其中，目标网络层指的是根据参数冗余度重新调整候选网络层的剪枝率后，并基于调整后的剪枝率确定的被剪枝的目标网络层，目标神经网络指的是针对原始神经网络中的目标网络层剪掉后的神经网络。
10

在本公开实施例中，根据候选网络层参数在不同剪枝率下的参数冗余度，确定原始神经网络中被剪枝的目标网络层的方式有多种，可以根据应用场景等进行选择，本实施例不做限制，示例说明如下：

15 一种可选的实施方式中，根据多个测试性能指标相对于参考性能指标的性能损失，绘制候选网络层与多个剪枝率对应的性能指标曲线，计算性能指标曲线中每个剪枝率的斜率，根据斜率变化确定候选网络层的最大剪枝率，其中，最大剪枝率对应的性能指标表示候选网络层参数的最大参数冗余度，根据目标剪枝率、每个候选网络层的最大参数冗余度对应的最大剪枝率，确定原始神经网络中被剪枝的目标网络层。
20

另一种可选的实施方式中，根据参数冗余度确定每个候选网络层的最大剪枝率，直接根据最大剪枝率确定原始神经网络中被剪枝的目标网络层。

25 基于上述实施例的描述，不同的应用场景对于网络具有不同压缩需求，从而剪枝率不同，并且需要针对候选网络层进行不同剪枝率的剪枝尝试，以剪枝之后的测试数据集的准确率作为候选网络层在不同剪枝率下的参数冗余度，因此需要不同的范数准则进行评估选择不同的剪枝器进行剪枝处理，以及原始神经网络中不同候选网络层之间通

道数量可能存在依赖关系，存在通道依赖关系的不同候选网络层的剪枝需要关联以进一步提高处理效率。

本公开实施例，从不同剪枝率得到子神经网络对测试数据集处理得到的测试性能指标和原始神经网络对测试数据集处理得到的参考性能指标之间的性能损失分析候选网络层的参数冗余度，从而得到原始神经网络中指定的候选网络层在指定的剪枝率下的相对参数冗余度，并且该参数冗余度是基于实际的测试数据集计算得到的，具有较高的可靠性。此外，在参数冗余度的求解过程中，为了进一步提高可靠性，针对不同的权重分布层选择不同的剪枝器；同时针对具有通道依赖关系的候选网络层针对每一个剪枝率综合考虑计算应该剪枝的候选网络层，然后再分别计算每个候选网络层的参数冗余度，最后以这些层的参数冗余度平均值作为所有层的参数冗余度，实现通道依赖感知的参数冗余计算，下面结合图 2 进行详细描述。

具体地，图 2 为本公开实施例提供的另一种数据处理方法的流程示意图，本实施例在上述实施例的基础上，进一步优化了上述数据处理方法。如图 2 所示，该方法包括：

步骤 201、获取网络压缩需求，根据网络压缩需求设置多个剪枝率，其中，多个剪枝率之间的差异与网络压缩度正相关。

具体地，在参数冗余度分析过程中，需要针对候选网络层进行不同的剪枝率的剪枝尝试，以剪枝之后的测试数据集准确率作为候选网络层在不同剪枝率下的参数冗余度。

在本公开实施例中，不同的应用场景对于网络压缩需求不同，比如音频处理平台对于网络压缩需求比较高，从而需要设置更多个数的剪枝率进行剪枝尝试，以需要更加精确的参数冗余度，从而进一步提高最终获取目标神经网络的处理精度；再比如图像处理平台对于网络压缩需求相对较低，从而需要设置相对少个数的剪枝率进行剪枝尝试，以提高原始神经网络调整效率。

其中，多个剪枝率之间的数据差与网络压缩度正相关，也就是说，

多个剪枝率之间的差异越大，网络压缩度越大；多个剪枝率之间的差异越小，网络压缩度越小。

步骤 202，检测原始神经网络中是否存在具有通道依赖特征的关联网络层，其中，通道依赖特征包括：相邻的网络层具有相加的数据操作和相乘的数据操作中的至少一种，如果存在关联网络层，将具有通道依赖特征的所有关联网络层设置为一个候选网络层。

具体地，由于原始神经网络中候选网络层之间通道数量可能存在依赖关系，具有通道依赖特征的关联网络层的剪枝需要对齐进行才能取得实际的加速效果，因此，在分析关联网络层的剪枝敏感性的时候应该将所有关联网络层设置为一个候选网络层，因此具有通道依赖特征的关联网络层具有相同的参数冗余度。

在本公开实施例中，根据某种卷积核评估准则选取第 n 候选网络层的剪枝率为 $p\%$ 的卷积核，即 $N \times p\%$ 个卷积核剪掉，该原始神经网络其余所有的层保持不变，直接测试此时原始神经网络在测试数据集上的性能为 B ，定义第 n 候选网络层在 $p\%$ 剪枝率下的剪枝性能损失为 S ， S 越大表示对该候选网络层进行剪枝造成的精度损失越大，该候选网络层的剪枝敏感性越大，剪枝越敏感就意味着该候选网络层包含了比较多的较为重要的卷积核/特征图，因此可以认为该候选网络层的参数冗余度越小，因此，参数冗余度与剪枝敏感性呈负相关的关系。

在本公开实施例中，多个性能指标为候选网络层与多个剪枝率分别对应的参数冗余度，包括：获取候选网络层中具有通道依赖特征的所有关联网络层的层数，将多个性能指标对层数取平均，获取每个关联网络层与多个剪枝率分别对应的参数冗余度。

举例而言，假设现在存在具有通道依赖特征的两个层 Conv1 和 Conv2，分析 Conv1 和 Conv2 在剪枝率 $p\%$ 情况下的参数冗余度，首先综合考虑 Conv1 和 Conv2 选取 $N \times p\%$ 个待剪掉的卷积核，然后根据在 Conv1 中剪掉这些卷积核计算第一参数冗余度和第二参数冗余度，最后计算第一参数冗余度和第二参数冗余度的平均值作为 Conv1 和

