(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
2022 年 4 月 21 日 (21.04.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/077907 A1

(51) 国际专利分类号:
G06N 3/04 (2006.01) *G06N 5/04* (2006.01)
G06N 3/08 (2006.01) *G06K 9/62* (2006.01)

(21) 国际申请号: PCT/CN2021/095175

(22) 国际申请日: 2021 年 5 月 21 日 (21.05.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202011094465.X 2020年10月14日 (14.10.2020) CN

(71) 申请人: 上海交通大学 (SHANGHAI JIAO TONG UNIVERSITY) [CN/CN]; 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。

(72) 发明人: 李照冬 (LI, Zhaodong); 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。 冷静文 (LENG, Jingwen); 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。 邱宇贤 (QIU, Yuxian); 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。 过敏意 (GUO, Minyi); 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。 朱禹皓 (ZHU, Yuhao); 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。 陈全 (CHEN, Quan); 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。 李超 (LI, Chao); 中国上海市闵行区东川路800号, Shanghai 200240 (CN)。

(74) 代理人: 上海光华专利事务所 (普通合伙) (J.Z.M.C. PATENT AND TRADEMARK LAW

WO 2022/077907 A1

(54) **Title:** ADVERSARIAL ATTACK DETECTION METHOD, SYSTEM AND DEVICE, AND COMPUTER-READABLE STORAGE MEDIUM

(54) 发明名称: 对抗攻击的检测方法、系统、设备、计算机可读存储介质

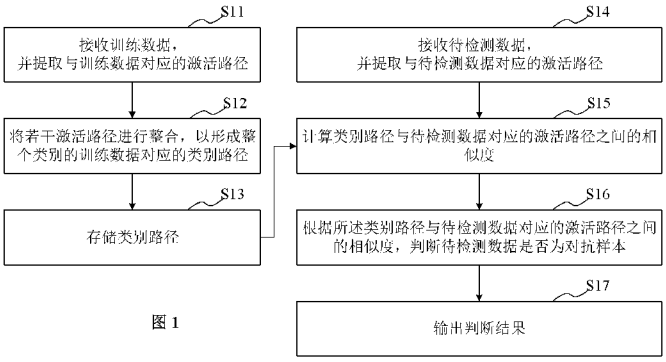


图 1

S11 Receive training data, and extract activation paths corresponding to the training data
S12 Integrate several activation paths to form a category path corresponding to the training data of the entire category
S13 Store the category path
S14 Receive data to be detected, and extract an activation path corresponding to said data
S15 Calculate the similarity between the category path and the activation path corresponding to said data
S16 Determine, according to the similarity between the category path and the activation path corresponding to said data, whether said data is an adversarial sample
S17 Output the determination result

(57) **Abstract:** An adversarial attack detection method, system and device, and a computer-readable storage medium. The adversarial attack detection method comprises: receiving training data, and extracting activation paths corresponding to the training data (S11); integrating several activation paths to form a category path corresponding to the training data of the entire category (S12); receiving data to be detected, and extracting an activation path corresponding to said data (S14); calculating the similarity between the category path and the activation path corresponding to said data (S15); and determining, according to the similarity between the category path

[见续页]

OFFICE)；中国上海市杨浦区国定路335号5022室余明伟, Shanghai 200433 (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))。

and the activation path corresponding to said data, whether said data is an adversarial sample (S16). Online adversarial attack detection can be realized in a reasoning process of a neural network, and thus the abnormality of a neural network model is detected; moreover, the detection method can achieve higher detection accuracy and lower overhead, and provides support for the robustness of a deep learning system.

(57) 摘要：一种对抗攻击的检测方法、系统、设备、计算机可读存储介质，所述对抗攻击的检测方法包括：接收训练数据，并提取与训练数据对应的激活路径（S11）；将若干激活路径进行整合，以形成整个类别的训练数据对应的类别路径（S12）；接收待检测数据，并提取与所述待检测数据对应的激活路径（S14）；计算所述类别路径与所述待检测数据对应的激活路径之间的相似度（S15）；根据所述类别路径与所述待检测数据对应的激活路径之间的相似度，判断所述待检测数据是否为对抗样本（S16）。能够在神经网络的推理过程中实现在线对抗攻击的检测，从而检测出神经网络模型的异常；且检测方法可以达到较高的检测准确率以及较低的开销，为深度学习系统的鲁棒性提供了支持。

对抗攻击的检测方法、系统、设备、计算机可读存储介质

技术领域

本发明属于神经网络技术领域，涉及一种检测方法、系统、设备，特别是涉及一种对抗攻击的检测方法、系统、设备、计算机可读存储介质。

背景技术

近年来，以深度学习为代表的新一代人工智能算法，在图像处理、语音识别、机器翻译等诸多领域取得一系列重要突破，极大地提高了当前人工智能算法的性能。然而，这类模型通常鲁棒性不足，很容易受到对抗攻击的影响，一个精心设计的对抗样本可能会使原本正常的模型产生错误的结果；对抗攻击对深度学习系统的可靠性、隐私性产生了较大的危害，如何提高模型的鲁棒性，是当前以及未来人工智能领域的研究热点和难点。

以下简单介绍一下现有技术：

现有技术一：对抗训练（adversarial training）是增强神经网络鲁棒性的重要方式。在对抗训练的过程中，样本会被混合一些微小的扰动，通过训练使得神经网络适应这种扰动，从而增强了神经网络对于对抗样本的鲁棒性。

现有技术一的缺陷：现这种通过训练来防御对抗攻击的方式不适合在推理过程中在线检测对抗攻击，并且构建训练数据和进行训练也需要较大的工作量。

现有技术二：通过构建多模块冗余系统。

现有技术二的缺陷：开销过大，使得无法实现在推理过程中在线检测对抗攻击。

所以，当前已有的对抗攻击检测方法存在诸多问题：一方面，这些检测方法无法实现在神经网络推理过程中在线进行对抗攻击检测；另一方面，这些检测方法也会带来很大的额外开销。对抗攻击会严重危害神经网络系统的可靠性、隐私性。在诸如自动驾驶等任务关键型场景下，实现在线的对抗攻击检测尤为重要。

因此，如何提供一种对抗攻击的检测方法、系统、设备、计算机可读存储介质，以解决现有技术无法实现在神经网络推理过程中在线进行对抗攻击检测，且会带来额外开销，导致神经网络系统的可靠性、隐私性受到严重危害等缺陷，实已成为本领域技术人员亟待解决的技术问题。

