

(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
2021 年 7 月 8 日 (08.07.2021)



(10) 国际公布号
WO 2021/135546 A1

(51) 国际专利分类号:
G06N 3/04 (2006.01)

(21) 国际申请号: PCT/CN2020/123196

(22) 国际申请日: 2020 年 10 月 23 日 (23.10.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010690459.4 2020年7月17日 (17.07.2020) CN

(71) 申请人: 平安科技 (深圳) 有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN];
中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼时梦如, Guangdong 518000 (CN)。

(72) 发明人: 陈筱(CHEN, Xiao); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼时梦如, Guangdong 518000 (CN)。 周细文(ZHOU, Xiwen); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼时梦如, Guangdong 518000 (CN)。 庄伯金(ZHUANG, Bojin); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼时梦如, Guangdong 518000 (CN)。 王少军(WANG, Shaojun); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼时梦如, Guangdong 518000 (CN)。

(74) 代理人: 深圳国新南方知识产权代理有限公司 (SHENZHEN CHINA INNOVATION SOUTH INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国广东省深圳市南山区

(54) Title: DEEP NEURAL NETWORK INTERPRETATION METHOD AND DEVICE, TERMINAL, AND STORAGE MEDIUM

(54) 发明名称: 深度神经网络的解释方法、装置、终端及存储介质

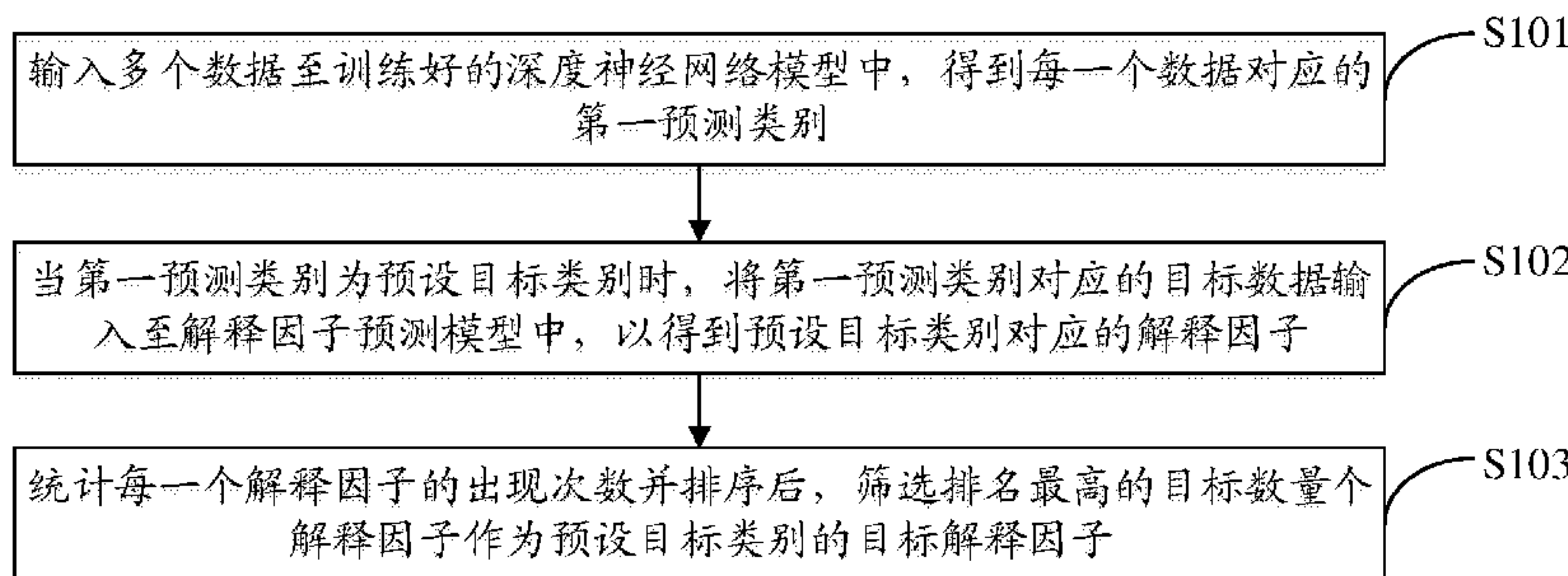


图 1

- S101 Input multiple pieces of data into a trained deep neural network model to obtain a first prediction category corresponding to each piece of data
- S102 When the first prediction category is a preset target category, input target data corresponding to the first prediction category into an interpretation factor prediction model to obtain an interpretation factor corresponding to the preset target category
- S103 After counting and sorting the number of occurrences of each interpretation factor, select a target number of highest-ranking interpretation factors as target interpretation factors of the preset target category

(57) Abstract: A deep neural network interpretation method and device, a terminal, and a storage medium, relating to the technical field of artificial intelligence. Specifically disclosed are: determining all pieces of target data among multiple input pieces of data according to a result output by a deep neural network, inputting the target data into an interpretation factor prediction model, separately to obtain a group of interpretation factors corresponding to each piece of target data, and then counting and sorting the number of occurrences of each interpretation factor in multiple groups of interpretation factors, and selecting a target number of highest-ranking interpretation factors as target interpretation factors of a preset target category. According to the method, an interpretation factor of

深圳湾科技生态园 11 栋 A 座 501C 姜宇,
Guangdong 518000 (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, IT, 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, 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, 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))。
- 在修改权利要求的期限届满之前进行, 在收到该修改后将重新公布 (细则48.2 (h))。
- 根据申请人的请求, 在条约第21条 (2) (a) 所规定的期限届满之前进行。

a prediction result is obtained by means of a pre-trained interpretation factor prediction model, so as to interpret a dominant factor affecting the prediction result of the deep neural network. The method can be applied to a scene of intelligent government affairs/smart city management/intelligent community/intelligent security/intelligent logistics/intelligent healthcare/intelligent education/intelligent environmental protection/intelligent transport, so as to promote the construction of smart city.

(57) 摘要: 一种深度神经网络的解释方法、装置、终端及存储介质, 涉及人工智能技术领域, 具体公开了通过深度神经网络输出的结果从输入的多个数据中确定所有的目标数据, 再将目标数据分别输入至解释因子预测模型中, 得到每个目标数据对应的一组解释因子, 再统计多组解释因子中每个解释因子的出现次数并排序后, 筛选排名最高的目标数量个解释因子作为预设目标类别的目标解释因子。该方法通过预先训练好的解释因子预测模型得到预测结果的解释因子, 以解释影响深度神经网络预测结果主要因素, 其可应用于智慧政务/智慧城管/智慧社区/智慧安防/智慧物流/智慧医疗/智慧教育/智慧环保/智慧交通场景中, 从而推动智慧城市的建设。

深度神经网络的解释方法、装置、终端及存储介质

本申请要求于2020年07月17日提交中国专利局、申请号为202010690459.4，发明名称为“深度神经网络的解释方法、装置、终端及存储介质”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

本申请涉及人工智能的深度学习技术领域，尤其涉及一种深度神经网络的解释方法、装置、终端及存储介质。

背景技术

神经网络算法，是20世纪80年代以来人工智能领域兴起的研究热点。它从信息处理角度对神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。其具有自主学习功能，可通过训练慢慢学会识别和预测；联想存储功能，具有很高算法健壮性；并行度高，具有高速寻找优化解的能力，可针对大数据复杂问题快速找到优化解；可塑性强，可充分逼近任意复杂的非线性关系；具有很强的信息综合能力，能同时处理定量和定性的信息，适用于多信息融合和多媒体技术。近年来，神经网络的研究工作不断深入，已经取得了很大的进展，其在图像处理、模式识别、智能机器人、自动控制、预测估计、生物、医学、经济等领域已成功地解决了许多传统计算机方法难以解决的实际问题，表现出了良好的智能特性，推动着信息处理技术和人工智能不断地发展。

深度神经网络模型的可解释性一直是业内的重点关注领域，模型的可解释性对于模型改进和模型可信性与透明度都有重要作用。特征可视化是目前常用的神经网络的解释方法，其寻找模型输入中各个区域对模型预测的影响并进行可视化表示，可以在一定程度上说明模型的关注点，但是在实际应用中，发明人发现仅通过特征可视化很难对模型进行定量的解释，也不能反映影响模型判断的主要因素，无法很好的解释模型是根据哪些因子得到的预测结果。