Conv2 在压缩率为 $p\%$ 时的参数冗余度。

步骤 203, 对候选网络层中的权重分布进行范数计算, 如果根据计算结果确定权重分布属于预设的第一区域分布的候选网络层, 则使用预设的第一剪枝器进行剪枝处理, 其中, 第一区域分布的范数区间大于预设的区间阈值, 且第一区域分布的范数最小值为零。

步骤 204, 如果根据计算结果确定权重分布属于预设的第二区域分布的候选网络层, 则使用预设的第二剪枝器进行剪枝处理, 其中, 第二区域分布的范数方差大于预设的方差阈值, 且第二区域分布的范数最小值不为零。

具体地, 在进行参数冗余度分析过程中, 需要针对候选网络层进行不同剪枝率的剪枝尝试, 以剪枝之后的测试数据集的准确率作为候选网络层在不同剪枝率下的参数冗余度。因此剪枝性能直接影响不同剪枝率的可信度。

具体地, 剪枝策略通常使用 L1 范数/L2 范数来评估卷积核的重要性, 基于范数的评估准则通常依赖于两个并不总是成立的假设: (1) 滤波器的范数分布广, 方差大; (2) 滤波器的最小范数应该很小, 逼近于 0。具体地, 当滤波器的范数偏差很小的时候, 也就是说滤波器的范数分布非常密集, 那么将会比较难找到一个合适的阈值来达到想要的目标稀疏率。同时当滤波器的最小范数很大的时候, 表示该候选网络层的所有的滤波器都非常重要, 此时再基于范数进行选择便会丧失准确性。

因此, 在上述两种情况下, 基于范数的评估准则变得不再适用, 本公开实施例中, 在剪枝之前会首先分析候选网络层的权重分布, 针对权重分布符合第一区域分布的候选网络层, 使用预设的第一剪枝器进行剪枝处理, 即采用第一范数的一次性 (one shot) 剪枝算法, 第一区域分布的范数大于预设范围且最小值为零, 其中, 预设范围根据应用场景需要设置; 针对权重分布符合第二区域分布的候选网络层, 使用预设的第二剪枝器进行剪枝处理, 即采用通过几何中值的过滤器剪

枝的一次性剪枝算法，第二区域的范数方差大于预设阈值且最小值不为零。其中，预设阈值根据应用场景需要设置。

5 步骤 205、将测试图像数据集分别输入原始神经网络和每个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出图像数据集与测试图像数据集之间的像素处理结果，获取与原始神经网络对应的峰值信噪比作为参考性能指标，以及与每个子神经网络对应的峰值信噪比作为测试性能指标。

10 在本公开实施例中，针对画面增强场景，需要对图像进行增强处理，测试数据集为测试图像数据集，将测试图像数据集分别输入原始神经网络和每个子神经网络进行处理获取输出图像数据集，通过输出图像数据集与测试图像数据集之间的像素处理结果，获取与原始神经网络对应的峰值信噪比作为参考性能指标，以及与每个子神经网络对应的峰值信噪比作为测试性能指标。

15 由此，在画面增强场景中，原始神经网络对应的参考性能指标，以及子神经网络对应的测试性能指标，从而基于测试性能指标相对于参考性能指标的性能损失确定原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度，从而基于该参数冗余度进行剪枝后的网络进行图像增强处理的具有更好处理效率和效果。

20 步骤 206，将测试音频数据集分别输入原始神经网络和每个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出的识别文字数据集与测试音频数据集的标注文字之间的对比结果，获取与原始神经网络对应的准确率作为参考性能指标，以及与每个子神经网络对应的准确率作为测试性能指标。

25 在本公开实施例中，针对语音识别场景，需要对语音进行识别处理，测试数据集为测试音频数据集，将测试音频数据集分别输入原始神经网络和每个子神经网络进行处理获取识别文字数据集，通过识别文字数据集与测试音频数据集的标注文字之间的对比结果，获取原始神经网络对应的准确率作为参考性能指标，以及与每个子神经网络对

应的准确率作为测试性能指标。

由此，在语音识别场景中，原始神经网络对应的参考性能指标，以及子神经网络对应的测试性能指标，从而基于测试性能指标相对于参考性能指标的性能损失确定原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度，从而基于该参数冗余度进行剪枝后的网络进行语音识别处理的具有更好处理效率和效果。

步骤 207，根据多个测试性能指标相对于参考性能指标的性能损失，绘制候选网络层与多个剪枝率对应的性能指标曲线，计算性能指标曲线中每个剪枝率的斜率，根据斜率变化确定候选网络层的最大剪枝率，其中，最大剪枝率对应的性能指标表示候选网络层参数的最大参数冗余度。

步骤 208，根据目标剪枝率、每个候选网络层的最大参数冗余度对应的最大剪枝率，确定原始神经网络中被剪枝的目标网络层。

在本公开实施例中，根据参数冗余度绘制候选网络层与多个剪枝率对应的性能指标曲线，也就是说，将多个剪枝率作为横坐标，参数冗余度，即多个测试性能指标相对于参考性能指标的性能损失作为纵坐标绘制性能指标曲线，从而可以得到每个剪枝率的斜率，根据斜率变化确定候选网络层的最大剪枝率，比如斜率变化最大时的剪枝率为候选网络层的最大剪枝率，最大剪枝率对应的性能指标表示候选网络层参数对应的最大参数冗余度。

进一步地，根据目标剪枝率、每个候选网络层的最大参数冗余度对应的最大剪枝率确定原始神经网络中被剪枝的目标网络层，以生成目标神经网络对目标数据集进行处理，也就是说，在确定最大剪枝率后还可以基于具体场景确定目标剪枝率，以及每个候选网络层的最大参数冗余度确定原始神经网络中被剪枝的目标网络层，以生成目标神经网络对目标数据集进行处理，从而获取的目标神经网络更符合个性化需求，进一步提高数据处理效率和精度。

