(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 2 月 18 日 (18.02.2021)

WIPO | PCT

(10) 国际公布号
WO 2021/027052 A1

(51) 国际专利分类号:
G06N 3/04 (2006.01)

(21) 国际申请号: PCT/CN2019/111612

(22) 国际申请日: 2019 年 10 月 17 日 (17.10.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910747317.4 2019年8月14日 (14.08.2019) CN

(71) 申请人: 南京大学 (NANJING UNIVERSITY) [CN/CN]; 中国江苏省南京市鼓楼区汉口路22号徐经纬, Jiangsu 210093 (CN)。

(72) 发明人: 徐经纬 (XU, Jingwei); 中国江苏省南京市栖霞区仙林大道163号, Jiangsu 210046 (CN)。 王慧妍 (WANG, Huiyan); 中国江苏省南京市栖霞区仙林大道163号徐经纬, Jiangsu 210046 (CN)。 许畅 (XU, Chang); 中国江苏省南京市江苏省南京市栖霞区仙林大道163号徐经纬, Jiangsu 210046 (CN)。 马晓星 (MA, Xiaoxing); 中国江苏省南京市栖霞区仙林大道163号徐经纬, Jiangsu 210046 (CN)。 吕建 (LYU, Jian); 中国江苏省南京市栖霞区仙林大道163号徐经纬, Jiangsu 210046 (CN)。

(74) 代理人: 南京乐羽知行专利代理事务所 (普通合伙) (NANJING LE YU ZHI XING PATENT AGENCY FIRM (GERERAL PARTHNERSHIP)); 中国江苏省南京市广州路37号江苏科技大厦26层, Jiangsu 210000 (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,

(54) **Title:** INTERLAYER PARSING-BASED INPUT INSTANCE VERFICATION METHOD FOR NEURAL NETWORK MODEL

(54) 发明名称: 面向神经网络模型的基于层间剖析的输入实例验证方法

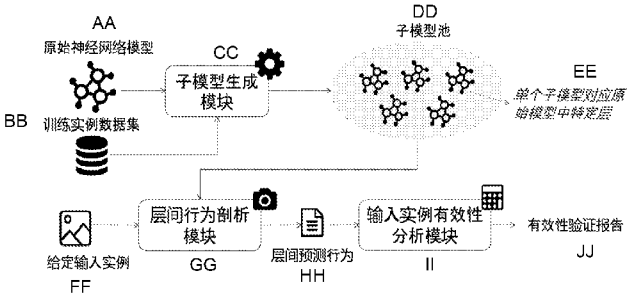


图 1

AA Original neural network model
BB Training instance data set
CC Sub-model generating module
DD Sub-model pool
EE A single sub-model corresponds to a specific layer in the original model
FF Given input instance
GG Interlayer behavior parsing module
HH Interlayer predicted behavior
II Input instance validity analysis module
JJ Validity verification report

(57) **Abstract:** Disclosed is an interlayer parsing-based input instance verification method for a neural network model, the method comprising: providing a neural network model and a training data set thereof, extracting intermediate information and generating a sub-model corresponding to each corresponding layer; for any input instance to be verified, inputting same into a sub-model to acquire a total behavior profile after interlayer parsing; analyzing the interlayer parsing profile of the input instance; verifying whether the input instance is valid; and providing a confidence score regarding the validity. In the present invention, the validity of a given input instance is analyzed by, on the basis of an interlayer parsing means inside of a training module, using the behavior of the input instance during the parsing of each layer of the model, which may avoid the disadvantage of verification consuming a lot of time which exists for the verification means in the prior art in which many different models must assist on another, and which may core accurately perform input

WO 2021/027052 A1

[见续页]

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, 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)。

根据细则4.17的声明：

— 关于申请人有权申请并被授予专利(细则4.17(ii))

— 发明人资格(细则4.17(iv))

本国际公布：

— 包括国际检索报告(条约第21条(3))。

verification, thus helping to distinguish the validity of an input instance of a given neural network, thereby improving the accuracy and security of the neural network during actual use.

(57) 摘要：本发明公开一种面向神经网络模型的基于层间剖析的输入实例验证方法，给定神经网络模型与其训练数据集，提取中间信息并生成对应每一层对应的子模型；对于任意待验证输入实例，输入子模型获取层间剖析后总行为profile；分析输入实例的层间剖析profile，验证该输入实例是否为有效并给出对于有效的置信度分数。本发明基于训练模型内部的层间剖析手段，利用输入实例在模型各层次剖析时的行为来分析给定输入实例的有效性，能避免已有技术需要较多不同模型相互借助的验证手段所存在的验证时间消耗巨大的弊端，能够更加准确的进行输入验证，从而能够帮助区分给定神经网络的输入实例的有效性，从而提高神经网络在实际使用中的准确性与安全性。