发明内容

本申请提供一种深度神经网络的解释方法、装置、终端及存储介质，以解决难以确定影响深度神经网络预测结果的因素的问题。

为实现上述目的，本申请提供一种深度神经网络的解释方法，包括：输入多个数据至训练好的深度神经网络模型中，得到每一个数据对应的第一预测类别；当第一预测类别为预设目标类别时，将第一预测类别对应的目标数据输入至解释因子预测模型中，以得到预设目标类别对应的解释因子，解释因子预测模型根据预先设定的目标数量和预设目标类别的第一预测值训练得到；统计每一个解释因子的出现次数并排序后，筛选排名最高的目标数量个解释因子作为预设目标类别的目标解释因子。

为实现上述目的，本申请还提供了一种深度神经网络的解释装置，包括：第一输入模块，用于输入多个数据至训练好的深度神经网络模型中，得到每一个数据对应的第一预测类别；第二输入模块，用于当第一预测类别为预设目标类别时，将第一预测类别对应的目标数据输入至解释因子预测模型中，以得到预设目标类别对应的解释因子，解释因子预测模型根据预先设定的目标数量和预设目标类别的第一预测值训练得到；筛选模块，用于统计每一个解释因子的出现次数并排序后，筛选排名最高的目标数量个解释因子作为预设目标类别的目标解释因子。

为实现上述目的，本申请还提供了一种终端，包括存储器、处理器以及存储在存储器上并可在处理器上运行的程序文件，其中，处理器执行程序文件时实现以下步骤：输入多

个数据至训练好的深度神经网络模型中，得到每一个数据对应的第一预测类别；当第一预测类别为预设目标类别时，将第一预测类别对应的目标数据输入至解释因子预测模型中，以得到预设目标类别对应的解释因子，解释因子预测模型根据预先设定的目标数量和预设目标类别的第一预测值训练得到；统计每一个解释因子的出现次数并排序后，筛选排名最高的目标数量个解释因子作为预设目标类别的目标解释因子。

为实现上述目的，本申请还提供了一种存储介质，其中，存储有能够实现深度神经网络的解释方法的程序文件，程序文件被处理器执行时实现以下步骤：输入多个数据至训练好的深度神经网络模型中，得到每一个数据对应的第一预测类别；当第一预测类别为预设目标类别时，将第一预测类别对应的目标数据输入至解释因子预测模型中，以得到预设目标类别对应的解释因子，解释因子预测模型根据预先设定的目标数量和预设目标类别的第一预测值训练得到；统计每一个解释因子的出现次数并排序后，筛选排名最高的目标数量个解释因子作为预设目标类别的目标解释因子。

本申请提出的深度神经网络的解释方法，通过当深度神经网络输出的第一预测类别为预设目标类别时，将第一预设类别对应的目标数据输入至训练好的解释因子预测模型中，得到预设目标类别对应的解释因子，通过输入不同的目标数据，从而得到多组解释因子，再统计每个解释因子的出现次数，从中筛选出排名最高的目标数量个解释因子作为该预设目标类别的目标解释因子，该目标解释因子将作为深度神经网络将目标数据的输出结果划分为预设目标类别的主要影响因素，方便用户找出影响深度神经网络模型进行分类预测的影响因子。此外，本申请的深度神经网络的解释方法还可应用于智慧政务/智慧城管/智慧社区/智慧安防/智慧物流/智慧医疗/智慧教育/智慧环保/智慧交通场景中，从而推动智慧城市的建设。

附图说明

图1为本申请深度神经网络的解释方法第一实施例的流程示意图；

图2为本申请深度神经网络的解释方法实施例的解释因子预测模型训练的流程示意图；

图3为深度神经网络的结构示意图；

图4为本申请深度神经网络的解释方法第二实施例的流程示意图；

图5为本申请深度神经网络的解释装置实施例的功能模块示意图；

图6为本申请终端实施例的结构示意图；

图7为本申请存储介质实施例的结构示意图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

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，所描述的实施例仅是本申请的一部分实施例，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。

本申请中的术语“第一”、“第二”、“第三”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”、“第三”的特征可以明示或者隐含地包括至少一个该特征。本申请的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。本申请实施例中所有方向性指示（诸如上、下、左、右、前、后……）仅用于解释在某一特定姿态（如附图所示）下各部件之间的相对位置关系、运动情况等，如果该特定姿态发生改变时，则该方向性指示也相应地随之改变。此外，术语“包括”和“具有”以及它们任何变形，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、产品或设备没有限定于已列出的步骤

或单元,而是可选地还包括没有列出的步骤或单元,或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。

在本文中提及“实施例”意味着,结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例,也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是,本文所描述的实施例可以与其它实施例相结合。

图1是本申请第一实施例的深度神经网络的解释方法的流程示意图。需注意的是,若有实质上相同的结果,本申请的方法并不以图1所示的流程顺序为限。如图1所示,该方法包括步骤:

10 步骤 S101: 输入多个数据至训练好的深度神经网络模型中,得到每一个数据对应的第一预测类别。

需要说明的是,深度神经网络是一种广泛应用于图像处理、模式识别、智能机器人、自动控制、预测估计、生物、医学、经济等领域神经网络算法,其通常包括输入层、隐藏层和输出层,请一并参阅图2,一般第一层(input layer)是输入层,最后一层(output layer)是输出层,而中间的层数(hidden layer)都是隐藏层,层与层之间是全连接的,即第*i*层的任意一个神经元一定与第*i*+1层的任意一个神经元相连,输出层的神经元可以为一个或多个,每个神经元对应一个输出结果。在使用深度神经网络对处理分类任务时,将数据输入至深度神经网络模型,得到每个类别的预测值,再将预测值最大的类别作为深度神经网络的输出结果,例如,以动物分类为例,使用训练好的深度神经网络识别一张图片中的动物时,其可得到该动物为猫、狗、牛、羊等动物时的预测值,然后从中选取预测值最大的动物作为输出结果。

步骤 S102: 当第一预测类别为预设目标类别时,将第一预测类别对应的目标数据输入至解释因子预测模型中,以得到预设目标类别对应的解释因子。

25 需要说明的是,解释因子预测模型根据预先设定的目标数量和预设目标类别的第一预测值训练得到。其中,目标数量是指目标解释因子的个数,可由用户预先设定;该预设目标类别为用户预先设定的类别;解释因子是指影响深度神经网络模型判断的因素,例如,通过深度神经网络模型识别一张图片中的动物是什么动物,其影响深度神经网络模型判断的因素可以包括动物的脚的数量、毛发生长情况、眼睛大小等,这些因素均可能对深度神经网络模型的输出造成影响。

30 需要理解的是,在使用深度神经网络模型对物品进行分类时,每一个类别均受到不同解释因子的影响,例如:以动物分类为例,通过深度神经网络模型识别动物时,“猫”这一预测类别的解释因子可以包括全身毛发、四条腿行走、眼睛大、长有胡须、隐藏爪子、脚掌有肉垫等,“狗”这一预测类别的解释因子可以包括全身毛发、四条腿行走、喜欢蹲坐、显示爪子、脚掌无肉垫等,本申请旨在使用深度神经网络时确定影响输出结果的主要因素。

35 在步骤 S102 中,每当第一预测类别为预设目标类别时,将该第一预测类别对应的目标数据输入至解释因子预测模型中,以得到预设目标类别对应的解释因子。通过输入多个目标数据至解释因子预测模型,即可得到多组预设目标类别对应的解释因子,而该多组预设目标类别对应的解释因子可能相同或不同,例如:以动物分类为例,A、B两个目标数据的预测结果均为猫,但是A目标数据是通过大眼睛、四条腿、全身毛发进行预测的,而B目标数据是通过四条腿、全身毛发、毛发上带有斑纹进行预测的,两者的解释因子只有四条腿、全身毛发相同。

进一步的,步骤 S102 具体包括:将预设目标类别对应的目标数据输入至与预设目标类别对应的解释因子预测模型中,得到每个解释因子的第二预测值;将第二预测值超过预设阈值的解释因子的值取1,并作为预设目标类别对应的解释因子。