作为一种场景举例，可使用相关工具对原始神经网络的指定的候

选网络层进行预设的剪枝率的性能损失分析，分析原理是分别对设定的候选网络层进行预设的剪枝率的结构化剪枝，然后将剪枝后的子神经网络对测试数据集处理以进行性能验证，作为该候选网络层当前剪枝率下的性能损失，分析结果如图 3 所示，各候选网络层之间的剪枝性能损失存在较大的差异，对于一些关键的候选网络层，比如 conv1 和 conv2 分别在剪枝率为 0.2 和 0.3 时，性能损失比较多，即 conv1 对剪枝率为 0.2 的性能损失比较大，此时的参数冗余度比较小，同理 conv1 对剪枝率为 0.3 的性能损失比较大，此时的参数冗余度比较小。

再比如个别层对剪枝处理极不敏感，可以考虑在原始神经网络设计中将其去除，或者增大其剪枝率，比如 conv3 在剪枝率为 0.1-0.9 对应的性能指标几乎没有变化，即 conv3 对剪枝率为 0.1-0.9 的性能损失都比较小，即 conv3 在剪枝率为 0.1-0.9 的参数冗余度都比较大。

作为另一种场景举例，采用注意力模块通过建模空间依赖关系使原始神经网络关注更为重要的空间特征，显示优异的性能。如图 4 所示，分析了注意力模块的参数冗余度，可以看到三个卷积层 c1-c3 的参数在各个剪枝率下的参数冗余度都比较靠前，因此证明其具有较高的参数冗余度，直接将三个卷积层 c1-c3 减少为一个卷积层，最后重新训练原始神经网络后显示性能无损。

本公开实施例提供的数据处理方案，获取网络压缩需求，根据网络压缩需求设置多个剪枝率，其中，多个剪枝率之间的差异与网络压缩度正相关，检测原始神经网络中是否存在具有通道依赖特征的关联网络层，其中，通道依赖特征包括：相邻的网络层具有相加的数据操作和相乘的数据操作中的至少一种，如果存在关联网络层，将具有通道依赖特征的所有关联网络层设置为一个候选网络层，对候选网络层中的权重分布进行范数计算，如果根据计算结果确定权重分布属于预设的第一区域分布的候选网络层，则使用预设的第一剪枝器进行剪枝处理，其中，第一区域分布的范数区间大于预设的区间阈值，且第一区域分布的范数最小值为零，如果根据计算结果确定权重分布属于预

5 设的第二区域分布的候选网络层，则使用预设的第二剪枝器进行剪枝处理，其中，第二区域分布的范数方差大于预设的方差阈值，且第二区域分布的范数最小值不为零，将测试图像数据集分别输入原始神经网络和每个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出图像数据集与测试图像数据集之间的像素处理结果，获取与原始神经网络对应的峰值信噪比作为参考性能指标，以及与每个子神经网络对应的峰值信噪比作为测试性能指标，或将测试音频数据集分别输入原始神经网络和每个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出的识别文字数据集与测试音频数据集的标注文字之间的对比结果，获取与原始神经网络对应的准确率作为参考性能指标，以及

10 与每个子神经网络对应的准确率作为测试性能指标，根据多个测试性能指标相对于参考性能指标的性能损失，绘制候选网络层与多个剪枝率对应的性能指标曲线，计算性能指标曲线中每个剪枝率的斜率，根据斜率变化确定候选网络层的最大剪枝率，其中，最大剪枝率对应的性能指标表示候选网络层参数的最大参数冗余度，根据目标剪枝率、每个候选网络层的最大参数冗余度对应的最大剪枝率，确定原始神经网络中被剪枝的目标网络层。采用上述技术方案，分析原始神经网络中每个候选网络层在指定的剪枝率下的相对参数冗余度，并且该参数冗余度是基于实际的测试数据集获取的，具有较高的可靠性，此外，在参数冗余度的求解过程中，为了进一步提高可靠性，针对候选网络层中的权重分布选择不同的剪枝器进行剪枝处理，同时将具有通道依赖特征的所有关联网络层设置为一个候选网络层，以及根据所有关联网络层的层数平均计算获取每个关联网络层与多个剪枝率分别对应的参数冗余度，进一步提高后续计算的精确度，从而提高目标神经网络的可靠性。

25

图 5 为本公开实施例提供的一种数据处理装置的结构示意图，该装置可由软件和/或硬件实现，一般可集成在电子设备中。如图 5 所示，该装置包括：

剪枝处理模块 301, 用于对原始神经网络中的每个候选网络层, 按照预设的多个剪枝率分别进行剪枝处理, 以获取对应的多个子神经网络;

5 处理获取模块 302, 用于将测试数据集分别输入所述原始神经网络和多个所述子神经网络进行处理, 基于所述原始神经网络和多个所述子神经网络的输出数据集, 获取所述原始神经网络对应的参考性能指标, 以及多个所述子神经网络对应的多个测试性能指标;

10 分析模块 303, 用于根据所述多个测试性能指标相对于所述参考性能指标的性能损失, 分析所述原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。

可选的, 所述测试数据集包括: 多媒体数据, 多媒体数据为音频数据、视频数据、图像数据中的一种或者多种组合。

可选的, 所述装置, 还包括:

15 获取模块, 用于获取网络压缩需求;

20 设置模块, 用于根据所述网络压缩需求设置所述多个剪枝率, 其中, 所述多个剪枝率之间的差异与网络压缩度正相关。

可选的, 所述剪枝处理模块 301 具体用于:

对所述候选网络层中的权重分布进行范数计算;

25 如果根据计算结果确定所述权重分布属于预设的第一区域分布的候选网络层, 则使用预设的第一剪枝器进行剪枝处理, 其中, 所述第一区域分布的范数区间大于预设的区间阈值, 且所述第一区域分布的范数最小值为零;

如果根据计算结果确定所述权重分布属于预设的第二区域分布的候选网络层, 则使用预设的第二剪枝器进行剪枝处理, 其中, 所述第二区域分布的范数方差大于预设的方差阈值, 且所述第二区域分布的范数最小值不为零。

可选的, 所述处理获取模块 302 具体用于:

将测试图像数据集分别输入所述原始神经网络和每个所述子神经

网络进行处理，基于所述原始神经网络和多个所述子神经网络的输出图像数据集与所述测试图像数据集之间的像素处理结果，获取与所述原始神经网络对应的峰值信噪比作为所述参考性能指标，以及与每个所述子神经网络对应的峰值信噪比作为所述测试性能指标；

5

或者，

将测试音频数据集分别输入所述原始神经网络和每个所述子神经网络进行处理，基于所述原始神经网络和多个所述子神经网络的输出的识别文字数据集与所述测试音频数据集的标注文字之间的对比结果，获取与所述原始神经网络对应的准确率作为所述参考性能指标，以及与所述子神经网络对应的准确率作为所述测试性能指标。

10

可选的，所述装置，还包括：

检测模块，用于检测所述原始神经网络中是否存在具有通道依赖特征的关联网络层，其中，所述通道依赖特征包括：相邻的网络层具有相加的数据操作和相乘的数据操作中的至少一种；

15

关联设置模块，用于如果存在所述关联网络层，将所述具有通道依赖特征的所有关联网络层设置为一个所述候选网络层。

可选的，所述装置还包括：

获取计算模块，用于获取所述候选网络层中具有通道依赖特征的所有关联网络层的层数，将所述多个性能指标对所述层数取平均，获取每个关联网络层与多个所述剪枝率分别对应的参数冗余度。

20

可选的，所述装置还包括确定模块，用于：

根据所述候选网络层参数在不同剪枝率下的参数冗余度，确定所述原始神经网络中被剪枝的目标网络层，以生成目标神经网络对目标数据集进行处理。

25

可选的，所述确定模块具体用于：

根据所述多个测试性能指标相对于所述参考性能指标的性能损失，绘制所述候选网络层与所述多个剪枝率对应的性能指标曲线；

计算所述性能指标曲线中每个剪枝率的斜率，根据斜率变化确定

所述候选网络层的最大剪枝率，其中，所述最大剪枝率对应的性能指标表示所述候选网络层参数的最大参数冗余度；

根据目标剪枝率、每个所述候选网络层的所述最大参数冗余度对应的所述最大剪枝率，确定所述原始神经网络中被剪枝的目标网络层。
5 本公开实施例所提供的数据处理装置可执行本公开任意实施例所提供的数据处理方法，具备执行方法相应的功能模块和有益效果。

本公开实施例提供的技术方案与相关技术相比具有如下优点：本公开实施例提供的数据处理方案，对原始神经网络中的候选网络层，按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络，
10 将测试数据集分别输入原始神经网络和多个子神经网络进行处理，基于原始神经网络和多个子神经网络的输出数据集，获取原始神经网络对应的参考性能指标，以及多个子神经网络对应的多个测试性能指标，根据多个测试性能指标相对于参考性能指标的性能损失，分析原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。
15 采用上述技术方案，基于实际的测试数据集获取参数冗余度，提高后续剪枝的可靠性，从而提高剪枝之后神经网络的高精度，提高数据处理的效率和精确性。

本公开实施例还提供了一种计算机程序产品，包括计算机程序/指令，
20 该计算机程序/指令被处理器执行时实现本公开任意实施例所提供的数据处理方法。

图 6 为本公开实施例提供的一种电子设备的结构示意图。下面具体参考图 6，其示出了适于用来实现本公开实施例中的电子设备 400 的结构示意图。本公开实施例中的电子设备 400 可以包括但不限于诸如移动电话、笔记本电脑、数字广播接收器、PDA（个人数字助理）、
25 PAD（平板电脑）、PMP（便携式多媒体播放器）、车载终端（例如车载导航终端）等等的移动终端以及诸如数字 TV、台式计算机等等的固定终端。图 6 示出的电子设备仅仅是一个示例，不应对本公开实施例

的功能和使用范围带来任何限制。

如图 6 所示，电子设备 400 可以包括处理装置（例如中央处理器、图形处理器等）401，其可以根据存储在只读存储器（ROM）402 中的程序或者从存储装置 408 加载到随机访问存储器（RAM）403 中的程序而执行各种适当的动作和处理。在 RAM 403 中，还存储有电子设备 400 操作所需的各种程序和数据。处理装置 401、ROM 402 以及 RAM 403 通过总线 404 彼此相连。输入/输出（I/O）接口 405 也连接至总线 404。

通常，以下装置可以连接至 I/O 接口 405：包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置 406；包括例如液晶显示器（LCD）、扬声器、振动器等的输出装置 407；包括例如磁带、硬盘等的存储装置 408；以及通信装置 409。通信装置 409 可以允许电子设备 400 与其他设备进行无线或有线通信以交换数据。虽然图 6 示出了具有各种装置的电子设备 400，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。

特别地，根据本公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在非暂态计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信装置 409 从网络上被下载和安装，或者从存储装置 408 被安装，或者从 ROM 402 被安装。在该计算机程序被处理装置 401 执行时，执行本公开实施例的数据处理方法中限定的上述功能。

需要说明的是，本公开上述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机

可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中，计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：电线、光缆、RF（射频）等等，或者上述的任意合适的组合。

在一些实施方式中，客户端、服务器可以利用诸如 HTTP（Hyper Text Transfer Protocol，超文本传输协议）之类的任何当前已知或未来研发的网络协议进行通信，并且可以与任意形式或介质的数字数据通信（例如，通信网络）互连。通信网络的示例包括局域网（“LAN”），广域网（“WAN”），网际网（例如，互联网）以及端对端网络（例如，ad hoc 端对端网络），以及任何当前已知或未来研发的网络。

上述计算机可读介质可以是上述电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。

上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该电子设备执行时，使得该电子设备：在视频的播放过程中，接收用户的信息展示触发操作；获取所述视频关联的至少两个目标信息；在所述视频的播放页面的信息展示区域中展示所述至少两个目标信息中的第一目标信息其中，所述信息展示区域的尺寸小于所述

