

模型测试的方法及装置

技术领域

[01] 本说明书一个或多个实施例涉及计算机技术领域，尤其涉及通过计算机进行模型测试的方法和装置。

5 背景技术

[02] 随着计算机和人工智能技术的发展，人工神经网络（ANN）的应用越来越多，例如模式识别、自动控制、信号处理、辅助决策等等。人工神经网络是一种运算模型，由大量的处理单元，或称神经元，之间相互连接构成。每个处理单元代表一种特定的输出函数，称为激励函数（activation function）。神经网络模型的性能测试往往通过生成神经网络模型的系统的代码覆盖率，或者其在样本集上的输出正确率以及在样本难易覆盖的特殊情况（如某地气温零下 70 摄氏度）下的输出结果准确性来衡量。

[03] 常规技术中，为了对模型的测试更充分，通常需要收集大量的训练样本；为了提高模型的准确性，也往往需要收集更多的训练样本来训练模型，对模型参数进行调节和优化。然而，这样的方式通过人工收集样本，浪费人力，同时，收集的训练样本毕竟是有限的，并不能保证对模型的充分测试，对模型准确性的提高也是有限的。例如，某地气温从来没有低到过零下 70 摄氏度，所采集的样本再多也无法覆盖到气温为零下 70 摄氏度的情况。特别地，在一些特殊情况下，例如无人驾驶车辆面对白色背景的卡车，如果所收集的测试样本中没有这样的情况，则可能判定为天空背景，影响决策，从而可能造成不可挽回的损失。

[04] 因此，希望能有改进的方案，通过产生更多差异化的样本，有效的进行模型测试，既能避免人工收集样本造成的人力浪费，又可以通过产生的这些差异化样本对模型进行进一步的测试，以及用于训练模型从而提升模型的准确性。

发明内容

[05] 本说明书一个或多个实施例描述了一种方法和装置，可以产生更多差异化的样本，有效的进行模型测试，既能避免人工收集样本造成的人力浪费，又可以通过产生的这些差异化样本对模型进行进一步的测试，以及用于训练模型从而提升模型的准确性。

- [06]根据第一方面, 提供了一种模型测试的方法, 包括:
- [07]从测试样本集中获取样本;
- [08]将所述样本输入模型集包括的多个待测试模型中, 以获得各个待测试模型的输出结果, 其中, 所述多个待测试模型包括至少一个神经网络模型;
- 5 [09]根据所述输出结果确定测试结果, 所述测试结果包括第一测试结果和第二测试结果中的至少一项, 所述第一测试结果包括各个待测试模型的输出结果的准确性, 所述第二测试结果包括所述至少一个神经网络模型的神经元覆盖率;
- [10]在所述测试结果未满足预定条件的情况下, 按照预定规则, 基于所述样本生成新样本, 并将所述新样本加入所述测试样本集。
- 10 [11]根据一方面的实施方式, 在所述测试结果包括第一测试结果的情况下, 所述根据所述输出结果确定测试结果包括:
- [12]对所述多个待测试模型中包括的功能相同的待测试模型, 投票确定正确输出结果, 其中, 票数最高的输出结果为正确输出结果;
- [13]确定输出结果为所述正确输出结果的待测试模型输出准确。
- 15 [14]在一个可能的设计中, 所述模型集中包括标准模型, 以及, 所述根据所述输出结果确定测试结果包括:
- [15]基于标准模型的输出结果, 确定与所述标准模型功能相同的待测试模型的正确输出结果;
- [16]确定输出结果为所述正确输出结果的待测试模型输出准确。
- 20 [17]进一步地, 在一个实施例中, 所述预定条件包括:
- [18]所述多个待测试模型的输出结果准确度均大于预设准确度阈值。
- [19]根据一方面的实施方式, 在所述测试结果包括第二测试结果的情况下, 所述根据所述输出结果确定测试结果包括:
- [20]根据所述至少一个神经网络模型中各个神经元的输出是否满足输出条件, 统计各个神经网络模型的神经元覆盖数, 其中, 所述神经元覆盖数包括, 测试过程中满足过输出条件的神经元个数;
- 25 [21]基于所述神经元覆盖数确定各个神经网络模型的神经元覆盖率, 其中, 所述神经元