发明内容

鉴于以上所述现有技术的缺点，本发明的目的在于提供一种对抗攻击的检测方法、系统、

设备、计算机可读存储介质，用于解决现有技术无法实现在神经网络推理过程中在线进行对抗攻击检测，且会带来额外开销，导致神经网络系统的可靠性、隐私性受到严重危害的问题。

为实现上述目的及其他相关目的，本发明一方面提供一种对抗攻击的检测方法，应用于神经网络；所述对抗攻击的检测方法包括： 接收训练数据，并提取与训练数据对应的激活路径；将若干激活路径进行整合，以形成整个类别的训练数据对应的类别路径；接收待检测数据，并提取与所述待检测数据对应的激活路径；计算所述类别路径与所述待检测数据对应的激活路径之间的相似度；根据所述相似度，判断所述待检测数据是否为对抗样本。

于本发明的一实施例中，所述提取与训练数据对应的激活路径的步骤包括：根据与训练数据对应的神经网络中最后一层的重要神经元，逐层反向提取每一层的重要神经元，以获取所述神经网络中针对一类别标签的所有重要神经元；通过每一层中确定的重要神经元形成若干所述激活路径。

于本发明的一实施例中，提取每一层的重要神经元的方式包括前向提取、阈值机制及选择性提取中的一种或任意组合方式。

于本发明的一实施例中，所述将若干激活路径进行整合，以形成整个类别的样本对应的类别路径的步骤包括：将属于同一类别的训练数据所对应的激活路径通过按位取或操作计算出整个类别对应的类别路径。

于本发明的一实施例中，所述提取与所述待检测数据对应的激活路径的步骤包括：根据与待检测数据对应的神经网络中最后一层的重要神经元，逐层反向提取每一层的重要神经元，以获取所述神经网络中针对一类别标签的所有重要神经元；通过每一层中确定的重要神经元形成与所述待检测数据对应的激活路径。

于本发明的一实施例中，计算所述类别路径与所述待检测数据对应的激活路径之间的相似度的计算公式如下：$S=\|P(x)\,\&\,Pc\|_1 / \|P(x)\|_1$ 其中，S 表示类别路径与待检测数据对应的激活路径；$P(x)$表示与待检测数据对应的激活路径；$\|P(x)\|_1$表示与待检测数据对应的激活路径中 1 的个数；&表示按位取或操作；Pc表示整个类别的训练数据对应的类别路径，c 为类别。

于本发明的一实施例中，所述根据所述相似度，判断所述待检测数据是否为对抗样本的步骤包括：通过预存分类器将所述类别路径与所述待检测数据对应的激活路径之间的相似度进行分类，以判断所述待检测数据是否为对抗样本。

本发明另一方面提供一种对抗攻击的检测系统，应用于神经网络；所述对抗攻击的检测系统包括：数据接收模块，用于接收训练数据和待检测数据；第一提取模块，用于提取与训练数据对应的激活路径；整合模块，用于将若干激活路径进行整合，以形成整个类别的训练

数据对应的类别路径；第二提取模块，用于提取与所述待检测数据对应的激活路径；处理模块，用于计算所述类别路径与所述待检测数据对应的激活路径之间的相似度；根据所述相似度，判断所述待检测数据是否为对抗样本。

本发明又一方面提供一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现所述对抗攻击的检测方法。

本发明最后一方面提供一种检测设备，包括：处理器及存储器；所述存储器用于存储计算机程序，所述处理器用于执行所述存储器存储的计算机程序，以使所述检测设备执行所述对抗攻击的检测方法。

于本发明的一实施例中，所述处理器包括神经网络加速器、路径构造器及控制器。

如上所述，本发明所述的对抗攻击的检测方法、系统、设备、计算机可读存储介质，具有以下有益效果：

本发明所述对抗攻击的检测方法、系统、设备、计算机可读存储介质能够在神经网络的推理过程中实现在线对抗攻击的检测，从而检测出神经网络模型的异常；且本实施例所述检测方法可以达到较高的检测准确率以及较低的开销，为深度学习系统的鲁棒性提供了支持。

附图说明

图 1 显示为本发明的对抗攻击的检测方法于一实施例中的流程示意图。

图 2 显示为本发明的神经网络的全连接层上重要神经元的计算示意图。

图 3 显示为本发明的神经网络的卷积层上重要神经元的计算示意图。

图 4 显示为本发明的神经网络的激活路径结构示意图。

图 5 显示为本发明的对抗攻击的检测系统于一实施例中的原理结构示意图。

图 6 显示为本发明的检测设备于一实施例中的原理结构示意图。

元件标号说明

5	对抗攻击的检测系统
51	数据接收模块
52	第一提取模块
53	整合模块
54	存储模块
55	第二提取模块

56	处理模块
6	检测设备
61	存储器
62	处理器
621	神经网络加速器
622	路径构造器
622	控制器
S11~S17	步骤

具体实施方式

以下通过特定的具体实例说明本发明的实施方式，本领域技术人员可由本说明书所揭露的内容轻易地了解本发明的其他优点与功效。本发明还可以通过另外不同的具体实施方式加以实施或应用，本说明书中的各项细节也可以基于不同观点与应用，在没有背离本发明的精神下进行各种修饰或改变。需说明的是，在不冲突的情况下，以下实施例及实施例中的特征可以相互组合。

需要说明的是，以下实施例中所提供的图示仅以示意方式说明本发明的基本构想，遂图式中仅显示与本发明中有关的组件而非按照实际实施时的组件数目、形状及尺寸绘制，其实际实施时各组件的型态、数量及比例可为一种随意的改变，且其组件布局型态也可能更为复杂。

本发明所述对抗攻击的检测方法、系统、设备、计算机可读存储介质的基本原理如下：

正常图片在经过扰动处理后会产生对抗样本。从肉眼观察上两者似乎没有区别，但深度神经网络却无法识别扰动处理后的图片。

在神经网络的推理过程中，神经网络的每个输入都会激活神经网络内部的若干神经元（这些被激活的神经元是指能对推理结果产生较大影响的神经元，后文中也将其称为“重要神经元”）；理论研究表明，具有同一类别标签的输入所激活的神经元往往比较类似，而属于不同类别的输入所激活的神经元往往有较大的差异。换言之，在神经网络的推理过程中，正常的输入所激活的神经元和对抗样本所激活的神经元之间会存在较大的差异。因此我们可以通过将一个输入所激活的神经元和正常样本所激活的神经元进行对比，来确定该输入是正常输入还是对抗样本，从而实现了对抗攻击检测。