面向神经网络模型的基于层间剖析的输入实例验证方法

技术领域

本发明涉及一种面向神经网络模型的基于层间剖析的输入实例验证方法，用于神经网络的测试和输入验证等技术领域任务。

背景技术

神经网络模型被广泛使用在现实生活的各个领域，例如图像处理、物体识别、自动驾驶等。然而，神经网络模型由于其结构复杂性与不可解释性，通常来说，对于给定训练完成的神经网络模型，现实生活中往往将其用于各式各样的场景并且得到好的结果。然而由于模型训练的特性，天然对于不同场景的输入会存在适用或不适用的情况，若不区分适用性进行现实部署，将有可能造成神经网络模型效果异常（例如，自动驾驶汽车若用于不使用的场景例如极端天气或处理极端曝光输入图片，有可能无法正常避障发生严重交通事故）。因此，对于神经网络模型对于输入的自动验证其是否适用（或是否有效）则至关重要。

当前已经有不少工作关注无效输入的问题，但是在使用上有一定的局限性。首先，不少工作核心是基于距离评估的方法，将未知输入与训练数据进行距离评测，从而判断其有效性，这一方式受制于训练数据的规模，难以应用到现实中往往要求大规模训练得到的神经网络模型中；其次，神经网络模型由于天然具备一定的泛化能力，其本身输入处理的能力与其训练数据并不严格等价，直接用后者帮助验证可能会带来一些精度的偏差；最后，当前方法往往属于线下验证的方式，效率难以满足实时验证的要求，从而难以在已部署的现实场景中使用。

发明内容

发明目的：针对现有技术中存在的问题与不足，本发明提供一种面向神经网络模型的基于层间剖析的输入实例验证方法，该方法具有易用性、有效性和高效性等特点。易用性是指该方法能够使用于现实生活中常见的大规模训练数据训练得到的神经网络模型中，使用场景并不严重受制于训练场景规模和神经网络模型的复杂程度。有效性是指该方法对于输入实例的有效性验证准确度较高，能够有效判别出有效与无效输入。高效性是指该方法对于验证输入实例所需时间代价较小，能够满足实时验证的要求，可部署于运行中的神经网络模型中进行输入验证。

技术方案：一种面向神经网络模型的基于层间剖析的输入实例验证方法，包括如下步骤：

步骤一：利用给定神经网络模型与其对应的训练数据，将训练数据输入给定神经网络模型，提取训练过程中数据在模型各中间层次的中间信息，并根据中间信息训练各层次对应的子模型，每一个子模型包含给定神经网络模型从输入层到对应中间层次的知识并模拟给定神经网络模型预测行为；

步骤二：利用步骤一获取的各中间层次对应的子模型，对于待验证的输入实例收集按照层次递增在各层次对应子模型上的预测行为 snapshot，并汇总形成输入实例在所有子模型中的总行为 profile；

步骤三：基于步骤二获取的给定输入实例对应的层次剖析所获取的总行为 profile，分析其层次预测行为 snapshot 的有效性以及总行为 profile 的有效性，并给出有效性置信度分数，并评估有效性。

为实现并优化上述技术方案，采取的具体措施还包括：

进一步的，所述神经网络是指一类利用神经元进行层次连接形成的进行大数据特征提取及预测的数据结构，包含输入层、隐含层、输出层，每一层包含大量神经元，层次间通过神经元相互连接，由输入层向输出层传递信息，例如常用的各类 DNN、CNN、RNN 模型等；所述神经元是对于神经元输入利用内置函数等对输入数据进行运算操作，并输出的数据结构；所述内置函数为固定常用的几种流行的激活核函数形式，例如 ReLU、Sigmoid、Softmax 等；所述输入实例是指神经网络模型的一次输入或批输入，例如：对于面向图片分类问题训练的神经网络，输入实例是指某一张图片文件或多张图片构成的批输入。

进一步的，所述步骤一中，提供给定神经网络模型及其训练数据集，提取训练过程各层次中间信息，其中中间信息包括中间层次每一层神经元在训练过程中得到的模型参数信息(如 CNN 模型中的 weight，bias 等)，各神经元输入值和输出值等。其中参数信息用于记录当前模型通过训练过程从训练数据集中学习到的知识，输入值和输出值用于为后续子模型的训练过程提供训练数据。

进一步的，所述步骤一中，所述每一层次如层次 k 对应的子模型是类似给定神经网络模型结构的神经网络模型，其包含两部分结构，第一部分继承给定神经网络在原训练过程后获得的后模型从输入层到对应层次 k 的所有得到的模型参

