

一种模型安全检测方法、装置以及电子设备

技术领域

本说明书涉及计算机软件技术领域，尤其涉及一种模型安全检测方法、装置以及电子设备。

5

背景技术

深度学习是目前比较热门的一个技术领域，基于深度学习模型，可以执行诸如风险识别、语音识别、图像识别等任务。

在现有技术中，深度学习模型通常处于服务器上，但其同样也可以处于诸如用户终端等相对开放的环境中。

对于后一种情况，需要能够检测模型安全的方案。

发明内容

本说明书实施例提供一种模型安全检测方法、装置以及电子设备，用以解决如下技术问题：需要能够检测处于相对开放的环境中的模型的安全的方案。

为解决上述技术问题，本说明书实施例是这样实现的：

本说明书实施例提供一种模型安全检测方法，包括：

获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据；

通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果；

其中，所述判别器是利用生成对抗网络(Generative Adversarial Nets, GAN)框架、所述待检测模型以及样本训练得到，在训练得到所述判别器的过程中，所述待检测模型处于可信环境中。

本说明书实施例提供一种模型安全检测装置，包括：

25

获取模块，获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据；

检测模块，通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果；

- 5 其中，所述判别器是利用 GAN 框架、所述待检测模型以及样本训练得到，在训练得到所述判别器的过程中，所述待检测模型处于可信环境中。

本说明书实施例提供的一种电子设备，包括：

至少一个处理器；以及，

与所述至少一个处理器通信连接的存储器；其中，

- 10 所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够：

获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据；

- 15 通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果；

其中，所述判别器是利用 GAN 框架、所述待检测模型以及样本训练得到，在训练得到所述判别器的过程中，所述待检测模型处于可信环境中。

- 20 本说明书实施例采用的上述至少一个技术方案能够达到以下有益效果：可以利于基于 GAN 框架训练得到的判别器，根据待检测模型计算出的中间结果和/或输出结果的数据进行模型安全检测，以了解所述计算过程或结果是否安全。

附图说明

- 25 为了更清楚地说明本说明书实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本说明书中记载的一些实施例，对于本领域普通技术人员来

讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图 1 为本说明书的方案在一种实际应用场景下涉及的一种整体架构示意图；

图 2 为本说明书实施例提供的一种模型安全检测方法的流程示意图；

5 图 3 为本说明书实施例提供的上述模型安全检测方法中用到的判别器的训练原理示意图；

图 4 为本说明书实施例提供的实际应用中所述模型安全检测方法的一种具体实施方案示意图；

10 图 5 为本说明书实施例提供的对应于图 2 的一种模型安全检测装置的结构示意图。

具体实施方式

本说明书实施例提供一种模型安全检测方法、装置以及电子设备。

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

本说明书中提到了模型安全这个概念，为了便于理解，在此进行解释。

20 所述模型一般用于分类预测，一种典型的应用场景是利用风控模型进行风险控制。具体地，将一些特征数据输入风控模型进行计算得到结果数据，进而根据结果数据，判定输入的特征数据对应的用户或者事件是否存在风险。

在实际应用中，可能有恶意用户对模型进行攻击，其攻击方式比如是篡改模型的输入数据、篡改模型的计算逻辑或者结果数据等，从而导致模型的计算
25 过程或者计算结果不再可信。在这种受到攻击的情况下，可以认为模型是不安全的。本说明书正是要提供检测模型是否安全的方案。

图 1 为本说明书的方案在一种实际应用场景下涉及的一种整体架构示意图。该整体架构中，主要涉及：待检测模型以及利用 GAN 框架训练得到的判别器所在设备。利用 GAN 框架、待检测模型以及样本训练得到判别器，通过判别器对待检测模型针对当前输入数据计算出的中间结果和/或输出结果的数
5 据进行判别，检测待检测模型当前是否安全。

该设备优选地可以是诸如用户终端等处于相对开放的环境中的设备。比如，在手机上部署风控引擎，通过风控引擎进行边缘计算（Edge Computing）以实现对于该手机的风险控制，则待检测模型可以是该风控引擎中的模型，手机可以为该设备。