覆盖率和以下比值正相关：所述神经元覆盖数与总神经元个数的比值。

[22]进一步地，在一个实施例中，所述预定条件包括：

[23]所述至少一个神经网络模型的神经元覆盖率均大于预设覆盖率阈值。

[24]在一个实施例中，按照预定规则，基于所述样本生成新样本包括：

5 [25]获取所述样本中各个特征的设定范围；

[26]基于各个特征的设定范围为所述样本添加随机扰动项，生成新样本。

[27]根据一种可能的设计，上述方法还包括以下中的一项或多项：

[28]剔除模型集中测试结果满足预定条件的待测试模型；

[29]在模型集中加入新的待测试模型。

10 [30]根据第二方面，提供一种模型测试的装置，包括：

[31]获取单元，配置为从测试样本集中获取样本；

[32]测试单元，配置为将所述样本输入模型集包括的多个待测试模型中，以获得各个待测试模型的输出结果，其中，所述多个待测试模型包括至少一个神经网络模型；

15 [33]确定单元，配置为根据所述输出结果确定测试结果，所述测试结果包括第一测试结果和第二测试结果中的至少一项，所述第一测试结果包括各个待测试模型的输出结果的准确性，所述第二测试结果包括所述至少一个神经网络模型的神经元覆盖率；

[34]生成单元，配置为在所述测试结果未满足预定条件的情况下，按照预定规则，基于所述样本生成新样本，并将所述新样本加入所述测试样本集。

20 [35]根据第三方面，提供了一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行第一方面的方法。

[36]根据第四方面，提供了一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现第一方面的方法。

25 [37]通过本说明书实施例提供的方法和装置，首先从测试样本集中获取样本，接着将样本输入模型集包括的多个待测试模型中，以获得各个待测试模型的输出结果，然后根据输出结果确定测试结果，进一步地，在该测试结果未满足预定条件的情况下，按照预定规则，基于上述样本生成新样本，并将所生成的新样本加入测试样本集。在该模型测试的方法被循环执行的情况下，一方面对待检测模型的准确性能和/或测试充分程度进行评

估, 另一方面将基于原样本生成的、与原样本具有差异化的新样本。如此既能避免人工收集样本造成的人力浪费, 又可以对模型进行进一步的测试, 以及将扩充后的测试样本集用于训练模型从而提升模型的准确性。

附图说明

- 5 [38] 为了更清楚地说明本发明实施例的技术方案, 下面将对实施例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发明的一些实施例, 对于本领域普通技术人员来讲, 在不付出创造性劳动的前提下, 还可以根据这些附图获得其它的附图。

[39] 图 1 示出本说明书披露的一个实施例的实施场景示意图;

- 10 [40] 图 2 示出根据一个实施例的模型测试的方法流程图;

[41] 图 3 示出神经网络模型的一个具体例子;

[42] 图 4 示出神经网络模型的神经元的一个具体例子;

[43] 图 5 示出根据一个实施例的用于模型测试的装置的示意性框图。

具体实施方式

- 15 [44] 下面结合附图, 对本说明书提供的方案进行描述。

[45] 图 1 为本说明书披露的一个实施例的实施场景示意图。如图所示, 用户 (如应用开发、测试人员) 在神经网络模型正式使用之前, 可以将这些模型通过计算平台进行测试。这里的计算平台可以是具有数据处理能力和数据输入功能的各种装置、设备, 例如台式计算机、服务器等等。可以理解, 计算平台还可以是上述电子设备组成的设备集群。用
20 户可以收集样本作为初始的测试样本集输入计算平台。

- [46] 测试时, 计算平台从测试样本集中获取一个样本, 如样本 1, 然后将样本 1 输入模型集包括的多个待测试模型中, 如模型 1、模型 2、模型 3 等等, 并获得各个测试模型的输出结果, 对应为输出 1、输出 2、输出 3 等等。可以理解, 这多个模型可以具有相似的功能 (预定功能), 以使一个样本 1 能够同时用于模型 1、模型 2、模型 3 等等的
25 测试, 例如都具有模式识别功能。如此, 样本 1 可以同时输入这多个模型中, 这多个模型分别给出各自的输出结果, 例如模式识别的输出结果。由于各个模型的准确性不同, 具有相同功能的待测试模型的输出结果可以相同, 也可以不相同。值得说明的是, 这些