数信息(如 weight、bias 等)及其对应模型结构，第二部分利用基础元模型连接层次 k 神经元与预测输出神经元，并利用步骤一记录的 k 层次神经元中间信息（原始训练数据输入给定神经网络后，在该层次 k 的输出集合）和原始训练集对应预测值标记进行重新训练，并获取该部分训练后的参数信息，两部分参数合并则得到带参数的子模型结构；所述基础元模型常指代线性回归模型，但并不仅限于该模型；所述重新训练通常只训练第二部分的参数，但并不仅限于此，重新训练根据不同的应用场景(如利用层间剖析方法对模型参数进行整体微调)训练第一部分和第二部分的参数。

进一步的，所述步骤二中，所述输入实例在各层次对应子模型上的预测行为 snapshot 是指利用输入实例传入各层次对应的子模型进行预测后，所得到的预测概率分布结果等信息，但并不仅限于预测概率分布结果；

进一步的，所述总行为 profile 是指各子模型得到的预测行为 snapshot 的集合，用于后续步骤三进行对于该待验证输入实例的验证评估，是基础材料。

进一步的，所述步骤三中，所述利用步骤二总行为 profile 进行对对应层次预测行为 snapshot 的有效性分析，可采用以下分析方法：

方法一：考虑当前层次预测行为中，预测最大值与最终预测值的概率差异，利用相对大小比例作为每一层次 snapshot 有效性分数；

方法二：考虑当前层次与之前层次直接预测行为差异后，利用预测行为各概率变化情况以及最终预测值的概率变化相对比例作为每一层次 snapshot 有效性分数；

进一步的，所述步骤三中， 所述利用步骤二总行为 profile 进行总 profile 有效性分析，可采用以下分析方法：

方法一：利用训练集在各层次实际在训练集上的预测准确性作为权重进行分析建模。该方法将各层次 snapshot 有效性分析的结果作为线性模型的输入，线性模型中包含的参数则基于训练集预测准确性进行设置，最终通过基于预测准确性加权的方式计算最终总 profile 有效性分数；

方法二：利用观察，采用常用增长函数曲线进行权重设置（线性、对数、指数）。该方法将各层次 snapshot 有效性分析的结果作为选取的增长函数的输入，并以人为设定的方式设置增长函数中包含的参数，用以计算最终总 profile 有效

性分数；

方法三：对训练集数据求得各层次 snapshot 及其总行为 profile，将 snapshot 有效性分析作为输入数据，对应验证结果作为标记数据，采用机器学习模型训练可计算最终 profile 有效性分数的模型；所述对应验证结果可由人为给出，或结合给定神经网络模型对该输入的预测准确性给出，但并不仅限于此；所述机器学习模型可采用如线性回归，逻辑回归，SVM，神经网络等经典机器学习模型，但并不仅限于此。

进一步的，所述步骤三中，所述有效性置信度分数是指给定待验证输入所计算得到的总 profile 有效性分数为 0 至 1 之间的某一数值，代表对于当前待验证输入有效性的置信度，其中越接近 0 代表越无效，越接近于 1 代表越有效。所述有效性置信度分数取值范围选取和大小关系设定并不仅限于此。

进一步的，所述步骤三中，所述评估有效性，是指利用计算得到的有效性置信度，通过设定阈值划分进行有效/无效划分，划分阈值可以由事先给定或经验得到，主要根据不同模型实际使用场景对于有效输入实例的容忍程度不同决定，一般来说安全性要求越严格的场景阈值越接近于 1。

有益效果：与现有技术相比，本发明能够弥补已有神经网络模型输入实例验证技术的不足，利用特定输入在模型中进行层间剖析的方法，高效地检测评估输入实例的有效性，并利用评估的有效性，从而进行输入实例的实时筛选，从而提升神经网络模型实际部署的效果。

附图说明

图 1 为本发明系统结构图；

图 2 为本发明提供的子模型结构细节图；

图 3 为本发明提供的子模型生成模块的工作流程图；

图 4 为本发明提供的层间行为剖析模块的工作流程图；

图 5 为本发明提供的有效性验证分析模块的工作流程图。

具体实施方式

下面结合具体实施例，进一步阐明本发明，应理解这些实施例仅用于说明本发明而不用于限制本发明的范围，在阅读了本发明之后，本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。

神经网络是指一类利用神经元进行层次连接形成的进行大数据特征提取及预测的数据结构，包含输入层、隐含层、输出层，每一层包含大量神经元，层次间通过神经元相互连接，由输入层向输出层传递信息，例如常用的各类DNN、CNN、RNN模型等；神经元是对于神经元输入利用内置函数等对输入数据进行运算操作，并输出的数据结构；内置函数为固定常用的几种流行的激活核函数形式，例如ReLU、Sigmoid、Softmax等；输入实例是指神经网络模型的一次输入或批输入，例如：对于面向图片分类问题训练的神经网络，输入实例是指某一张图片文件或多张图片构成的批输入。

面向神经网络模型的基于层间剖析的输入实例验证方法，包括如下步骤：

步骤一：利用给定神经网络模型与其对应的训练数据，将训练数据输入给定神经网络模型，提取训练过程中数据在模型各中间层次的中间信息，并根据中间信息训练各层次对应的子模型，每一个子模型包含给定神经网络模型从输入层到对应中间层次的知识并模拟其预测行为；