10 基于以上整体架构，下面对本说明书的方案进行详细说明。

图 2 为本说明书实施例提供的一种模型安全检测方法的流程示意图。该流程可能的执行主体包括但不限于可作为终端或者服务器的以下设备：手机、平板电脑、智能可穿戴设备、车机、个人计算机、中型计算机、计算机集群等。

图 2 中的流程可以包括以下步骤：

15 S202：获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据。

在本说明书实施例中，待检测模型处于执行主体上。本申请对待检测模型所属领域不做限定。比如，其可以属于人工智能领域，具体为风控模型，根据输入数据（如行为特征、环境信息等数据）进行计算，根据计算得到的输出结果
20 的数据判定当前执行操作的是否是用户本人，在这类用途下，待检测模型本身的安全显得尤为重要。

所述结果数据的形式可能有多种。其中一种常见的形式是向量形式。

以深度学习模型为例，其主体是一个深度神经网络。一般地，该深度神经网络的输入数据是向量，该深度神经网络的输出层输出的上述的输出结果的数
25 据也是向量；进一步地，若有需要，该深度神经网络的中间层（也即，隐层）计算出的数据也是可以输出的，中间层输出的数据通常也是向量，其属于上述

的中间结果的数据。

除了向量形式以外，结果数据也可能是数值、布尔值、或者字符串等其他形式。

S204: 通过判别器对所述结果数据进行判别，确定所述待检测模型的安全
5 检测结果；其中，所述判别器是利用 GAN 框架、所述待检测模型以及样本训练得到，在训练得到所述判别器的过程中，所述待检测模型处于可信环境中。

在本说明书实施例中，所述结果数据是通过将输入数据输入待检测模型进行计算得到的，所述输入数据可以不带标签，通常用于模型预测阶段。

所述样本有多个而且可以带有标签，通常用于模型训练阶段。在训练判别
10 器时，需要将样本输入待检测模型进行计算，得到第二结果数据，所述第二结果数据包括第二中间结果和/或第二输出结果的数据。

需要说明的是，这里的“第二”只是为了对所述输入数据、所述样本进行区分，而待检测模型计算结果数据和第二结果数据所采用的方式可以是一致的。

在本说明书实施例中，基于 GAN 框架，以第二结果数据为训练数据，可
15 以训练得到对应的 GAN，该 GAN 由生成器和所述判别器构成。其中，生成器也可以称为生成 (Generative, G) 网络，判别器也可以称为判别 (Discriminative, D) 网络。

在本说明书实施例中，在训练得到判别器的过程中，也需要待检测模型针对输入的样本进行计算，并将计算得到的第二结果数据作为判别器训练依据。

20 那么，若训练过程本身不可信，则难以确保训练得到的判别器是可信的，针对这个问题，可以优选地确保在判别器训练过程中，待检测模型处于可信环境中，从而保证第二结果数据是可信的。

通过图 2 的方法，可以利于基于 GAN 框架训练得到的判别器，根据待检测模型计算出的中间结果和/或输出结果的数据进行模型安全检测，以了解所述
25 计算过程或结果是否安全，所述计算过程或结果是否安全能够直接反应该待检测模型是否安全。

基于图 2 的方法，本说明书实施例还提供了该方法的一些具体实施方案，以及扩展方案，下面进行说明。

在本说明书实施例中，对于步骤 S204，利用所述第二结果数据和 GAN 框架，训练得到相应的 GAN，具体可以包括：

- 5 获取所述待检测模型针对样本计算出的第二结果数据，所述第二结果数据包括第二中间结果和/或第二输出结果的数据；利用所述第二结果数据和 GAN 框架，训练得到相应的 GAN（也即，训练得到了所述判别器），所述 GAN 由生成器和所述判别器构成。

10 进一步地，对于诸如深度学习模型等使用了神经网络的模型，中间结果相比于对应的输出结果维度通常更高，相应地，包含的信息量往往也更丰富更精细。比如，中间结果为一个 100 维的向量，对应的输出结果为一个 10 维的向量；再比如，中间结果为一个 100 维的向量，对应的输出结果为一个 1 维的布尔值；等等，可以看到，从中间结果到输出结果是一个信息收敛的过程。