播放页面的尺寸；接收用户的第一切换触发操作，将所述信息展示区域中展示的所述第一目标信息切换为所述至少两个目标信息中的第二目标信息。

可以以一种或多种程序设计语言或其组合来编写用于执行本公开的
5 操作的计算机程序代码，上述程序设计语言包括但不限于面向对象的
程序设计语言—诸如 Java、Smalltalk、C++，还包括常规的过程式程
序设计语言—诸如“C”语言或类似的设计语言。程序代码可以完全
地地在用户计算机上执行、部分地在用户计算机上执行、作为一个独
立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或
10 者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，
远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网
(WAN)—连接到用户计算机，或者，可以连接到外部计算机（例如利
用因特网服务提供商来通过因特网连接）。

附图中的流程图和框图，图示了按照本公开各种实施例的系统、
15 方法和计算机程序产品的可能实现的体系架构、功能和操作。在这
点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代
码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于
实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的
实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发
20 生。例如，两个接连地表示的方框实际上可以基本并行地执行，它
们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注
意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的
方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统
来实现，或者可以用专用硬件与计算机指令的组合来实现。

25 描述于本公开实施例中所涉及到的单元可以通过软件的方式实
现，也可以通过硬件的方式来实现。其中，单元的名称在某种情况
下并不构成对该单元本身的限定。

本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部

件来执行。例如，非限制性地，可以使用的示范类型的硬件逻辑部件包括：现场可编程门阵列（FPGA）、专用集成电路（ASIC）、专用标准产品（ASSP）、片上系统（SOC）、复杂可编程逻辑设备（CPLD）等等。

5 在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者
10 上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦除可编程只读存储器（EPROM或快闪存储器）、光纤、便捷式紧凑盘只读存储器（CD-ROM）、光学储存设备、磁储存设备、或上述内容的任何合适组合。

15 根据本公开的一个或多个实施例，本公开提供了一种电子设备，包括：

处理器；

用于存储可执行指令的存储器；

20 其中所述可执行指令能够由所述处理器从所述存储器中读取，并执行以实现如本公开提供的任一所述的数据处理方法。

根据本公开的一个或多个实施例，本公开提供了一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序用于执行如本公开提供的任一所述的数据处理方法。

25 根据本公开的一个或多个实施例，本公开提供了一种计算机程序，包括：指令，所述指令当由处理器执行时使所述处理器执行如本公开提供的任一所述的数据处理方法。

以上描述仅为本公开的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本公开中所涉及的公开范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述公开构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开中公开的（但不限于）具有类似功能的技术特征进行互相替换而形成的技术方案。

此外，虽然采用特定次序描绘了各操作，但是这不应理解为要求这些操作以所示出的特定次序或以顺序次序执行来执行。在一定环境下，多任务和并行处理可能是有利的。同样地，虽然在上面论述中包含了若干具体实现细节，但是这些不应被解释为对本公开的范围的限制。在单独的实施例的上下文中描述的某些特征还可以组合地实现在单个实施例中。相反地，在单个实施例的上下文中描述的各种特征也可以单独地或以任何合适的子组合的方式实现在多个实施例中。

尽管已经采用特定于结构特征和/或方法逻辑动作的语言描述了本主题，但是应当理解所附权利要求书中所限定的主题未必局限于上面描述的特定特征或动作。相反，上面所描述的特定特征和动作仅仅是实现权利要求书的示例形式。

权 利 要 求

1.一种数据处理方法，包括：

对原始神经网络中的候选网络层，按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络；

将测试数据集分别输入所述原始神经网络和多个所述子神经网络进行处理，基于所述原始神经网络和多个所述子神经网络的输出数据集，获取所述原始神经网络对应的参考性能指标，以及多个所述子神经网络对应的多个测试性能指标；

根据所述多个测试性能指标相对于所述参考性能指标的性能损失，分析所述原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。

2.根据权利要求1所述的数据处理方法，还包括：

获取网络压缩需求；

根据所述网络压缩需求设置所述多个剪枝率，其中，所述多个剪枝率之间的差异与网络压缩度正相关。

3.根据权利要求1所述的数据处理方法，其中，所述按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络，包括：

对所述候选网络层中的权重分布进行范数计算；

如果根据计算结果确定所述权重分布属于预设的第一区域分布的候选网络层，则使用预设的第一剪枝器进行剪枝处理，其中，所述第一区域分布的范数区间大于预设的区间阈值，且所述第一区域分布的范数最小值为零；

如果根据计算结果确定所述权重分布属于预设的第二区域分布的候选网络层，则使用预设的第二剪枝器进行剪枝处理，其中，所述第二区域分布的范数方差大于预设的方差阈值，且所述第二区域分布的范数最小值不为零。

4.根据权利要求1所述的数据处理方法，其中，所述测试数据集包括：多媒体数据，其中，所述多媒体数据为音频数据、视频数据、图

像数据中的一种或者多种组合。

5.根据权利要求4所述的数据处理方法,其中,所述将测试数据集分别输入所述原始神经网络和多个所述子神经网络进行处理,基于所述原始神经网络和多个所述子神经网络的输出数据集,获取所述原始神经网络对应的参考性能指标,以及多个所述子神经网络对应的多个测试性能指标,包括:

将测试图像数据集分别输入所述原始神经网络和每个所述子神经网络进行处理,基于所述原始神经网络和多个所述子神经网络的输出图像数据集与所述测试图像数据集之间的像素处理结果,获取与所述原始神经网络对应的峰值信噪比作为所述参考性能指标,以及与每个所述子神经网络对应的峰值信噪比作为所述测试性能指标;

或者,

将测试音频数据集分别输入所述原始神经网络和每个所述子神经网络进行处理,基于所述原始神经网络和多个所述子神经网络的输出的识别文字数据集与所述测试音频数据集的标注文字之间的对比结果,获取与所述原始神经网络对应的准确率作为所述参考性能指标,以及与每个所述子神经网络对应的准确率作为所述测试性能指标。

6.根据权利要求1所述的数据处理方法,还包括:

检测所述原始神经网络中是否存在具有通道依赖特征的关联网络层,其中,所述通道依赖特征包括:相邻的网络层具有相加的数据操作和相乘的数据操作中的至少一种;

如果存在所述关联网络层,将所述具有通道依赖特征的所有关联网络层设置为一个所述候选网络层。

7.根据权利要求1-6任一项所述的数据处理方法,还包括:

根据所述候选网络层参数在不同剪枝率下的参数冗余度,确定所述原始神经网络中被剪枝的目标网络层,以生成目标神经网络对目标数据集进行处理。

8.根据权利要求7所述的数据处理方法,其中,所述根据所述候选

网络层参数在不同剪枝率下的参数冗余度，确定所述原始神经网络中被剪枝的目标网络层，包括：

根据所述多个测试性能指标相对于所述参考性能指标的性能损失，绘制所述候选网络层与所述多个剪枝率对应的性能指标曲线；

计算所述性能指标曲线中每个剪枝率的斜率，根据斜率变化确定所述候选网络层的最大剪枝率，其中，所述最大剪枝率对应的性能指标表示所述候选网络层参数的最大参数冗余度；

根据目标剪枝率、每个所述候选网络层的所述最大参数冗余度对应的所述最大剪枝率，确定所述原始神经网络中被剪枝的目标网络层。

9.一种数据处理装置，包括：

剪枝处理模块，用于对原始神经网络中的候选网络层，按照预设的多个剪枝率分别进行剪枝处理，以获取对应的多个子神经网络；

处理获取模块，用于将测试数据集分别输入所述原始神经网络和多个所述子神经网络进行处理，基于所述原始神经网络和多个所述子神经网络的输出数据集，获取所述原始神经网络对应的参考性能指标，以及多个所述子神经网络对应的多个测试性能指标；

分析模块，用于根据所述多个测试性能指标相对于所述参考性能指标的性能损失，分析所述原始神经网络中的候选网络层参数在不同剪枝率下的参数冗余度。

10.一种电子设备，所述电子设备包括：

处理器；

用于存储可执行指令的存储器；

其中所述可执行指令能够由所述处理器从所述存储器中读取，并执行以实现上述权利要求 1-8 中任一项所述的数据处理方法。

11.一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序用于执行上述权利要求 1-8 中任一项所述的数据处理方法。