基于上述思想，本发明利用重要神经元的概念及其计算方法。重要神经元是指针对某个

输入，在推理过程中对推理结果会产生较大影响的神经元，即上文中提到的被输入激活的神经元；通过比较在推理过程中不同输入所对应的重要神经元，能够实现在推理过程中的对抗攻击检测。

实施例一

本实施例提供一种对抗攻击的检测方法，应用于神经网络；所述对抗攻击的检测方法包括：

接收训练数据，并提取与训练数据对应的激活路径；

将若干激活路径进行整合，以形成整个类别的训练数据对应的类别路径；

接收待检测数据，并提取与所述待检测数据对应的激活路径；

计算所述类别路径与所述待检测数据对应的激活路径之间的相似度；

根据所述相似度，判断所述待检测数据是否为对抗样本。

以下将结合图示对本实施例所提供的对抗攻击的检测方法进行详细说明。请参阅图 1，显示为对抗攻击的检测方法于一实施例中的流程示意图。如图 1 所示，所述对抗攻击的检测方法具体包括以下步骤：

S11，接收训练数据，并提取与训练数据对应的激活路径。

在本实施例中，所述 S11 具体包括：

根据与训练数据对应的神经网络中最后一层的重要神经元，逐层反向提取每一层的重要神经元，以获取所述神经网络中针对一类别标签的所有重要神经元。

通过每一层中确定的重要神经元形成若干所述激活路径。

具体地，以图 2 所示，说明神经网络的全连接层上重要神经元的计算方法：

假设后一层中的重要神经元为第二个神经元（值为 0.46），根据全连接层的定义，0.46 = 0.1 x 2.1 + 1.0 x 0.09 + 0.4 x 0.2 + 0.3 x 0.2 + 0.2 x 0.1（每项按值的降序排列）；确定一个比例系数 θ，θ 控制着前一层重要神经元在后一层重要神经元中所占的比例，进而控制了前一层重要神经元的个数；例如当 θ=0.6 时，有 0.1 x 2.1 + 1.0 x 0.09 > 0.6 x 0.46 成立，这表明最少仅需要前一层中第四和第五两个神经元（值分别为 1.0、0.1）即可超过后一层的重要神经元的值的 60%，因此可认为前一层中的第四个神经元和第五个神经元对后一层的影响最大，这两个神经元即为前一层中的重要神经元；若前一层还有其他的重要神经元，按同样的方式计算。初始时，我们人为设定神经网络最后一层中值最大的神经元（即对应输出类别的神经元）为最后一层唯一的重要神经元。

具体地，以图 3 所示，说明神经网络的卷积层上重要神经元的计算方法：

在卷积层中，首先根据后一层的重要神经元确定其在前一层的接受域，例如 5.47 = 2.0 x 0.7 + 1.4 x 0.9 + 1.5 x 0.8 + 1.0 x 0.9 + 1.1 x 0.2 +1.0 x 0.2 + 1.9 x 0.1 + 0.2 x 0.3 + 0.2 x 0.2（每项按值的降序排列）；然后根据 θ 的值找到前一层的重要神经元，例如当 θ=0.6 时，有 2.0 x 0.7 + 1.4 x 0.9 + 1.5 x 0.8 > 0.6 x 5.47 成立，这表明最少仅需要接受域中值为 2.0、1.5、1.4 这三个神经元即可超过后一层的重要神经元的值的 60%，这三个神经元即为前一层的重要神经元，若前一层还有其他的重要神经元，按同样的方式计算。

本实施例从对抗攻击检测准确率以及开销两方面出发，权衡两者的影响，对激活路径的提取方法。

在本实施例中，提取每一层的重要神经元的方式包括前向提取、阈值机制及选择性提取中的一种或任意组合方式。

前向提取重要神经元：

在全连接层上重要神经元的计算方法和卷积层上重要神经元的计算方法中，重要神经元的提取是反向进行的，该方法使得推理和提取两个过程之间是串行关系，虽然重要神经元的计算非常准确，却带来了较大的时间开销。因此本发明提出了前向提取法：当推理进行到第 Li+1 层时，可直接从第 Li+1 层中挑选值最大的神经元作为该层的重要神经元，然后再用它来计算第 Li 层的重要神经元。这样第 Li+1 层的推理和第 Li 层的提取可以同时进行，这虽然会牺牲一部分对抗检测的准确率，但大大减少了整个激活路径的提取所需的时间开销。

阈值机制

在全连接层上重要神经元的计算方法和卷积层上重要神经元的计算方法中，重要神经元的提取是通过将和式中的每一项进行排序，再从中依次挑选重要神经元，以使得前一层的部分和在后一层重要神经元中所占比例不低于 θ，这种方法有着较大的计算开销以及存储开销。因此本发明提出了阈值机制：在神经网络中的第 Li 层可确定一个阈值 φi，该层中的所有值大于阈值 φi 的神经元都被判定为重要神经元，这大大简化了提取重要神经元的过程，减少了计算和存储开销。

选择性提取：

在全连接层上重要神经元的计算方法和卷积层上重要神经元的计算方法中，重要神经元的提取是将整个网络的所有层的重要神经元都进行提取，然而理论研究表明[1]，在许多神经网络中，靠后的层往往比前边的层对推理结果有着更大的影响。因此，本申请提出仅对神经网络后面的某些层提取重要神经元，前面层的重要神经元可以直接忽略。选择性提取策略进一步降低了提取重要神经元的开销。

本实施例采用应用程序接口（API），使得用户可通过调用 API 来实现上述不同优化方案的组合，以适应不同应用场景下不同的准确率要求和开销要求。例如，可将前向提取、采用阈值机制以及仅提取后三层重要神经元这三者进行组合来提取激活路径。不同优化方案的组合使得本发明的灵活性大大增强。同时本实施例针对对抗攻击检测过程的计算特性，设计了一套指令集并进行了相应的编译器优化，以进一步提升系统的性能。

根据上述重要神经元计算方法，逐层反向计算每一层的重要神经元，最终得到整个神经网络针对某个输入数据的所有重要神经元,这些重要神经元对神经网络的输入有着重要的影响，它们构成了整个神经网络的激活路径，如图 4 所示。

具体地，本实施例采用比特掩码（bitmask）表示激活路径，对于输入的训练数据 x，设其在推理过程中的激活路径为 P(x)，则 P(x)中的 $m_{i,j}$ 位的含义为神经网络第 i 层的第 j 个神经元是否为重要神经元；具有同一类别标签的输入数据所对应的激活路径往往比较类似，而属于不同类别的输入所对应的激活路径往往存在较大的差异。因此在推导过程中，通过计算当前输入所对应的激活路径，再将该激活路径与类别路径相比较，就能在推理过程中检测对抗攻击；在本实施例中，类别路径是指属于同一类别的所有正常输入所对应的激活路径按位取或得到的路径，若激活路径和类别路径比较相似，则可认为输入是正常输入。