45 具体地,将目标数据输入至解释因子预测模型,从而得到每个解释因子的第二预测值,

再根据第二预测值的大小，将第二预测值超过预设阈值的解释因子的值设置为 1，其余解释因子的值设置为 0。

进一步的，需要理解的是，不同的预设目标类通过利用不同的样本训练得到，其对应的解释因子预测模型各不相同，因此，步骤 S102 中，在将第一预测类别对应的目标数据输入至解释因子预测模型之前，还包括：获取与预设目标类别对应的解释因子预测模型。

步骤 S103：统计每一个解释因子的出现次数并排序后，筛选排名最高的目标数量个解释因子作为预设目标类别的目标解释因子。

在步骤 S103 中，通过统计多个目标数据对应的解释因子中，每个解释因子的出现次数，并进行排序，从中筛选出排名最高的目标数量个解释因子作为预设目标类别的目标解释因子。例如，以动物分类为例，解释因子的总数量为 100 个，通过上述解释因子预测模型，从 100 个解释因子中筛选将动物划分为猫的 10 个相关性最高的解释因子作为目标解释因子。

本实施例通过当深度神经网络输出的第一预测类别为预设目标类别时，将第一预设类别对应的目标数据输入至训练好的解释因子预测模型中，得到预设目标类别对应的解释因子，通过输入不同的目标数据，从而得到多组解释因子，再统计每个解释因子的出现次数，从中筛选出排名最高的目标数量个解释因子作为该预设目标类别的目标解释因子，该目标解释因子将作为深度神经网络将目标数据的输出结果划分为预设目标类别的主要影响因素，方便用户找出影响深度神经网络模型进行分类预测的影响因子。此外，本申请的深度神经网络的解释方法还可应用于智慧城管/智慧社区/智慧安防/智慧物流等场景中，从而推动智慧城市的建设。

进一步的，上述实施例中，每个预设目标类别对应的解释因子预测模型预先训练得到，如图 3 所示，解释因子预测模型的训练步骤具体包括：

步骤 S201：逐个输入样本至训练好的深度神经网络模型中，得到每个样本对应的预测类别以及第一预测值。

需要说明的是，深度神经网络模型的输出通常包括多个预测类别，每个预测类别对应一个预测值，其中预测值最高的结果作为最终输出结果。

进一步的，在步骤 S201 之前，还包括：获取用户预先输入的至少一个预设目标类别，以及每个预设目标类别对应的目标解释因子的目标数量。

本实施例中，用户可根据需求选择需要训练类别，并且，用户还可设定最终输出的解释因子的目标数量。例如：用户可以设定动物分类中猫这一类别的目标解释因子的目标数量为 10 个。

步骤 S202：当预测类别为预设目标类别时，将预测类别对应的目标样本输入至解释因子预测模型中，得到每个解释因子的第二预测值。

步骤 S203：根据第一预测值、第二预测值和目标数量构建解释因子标签。

本实施例中，该构建解释因子标签的步骤具体包括：

1、根据第一预测值和目标数量计算所有的解释因子中取值为 1 的解释因子的数目 w 。计算公式为：

$$w = \text{ceil} (m \times S_k) ;$$

其中， ceil 表示向上取整， m 表示目标数量， S_k 表示第一预测值。

2、根据每个解释因子的第二预测值的大小，将第二预测值最大的 w 个解释因子的值设置为 1，将其余的解释因子的值设置为 0，得到解释因子标签。

具体地，该解释因子标签具体如下：

$$A_{q_1} = A_{q_2} = \dots = A_{q_w} = 1;$$

$$A_i = 0, i \leq m \text{ 且 } i \notin \{q_1, q_2, \dots, q_w\};$$

其中， $q_1, q_2, \dots, q_w$ 是指第二预测值最大的 w 个解释因子的角标。

例如：以一个例子进行说明， m 为 10， S_k 为 80%，根据上述公式计算得到 w 为 8，假设所有的解释因子共有 30 个，将目标样本输入至解释因子预测模型中，即可得到该 30 个解释因子的第二预测值，再从中挑选出第二预测值最高的 8 个解释因子，将其值设置为 1，其余的 22 个解释因子的值设置为 0，从而得到该预测类别对应的解释因子标签。

5 需要理解的是，每一个预测类别的解释因子标签均是基于预测类别当前的第一预测值构建，因此，每次输入目标样本得到预测类别时，均需构建一次解释因子标签。

步骤 S204：基于解释因子标签和第二预测值反向传播更新解释因子预测模型。

在步骤 S204 中，通过比对解释因子标签与每个解释因子的第二预测值反向传播更新解释因子预测模型的方式与深度神经网络训练时的反向传播更新过程相同。

10 步骤 S205：循环执行上述步骤直至解释因子预测模型收敛，并将收敛的解释因子预测模型与预设目标类别绑定。

具体地，通过循环输入样本至训练好的深度神经网络模型，从而对解释因子预测模型进行反复迭代训练，直至解释因子预测模型收敛时，停止训练并将收敛的解释因子预测模型，并将收敛的解释因子预测模型与预设目标类别绑定。

15 本实施例中，通过结合预测类别的第一预测值动态得到解释因子的标签，并利用该标签对解释因子预测模型进行训练，从而使得预设目标类别对应的解释因子将对于预设目标类别的样本取较大值 1，对于其他类别样本取较小值 0。

20 图 4 展示了本申请第二实施例的深度神经网络的解释方法的流程示意图。需注意的是，若有实质上相同的结果，本申请的方法并不以图 4 所示的流程顺序为限。如图 4 所示，该方法包括步骤：

步骤 S301：获取用户预先输入的至少一个预设目标类别，每个预设目标类别均对应预先训练至收敛状态的解释因子预测模型。

25 在步骤 S301 中，预设目标类别可由用户预先指定，用户可以一次性指定一个或多个预设目标类别，而每个预设目标类别在分别利用解释因子预测模型确认目标解释因子时，分别通过每个预设目标类别对应的解释因子预测模型得到，相互之间互不干扰。

步骤 S302：输入多个数据至训练好的深度神经网络模型中，得到每一个数据对应的第一预测类别。

在本实施例中，图 4 中的步骤 S302 和图 1 中的步骤 S101 类似，为简约起见，在此不再赘述。

30 步骤 S303：当第一预测类别为预设目标类别时，将第一预测类别对应的目标数据输入至解释因子预测模型中，以得到预设目标类别对应的解释因子。

在本实施例中，图 4 中的步骤 S303 和图 1 中的步骤 S102 类似，为简约起见，在此不再赘述。

35 步骤 S304：统计每一个解释因子的出现次数并排序后，筛选排名最高的目标数量个解释因子作为预设目标类别的目标解释因子。

在本实施例中，图 4 中的步骤 S304 和图 1 中的步骤 S103 类似，为简约起见，在此不再赘述。

40 本实施例中，预设目标类别可以由用户自行指定，通过获取用户预先输入预设目标类别即可为用户确定该深度神经网络模型输出该预设目标类别的主要影响因素，并且，用户可以一次性输入多个预设目标类别，帮助用户快速确认每个预设目标类别的主要影响因素。

图 5 展示了本申请实施例深度神经网络的解释装置的功能模块示意图。如图 5 所示，该深度神经网络的解释装置 10 包括第一输入模块 11、第二输入模块 12 和筛选模块 13。

第一输入模块 11，用于输入多个数据至训练好的深度神经网络模型中，得到每一个数据对应的第一预测类别；

45 第二输入模块 12，用于当第一预测类别为预设目标类别时，将第一预测类别对应的

目标数据输入至解释因子预测模型中，以得到预设目标类别对应的解释因子，解释因子预测模型根据预先设定的目标数量和预设目标类别的第一预测值训练得到；

筛选模块 13，用于统计每一个解释因子的出现次数并排序后，筛选排名最高的目标数量个解释因子作为预设目标类别的目标解释因子。

5 可选地，第二输入模块 12 将第一预测类别对应的目标数据输入至解释因子预测模型之前，还包括预先训练解释因子预测模型，包括：逐个输入样本至训练好的深度神经网络模型中，得到每个样本对应的预测类别以及第一预测值；当预测类别为预设目标类别时，将预测类别对应的目标样本输入至解释因子预测模型中，得到每个解释因子的第二预测值；根据第一预测值、第二预测值和目标数量构建解释因子标签；基于解释因子标签和第二预
10 测值反向传播更新解释因子预测模型；循环执行上述步骤直至解释因子预测模型收敛，并将收敛的解释因子预测模型与预设目标类别绑定。