12.一种计算机程序产品，所述计算机程序产品包括计算机程序，

所述计算机程序被处理器执行时实现如权利要求 1-8 中任一项所述的数据处理方法。

13、一种计算机程序，包括：

指令，所述指令当由处理器执行时使所述处理器执行根据权利要求 1-8 中任一项所述的数据处理方法。

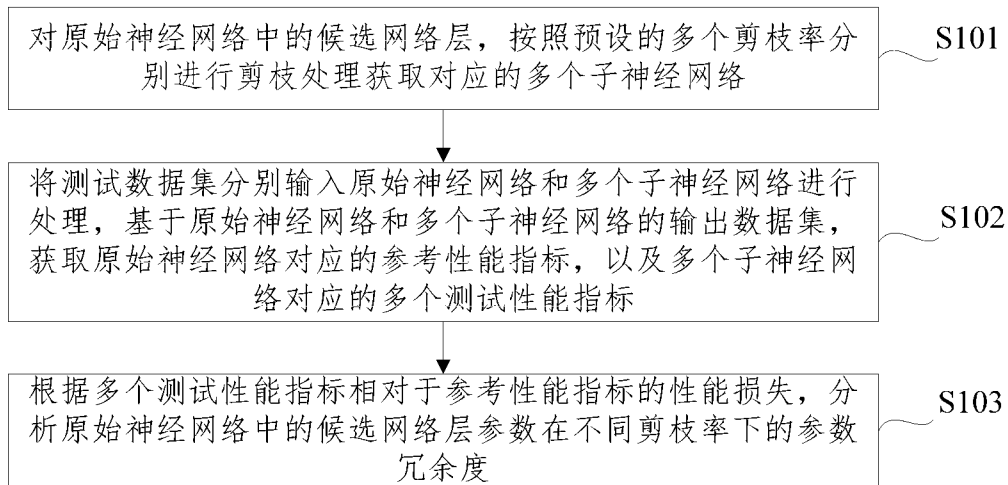


图 1



图 2

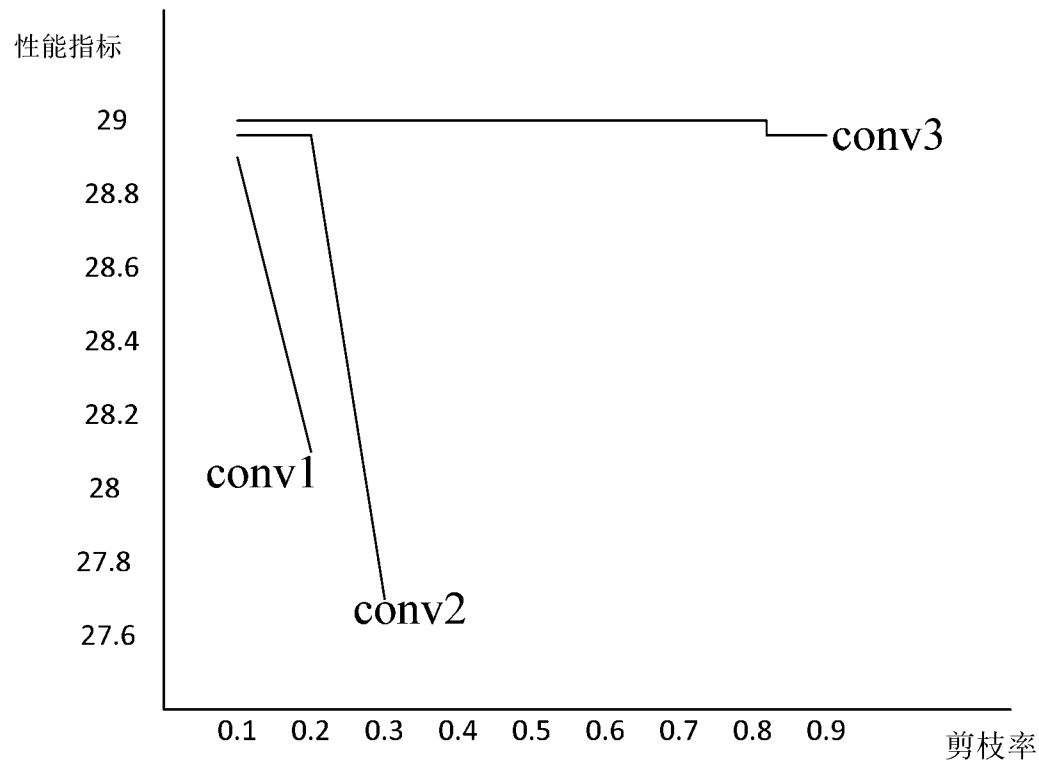


图 3

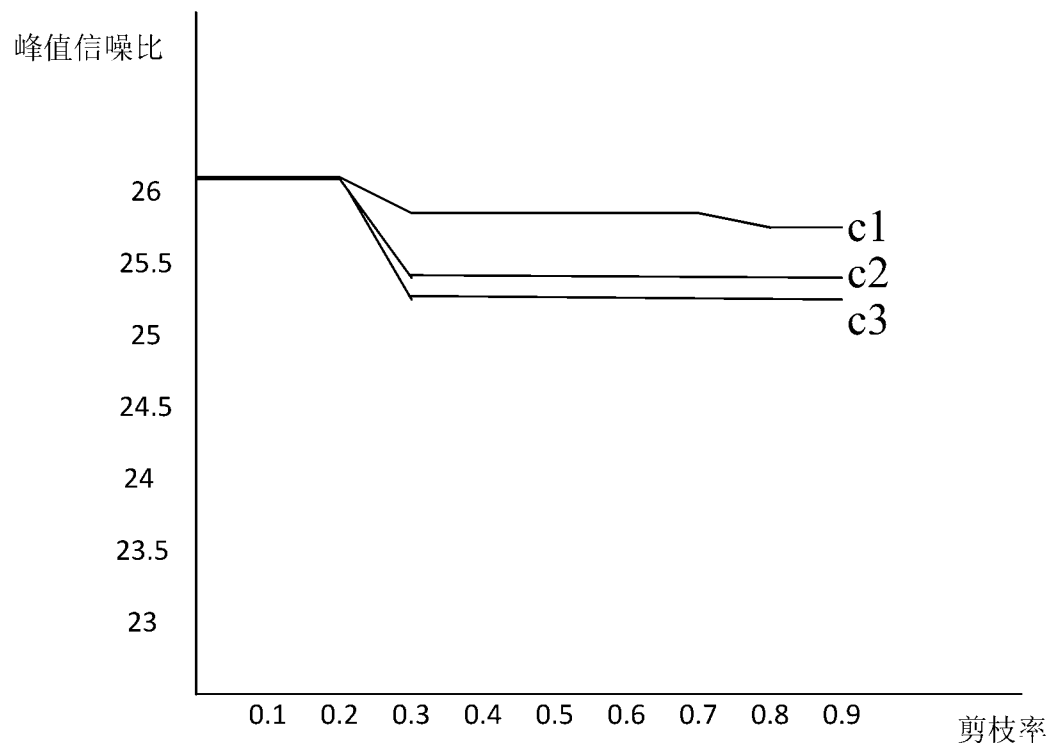


图 4



图 5

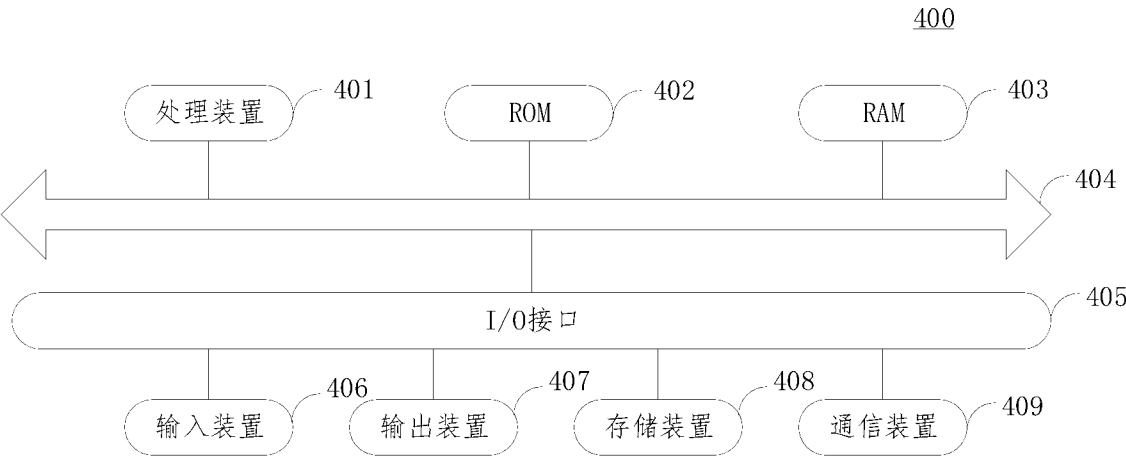


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2023/093805

A. CLASSIFICATION OF SUBJECT MATTER

G06N3/08(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

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)

VEN, CNABS, CNTXT, CNKI, IEEE: 剪枝, 神经网络, 指标, 损失, 冗余度, 压缩, 范数, 峰值信噪比, pruning, neural network, indicator, loss, redundancy, compression, norm, peak signal-to-noise ratio, PSNR