S12，将若干激活路径进行整合，以形成整个类别的训练数据对应的类别路径。

具体地，S12 包括将属于同一类别的训练数据所对应的激活路径通过按位或操作计算出整个类别对应的类别路径，即把用比特掩码表示的激活路径的每一位，依次进行或（OR）运算，最后得到一个用比特掩码表示的整个类别对应的类别路径；例如，两个激活路径的比特掩码分别为 0011 和 0101，则计算出的类别路径的比特掩码为 0111。

S13，存储所述类别路径，以便在推理时，直接读取已存储的类别路径。

S14，接收待检测数据，并提取与所述待检测数据对应的激活路径。

具体地，S14 包括：

根据与待检测数据对应的神经网络中最后一层的重要神经元，逐层反向提取每一层的重要神经元，以获取所述神经网络中针对一类别标签的所有重要神经元。

通过每一层中确定的重要神经元形成与所述待检测数据对应的激活路径。

在本实施例中，所述待检测数据对应的激活路径的提取方法与所述训练数据的激活路径的提取方法一样，此处不再赘述。

以上 S14 与 S11 可以同时执行，也可以分先后顺序执行。在本实施例中，所述 S14 与所述 S11 为同时执行。

S15，计算所述类别路径与所述待检测数据对应的激活路径之间的相似度。

于本实施例中，计算所述类别路径与所述待检测数据对应的激活路径之间的相似度的计算公式如下：

$$S = \left\| P(x) \,\&\, Pc \right\|_1 / \left\| P(x) \right\|_1$$

其中，S 表示类别路径与待检测数据对应的激活路径；$P(x)$表示与待检测数据对应的激活路径；$\left\| P(x) \right\|_1$表示与待检测数据对应的激活路径中年 1 的个数；&表示按位取或操作；Pc表示整个类别的训练数据对应的类别路径，c 为类别。

S16，根据所述类别路径与所述待检测数据对应的激活路径之间的相似度，判断所述待检测数据是否为对抗样本。

具体地，通过预存分类器将所述类别路径与所述待检测数据对应的激活路径之间的相似度进行分类，以判断所述待检测数据是否为对抗样本。

于本实施例中，所述分类器采用机器学习领域中的随机森林方法。

S17，输出所述 S16 的判断结果，即所述待检测数据为对抗样本或为非对抗样本。

本实施例所述对抗攻击的检测方法能够在神经网络的推理过程中实现在线对抗攻击的检测，从而检测出神经网络模型的异常；且本实施例所述检测方法可以达到较高的检测准确率以及较低的开销，为深度学习系统的鲁棒性提供了支持。

本实施例还提供一种计算机可读存储介质，其上存储有计算机程序，该程序被处理器执行时实现所述对抗攻击的检测方法。

本领域普通技术人员可以理解计算机可读存储介质为：实现上述各方法实施例的全部或部分步骤可以通过计算机程序相关的硬件来完成。前述的计算机程序可以存储于一计算机可读存储介质中。该程序在执行时，执行包括上述各方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

实施例二

本实施例提供一种对抗攻击的检测系统，应用于神经网络；所述对抗攻击的检测系统包括：

数据接收模块，用于接收训练数据和待检测数据；

第一提取模块，用于提取与训练数据对应的激活路径；

整合模块，用于将若干激活路径进行整合，以形成整个类别的训练数据对应的类别路径；

第二提取模块，用于提取与所述待检测数据对应的激活路径；

处理模块，用于计算所述类别路径与所述待检测数据对应的激活路径之间的相似度；根

据所述相似度，判断所述待检测数据是否为对抗样本。

以下将结合图示对本实施例所提供的对抗攻击的检测系统进行详细描述。请参阅图 5，显示为对抗攻击的检测系统于一实施例中的原理结构示意图。如图 5 所示，所述对抗攻击的检测系统 5 包括数据接收模块 51、第一提取模块 52、整合模块 53、存储模块 54、第二提取模块 55 及处理模块 56。

所述数据接收模块 51 用于接收训练数据和待检测数据。

与所述数据接收模块 51 连接的所述第一提取模块 52 用于提取与训练数据对应的激活路径。

在本实施例中，所述第一提取模块 52 根据与训练数据对应的神经网络中最后一层的重要神经元，逐层反向提取每一层的重要神经元，以获取所述神经网络中针对一类别标签的所有重要神经元。通过每一层中确定的重要神经元形成若干所述激活路径。

在本实施例中，所述第一提取模块 52 提取每一层的重要神经元的方式包括前向提取、阈值机制及选择性提取中的一种或任意组合方式。具体提取过程与上述方法中的提取方案相同，此处不再赘述。

与所述第一提取模块 52 连接的所述整合模块 53 用于将属于同一类别的训练数据所对应的激活路径通过按位取或操作计算出整个类别对应的类别路径。

与所述整合模块 53 连接的存储模块 54 用于存储所述类别路径，以便所述处理模块 56 在推理时，直接读取已存储的类别路径。

与所述数据接收模块 51 连接的第二提取模块 55 用于根据与待检测数据对应的神经网络中最后一层的重要神经元，逐层反向提取每一层的重要神经元，以获取所述神经网络中针对一类别标签的所有重要神经元。

所述第二提取模块 55 提取与所述待检测数据对应的激活路径的过程与第一提取模块 52 提取的过程相同，此处不再赘述。

与所述存储模块 54 和所述第二提取模块 55 连接的处理模块 56 用于计算所述类别路径与所述待检测数据对应的激活路径之间的相似度；根据所述相似度，判断所述待检测数据是否为对抗样本，并输出判断结果。所述判断结果包括所述待检测数据为对抗样本或为非对抗样本。

具体地，所述处理模块 56 通过公式 $S=\|P(x)\&Pc\|_1/\|P(x)\|_1$ 计算所述类别路径与所述待检测数据对应的激活路径之间的相似度。其中，S 表示类别路径与待检测数据对应的激活路径；$P(x)$ 表示与待检测数据对应的激活路径；$\|P(x)\|_1$ 表示与待检测数据对应的激活路径中年