可选地，利用第一预测值、每个解释因子的第二预测值和目标数量构建解释因子标签操作还可以为：根据第一预测值和目标数量计算所有的解释因子中取值为 1 的解释因子的数目 w ， $w = \text{ceil}(m \times S_k)$ ，其中， ceil 表示向上取整， m 表示目标数量， S_k 表示第一预
15 测值；根据每个解释因子的第二预测值的大小，将第二预测值最大的 w 个解释因子的值设置为 1，将其余的解释因子的值设置为 0，得到解释因子标签。

可选地，逐个输入样本至训练好的深度神经网络模型的操作之前还包括：获取用户预先输入的至少一个预设目标类别，以及每个预设目标类别对应的目标解释因子的目标数量。

20 可选地，第一输入模块 11 输入多个数据至训练好的深度神经网络模型中之前，还用于获取用户预先输入的至少一个预设目标类别，每个预设目标类别均对应预先训练至收敛状态的解释因子预测模型。

可选地，第二输入模块 12 将第一预测类别对应的目标数据输入至解释因子预测模型中，以得到预设目标类别对应的解释因子的操作还可以为：将预设目标类别对应的目标数据输入至与预设目标类别对应的解释因子预测模型中，得到每个解释因子的第二预测值；
25 将第二预测值超过预设阈值的解释因子的值取 1，并作为预设目标类别对应的解释因子。

可选地，第一输入模块 11 将第一预测类别对应的目标数据输入至解释因子预测模型的操作之前，还包括：获取与预设目标类别对应的解释因子预测模型。

请参阅图 6，图 6 为本申请实施例的终端的结构示意图。如图 6 所示，该终端 20 包括处理器 200 及和处理器 200 耦接的存储器 201。

30 存储器 201 存储有用于实现上述任一实施例所述的深度神经网络的解释方法的程序指令。

处理器 200 用于执行存储器 201 存储的程序指令以训练得到预设目标类别的目标解释因子。

其中，处理器 200 还可以称为 CPU (Central Processing Unit, 中央处理单元)。处理器
35 200 可能是一种集成电路芯片，具有信号的处理能力。处理器 200 还可以是通用处理器、数字信号处理器 (DSP)、专用集成电路 (ASIC)、现场可编程门阵列 (FPGA) 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

参阅图 7，图 7 为本申请实施例的存储介质的结构示意图。本申请实施例的存储介质
40 存储有能够实现上述所有方法的程序文件 30，其中，该程序文件 30 可以以软件产品的形式存储在上述存储介质中，包括若干指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）或处理器（processor）执行本申请各个实施方式所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器 (ROM, Read-Only Memory)、随机存取存储器 (RAM, Random Access Memory)、磁碟或者光盘等各种可以
45 存储程序代码的介质，或者是计算机、服务器、手机、平板等终端设备。所述计算机可读

存储介质可以是非易失性，也可以是易失性

在本申请所提供的几个实施例中，应该理解到，所揭露的终端，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，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，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

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。以上仅为本申请的实施方式，并非因此限制本申请的专利范围，凡是利用本申请说明书及附图内容所作的等效结构或等效流程变换，或直接或间接运用在其他相关的技术领域，均同理包括在本申请的专利保护范围内。

权 利 要 求 书

1、一种深度神经网络的解释方法，其中，包括：

输入多个数据至训练好的深度神经网络模型中，得到每一个数据对应的第一预测类别；

5 当所述第一预测类别为预设目标类别时，将所述第一预测类别对应的目标数据输入至解释因子预测模型中，以得到所述预设目标类别对应的解释因子，所述解释因子预测模型根据预先设定的目标数量和所述预设目标类别的第一预测值训练得到；

统计每一个解释因子的出现次数并排序后，筛选排名最高的所述目标数量个解释因子作为所述预设目标类别的目标解释因子。

10 2、根据权利要求1所述的方法，其中，还包括预先训练所述解释因子预测模型，包括：

逐个输入样本至所述训练好的深度神经网络模型中，得到每个样本对应的预测类别以及第一预测值；

15 当所述预测类别为预设目标类别时，将所述预测类别对应的目标样本输入至所述解释因子预测模型中，得到每个解释因子的第二预测值；

根据所述第一预测值、所述第二预测值和所述目标数量构建解释因子标签；

基于所述解释因子标签和所述第二预测值反向传播更新所述解释因子预测模型；

循环执行上述步骤直至所述解释因子预测模型收敛，并将收敛的解释因子预测模型与所述预设目标类别绑定。

20 3、根据权利要求2所述的方法，其中，所述利用所述第一预测值、所述每个解释因子的第二预测值和所述目标数量构建解释因子标签，包括：

根据所述第一预测值和所述目标数量计算所有的解释因子中取值为1的解释因子的数目 w ， $w = \text{ceil}(m \times S_k)$ ，其中， ceil 表示向上取整， m 表示所述目标数量， S_k 表示所述第一预测值；

25 根据所述每个解释因子的第二预测值的大小，将所述第二预测值最大的所述 w 个解释因子的值设置为1，将其余的解释因子的值设置为0，得到所述解释因子标签。

4、根据权利要求2所述的方法，其中，所述逐个输入样本至所述训练好的深度神经网络模型中之前，还包括：

30 获取用户预先输入的至少一个预设目标类别，以及每个预设目标类别对应的目标解释因子的目标数量。

5、根据权利要求1所述的方法，其中，所述输入多个数据至训练好的深度神经网络模型中之前，还包括：

获取用户预先输入的至少一个预设目标类别，每个所述预设目标类别均对应预先训练至收敛状态的解释因子预测模型。

35 6、根据权利要求1所述的方法，其中，所述将所述第一预测类别对应的目标数据输入至解释因子预测模型中，以得到所述预设目标类别对应的解释因子，包括：

将所述预设目标类别对应的目标数据输入至与所述预设目标类别对应的解释因子预测模型中，得到所述每个解释因子的第二预测值；

40 将所述第二预测值超过预设阈值的解释因子的值取1，并作为所述预设目标类别对应的解释因子。

7、根据权利要求1所述的方法，其中，所述将所述第一预测类别对应的目标数据输入至解释因子预测模型中之前，还包括：

获取与所述预设目标类别对应的解释因子预测模型。

8、一种深度神经网络的解释装置，其中，包括：

45 第一输入模块，用于输入多个数据至训练好的深度神经网络模型中，得到每一个数据

对应的第一预测类别;

第二输入模块,用于当所述第一预测类别为预设目标类别时,将所述第一预测类别对应的目标数据输入至解释因子预测模型中,以得到所述预设目标类别对应的解释因子,所述解释因子预测模型根据预先设定的目标数量和所述预设目标类别的第一预测值训练得到;

筛选模块,用于统计每一个解释因子的出现次数并排序后,筛选排名最高的所述目标数量个解释因子作为所述预设目标类别的目标解释因子。

9、一种终端,包括存储器、处理器以及存储在存储器上并可在处理器上运行的程序文件,其中,所述处理器执行所述程序文件时实现以下步骤:

输入多个数据至训练好的深度神经网络模型中,得到每一个数据对应的第一预测类别;当所述第一预测类别为预设目标类别时,将所述第一预测类别对应的目标数据输入至解释因子预测模型中,以得到所述预设目标类别对应的解释因子,所述解释因子预测模型根据预先设定的目标数量和所述预设目标类别的第一预测值训练得到;

统计每一个解释因子的出现次数并排序后,筛选排名最高的所述目标数量个解释因子作为所述预设目标类别的目标解释因子。

10、根据权利要求9所述的终端,其中,所述处理器执行所述程序文件时还用于预先训练所述解释因子预测模型,包括:

逐个输入样本至所述训练好的深度神经网络模型中,得到每个样本对应的预测类别以及第一预测值;

当所述预测类别为预设目标类别时,将所述预测类别对应的目标样本输入至所述解释因子预测模型中,得到每个解释因子的第二预测值;

根据所述第一预测值、所述第二预测值和所述目标数量构建解释因子标签;

基于所述解释因子标签和所述第二预测值反向传播更新所述解释因子预测模型;

循环执行上述步骤直至所述解释因子预测模型收敛,并将收敛的解释因子预测模型与所述预设目标类别绑定。

11、根据权利要求10所述的终端,其中,所述利用所述第一预测值、所述每个解释因子的第二预测值和所述目标数量构建解释因子标签,包括:

根据所述第一预测值和所述目标数量计算所有的解释因子中取值为1的解释因子的数目 w , $w = \text{ceil}(m \times S_k)$,其中, ceil 表示向上取整, m 表示所述目标数量, S_k 表示所述第一预测值;

根据所述每个解释因子的第二预测值的大小,将所述第二预测值最大的所述 w 个解释因子的值设置为1,将其余的解释因子的值设置为0,得到所述解释因子标签。

12、根据权利要求10所述的终端,其中,所述逐个输入样本至所述训练好的深度神经网络模型中之前,还包括:

获取用户预先输入的至少一个预设目标类别,以及每个预设目标类别对应的目标解释因子的目标数量。

13、根据权利要求9所述的终端,其中,所述输入多个数据至训练好的深度神经网络模型中之前,还包括:

获取用户预先输入的至少一个预设目标类别,每个所述预设目标类别均对应预先训练至收敛状态的解释因子预测模型。

14、根据权利要求9所述的终端,其中,所述将所述第一预测类别对应的目标数据输入至解释因子预测模型中,以得到所述预设目标类别对应的解释因子,包括:

将所述预设目标类别对应的目标数据输入至与所述预设目标类别对应的解释因子预测模型中,得到所述每个解释因子的第二预测值;

将所述第二预测值超过预设阈值的解释因子的值取1,并作为所述预设目标类别对应

的解释因子。

15、根据权利要求 9 所述的终端，其中，所述将所述第一预测类别对应的目标数据输入至解释因子预测模型中之前，还包括：

获取与所述预设目标类别对应的解释因子预测模型。

5 16、一种存储介质，其中，存储有能够实现深度神经网络的解释方法的程序文件，所述程序文件被处理器执行时实现以下步骤：

输入多个数据至训练好的深度神经网络模型中，得到每一个数据对应的第一预测类别；

10 当所述第一预测类别为预设目标类别时，将所述第一预测类别对应的目标数据输入至解释因子预测模型中，以得到所述预设目标类别对应的解释因子，所述解释因子预测模型根据预先设定的目标数量和所述预设目标类别的第一预测值训练得到；

统计每一个解释因子的出现次数并排序后，筛选排名最高的所述目标数量个解释因子作为所述预设目标类别的目标解释因子。

17、根据权利要求 16 所述的存储介质，其中，所述程序文件被执行时还用于预先训练所述解释因子预测模型，包括：

15 逐个输入样本至所述训练好的深度神经网络模型中，得到每个样本对应的预测类别以及第一预测值；

当所述预测类别为预设目标类别时，将所述预测类别对应的目标样本输入至所述解释因子预测模型中，得到每个解释因子的第二预测值；

根据所述第一预测值、所述第二预测值和所述目标数量构建解释因子标签；

20 基于所述解释因子标签和所述第二预测值反向传播更新所述解释因子预测模型；

循环执行上述步骤直至所述解释因子预测模型收敛，并将收敛的解释因子预测模型与所述预设目标类别绑定。

18、根据权利要求 17 所述的存储介质，其中，所述利用所述第一预测值、所述每个解释因子的第二预测值和所述目标数量构建解释因子标签，包括：

25 根据所述第一预测值和所述目标数量计算所有的解释因子中取值为 1 的解释因子的数目 w ， $w = \text{ceil}(m \times S_k)$ ，其中， ceil 表示向上取整， m 表示所述目标数量， S_k 表示所述第一预测值；