待测试模型包括至少一个神经网络模型。

[47]接着，计算平台可以根据各个模型的输出结果确定测试结果。这里，测试结果可以包括第一测试结果和第二测试结果中的至少一项。其中，第一测试结果包括各个待测试模型的输出结果的准确性，第二测试结果包括至少一个神经网络模型的神经元覆盖率。

5 这里，神经元覆盖率可以用于表示模型的神经元激活情况。如此，对每个待测试模型而言，通过输出结果的准确性可以对该模型的性能进行评判；另一方面，通过神经元覆盖率，可以确定对该模型的测试是否充分。

[48]可以理解，测试之前，用户（如应用开发、测试人员）还可以对测试结果设定相关条件作为预定条件，用来控制测试的结束时机，例如，可以设定所有待测试模型准确性
10 均合格的预定条件，来控制测试过程结束等等。如果测试过程没有结束，则计算平台可以再从测试样本集中获取其他样本，循环执行以上测试过程。同时，由于原始的测试样本集的样本数量具有一定的局限性，如果测试过程还要继续循环执行，则计算平台可以按照预定规则，基于用于此次的测试过程的样本 1 生成一个新样本，并将该新样本加入到测试样本集中，以用于后续的测试过程。用户（如应用开发、测试人员）可以通过设
15 定预定规则，控制样本的生成，既能保证涵盖各种情况，又不会与实际情况相差甚远。

[49]如此，一方面对待检测模型的准确性和/或测试充分程度进行了评估，另一方面将基于原样本生成的、与原样本具有差异化的新样本加入测试样本集，可以使测试样本集中的样本数量和所涵盖的情况得以扩充。通过自动生成这些差异化样本，既能避免人工收集样本造成的人力浪费，又可以对模型进行进一步的测试，以及用于训练模型从而提升
20 模型的准确性。下面描述上述场景的具体执行过程。

[50]图 2 示出根据一个实施例的模型测试的方法流程图。该方法的执行主体可以是任何具有计算、处理能力的系统、设备、装置、平台或服务器，例如图 1 所示的计算平台等。如图 2 所示，该方法包括以下步骤：步骤 21，从测试样本集中获取样本；步骤 22，将上述样本输入模型集包括的多个待测试模型中，以获得各个待测试模型的输出结果，其
25 中，上述多个待测试模型包括至少一个神经网络模型；步骤 23，根据上述输出结果确定测试结果，该测试结果包括第一测试结果和第二测试结果中的至少一项，第一测试结果包括各个待测试模型的输出结果的准确性，第二测试结果包括上述至少一个神经网络模型的神经元覆盖率；步骤 24，在上述测试结果不满足预定条件的情况下，按照预定规则，基于以上样本生成新样本，并将该新样本加入上述测试样本集。

30 [51]首先，在步骤 21，从测试样本集中获取样本。可以理解，测试样本集用于存储模型

测试用的样本，其可以是一个集合，也可以是数据库，本申请对此不作限定。

[52]模型测试的方法首次执行时，测试样本集为初始的测试样本集。初始的测试样本集存储的样本可以包括用户收集的样本，也可以包括用户收集的样本和以往的模型测试过程中生成的样本。

- 5 [53]根据一种实施方式，用户可以对收集的样本进行筛选，挑选一些在模型中容易出错的样本，或者样本特征所占比例比较低的样本，放入初始的测试样本集。举例而言，一个样本在输入 10 个不同的用于图像识别的模型时，5 个模型的输出结果为卡车，5 个模型的输出结果为蓝天，则该样本容易出错，可以挑选出来；再例如，所收集的 100 个样本中有 90 个的温度在零至 20 摄氏度，有 4 个样本的温度在零下较接近零摄氏度（比如都高于零下三摄氏度），有 1 个样本的温度在零下 10 摄氏度，则这 1 个温度在零下 10 摄氏度的样本所占比例比较低，可以挑选出来，对于这 4 个温度在零下较接近零摄氏度的样本，可以全部挑选出来，也可以择一或择二挑选。如此，可以避免漏掉一些较特殊的情况，有利于测试样本集中的样本均衡地涵盖各种情况。