中间信息包括中间层次每一层神经元在训练过程中得到的模型参数信息(如CNN模型中的weight，bias等)，各神经元输入值和输出值等。其中参数信息用于记录当前模型通过训练过程从训练数据集中学习到的知识，输入值和输出值用于为后续子模型的训练过程提供训练数据。

每一层次如层次k对应的子模型是类似给定神经网络模型结构的神经网络模型，其包含两部分结构，第一部分继承给定神经网络在原训练过程后获得的后模型从输入层到对应层次k的所有得到的模型参数信息(如weight、bias等)及其对应模型结构，第二部分利用基础元模型连接层次k神经元与预测输出神经元，并利用步骤一记录的k层次神经元中间信息（原始训练数据输入给定神经网络后，在该层次k的输出集合）和原始训练集对应预测值标记进行重新训练，并获取该部分训练后的参数信息，两部分参数合并则得到带参数的子模型结构；基础元模型常指代线性回归模型，但并不仅限于该模型；重新训练通常只训练第二部分的参数，但并不仅限于此，重新训练根据不同的应用场景(如利用层间剖析方法对模型参数进行整体微调)训练第一部分和第二部分的参数。

步骤二：利用步骤一获取的各中间层次对应的子模型，对于待验证的输入实例收集按照层次递增在各层次对应子模型上的预测行为snapshot，并汇总形成输

入实例在所有子模型中的总行为 profile；

输入实例在各层次对应子模型上的预测行为 snapshot 是指利用输入实例传入各层次对应的子模型进行预测后，所得到的预测概率分布结果等信息；

总行为 profile 是指各子模型得到的预测行为 snapshot 的集合，用于后续步骤三进行对于该待验证输入实例的验证评估，是基础材料。

步骤三：基于步骤二获取的给定输入实例对应的层次剖析所获取的总行为 profile，分析其层次预测行为 snapshot 的有效性以及总行为 profile 的有效性，并给出有效性置信度分数，并评估有效性。

利用步骤二总行为 profile 进行对对应层次预测行为 snapshot 的有效性分析，可采用以下分析方法：

方法一：考虑当前层次预测行为中，预测最大值与最终预测值的概率差异，利用相对大小比例作为每一层次 snapshot 有效性分数；

方法二：考虑当前层次与之前层次直接预测行为差异后，利用预测行为各概率变化情况以及最终预测值的概率变化相对比例作为每一层次 snapshot 有效性分数；

利用步骤二总行为 profile 进行总 profile 有效性分析，可采用以下分析方法：

方法一：利用训练集在各层次实际在训练集上的预测准确性作为权重，并综合各层次 snapshot 有效性分析进行计算最终 profile 有效性分数；

方法二：利用观察，采用常用增长函数曲线进行权重设置（线性、对数、指数），并综合各层次 snapshot 有效性分析进行计算最终 profile 有效性分数；

方法三：对训练集数据求得各层次 snapshot 及其总行为 profile，将 snapshot 有效性分析作为输入数据，对应验证结果作为标记数据，采用机器学习模型训练可计算最终 profile 有效性分数的模型；对应验证结果可由人为给出，或结合给定神经网络模型对该输入的预测准确性给出，但并不仅限于此；所述机器学习模型可采用如线性回归，逻辑回归，SVM，神经网络等经典机器学习模型。

有效性置信度分数是指给定待验证输入所计算得到的总 profile 有效性分析，为 0 至 1 之间的某一数值，代表对于当前待验证输入有效性的置信度，其中越接近 0 代表越无效，越接近于 1 代表越有效。

评估有效性，是指利用计算得到的有效性置信度，通过设定阈值划分进行有

效/无效划分，划分阈值可以由事先给定或经验得到，主要根据不同模型实际使用场景对于有效输入实例的容忍程度不同决定，一般来说安全性要求越严格的场景阈值越接近于 1。

如图 1 所示，本发明实施例提供的面向神经网络模型的基于层间剖析的输入实例验证方法，首先可事先/线下利用原始数据网络模型与其训练实例数据集，生成各层次对应子模型并组成子模型池，池中每一个单个子模型对应原始模型中特定层包括的知识并能够用于预测。其次，对于任意给定待验证的输入实例，通过输入子模型池中各个模型进行预测，剖析其对于原始神经网络各层次按层次预测下的层间行为，并输出层间剖析总行为（profile），其中包括对应各层次的层间预测行为（snapshot）。最后，综合层间预测总行为 profile 与其包含的各层间预测行为 snapshot 进行有效性分析模块，并输出有效性分析报告。整个方法框架包含三个模块对应三个步骤：子模型生成模块，层间行为剖析模块，与输入实例有效性分析模块。