15 对于上一段中的这类模型，相比于利用第二输出结果的数据，优选地可以利用第二中间结果的数据，训练得到判别器，如此有利于提高判别器的信息分析能力，进而有利于得到更准确的判别结果。

20 另外，在实际应用中，可能存在这样的场景：待检测模型是从一个更完整的模型中部分拆分出来的。比如，训练一个共有 10 层的神经网络模型，训练完毕后，将该神经网络模型的前 9 层抽取出来，作为待检测模型，在这种情况下，待检测模型的输出结果实际上就是神经网络模型的中间结果，那么，利用待检测模型的输出结果的数据训练得到判别器也是比较合适的。

图 3 为本说明书实施例提供的上述模型安全检测方法中用到的判别器的训练原理示意图。

25 在图 3 中，所训练的 GAN 由判别器和生成器构成，以第二结果数据为标准，训练 GAN 使得生成器能够生成与第二结果数据的数据分布一致的生成结果。

在训练过程中，认为第二结果数据应当为真（Label=1），认为生成结果应当为假（Label=0），第二结果数据、生成结果会分别被输入判别器进行判别是真还是假。对于生成器的训练而言，训练目标是使针对生成结果的判别误差最大化，对于判别器的训练而言，训练目标是使针对生成结果以及第二结果数据的判别误差最小化。

一般地，在训练过程中，输入生成器的数据是随机数据，比如，随机数或者随机向量等。进一步地，为了使得训练过程尽快收敛，可以将第二结果数据叠加到随机数据，作为生成器的输入，如此，作为生成器的输入数据本身已经携带了第二结果数据的至少部分数据分布特征，从而使得生成结果的数据分布更容易逼近第二结果数据的数据分布。

当然，也可以不依赖于随机数据，针对第二结果数据进行一定的非线性变换后，作为生成器的输入，等等。

根据上一段的分析，所述利用所述第二结果数据和 GAN 框架，训练得到相应的 GAN，具体可以包括：

根据随机数据和/或所述第二结果数据，生成用于输入生成器的数据；根据所述第二结果数据、所述用于输入生成器的数据和 GAN 框架，训练得到相应的 GAN。

在本说明书实施例中，所述样本优选地是没有受到攻击的样本，这可以通过上述的可信环境保证，进而可以认为所述第二结果数据是可信的。否则，若第二结果数据本身不可信，则会对基于第二结果数据训练得到的判别器的可靠性产生不利影响，进而也会降低模型安全检测结果的可信性。

在本说明书实施例中，基于上述训练过程，对实际应用中所述模型安全检测方法的一种具体实施方案进行说明，图 4 为该方案示意图。

在图 4 中，安全检测模型至少包含判别器，除此之外，还可以包含获得判别结果后进一步的处理逻辑。输入数据被输入待检测模型进行处理，计算得到结果数据；将中间结果和/或输出结果的数据输入安全检测模型，由判别器进行

判别, 在获得判别结果后, 可以进一步地由安全检测模型根据输出结果的数据和判别结果生成安全检测结果。

例如, 假定输出结果的数据表示待检测模型判定输入数据对应的操作是本人进行的, 而判别结果却为假 (也即, 数据分布与预期不一致), 由此可以推测计算过程很有可能遭到攻击, 导致中间结果和/或输出结果的数据已被篡改, 则对应生成的安全检测结果比如可以反映这样的信息: “操作是本人 (代表待检测模型计算的输出结果的数据), 不安全已被篡改 (代表判别器的判别结果)”; 安全检测结果的具体格式和表现形式不做限定。

当然, 也可以只根据判别结果生成安全检测结果, 而不考虑输出结果的数据。

待检测模型处于不安全环境 (比如背景技术中所述的开放环境) 中, 可能受到攻击, 为了防止安全检测模块本身也受到攻击, 安全检测模块可以处于预定的安全环境中。所述安全环境可以基于软件和/或硬件构建, 其比如可以是可信执行环境 (Trusted Execution Environment, TEE) 等。

根据上例进行总结, 对于步骤 S204, 所述判别器对所述结果数据进行判别, 具体可以包括:

判别器判别所述结果数据是真还是假, 其中, 所述判别结果反映所述结果数据与所述第二结果数据的数据分布是否一致。

进一步地, 所述通过判别器对所述结果数据进行判别, 确定所述待检测模型的安全检测结果, 具体可以包括:

通过判别器对所述结果数据进行判别获得判别结果, 根据所述判别结果确定所述待检测模型的安全检测结果; 或者, 获取所述输出结果的数据, 根据所述输出结果的数据和所述判别结果, 确定所述待检测模型的安全检测结果。

基于同样的思路, 本说明书实施例还提供了对应的装置, 如图 5 所示。

图 5 为本说明书实施例提供的对应于图 2 的一种模型安全检测装置的结构示意图, 包括:

获取模块 501，获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据；

检测模块 502，通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果；

5 其中，所述判别器是利用 GAN 框架、所述待检测模型以及样本训练得到。

可选地，所述待检测模型为深度学习模型，所述中间结果的数据为所述待检测模型的中间层计算出的数据。

可选地，利用 GAN 框架、所述待检测模型以及样本训练得到所述判别器，具体包括：

10 获取所述待检测模型针对样本计算出的第二结果数据，所述第二结果数据包括第二中间结果和/或第二输出结果的数据；

利用所述第二结果数据和 GAN 框架，训练得到相应的 GAN，所述 GAN 由生成器和所述判别器构成。

可选地，所述第二结果数据是可信的。

15 可选地，所述利用所述第二结果数据和 GAN 框架，训练得到相应的 GAN，具体包括：

根据随机数据和/或所述第二结果数据，生成用于输入生成器的数据；

根据所述第二结果数据、所述用于输入生成器的数据和 GAN 框架，训练得到相应的 GAN。

20 可选地，所述检测模块 502 通过判别器对所述结果数据进行判别，具体包括：

所述检测模块 502 通过判别器判别所述结果数据是真还是假，其中，所述判别结果反映所述结果数据与所述第二结果数据的数据分布是否一致。

25 可选地，所述检测模块 502 通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果，具体包括：

所述检测模块 502 通过判别器对所述结果数据进行判别获得判别结果，根

据所述判别结果确定所述待检测模型的安全检测结果；或者，

所述检测模块 502 获取所述输出结果的数据，根据所述输出结果的数据和所述判别结果，确定所述待检测模型的安全检测结果。

可选地，所述判别器处于预定的安全环境中。

5 可选地，所述判别器处于用户终端。

基于同样的思路，本说明书实施例还提供了对应的一种电子设备，包括：

至少一个处理器；以及，

与所述至少一个处理器通信连接的存储器；其中，

10 所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够：

获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据；

通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果；

15 其中，所述判别器是利用 GAN 框架、所述待检测模型以及样本训练得到。

基于同样的思路，本说明书实施例还提供了对应的一种非易失性计算机存储介质，存储有计算机可执行指令，所述计算机可执行指令设置为：

获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据；

20 通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果；

其中，所述判别器是利用 GAN 框架、所述待检测模型以及样本训练得到。

25 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施

方式中，多任务处理和并行处理也是可以的或者可能是有利的。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于装置、电子设备、非易失性计算机存储介质实施例而言，由于
5 其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

本说明书实施例提供的装置、电子设备、非易失性计算机存储介质与方法是相对应的，因此，装置、电子设备、非易失性计算机存储介质也具有与对应方法类似的有益技术效果，由于上面已经对方法的有益技术效果进行了详细说
10 明，因此，这里不再赘述对应装置、电子设备、非易失性计算机存储介质的有益技术效果。