- 15 [54]从测试样本集中获取样本可以是任意可行的方式。在一个实施例中，可以按照测试样本集中样本的排列顺序获取样本，例如第一遍执行该模型测试的方法时获取排在第一的样本，第二遍执行该模型测试的方法时获取排在第二的样本，以此类推。

- [55]在另一个实施例中，可以随机获取测试样本集中的样本。在随机获取样本的情况下，该模型测试的方法被执行多遍后，可能会出现一些样本被使用的次数较多，一些样本未被使用过或使用较少，还可以设置样本调取权重，使用次数越少，调取权重越高，以增加未被使用过或使用较少的样本的使用概率。
- 20

[56]在更多实施例中，用户还可以通过更多可行方式从测试样本集中获取样本。

- [57]接着，在步骤 22，将上述样本输入模型集包括的多个待测试模型中，以获得各个待测试模型的输出结果。值得说明的是，这里的模型是用来揭示变量之间因果关系的运算模型，其可以包括一个或一系列函数组成的算法，通过这样的算法，可以根据输入的变量得到一个由这些变量确定的输出结果。如误差反向传播算法 BP（Back Propagation）算法、卷积神经网络 CNNs（Convolutional neural networks）等等。
- 25

[58]对于较复杂的情况，可以通过神经网络模型来实现。神经网络模型由多个神经元构成。如图 3 所示，是一个三层结构的神经网络模型的一个具体例子。图 3 中的神经网络模型包括多个神经元，如神经元 31、32 等。第一层各个神经元接收不同的输入特征参

数 x_1 、 x_2 、 x_3 、 x_4 传递给第二层，第二层的神经元 31 分别对应运算单元，各个运算单元可以以第一层的 x_1 、 x_2 、 x_3 、 x_4 为输入。其中，各个特征 x_1 、 x_2 、 x_3 、 x_4 由样本给出，或者从样本中提取，例如分别是水果的颜色、形状、大小等参数。同理，第三层的神经元 32 是以第二层的神经元 31 等的各个输出值作为输入的运算单元。每两个神经元间的连接都代表一个对于通过该连接信号的加权值，称之为权重。该神经网络模型经过神经元 32，可以得到一个输出结果，例如桔子、橙子等。

[59]进一步地，如图 4 所示，每个神经元代表一种特定的输出函数，称为激励函数（activation function，例如为 sigmoid 函数 $f(z)=1/(1+e^{-z})$ ）。在图 4 示出的神经元中，输出函数可以为 $H = f(\sum_{i=1}^3 W_i x_i)$ ，其中，对于 W_i ， i 值从 1 取值到 3，分别代表权值 1、权值 2、权值 3， x_i 中， i 值从 1 取值到 3，分别代表特征参数 1、特征参数 2、特征参数 3。对特征参数 1、特征参数 2、特征参数 3 加权求和后通过激励函数映射得到神经元的输出。

[60]其中，本实施例模型集中的这多个待测试模型中包括至少一个神经网络模型。待测试模型是预先通过训练样本集训练好的。训练样本集可以包括样本和对应的已知输出结果。根据训练样本集中样本确定的输入特征参数和已知输出结果调整模型参数（如图 4 中的权重 1 等），直到模型对于训练样本集达到所要求的准确率，确定各个模型参数。尽管如此，由于训练样本集的局限性，还需要进一步通过测试样本集对模型进行测试。

[61]实践中，模型集可以是一个模型集合。在模型测试的方法的执行主体是一个测试类平台的情况下，该平台中可以包括模型操作模块，用户可以通过该模块添加待测试的模型，从而形成上述模型集合。虽然测试样本集中的样本可以适用于所有待测试模型，但是这些待测试模型可以具有相同的功能，也可以具有不相同的功能，根据测试的目的和具体情况而定。例如，对于同样的图片样本，一个待测试模型可以根据树木、花草的各种颜色外观特征识别品类等，另一个待测试模型可以根据树木、花草的颜色特征识别季节等。