步骤一：子模型生成模块，生成各层次对应子模型。

如图 2 所示，对于选定层次 k 对应的子模型结构设计包括两个部分，第一部分为原始模型输入层到当前选定层之间的原始神经网络模型结构的拷贝，详细包括模型上结构信息、参数信息等；第二部分为利用元模型结构基于输入数据在当前层 k 的输出值与最终预测值的重新训练模型，其中元模型图示中为单层线性全连接，即线性回归模型，但不限于此。

如图 3 所示，图 3 表示子模型生成模块的工作流程图，输入原始神经网络模型与训练数据集，首先保存所有后续所需中间结果，如各层神经元的中间输入输出值。然后迭代选择第 k 层进行子模型生成，单独生成子模型第一第二部分最终拼接完成子模型生成。最终，将所有选定层对应生成的子模型综合输出成子模型池。

步骤二：层间行为剖析模块，剖析待验证输入实例的层间行为。

如图 4 所示，图 4 表示层间行为剖析模块的工作流程图，输入待验证输入实例与步骤一获取的子模型池，以此将待验证输入实例输入子模型池中各层次对应的子模型进行层间行为 snapshot 获取，各 snapshot 对应反映了子模型对应的原始模型的某一特定层的行为信息，最后汇总形成此输入实例的总行为 profile，从而

反映此输入实例在原模型上层次直接传输的剖析行为。

步骤三：输入实例有效性分析模块，分析待测输入实例有效性并报告。

如图 5 所示，图 5 表示输入实例有效性分析模块，对于待验证输入实例在步骤二中获取的总层间剖析行为 profile，进行剖析方法选择（权重-based 或学习-based）和有效性程度计算。细节上，本方法首先选择 snapshot 分析方法（预测最大值与最终预测值的概率差异的相对大小比例，或概率变化情况以及最终预测值的概率变化相对比例，如上文所述），对于单个 snapshot 进行评分的，并选择特定 profile 分析方法（两种权重-based 或一种学习-based 的分析方法，如上文所述），汇总对于各 snapshot 的评分作为对于整个层间剖析总行为 profile 的有效性程度评估。通过对于与实际应用场景要求的安全性阈值判别，可以进一步获取对于待验证输入实例是否有效的决策，并报告。在实际场景中，安全性要求更高的场景通常伴随更高的阈值，使得相同情况下无效输入实例比重相对增加。本方法可以后续与过滤无效输入实例的手段总结，达到在实际场景中合理挑选输入实例输入神经网络模型进行判别的方式，增加实际使用神经网络模型的准确程度。

权 利 要 求 书

1. 一种面向神经网络模型的基于层间剖析的输入实例验证方法，其特征在于，包括如下步骤：

步骤一：利用给定神经网络模型与其对应的训练数据，将训练数据输入给定神经网络模型，提取训练过程中数据在模型各中间层次的中间信息，并根据中间信息训练各层次对应的子模型，每一个子模型包含给定神经网络模型从输入层到对应中间层次的知识并模拟其预测行为；

步骤二：利用步骤一获取的各中间层次对应的子模型，对于待验证的输入实例收集按照层次递增在各层次对应子模型上的预测行为 snapshot，并汇总形成输入实例在所有子模型中的总行为 profile；

步骤三：基于步骤二获取的给定输入实例对应的层次剖析所获取的总行为 profile，分析其层次预测行为 snapshot 的有效性以及总行为 profile 的有效性，并给出有效性置信度分数，并评估有效性。

2. 如权利要求 1 所述的面向神经网络模型的基于层间剖析的输入实例验证方法，其特征在于，所述步骤一中，提供给定神经网络模型及其训练数据集，提取训练过程各层次中间信息，其中中间信息包括中间层次每一层神经元在训练过程中得到的模型参数信息、各神经元输入值和输出值；其中参数信息用于记录当前模型通过训练过程从训练数据集中学习到的知识，输入值和输出值用于为后续子模型的训练过程提供训练数据。

3. 如权利要求 1 所述的面向神经网络模型的基于层间剖析的输入实例验证方法，其特征在于，所述步骤一中，层次 k 对应的子模型是类似给定神经网络模型结构的神经网络模型，其包含两部分结构，第一部分继承给定神经网络在原训练过程后获得的后模型从输入层到对应层次 k 的所有得到的模型参数信息及其对应模型结构，第二部分利用基础元模型连接层次 k 神经元与预测输出神经元，并利用步骤一记录的 k 层次神经元中间信息和原始训练集对应预测值标记进行重新训练，并获取该部分训练后的参数信息，两部分参数合并则得到带参数的子模型结构；所述基础元模型常指代线性回归模型。

4. 如权利要求 1 所述的面向神经网络模型的基于层间剖析的输入实例验证方法，其特征在于，所述步骤二中，所述输入实例在各层次对应子模型上的预测行为 snapshot 是指利用输入实例传入各层次对应的子模型进行预测后，所得到的