根据所述每个解释因子的第二预测值的大小，将所述第二预测值最大的所述 w 个解释因子的值设置为 1，将其余的解释因子的值设置为 0，得到所述解释因子标签。

30 19、根据权利要求 17 所述的存储介质，其中，所述逐个输入样本至所述训练好的深度神经网络模型中之前，还包括：

获取用户预先输入的至少一个预设目标类别，以及每个预设目标类别对应的目标解释因子的目标数量。

35 20、根据权利要求 16 所述的存储介质，其中，所述输入多个数据至训练好的深度神经网络模型中之前，还包括：

获取用户预先输入的至少一个预设目标类别，每个所述预设目标类别均对应预先训练至收敛状态的解释因子预测模型。

21、根据权利要求 16 所述的存储介质，其中，所述将所述第一预测类别对应的目标数据输入至解释因子预测模型中，以得到所述预设目标类别对应的解释因子，包括：

40 将所述预设目标类别对应的目标数据输入至与所述预设目标类别对应的解释因子预测模型中，得到所述每个解释因子的第二预测值；

将所述第二预测值超过预设阈值的解释因子的值取 1，并作为所述预设目标类别对应的解释因子。

45 22、根据权利要求 16 所述的存储介质，其中，所述将所述第一预测类别对应的目标数据输入至解释因子预测模型中之前，还包括：

获取与所述预设目标类别对应的解释因子预测模型。

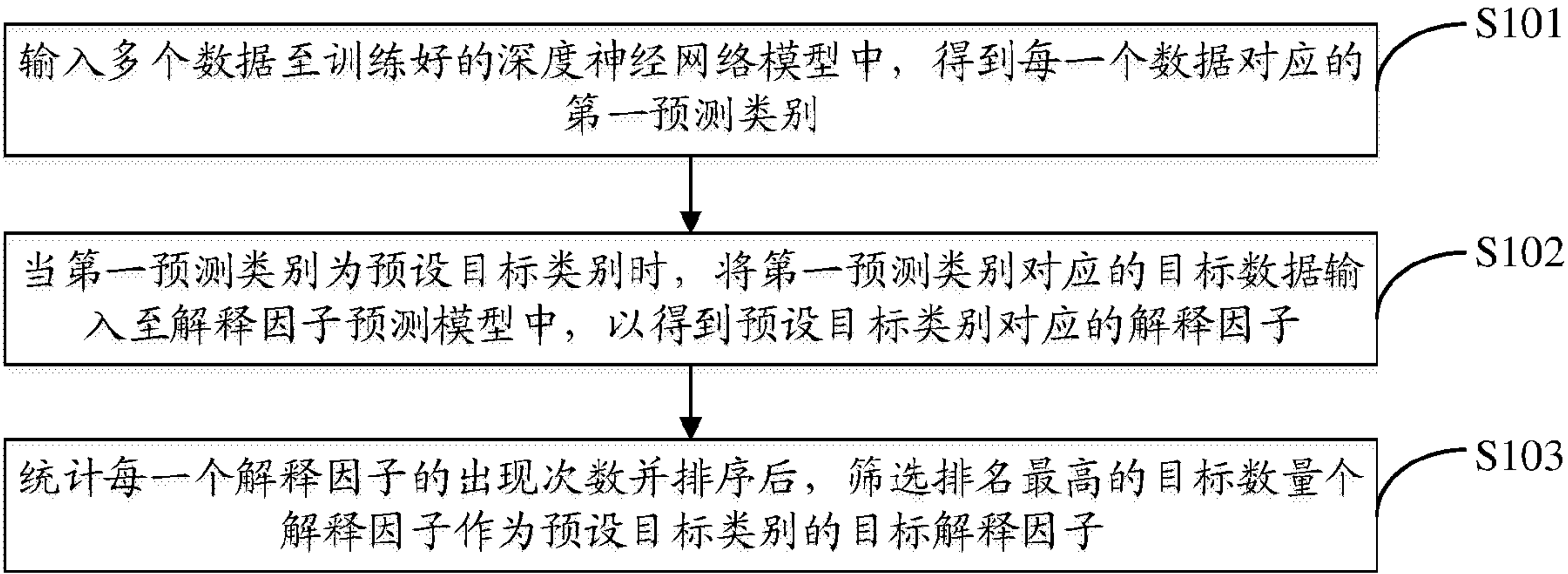


图 1

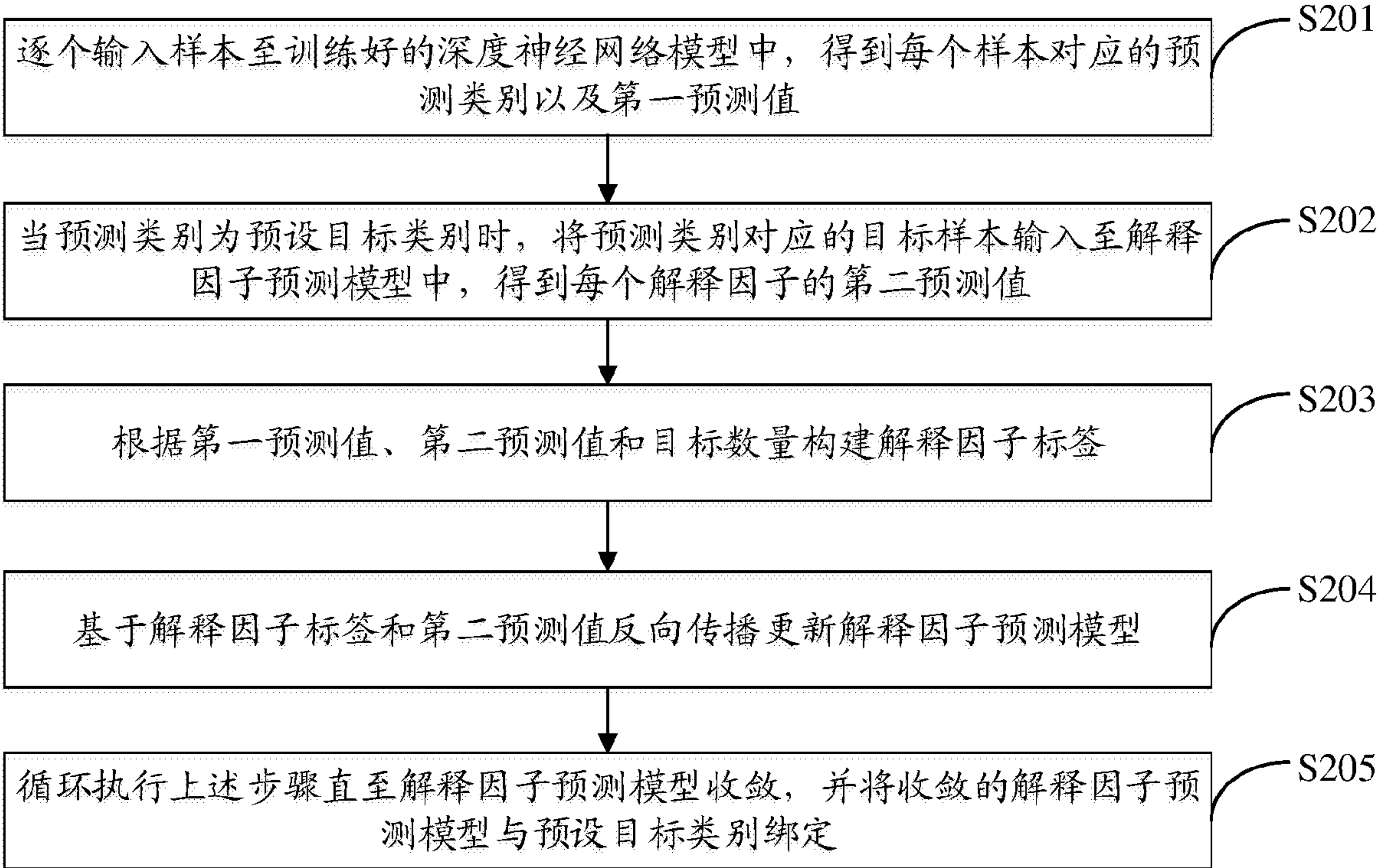


图 2

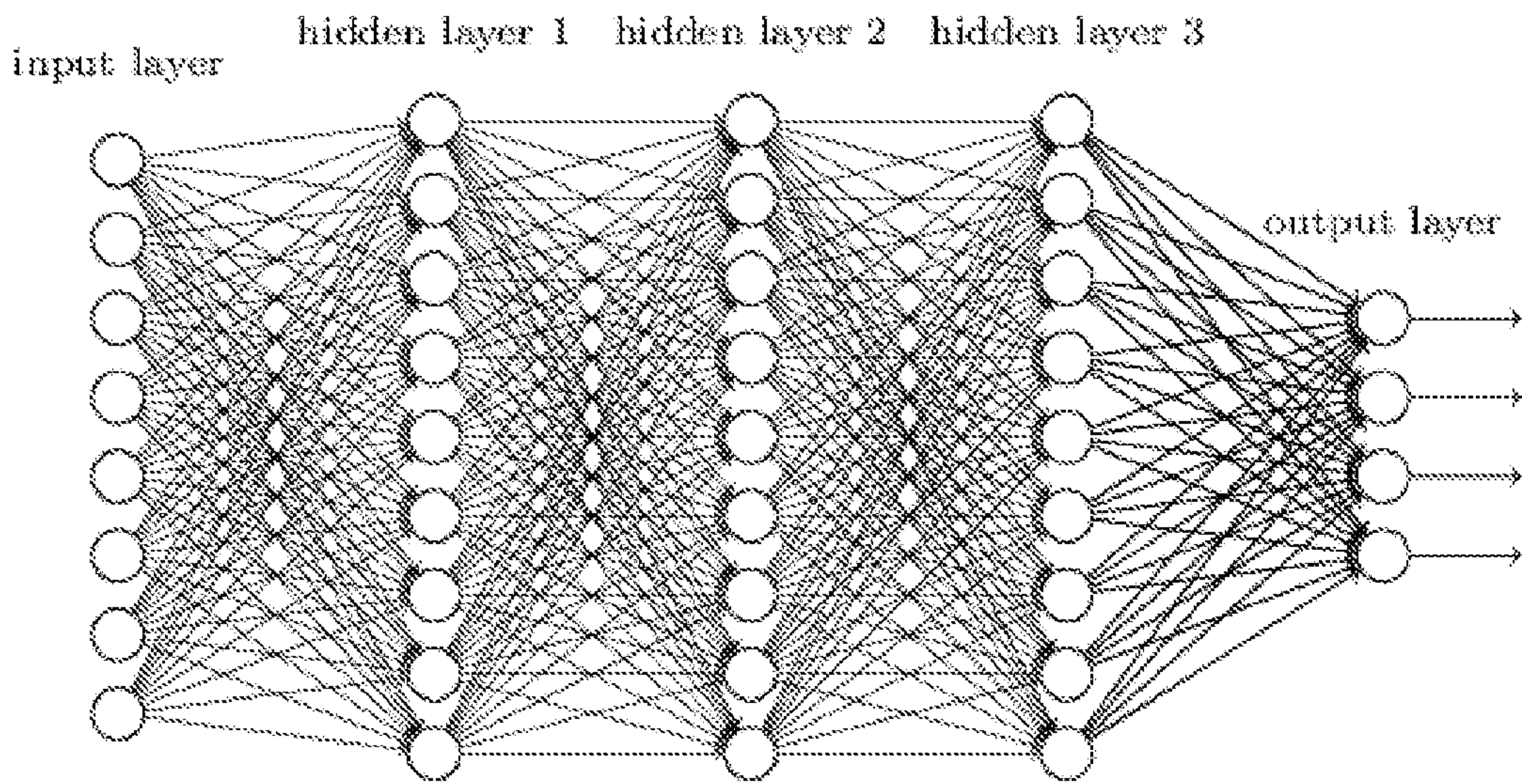


图 3

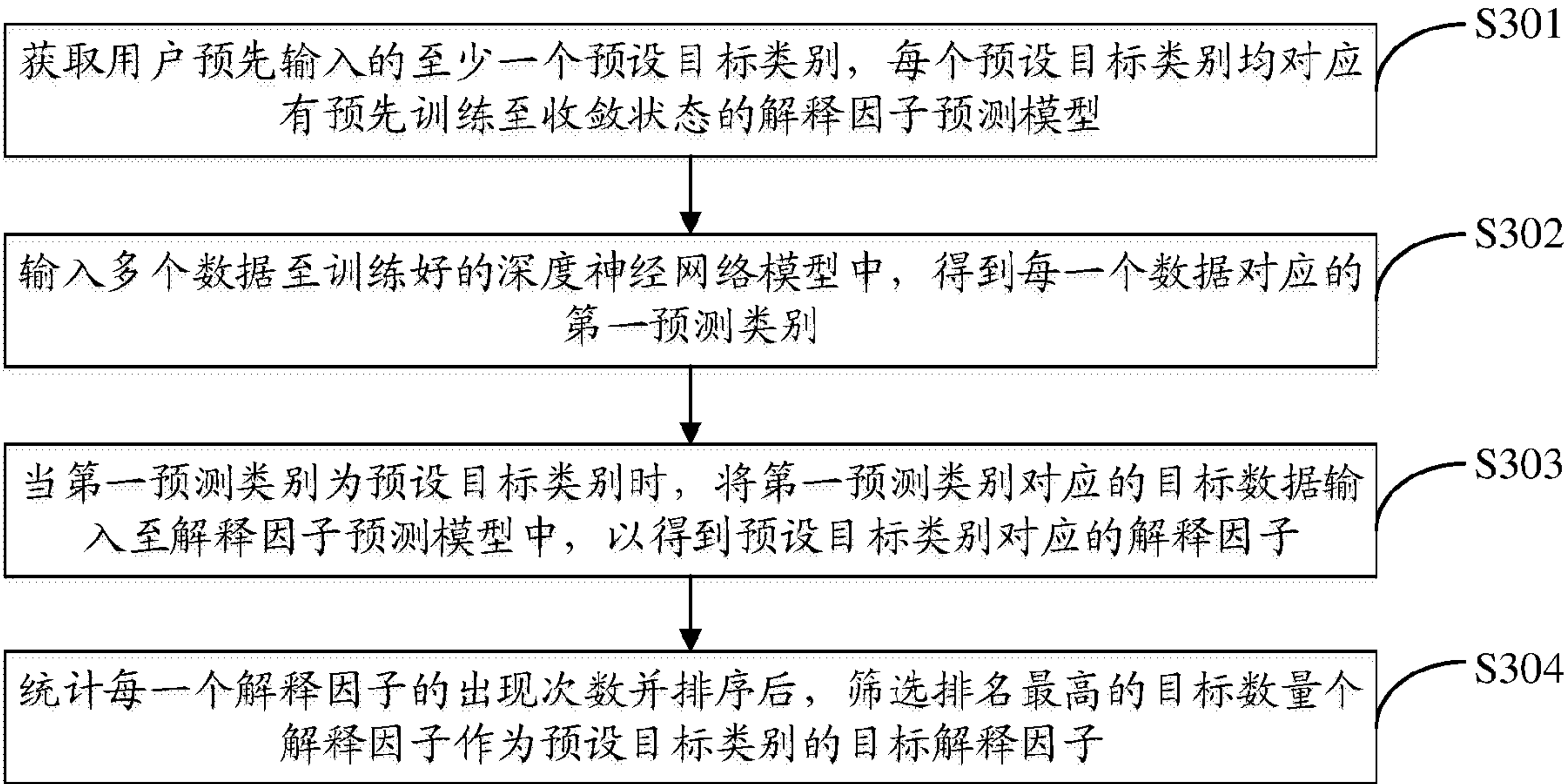


图 4

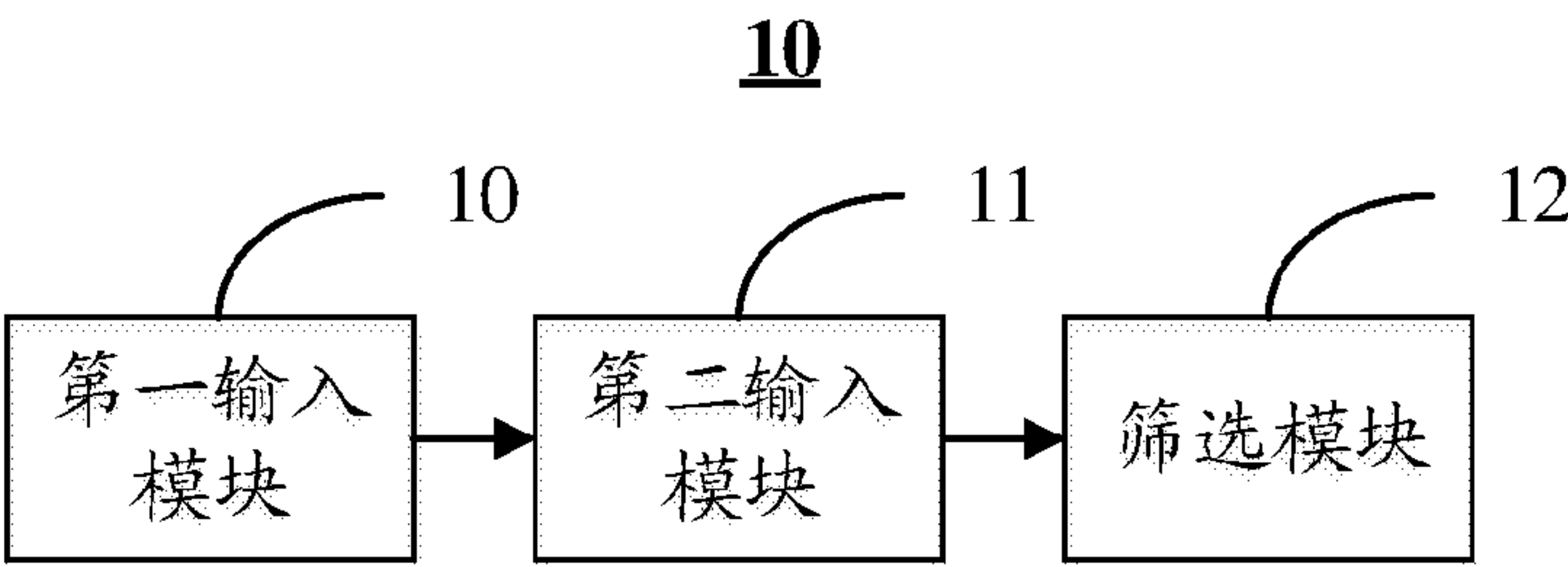


图 5

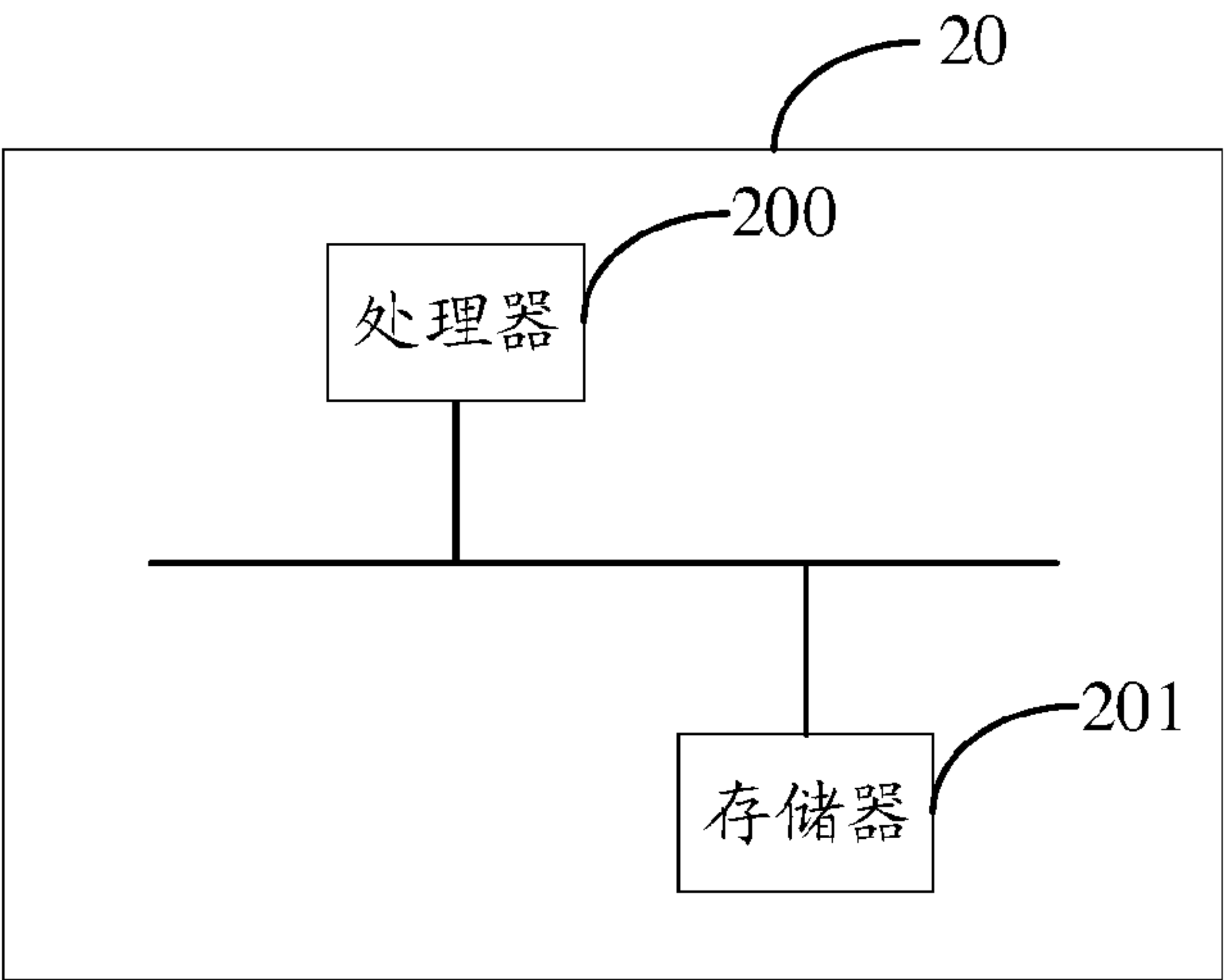


图 6

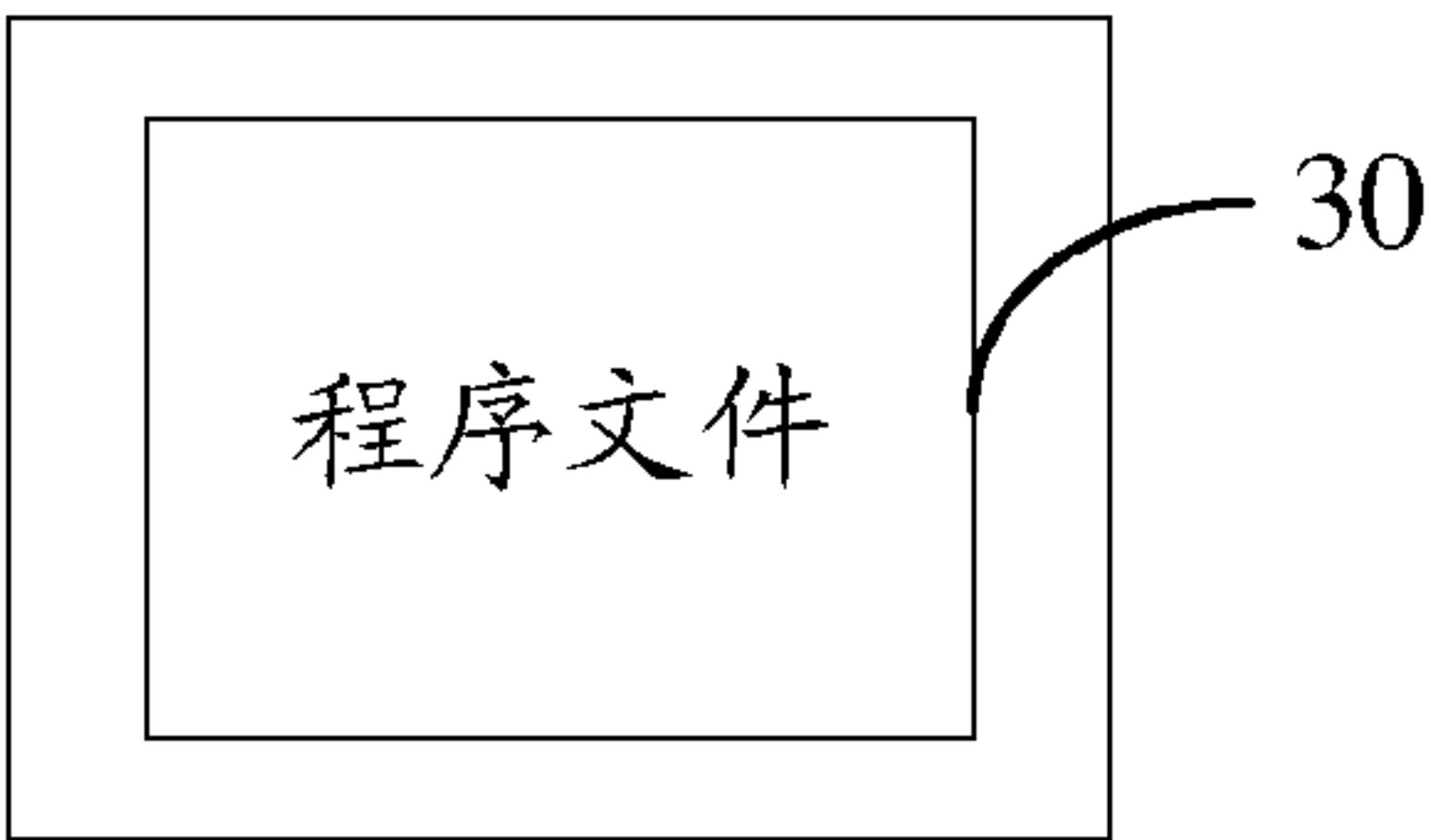


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/123196

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)

WPI, EPODOC, CNPAT, CNKI, IEEE: 深度神经网络, 解释, 预测, 类别, 因子, 因素, 次数, 频率, 排序, 标签, deep neural network, explain, interpret, prediction, forecast, category, factor, determiner, times, number, frequency, rank, sort, label, tag