在 20 世纪 90 年代，对于一个技术的改进可以很明显地区分是硬件上的改进（例如，对二极管、晶体管、开关等电路结构的改进）还是软件上的改进（对于方法流程的改进）。然而，随着技术的发展，当今的很多方法流程的改进已经可以视为硬件电路结构的直接改进。设计人员几乎都通过将改进的方法流程
15 编程到硬件电路中来得到相应的硬件电路结构。因此，不能说一个方法流程的改进就不能用硬件实体模块来实现。例如，可编程逻辑器件（Programmable Logic Device, PLD）（例如现场可编程门阵列（Field Programmable Gate Array, FPGA））就是这样一种集成电路，其逻辑功能由用户对器件编程来确定。由设计人员自行编程来把一个数字系统“集成”在一片 PLD 上，而不需要请芯片制造
20 厂商来设计和制作专用的集成电路芯片。而且，如今，取代手工地制作集成电路芯片，这种编程也多半改用“逻辑编译器（logic compiler）”软件来实现，它与程序开发撰写时所用的软件编译器相类似，而要编译之前的原始代码也得用特定的编程语言来撰写，此称之为硬件描述语言（Hardware Description
25 Language, HDL），而 HDL 也并非仅有一种，而是有许多种，如 ABEL（Advanced Boolean Expression Language）、AHDL（Altera Hardware Description Language）、

Confluence、CUPL (Cornell University Programming Language)、HDCal、JHDL (Java Hardware Description Language)、Lava、Lola、MyHDL、PALASM、RHDL (Ruby Hardware Description Language) 等，目前最普遍使用的是 VHDL (Very-High-Speed Integrated Circuit Hardware Description Language)与 Verilog。

- 5 本领域技术人员也应该清楚，只需要将方法流程用上述几种硬件描述语言稍作逻辑编程并编程到集成电路中，就可以很容易得到实现该逻辑方法流程的硬件电路。

控制器可以按任何适当的方式实现，例如，控制器可以采取例如微处理器或处理器以及存储可由该（微）处理器执行的计算机可读程序代码（例如软件
10 或固件）的计算机可读介质、逻辑门、开关、专用集成电路（Application Specific Integrated Circuit，ASIC）、可编程逻辑控制器和嵌入微控制器的形式，控制器的例子包括但不限于以下微控制器：ARC 625D、Atmel AT91SAM、Microchip PIC18F26K20 以及 Silicone Labs C8051F320，存储器控制器还可以被实现为存储器的控制逻辑的一部分。本领域技术人员也知道，除了以纯计算机可读程序
15 代码方式实现控制器以外，完全可以通过将方法步骤进行逻辑编程来使得控制器以逻辑门、开关、专用集成电路、可编程逻辑控制器和嵌入微控制器等的形式来实现相同功能。因此这种控制器可以被认为是一种硬件部件，而对其内包括的用于实现各种功能的装置也可以视为硬件部件内的结构。或者甚至，可以将用于实现各种功能的装置视为既可以是实现方法的软件模块又可以是硬件
20 部件内的结构。

上述实施例阐明的系统、装置、模块或单元，具体可以由计算机芯片或实体实现，或者由具有某种功能的产品来实现。一种典型的实现设备为计算机。具体的，计算机例如可以为个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件设备、游戏控制
25 台、平板计算机、可穿戴设备或者这些设备中的任何设备的组合。

为了描述的方便，描述以上装置时以功能分为各种单元分别描述。当然，

在实施本说明书时可以把各单元的功能在同一个或多个软件和/或硬件中实现。

本领域内的技术人员应明白，本说明书实施例可提供为方法、系统、或计算机程序产品。因此，本说明书实施例可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本说明书实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

本说明书是参照根据本说明书实施例的方法、设备（系统）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

在一个典型的配置中，计算设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。内

存是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存
5 (PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他
10 磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体(transitory media)，如调制的数据信号和载波。

还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、
15 方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

本说明书可以在由计算机执行的计算机可执行指令的一般上下文中描述，例如程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型
20 的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本说明书，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的
25

比较简单，相关之处参见方法实施例的部分说明即可。

以上所述仅为本说明书实施例而已，并不用于限制本申请。对于本领域技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等，均应包含在本申请的权利要求范围之内。