预测信息。

5. 如权利要求 1 所述的面向神经网络模型的基于层间剖析的输入实例验证方法，其特征在于，所述总行为 profile 是指各子模型得到的预测行为 snapshot 的集合，用于后续步骤三进行对于该待验证输入实例的验证评估。

6. 如权利要求 1 所述的面向神经网络模型的基于层间剖析的输入实例验证方法，其特征在于，所述步骤三中，所述利用步骤二总行为 profile 进行对对应层次预测行为 snapshot 的有效性分析，可采用以下分析方法：

方法一：考虑当前层次预测行为中，预测最大值与最终预测值的概率差异，利用相对大小比例作为每一层次 snapshot 有效性分数；

方法二：考虑当前层次与之前层次直接预测行为差异后，利用预测行为各概率变化情况以及最终预测值的概率变化相对比例作为每一层次 snapshot 有效性分数。

7. 如权利要求 1 所述的面向神经网络模型的基于层间剖析的输入实例验证方法，其特征在于，所述步骤三中， 所述利用步骤二总行为 profile 进行总 profile 有效性分析，可采用以下分析方法：

方法一：利用训练集在各层次实际在训练集上的预测准确性作为权重，并综合各层次 snapshot 有效性分析进行计算最终 profile 有效性分数；

方法二：利用观察，采用常用增长函数曲线进行权重设置，并综合各层次 snapshot 有效性分析进行计算最终 profile 有效性分数；

方法三：对训练集数据求得各层次 snapshot 及其总行为 profile，将 snapshot 有效性分析作为输入数据，对应验证结果作为标记数据，采用机器学习模型训练可计算最终 profile 有效性分数的模型。

8. 如权利要求 1 所述的面向神经网络模型的基于层间剖析的输入实例验证方法，其特征在于，所述步骤三中，所述有效性置信度分数是指给定待验证输入所计算得到的总 profile 有效性分数，为 0 至 1 之间的某一数值，代表对于当前待验证输入有效性的置信度。

9. 如权利要求 1 所述的面向神经网络模型的基于层间剖析的输入实例验证方法，其特征在于，所述步骤三中，所述评估有效性，是指利用计算得到的有效性置信度，通过设定阈值划分进行有效/无效划分。

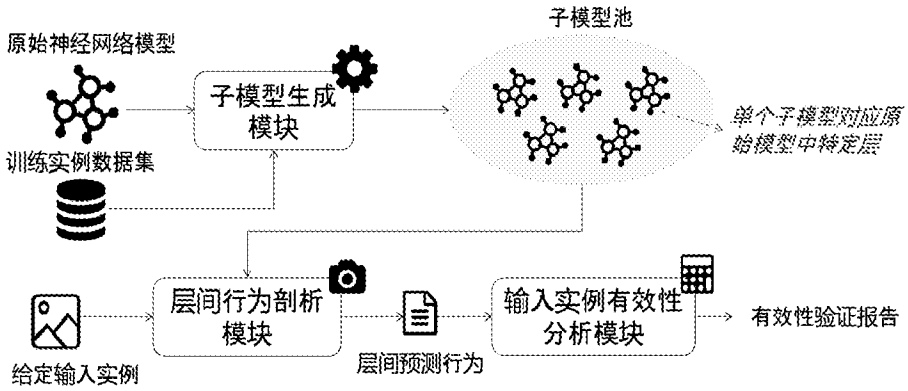

原始神经网络模型
训练实例数据集
子模型生成模块
子模型池
单个子模型对应原始模型中特定层
层间行为剖析模块
给定输入实例
层间预测行为
输入实例有效性分析模块
有效性验证报告