1 的个数；&表示按位取或操作；*Pc* 表示整个类别的训练数据对应的类别路径，c 为类别。

需要说明的是，应理解以上系统的各个模块的划分仅仅是一种逻辑功能的划分，实际实现时可以全部或部分集成到一个物理实体上，也可以物理上分开。且这些模块可以全部以软件通过处理元件调用的形式实现，也可以全部以硬件的形式实现，还可以部分模块通过处理元件调用软件的形式实现，部分模块通过硬件的形式实现。例如：x 模块可以为单独设立的处理元件，也可以集成在上述系统的某一个芯片中实现。此外，x 模块也可以以程序代码的形式存储于上述系统的存储器中，由上述系统的某一个处理元件调用并执行以上 x 模块的功能。其它模块的实现与之类似。这些模块全部或部分可以集成在一起，也可以独立实现。这里所述的处理元件可以是一种集成电路，具有信号的处理能力。在实现过程中，上述方法的各步骤或以上各个模块可以通过处理器元件中的硬件的集成逻辑电路或者软件形式的指令完成。以上这些模块可以是被配置成实施以上方法的一个或多个集成电路，例如：一个或多个特定集成电路（Application Specific Integrated Circuit，简称 ASIC），一个或多个微处理器（Digital Singnal Processor，简称 DSP），一个或者多个现场可编程门阵列（Field Programmable Gate Array，简称 FPGA）等。当以上某个模块通过处理元件调度程序代码的形式实现时，该处理元件可以是通用处理器，如中央处理器（Central Processing Unit，简称 CPU）或其它可以调用程序代码的处理器。这些模块可以集成在一起，以片上系统（System-on-a-chip，简称 SOC）的形式实现。

实施例三

本实施例提供一种检测设备，请参阅图 6，显示为检测设备于一实施例中的原理结构示意图。如图 6 所示，所述检测设备 6 包括存储器 61 和处理器 62。

在本实施例中，所述存储器 61 用于存储计算机程序，所述处理器 62 用于执行所述存储器 61 存储的计算机程序，以使所述检测设备执行实施例一所述对抗攻击的检测方法。

在本实施例中，所述存储器 61 包括 SRAM 和 DRAM。

在本实施例中，所述处理器 62 包括神经网络加速器 621、路径构造器 622 及控制器 623。

上述的处理器可以是通用处理器，包括中央处理器（Central Processing Unit，简称 CPU）、网络处理器（Network Processor，简称 NP）等；还可以是数字信号处理器（Digital Signal Processing，简称 DSP）、专用集成电路（Application Specific Integrated Circuit，简称 ASIC）、现场可编程门阵列（Field Programmable Gate Array，简称 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。

于本实施例中，所述神经网络加速器 621 用于推理计算以及推理时部分和的重计算。

所述路径构造器 622 用于提取神经网络推理时的重要神经元，构造激活路径，以及计算路径之间的相似度。所述控制器 623 用于指令的调度以及运行随机森林分类器。

其中，神经网络加速器 621、路径构造器 622 和控制器 623 的 SRAM 均采用双缓冲，通过总线和 DRAM 进行数据传输。

本发明所述的对抗攻击的检测方法的保护范围不限于本实施例列举的步骤执行顺序，凡是根据本发明的原理所做的现有技术的步骤增减、步骤替换所实现的方案都包括在本发明的保护范围内。

本发明还提供一种对抗攻击的检测系统，所述对抗攻击的检测系统可以实现本发明所述的对抗攻击的检测方法，但本发明所述的对抗攻击的检测方法的实现装置包括但不限于本实施例列举的对抗攻击的检测系统的结构，凡是根据本发明的原理所做的现有技术的结构变形和替换，都包括在本发明的保护范围内。

综上所述，本发明所述对抗攻击的检测方法、系统、设备、计算机可读存储介质能够在神经网络的推理过程中实现在线对抗攻击的检测，从而检测出神经网络模型的异常；且本实施例所述检测方法可以达到较高的检测准确率以及较低的开销，为深度学习系统的鲁棒性提供了支持。本发明有效克服了现有技术中的种种缺点而具高度产业利用价值。

上述实施例仅例示性说明本发明的原理及其功效，而非用于限制本发明。任何熟悉此技术的人士皆可在不违背本发明的精神及范畴下，对上述实施例进行修饰或改变。因此，举凡所属技术领域中具有通常知识者在未脱离本发明所揭示的精神与技术思想下所完成的一切等效修饰或改变，仍应由本发明的权利要求所涵盖。