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 114282670 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 05 April 2022 (2022-04-05) description, paragraphs [0023]-[0070]	1-13
A	CN 111931930 A (GUANGDONG OPPO MOBILE COMMUNICATIONS CO., LTD.) 13 November 2020 (2020-11-13) entire document	1-13
A	CN 113255910 A (ZHEJIANG UNIVIEW TECHNOLOGIES CO., LTD.) 13 August 2021 (2021-08-13) entire document	1-13
A	US 2021264278 A1 (ADOBE INC.) 26 August 2021 (2021-08-26) entire document	1-13
A	WO 2022057262 A1 (INSPUR SUZHOU INTELLIGENT TECHNOLOGY CO., LTD.) 24 March 2022 (2022-03-24) 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

“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

14 June 2023

Date of mailing of the international search report

23 June 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/CN2023/093805

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	114282670	A	05 April 2022	None	
CN	111931930	A	13 November 2020	None	
CN	113255910	A	13 August 2021	None	
US	2021264278	A1	26 August 2021	None	
WO	2022057262	A1	24 March 2022	None	

<p>A. 主题的分类</p> <p>G06N3/08 (2023.01) i</p> <p>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</p>																						
<p>B. 检索领域</p> <p>检索的最低限度文献 (标明分类系统和分类号)</p> <p>IPC: G06N</p> <p>包含在检索领域中的除最低限度文献以外的检索文献</p> <p>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))</p> <p>VEN, CNABS, CNTXT, CNKI, IEEE: 剪枝, 神经网络, 指标, 损失, 冗余度, 压缩, 范数, 峰值信噪比, pruning, neural network, indicator, loss, redundancy, compression, norm, peak signal-to-noise ratio, PSNR</p>																						
<p>C. 相关文件</p> <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>X</td> <td>CN 114282670 A (北京百度网讯科技有限公司) 2022年4月5日 (2022 - 04 - 05) 说明书第[0023]-[0070]段</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>CN 111931930 A (OPPO广东移动通信有限公司) 2020年11月13日 (2020 - 11 - 13) 全文</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>CN 113255910 A (浙江宇视科技有限公司) 2021年8月13日 (2021 - 08 - 13) 全文</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>US 2021264278 A1 (AD0BE INC.) 2021年8月26日 (2021 - 08 - 26) 全文</td> <td>1-13</td> </tr> <tr> <td>A</td> <td>WO 2022057262 A1 (INSUR SUZHOU INTELLIGENT TECHNOLOGY CO., LTD.) 2022年3月24日 (2022 - 03 - 24) 全文</td> <td>1-13</td> </tr> </tbody> </table> <p>☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。</p> <table> <tr> <td> <p>* 引用文件的具体类型:</p> <p>“A” 认为不特别相关的表示了现有技术一般状态的文件</p> <p>“D” 申请人在国际申请中引证的文件</p> <p>“E” 在国际申请日的当天或之后公布的在先申请或专利</p> <p>“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)</p> <p>“O” 涉及口头公开、使用、展览或其他方式公开的文件</p> <p>“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</p> </td> <td> <p>“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</p> <p>“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</p> <p>“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</p> <p>“&” 同族专利的文件</p> </td> </tr> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 114282670 A (北京百度网讯科技有限公司) 2022年4月5日 (2022 - 04 - 05) 说明书第[0023]-[0070]段	1-13	A	CN 111931930 A (OPPO广东移动通信有限公司) 2020年11月13日 (2020 - 11 - 13) 全文	1-13	A	CN 113255910 A (浙江宇视科技有限公司) 2021年8月13日 (2021 - 08 - 13) 全文	1-13	A	US 2021264278 A1 (AD0BE INC.) 2021年8月26日 (2021 - 08 - 26) 全文	1-13	A	WO 2022057262 A1 (INSUR SUZHOU INTELLIGENT TECHNOLOGY CO., LTD.) 2022年3月24日 (2022 - 03 - 24) 全文	1-13	<p>* 引用文件的具体类型:</p> <p>“A” 认为不特别相关的表示了现有技术一般状态的文件</p> <p>“D” 申请人在国际申请中引证的文件</p> <p>“E” 在国际申请日的当天或之后公布的在先申请或专利</p> <p>“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)</p> <p>“O” 涉及口头公开、使用、展览或其他方式公开的文件</p> <p>“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</p>	<p>“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</p> <p>“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</p> <p>“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</p> <p>“&” 同族专利的文件</p>
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																				
X	CN 114282670 A (北京百度网讯科技有限公司) 2022年4月5日 (2022 - 04 - 05) 说明书第[0023]-[0070]段	1-13																				
A	CN 111931930 A (OPPO广东移动通信有限公司) 2020年11月13日 (2020 - 11 - 13) 全文	1-13																				
A	CN 113255910 A (浙江宇视科技有限公司) 2021年8月13日 (2021 - 08 - 13) 全文	1-13																				
A	US 2021264278 A1 (AD0BE INC.) 2021年8月26日 (2021 - 08 - 26) 全文	1-13																				
A	WO 2022057262 A1 (INSUR SUZHOU INTELLIGENT TECHNOLOGY CO., LTD.) 2022年3月24日 (2022 - 03 - 24) 全文	1-13																				
<p>* 引用文件的具体类型:</p> <p>“A” 认为不特别相关的表示了现有技术一般状态的文件</p> <p>“D” 申请人在国际申请中引证的文件</p> <p>“E” 在国际申请日的当天或之后公布的在先申请或专利</p> <p>“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)</p> <p>“O” 涉及口头公开、使用、展览或其他方式公开的文件</p> <p>“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</p>	<p>“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</p> <p>“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</p> <p>“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</p> <p>“&” 同族专利的文件</p>																					
<p>国际检索实际完成的日期</p> <p>2023年6月14日</p>		<p>国际检索报告邮寄日期</p> <p>2023年6月23日</p>																				
<p>ISA/CN的名称和邮寄地址</p> <p>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</p>		<p>授权官员</p> <p>吴雪</p> <p>电话号码 (+86) 010-53961406</p>																				

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2023/093805

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	114282670	A	2022年4月5日	无	
CN	111931930	A	2020年11月13日	无	
CN	113255910	A	2021年8月13日	无	
US	2021264278	A1	2021年8月26日	无	
WO	2022057262	A1	2022年3月24日	无	