权利要求书

1、一种模型安全检测方法，包括：

获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据；

5 通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果；

其中，所述判别器是利用生成对抗网络 GAN 框架、所述待检测模型以及样本训练得到，在训练得到所述判别器的过程中，所述待检测模型处于可信环境中。

10 2、如权利要求 1 所述的方法，所述待检测模型为深度学习模型，所述中间结果的数据为所述待检测模型的中间层计算出的数据。

3、如权利要求 1 所述的方法，利用 GAN 框架、所述待检测模型以及样本训练得到所述判别器，具体包括：

15 获取所述待检测模型针对样本计算出的第二结果数据，所述第二结果数据包括第二中间结果和/或第二输出结果的数据；

利用所述第二结果数据和 GAN 框架，训练得到相应的 GAN，所述 GAN 由生成器和所述判别器构成。

4、如权利要求 3 所述的方法，所述第二结果数据是可信的。

20 5、如权利要求 3 所述的方法，所述利用所述第二结果数据和 GAN 框架，训练得到相应的 GAN，具体包括：

根据随机数据和/或所述第二结果数据，生成用于输入生成器的数据；

根据所述第二结果数据、所述用于输入生成器的数据和 GAN 框架，训练得到相应的 GAN。

25 6、如权利要求 1 所述的方法，所述判别器对所述结果数据进行判别，具体包括：

判别器判别所述结果数据是真还是假，其中，所述判别结果反映所述结果

数据与所述第二结果数据的数据分布是否一致。

7、如权利要求 6 所述的方法，所述通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果，具体包括：

通过判别器对所述结果数据进行判别获得判别结果，根据所述判别结果确定所述待检测模型的安全检测结果；或者，

获取所述输出结果的数据，根据所述输出结果的数据和所述判别结果，确定所述待检测模型的安全检测结果。

8、如权利要求 1~7 任一项所述的方法，所述判别器处于预定的安全环境中。

9、如权利要求 8 所述的方法，所述判别器处于用户终端。

10、一种模型安全检测装置，包括：

获取模块，获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据；

检测模块，通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果；

其中，所述判别器是利用生成对抗网络 GAN 框架、所述待检测模型以及样本训练得到，在训练得到所述判别器的过程中，所述待检测模型处于可信环境中。

11、如权利要求 10 所述的装置，所述待检测模型为深度学习模型，所述中间结果的数据为所述待检测模型的中间层计算出的数据。

12、如权利要求 10 所述的装置，利用 GAN 框架、所述待检测模型以及样本训练得到所述判别器，具体包括：

获取所述待检测模型针对样本计算出的第二结果数据，所述第二结果数据包括第二中间结果和/或第二输出结果的数据；

利用所述第二结果数据和 GAN 框架，训练得到相应的 GAN，所述 GAN 由生成器和所述判别器构成。

13、如权利要求 12 所述的装置，所述第二结果数据是可信的。

14、如权利要求 12 所述的装置，所述利用所述第二结果数据和 GAN 框架，训练得到相应的 GAN，具体包括：

根据随机数据和/或所述第二结果数据，生成用于输入生成器的数据；

5 根据所述第二结果数据、所述用于输入生成器的数据和 GAN 框架，训练得到相应的 GAN。

15、如权利要求 10 所述的装置，所述检测模块通过判别器对所述结果数据进行判别，具体包括：

所述检测模块通过判别器判别所述结果数据是真还是假，其中，所述判别
10 结果反映所述结果数据与所述第二结果数据的数据分布是否一致。

16、如权利要求 15 所述的装置，所述检测模块通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果，具体包括：

所述检测模块通过判别器对所述结果数据进行判别获得判别结果，根据所述判别结果确定所述待检测模型的安全检测结果；或者，

15 所述检测模块获取所述输出结果的数据，根据所述输出结果的数据和所述判别结果，确定所述待检测模型的安全检测结果。

17、如权利要求 10~16 任一项所述的装置，所述判别器处于预定的安全环境中。

18、如权利要求 17 所述的装置，所述判别器处于用户终端。

20 19、一种电子设备，包括：

至少一个处理器；以及，

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够：

25 获取待检测模型计算出的结果数据，所述结果数据包括中间结果和/或输出结果的数据；

通过判别器对所述结果数据进行判别，确定所述待检测模型的安全检测结果；

其中，所述判别器是利用生成对抗网络 GAN 框架、所述待检测模型以及样本训练得到，在训练得到所述判别器的过程中，所述待检测模型处于可信环
5 境中。