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 111353603 A (SHENZHEN SUOXINDA DATA TECHNOLOGY CO., LTD.) 30 June 2020 (2020-06-30) description, paragraphs [0023]-[0033]	1-22
A	CN 107688870 A (INSTITUTE OF SOFTWARE, CHINESE ACADEMY OF SCIENCES) 13 February 2018 (2018-02-13) entire document	1-22
A	CN 110414627 A (GOODWILL HESSIAN HEALTH TECHNOLOGY CO., LTD.) 05 November 2019 (2019-11-05) entire document	1-22
A	CN 109492760 A (BEIHANG UNIVERSITY) 19 March 2019 (2019-03-19) entire document	1-22
A	US 2017315960 A1 (NEC CORPORATION) 02 November 2017 (2017-11-02) entire document	1-22

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

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

06 April 2021

Date of mailing of the international search report

15 April 2021

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/CN2020/123196

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	111353603	A	30 June 2020	None			
CN	107688870	A	13 February 2018	None			
CN	110414627	A	05 November 2019	None			
CN	109492760	A	19 March 2019	None			
US	2017315960	A1	02 November 2017	WO	2016079972	A1	26 May 2016
				JP	WO2016079972	A1	14 September 2017

国际检索报告		国际申请号 PCT/CN2020/123196