[62]可以理解，每个待测试模型会根据输入的样本得到一个输出结果。对于同一个样本，每个待测试模型的输出结果可能会不同。例如，对于同一个图片样本，具有不同功能的模型，一个输出品类枫叶，一个输出季节秋季。对于功能相同的模型，由于模型准确性不一样，输出结果也不一样，例如对于同一个图片样本，一个输出春季，一个输出夏季。在一个实施例中，该多个待测试模型均包括相同功能。

[63]然后，在步骤 23，根据上述输出结果确定测试结果。可以理解，模型的测试结果通

过模型在输入样本后的输出结果来判定。而具体的测试结果是什么，根据测试内容是什么而定。

[64]具体地，根据本实施例的一个方面，测试内容可以是各个待测试模型的输出结果的准确性。为了便于描述，把各个待测试模型的输出结果的准确性称为第一测试结果。第一测试结果可以是对输出结果准确性的具体描述。例如，第一测试结果可以是各个待测试模型的结果准确、不准确中的一项，也可以是对待测试模型多次输出结果是否正确进行统计所计算的准确率（输出结果准确次数/总判断次数），等等。

[65]在一个实施例中，对上述多个待测试模型中包括的功能相同的待测试模型，投票确定正确输出结果，每个输出结果就是相应模型的投票结果，其中，票数最多的输出结果为正确输出结果；然后确定输出结果为所述正确输出结果的待测试模型输出结果准确。作为示例，假设模型集中有 10 个待检测模型，其中 7 个待检测模型可以用于树木、花草的颜色特征识别季节，则这 7 个待检测模型功能相同。对某个图片样本，在这 7 个待检测模型的输出结果分别是“春季”、“春季”、“夏季”、“秋季”、“春季”、“秋季”、“春季”的情况下，投票结果为“春季”4 票，“夏季”1 票，“秋季”2 票，则票数最多的“春季”为正确输出结果。然后，确定第一、第二、第五、第七个模型的输出结果准确，第三、第四、第六个模型的输出结果不准确。对另外 3 个待检测模型，如果他们的功能相同，在测试内容为输出结果的准确性的情况下，可以按照类似的方法处理，在此不再赘述。在一个可能的设计里，模型集中的全部待测试模型具有的功能相同。

[66]在另一个实施例中，模型集中可以包括有标准模型。此时，可以基于标准模型的输出结果，确定与标准模型功能相同的待测试模型的正确输出结果；确定输出结果为正确输出结果的待测试模型的输出结果准确。其中，标准模型可以是经过测试的准确度较高的模型，例如准确度 99.99%。由此，可以将标准模型的输出结果确定为正确输出结果。此时，对于人物图片样本，在标准模型确定的输出结果为“男士”的情况下，7 个与该标准模型功能都相同的待测试模型的输出结果有 4 个为“女士”，3 个为“男士”，则确定该输出结果为“男士”的 3 个待测试模型的输出结果准确。

[67]在一个可能的设计中，在与标准模型功能相同的待测试模型的输出结果与标准模型的输出结果都不一致时，还可以检测这些待测试模型的输出结果是否一致，如果一致，也可以将该一致的结果确定为正确输出结果。例如，对于人物图片样本，标准模型确定的输出结果为“男士”，7 个与该标准模型功能都相同的待测试模型的输出结果均为“女士”，则确定正确输出结果为“女士”。由于标准模型不是必然准确的模型，如此，可

以防止标准模型输出结果错误的情况下，导致对待测试模型的评判错误。

[68]对于与标准模型功能不同的其他待测试模型，也可以按照上述投票确定正确输出结果的方式确定测试结果，在此不再赘述。可选地，模型集中的全部待测试模型与该标准模型具有的功能相同。