图 1

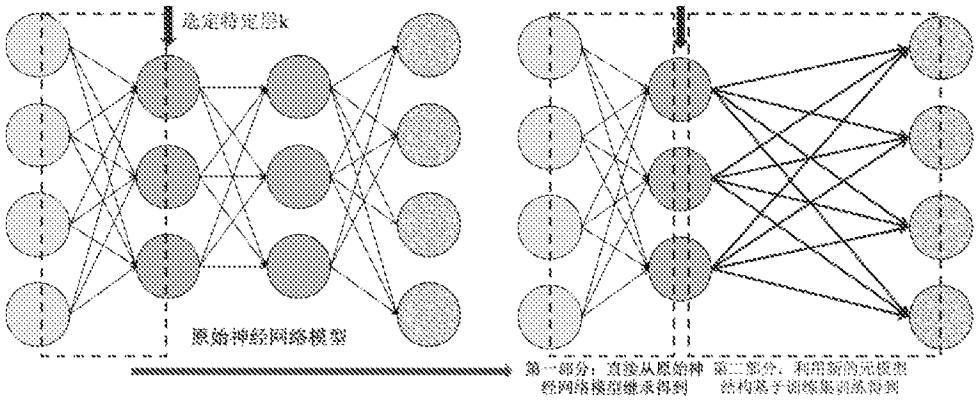

给定特定层k
原始神经网络模型

图 2

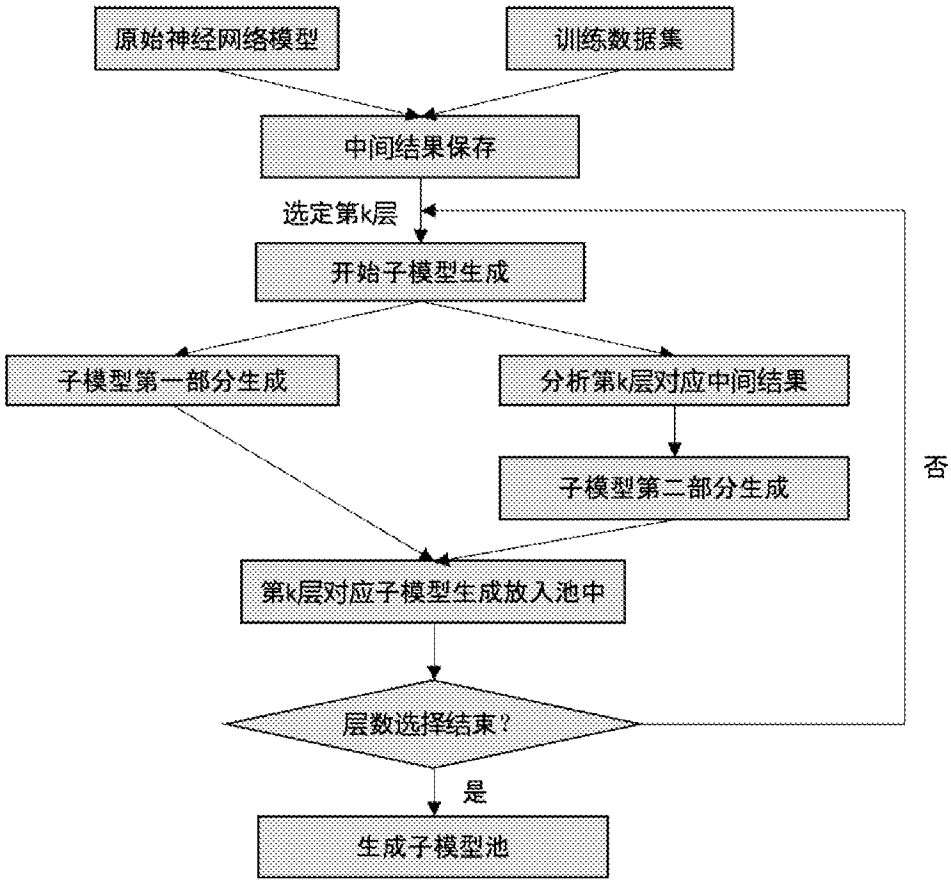

原始神经网络模型
训练数据集
中间结果保存
选定第k层
开始子模型生成
子模型第一部分生成
分析第k层对应中间结果
子模型第二部分生成
第k层对应子模型生成放入池中
层数选择结束？
否
是
生成子模型池

图 3

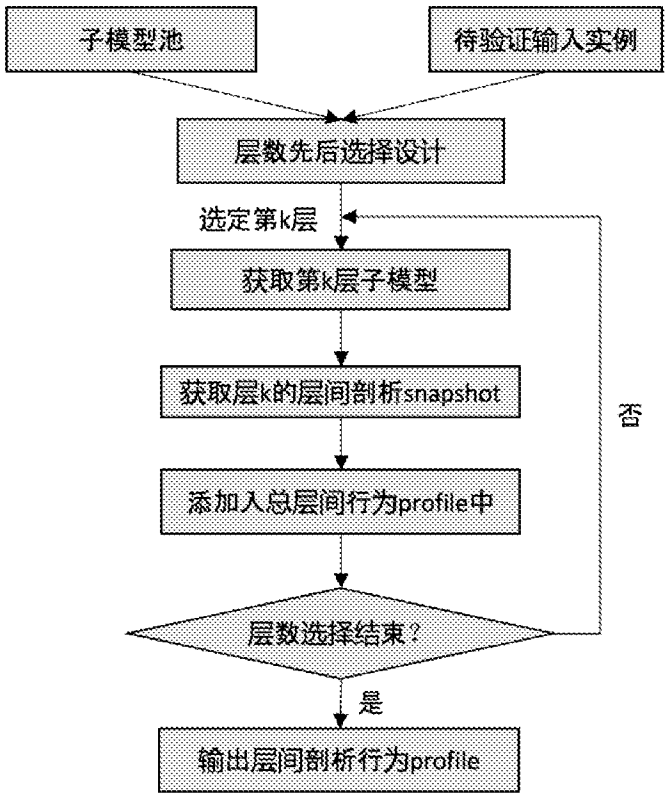

子模型池
待验证输入实例
层数先后选择设计
选定第k层
获取第k层子模型
获取层k的层间剖析snapshot
添加入总层间行为profile中
层数选择结束？
否
是
输出层间剖析行为profile