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)) WPI, EPODOC, CNPAT, CNKI, IEEE: 深度神经网络, 解释, 预测, 类别, 因子, 因素, 次数, 频率, 排序, 标签, deep neural network, explain, interpret, prediction, forecast, category, factor, determiner, times, number, frequency, rank, sort, label, tag		
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
A	CN 111353603 A (深圳索信达数据技术有限公司) 2020年 6月 30日 (2020 - 06 - 30) 说明书第[0023]-[0033]段	1-22
A	CN 107688870 A (中国科学院软件研究所) 2018年 2月 13日 (2018 - 02 - 13) 全文	1-22
A	CN 110414627 A (北京嘉和海森健康科技有限公司) 2019年 11月 5日 (2019 - 11 - 05) 全文	1-22
A	CN 109492760 A (北京航空航天大学) 2019年 3月 19日 (2019 - 03 - 19) 全文	1-22
A	US 2017315960 A1 (NEC CORPORATION) 2017年 11月 2日 (2017 - 11 - 02) 全文	1-22
☐ 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 2021年 4月 6日		国际检索报告邮寄日期 2021年 4月 15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 吴雪 电话号码 86-(10)-53961406

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2020/123196

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	111353603	A	2020年 6月 30日	无			
CN	107688870	A	2018年 2月 13日	无			
CN	110414627	A	2019年 11月 5日	无			
CN	109492760	A	2019年 3月 19日	无			
US	2017315960	A1	2017年 11月 2日	W0	2016079972	A1	2016年 5月 26日
				JP	W02016079972	A1	2017年 9月 14日