权 利 要 求 书

1. 一种对抗攻击的检测方法，其特征在于，应用于神经网络；所述对抗攻击的检测方法包括：

接收训练数据，并提取与训练数据对应的激活路径；

将若干激活路径进行整合，以形成整个类别的训练数据对应的类别路径；

接收待检测数据，并提取与所述待检测数据对应的激活路径；

计算所述类别路径与所述待检测数据对应的激活路径之间的相似度；

根据所述相似度，判断所述待检测数据是否为对抗样本。

2. 根据权利要求 1 所述的对抗攻击的检测方法，其特征在于，所述提取与训练数据对应的激活路径的步骤包括：

根据与训练数据对应的神经网络中最后一层的重要神经元，逐层反向提取每一层的重要神经元，以获取所述神经网络中针对一类别标签的所有重要神经元；

通过每一层中确定的重要神经元形成若干所述激活路径。

3. 根据权利要求 2 所述的对抗攻击的检测方法，其特征在于，提取每一层的重要神经元的方式包括前向提取、阈值机制及选择性提取中的一种或任意组合方式。

4. 根据权利要求 2 所述的对抗攻击的检测方法，其特征在于，所述将若干激活路径进行整合，以形成整个类别的样本对应的类别路径的步骤包括：

将属于同一类别的训练数据所对应的激活路径通过按位取或操作计算出整个类别对应的类别路径。

5. 根据权利要求 2 所述的对抗攻击的检测方法，其特征在于，所述提取与所述待检测数据对应的激活路径的步骤包括：

根据与待检测数据对应的神经网络中最后一层的重要神经元，逐层反向提取每一层的重要神经元，以获取所述神经网络中针对一类别标签的所有重要神经元；

通过每一层中确定的重要神经元形成与所述待检测数据对应的激活路径。

6. 根据权利要求 5 所述的对抗攻击的检测方法，其特征在于，计算所述类别路径与所述待检测数据对应的激活路径之间的相似度的计算公式如下：

$$S = \left\| P(x) \,\&\, Pc \right\|_1 / \left\| P(x) \right\|_1$$

其中，S 表示类别路径与待检测数据对应的激活路径；$P(x)$ 表示与待检测数据对应的激

活路径；$\|P(x)\|_1$表示与待检测数据对应的激活路径中 1 的个数；&表示按位取或操作；Pc表示整个类别的训练数据对应的类别路径，c 为类别。

7. 根据权利要求 1 或 6 所述的对抗攻击的检测方法，其特征在于，所述根据所述相似度，判断所述待检测数据是否为对抗样本的步骤包括：

通过预存分类器将所述类别路径与所述待检测数据对应的激活路径之间的相似度进行分类，以判断所述待检测数据是否为对抗样本。

8. 一种对抗攻击的检测系统，其特征在于，应用于神经网络；所述对抗攻击的检测系统包括：

数据接收模块，用于接收训练数据和待检测数据；

第一提取模块，用于提取与训练数据对应的激活路径；

整合模块，用于将若干激活路径进行整合，以形成整个类别的训练数据对应的类别路径；

第二提取模块，用于提取与所述待检测数据对应的激活路径；

处理模块，用于计算所述类别路径与所述待检测数据对应的激活路径之间的相似度；根据所述相似度，判断所述待检测数据是否为对抗样本。

9. 一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现权利要求 1 至 7 中任一项所述对抗攻击的检测方法。

10. 一种检测设备，其特征在于，包括：处理器及存储器；

所述存储器用于存储计算机程序，所述处理器用于执行所述存储器存储的计算机程序，以使所述检测设备执行如权利要求 1 至 7 中任一项所述对抗攻击的检测方法。

11. 根据权利要求 11 所述的检测设备，其特征在于，所述处理器包括神经网络加速器、路径构造器及控制器。

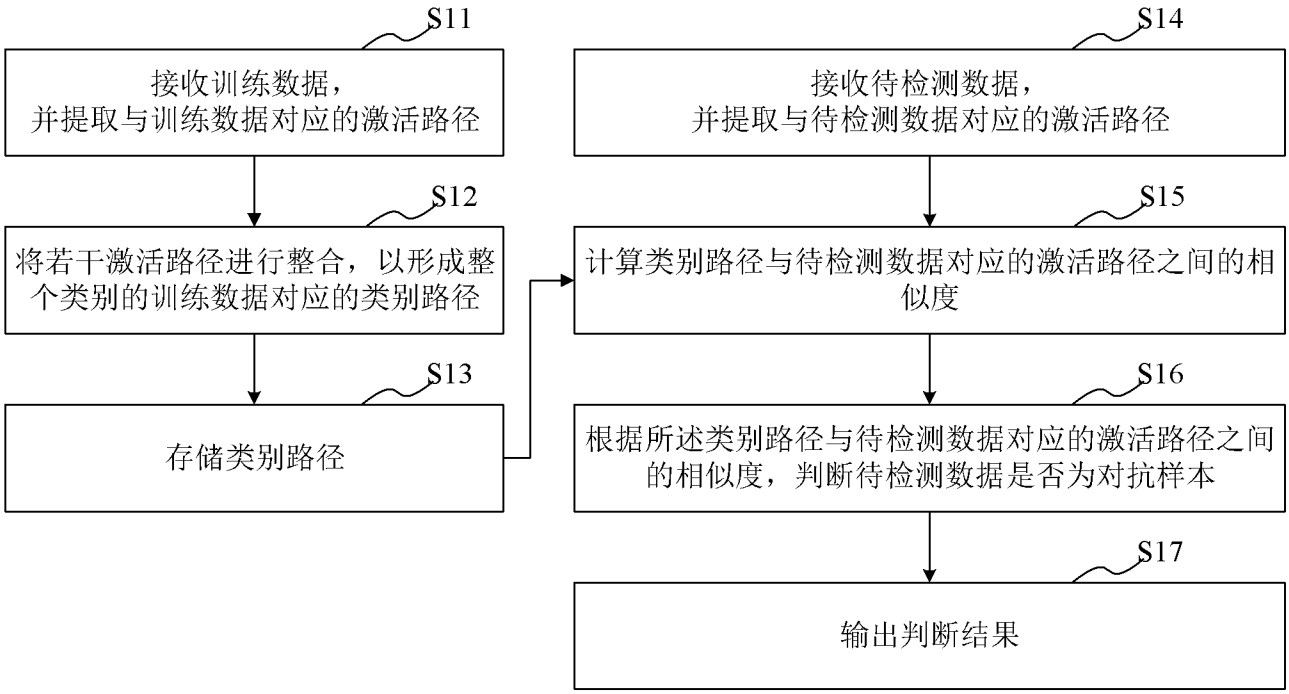

S11
接收训练数据，
并提取与训练数据对应的激活路径
S12
将若干激活路径进行整合，以形成整个类别的训练数据对应的类别路径
S13
存储类别路径
S14
接收待检测数据，
并提取与待检测数据对应的激活路径
S15
计算类别路径与待检测数据对应的激活路径之间的相似度
S16
根据所述类别路径与待检测数据对应的激活路径之间的相似度，判断待检测数据是否为对抗样本
S17
输出判断结果

图 1

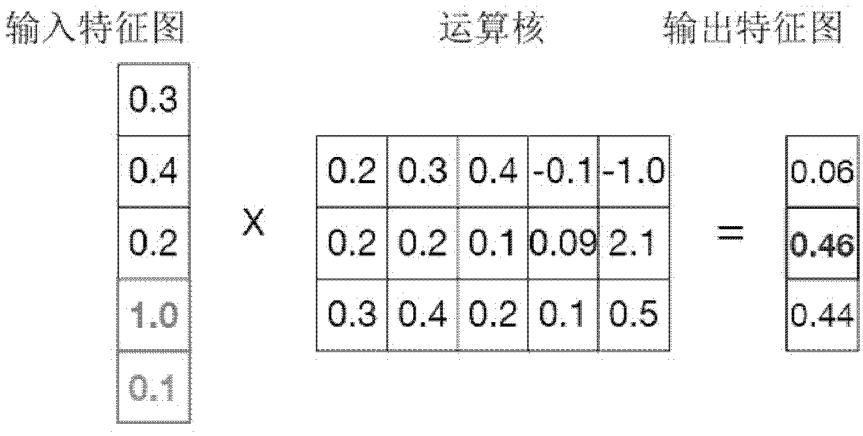

输入特征图
运算核
输出特征图
0.3
0.4
0.2
1.0
0.1
X
0.2 0.3 0.4 -0.1 -1.0
0.2 0.2 0.1 0.09 2.1
0.3 0.4 0.2 0.1 0.5
=
0.06
0.46
0.44