图 4

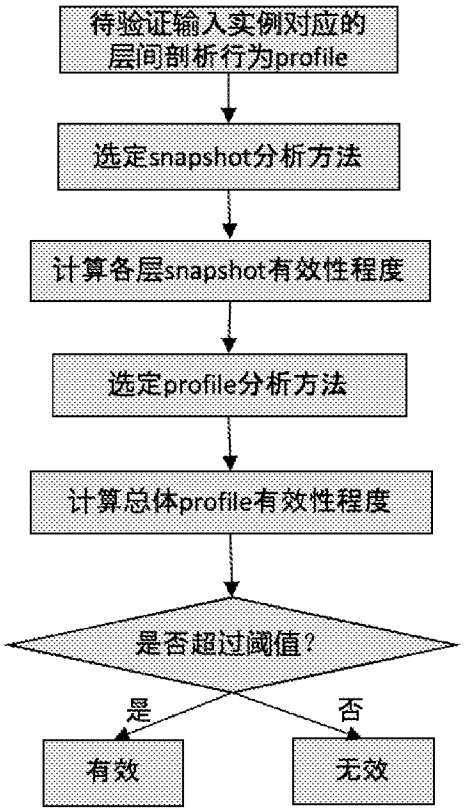

待验证输入实例对应的层间剖析行为profile
选定snapshot分析方法
计算各层snapshot有效性程度
选定profile分析方法
计算总体profile有效性程度
是否超过阈值？
是
否
有效
无效

图 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/111612

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,G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, CNABS, CNKI, DWPI, SIPOABS, IEEE: 神经网络, 层间, 剖析, 输入, 实例, 验证, 子模型, 模拟, 预测, 总行为, 置信度, 训练, 权重, 阈值, 中间层, neural network model, interlayer, profiling, input, example, verify, hierarchy, submodel, simulate, predict, profile, confidence, train, weight, threshold, intermediate layer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 109344806 A (4PARADIGM (BEIJING) TECHNOLOGY CO., LTD.) 15 February 2019 (2019-02-15) see entire document	1-9
A	CN 105184678 A (QIQIHAR UNIVERSITY) 23 December 2015 (2015-12-23) see entire document	1-9
A	US 2019156933 A1 (VERISIM LIFE INC) 23 May 2019 (2019-05-23) see entire document	1-9
A	WO 2019098418 A1 (SAMSUNG ELECTRONICS CO., LTD. et al.) 23 May 2019 (2019-05-23) see entire document	1-9

☐ 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
11 May 2020	**20 May 2020**

Name and mailing address of the ISA/CN	Authorized officer
China National Intellectual Property Administration (ISA/CN) **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/111612

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109344806 A	15 February 2019	CN 109344806 B	23 August 2019
CN 105184678 A	23 December 2015	None	
US 2019156933 A1	23 May 2019	WO 2019104101 A1	31 May 2019
WO 2019098418 A1	23 May 2019	None	

国际检索报告

国际申请号
PCT/CN2019/111612

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,G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNTXT,CNABS,CNKI,DWPI,SIPOABS,IEEE:神经网络，层间，剖析，输入，实例，验证，子模型，模拟，预测，总行为，置信度，训练，权重，阈值，中间层，neural network model, interlayer, profiling, input, example, verify, hierarchy, submodel, simulate, predict, profile, confidence, train, weight, threshold, intermediate layer

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 109344806 A (第四范式北京技术有限公司) 2019年 2月 15日 (2019-02-15) 参见全文	1-9
A	CN 105184678 A (齐齐哈尔大学) 2015年 12月 23日 (2015-12-23) 参见全文	1-9
A	US 2019156933 A1 (VERISIM LIFE INC) 2019年 5月 23日 (2019-05-23) 参见全文	1-9
A	WO 2019098418 A1 (SAMSUNG ELECTRONICS CO LTD等) 2019年 5月 23日 (2019-05-23) 参见全文	1-9

☐ 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 5月 11日	2020年 5月 20日

ISA/CN的名称和邮寄地址	受权官员
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	李劲娴
传真号 (86-10)62019451	电话号码 62411739

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/111612

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 109344806 A	2019年 2月 15日	CN 109344806 B	2019年 8月 23日
CN 105184678 A	2015年 12月 23日	无	
US 2019156933 A1	2019年 5月 23日	WO 2019104101 A1	2019年 5月 31日
WO 2019098418 A1	2019年 5月 23日	无	