- 5 [69]在一个实施例中，第一测试结果还可以是某个样本在测试中输出结果的准确性。例如某个样本，测试中有4个模型输出结果准确，3个模型输出结果不准确等等。可选地，还可以计算该样本的准确度。对于单次测试过程，样本的准确度=输出结果正确的模型数/总模型数。对于该样本的多次测试过程，准确度可以是单次准确度的平均值，也可以是被待测试模型输出正确结果的总次数/总共被待测试模型执行并给出输出结果的总次数。如此，可以对样本在模型测试过程中的有用程度进行评估。在一个样本每次进行测试，各个模型的输出结果都是正确的情况下，该样本对模型测试的意义不大。

- 15 [70]另一方面，测试内容可以是上述至少一个神经网络模型的神经元覆盖率。为了便于描述，把上述至少一个神经网络模型的神经元覆盖率称为第二测试结果。容易理解，模型的神经元覆盖率可以用来表示被激活的神经元数在总神经元数中的比例。此时，可以检测至少一个神经网络模型中各个神经元的输出是否满足输出条件，统计各个神经网络模型的神经元覆盖数，其中，神经元覆盖数包括，测试过程中满足过输出条件的神经元个数；基于神经元覆盖数确定各个神经网络模型的神经元覆盖率，其中，神经元覆盖率和神经元覆盖数与总神经元个数的比值正相关。

- 20 [71]可以理解，神经网络模型中，激励函数也叫点火规则，与人脑的工作类似，只要输入超过一定标准时才会产生输出。因此，在一个样本中，如果该样本的某些特征参数值比较小，则仅依靠这些特征参数的神经元将不会产生输出，即神经元未被激活。如此，可以通过设置阈值来作为神经元的输出条件，例如输出条件设置为神经元的输出大于0。

- 25 [72]值得说明的是，神经元覆盖数是测试过程中满足过输出条件的神经元个数。例如，模型1共有50个神经元，在样本1输入模型1的情况下，第1-5个神经元满足输出条件，此时，神经元覆盖数是5，神经元覆盖率为与5/50正相关的数，例如百分数10%。在样本2输入模型1的情况下，第2-6个神经元满足输出条件，因此满足过输出条件的神经元为第1-6个，此时，神经元覆盖数是6，神经元覆盖率为与6/50正相关的数，例如百分数12%。

[73]综上，该测试结果可以包括第一测试结果和第二测试结果中的至少一项，其中，第

一测试结果包括各个待测试模型的输出结果的准确性，第二测试结果包括上述至少一个神经网络模型的神经元覆盖率。

- [74]接着，在步骤 24，在上述测试结果未满足预定条件的情况下，按照预定规则，基于以上样本生成新样本，并将该新样本加入上述测试样本集。这里，可以存储有预定条件，
- 5 用于限定模型测试流程的结束时机。在该步骤中，可以进一步检测步骤 23 中的测试结果是否满足预定条件，满足预定条件则停止模型检测过程。在上述测试结果未满足预定条件的情况下，可以按照预定规则对样本进行处理，生成新样本，并将该新样本加入上述的测试样本集，以扩充测试样本集，用于后续的测试过程。其中，对样本的处理例如是按照预定规则（如梯度上升算法）改变各个特征的取值。
- 10 [75]在一个实施例中，上述预定条件基于神经元覆盖率确定。具体地，在测试结果包括第二测试结果的情况下，预定条件可以包括：至少一个神经网络模型的神经元覆盖率均大于预设覆盖率阈值，即如果待测试模型中包括一个神经网络模型，则要求该神经网络模型的神经元覆盖率大于预设覆盖率阈值；如果待测试模型中包括多个神经网络模型，则要求每个神经网络模型的神经元覆盖率均大于预设覆盖率阈值。预设覆盖率阈值可以
- 15 用来限定样本对于模型测试的覆盖程度，例如是 99%。当神经元覆盖率超过预设覆盖率阈值时，说明测试样本集中的样本已经可以充分地对该模型的绝大部分神经元进行测试。
- [76]在另一些实现中，还可以进一步基于测试结果准确度设定上述预定条件。具体地，预定条件还可以包括，多个待测试模型的输出结果准确度均大于预设准确度阈值，其中，准确度和历史测试结果准确的次数与总测试次数的比值正相关。作为示例，当前模型测
- 20 试中，模型 1 通过了 100 个样本的测试，其中输出结果准确的次数为 50，则其准确度与 50/100 正相关，例如是一个百分数 50%。假定预设准确度阈值为 90%，则在模型集中的各个待检测模型的准确度都大于 90%的情况下，可以结束测试。由于次数较少的测试对模型准确度的评估意义不大，比如该模型测试方法执行 1 次的时候，对于单个样本，模型的准确度要么是 0，要么是 100%。因此，在可选的实施例中，还可以限定准确度统
- 25 计的次数，例如对某个待测试模型，至少执行 100 次循环之后开始计算其准确度。
- [77]在一个实施例中，还可以对每个待测试模型，判断是否满足预定条件，如准确度达到预设准确度阈值，或者神经元覆盖率达到预设覆盖率阈值。在准确度达到预设准确度阈值的情况下，说明待测试模型的性能已经比较好了，继续测试的意义不大，可以直接从模型集中剔除。神经元覆盖率达到预设覆盖率阈值，则说明待测试模型通过当前的测
- 30 试已经可以充分发现其中缺陷或问题，也可以从模型集中剔除。如此，可以及时剔除测