图 2

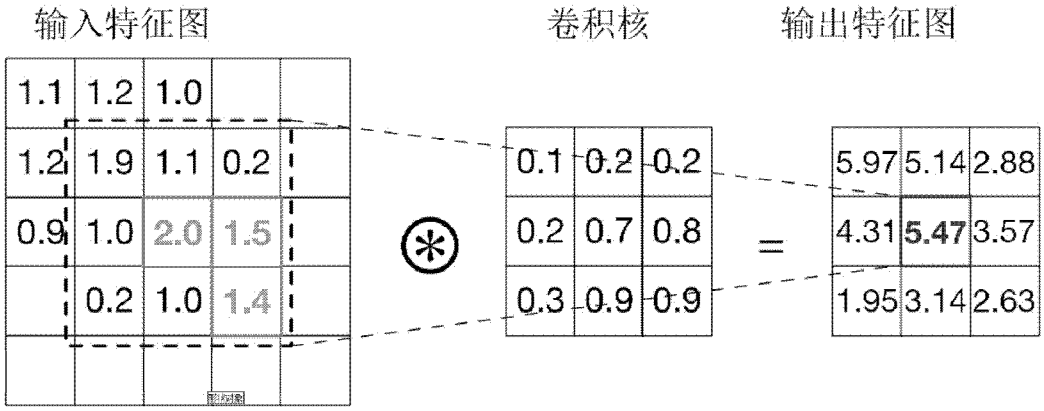

输入特征图
卷积核
输出特征图
1.1 1.2 1.0
1.2 1.9 1.1 0.2
0.9 1.0 2.0 1.5
0.2 1.0 1.4
0.1 0.2 0.2
0.2 0.7 0.8
0.3 0.9 0.9
=
5.97 5.14 2.88
4.31 5.47 3.57
1.95 3.14 2.63

图 3

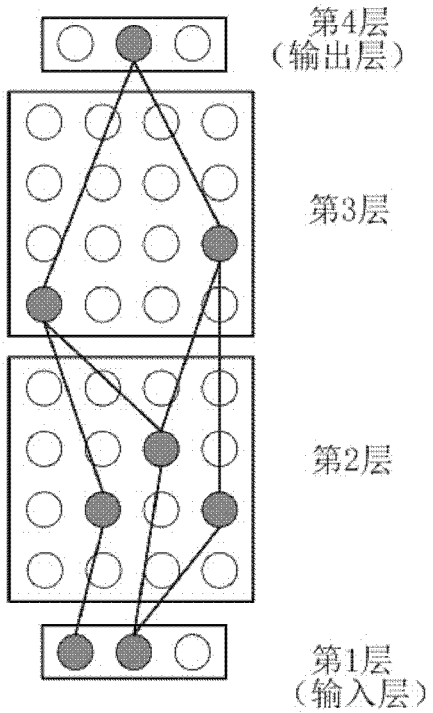

第4层
（输出层）
第3层
第2层
第1层
（输入层）

图 4

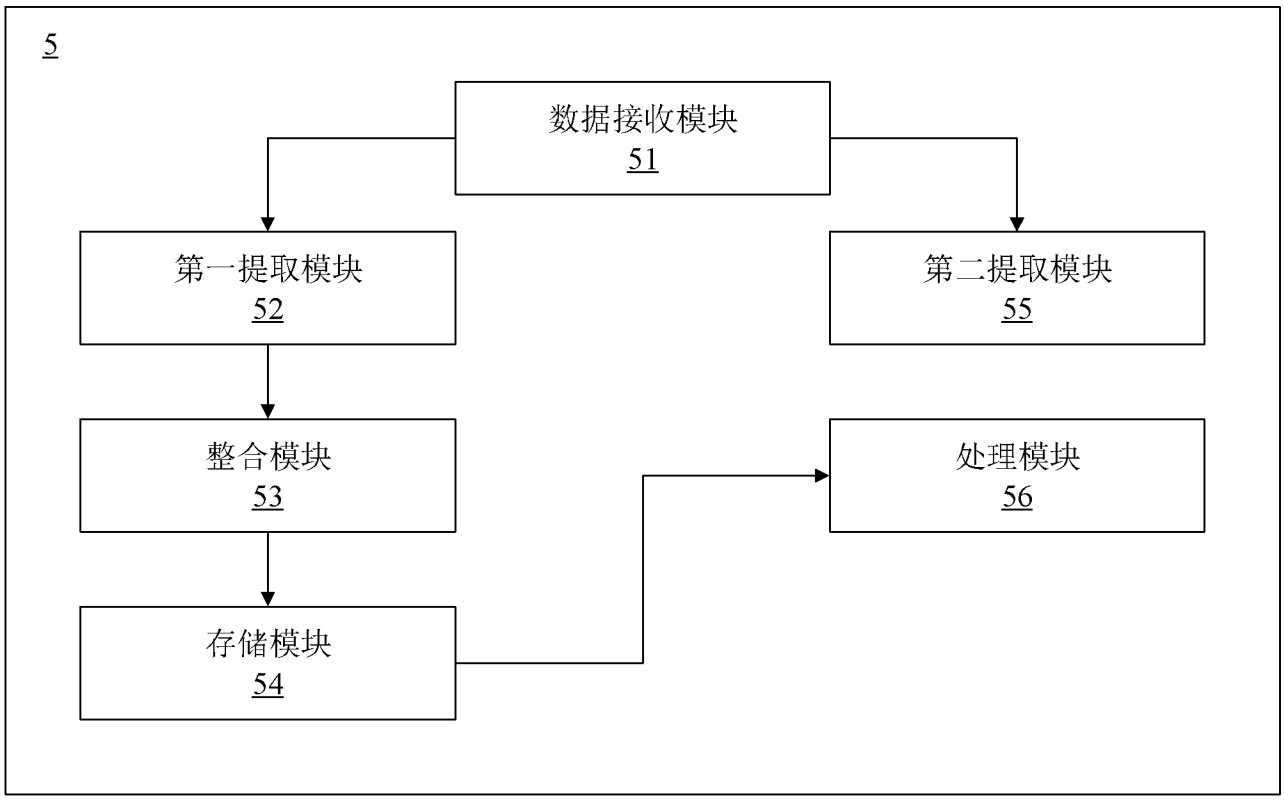

5
数据接收模块
51
第一提取模块
52
第二提取模块
55
整合模块
53
处理模块
56
存储模块
54

图 5

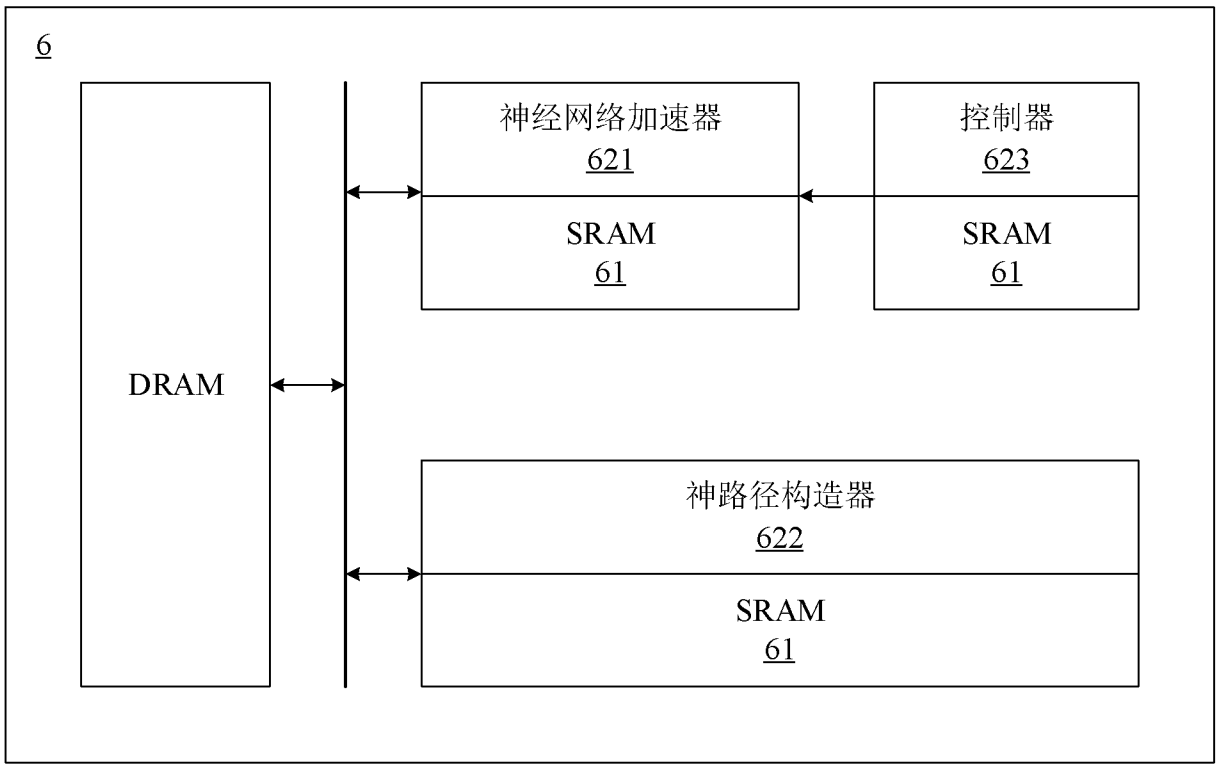

6
DRAM
神经网络加速器
621
SRAM
61
控制器
623
SRAM
61
神路径构造器
622
SRAM
61

图 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/095175

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/04(2006.01)i; G06N 3/08(2006.01)i; G06N 5/04(2006.01)i; G06K 9/62(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; G06K

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)

CNKI, CNPAT, WPI, EPODOC, IEEE: 对抗, 攻击, 网络, 路径, 激活, 类别, 分类, 种类, 样本, 训练, 相似, counter+, attack, network, path, activation, category, type, classification, sample, training, similarity

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112163670 A (SHANGHAI JIAO TONG UNIVERSITY) 01 January 2021 (2021-01-01) claims 1-11	1-11
A	CN 111723927 A (INTEL CORPORATION) 29 September 2020 (2020-09-29) description, paragraphs 0021-0086, figures 1-17B	1-11
A	CN 111079930 A (SHENZHEN SENSETIME TECHNOLOGY CO., LTD.) 28 April 2020 (2020-04-28) entire document	1-11
A	CN 110046712 A (TIANJIN UNIVERSITY OF SCIENCE & TECHNOLOGY) 23 July 2019 (2019-07-23) entire document	1-11
A	CN 111651765 A (SHANGHAI JIAO TONG UNIVERSITY) 11 September 2020 (2020-09-11) entire document	1-11
A	US 2019220605 A1 (INTEL CORPORATION) 18 July 2019 (2019-07-18) entire document	1-11

☐ 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
09 August 2021	**20 August 2021**

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/CN2021/095175

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 112163670 A	01 January 2021	None	
CN 111723927 A	29 September 2020	None	
CN 111079930 A	28 April 2020	None	
CN 110046712 A	23 July 2019	None	
CN 111651765 A	11 September 2020	None	
US 2019220605 A1	18 July 2019	EP 3712822 A1	23 September 2020

国际检索报告

国际申请号
PCT/CN2021/095175

A. 主题的分类

G06N 3/04(2006.01)i; G06N 3/08(2006.01)i; G06N 5/04(2006.01)i; G06K 9/62(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; G06K

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））

CNKI, CNPAT, WPI, EPODOC, IEEE: 对抗，攻击，网络，路径，激活，类别，分类，种类，样本，训练，相似，counter+, attack, network, path, activation, category, type, classification, sample, training, similarity

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
PX	CN 112163670 A (上海交通大学) 2021年 1月 1日 (2021-01-01) 权利要求1-11	1-11
A	CN 111723927 A (英特尔公司) 2020年 9月 29日 (2020-09-29) 说明书第0021-0086段、附图1-17B	1-11
A	CN 111079930 A (深圳市商汤科技有限公司) 2020年 4月 28日 (2020-04-28) 全文	1-11
A	CN 110046712 A (天津科技大学) 2019年 7月 23日 (2019-07-23) 全文	1-11
A	CN 111651765 A (上海交通大学) 2020年 9月 11日 (2020-09-11) 全文	1-11
A	US 2019220605 A1 (INTEL CORPORATION) 2019年 7月 18日 (2019-07-18) 全文	1-11

☐ 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
2021年 8月 9日	2021年 8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	张雯
传真号 (86-10)62019451	电话号码 86-(10)-53961314

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/095175

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 112163670 A	2021年 1月 1日	无	
CN 111723927 A	2020年 9月 29日	无	
CN 111079930 A	2020年 4月 28日	无	
CN 110046712 A	2019年 7月 23日	无	
CN 111651765 A	2020年 9月 11日	无	
US 2019220605 A1	2019年 7月 18日	EP 3712822 A1	2020年 9月 23日