试完毕的待测试模型，减少模型测试主体的数据处理压力。

[78]在一个实施例，模型测试的流程执行过程中，还可以随时加入新的待测试模型进行测试。如此，可以节约模型测试的执行资源。

[79]在继续执行测试的情况下，如前所述，可以按照预定规则对当前样本进行处理，生成新样本，并将其加入样本集。上述预定规则可以由测试人员根据需要而设定。这是考虑到，一方面，样本的各个特征并不是可以任意取值的，一些过于极端的取值会使样本失去意义，比如，人脸宽度超过人脸长度的2倍、人眼面积占人脸面积的四分之三，等等。另一方面测试人员还可以有针对性的产生特定条件下的样本，以进行更有针对性的测试。例如，通过设置预定规则，将当前图片样本中的卡车颜色设置为特定颜色，例如白色，从而测试无人驾驶识别时对于各种车辆颜色与背景的识别情况。

[80]在一个实施例，根据预定规则设置样本特征的约束范围。例如，将像素范围设定为 128×128 - 16384×16384 等。进一步地，基于各个特征的设定范围为样本添加随机扰动项，生成新样本。其中，随机扰动项可以只改变样本原有的特征，也可以为样本添加新的、与待测试模型的输入相关的特征。例如，一个待测试模型根据图片颜色识别季节，一个图片样本的各个像素点的颜色通过 RGB 值表示为 $[(235,108,0),(199, 98,3),(153,72,20), \dots]$ ，由于各个像素点的取值范围都在 0-255 之间，随机扰动项可以使每个像素点改变后的值都在 0-255 范围内，如 $[(0,1,30),(26, 0,3),(-153,72,20), \dots]$ ，将随机扰动项叠加到原样本生成新样本为 $[(235,109,30),(255, 98,6),(0,144,40), \dots]$ 。随机扰动项可以在像素范围内增加样本的像素点，例如原样本有 128×128 个像素点，随机扰动项可以增加 128×256 个像素点，生成新样本为 128×384 个像素点。

[81]如此，可以自动生成与人工收集的样本具有差异性的新样本，避免人工搜集的繁琐。进一步地，由于可以人工设定特征取值范围，从而生成的新样本比较符合真实情况。

[82]在一个实施例，对于已经用于对模型进行测试的样本，可以保留在测试样本集中以供后续测试过程中使用。在另一些可能的实施例，对于各个样本测试准确性都比较高的样本，用于模型测试的意义不大，还可以从测试样本集中移除。例如某个样本，在每次使用该样本进行测试的情况下，所有待测试模型都能给出正确输出结果，则该样本可以从测试样本集中移除。实践中，可以设定阈值，在准确度大于该设定阈值的情况下，将该样本从测试样本集中移除。

[83]回顾以上过程，首先从测试样本集中获取样本，接着将样本输入模型集包括的多个

待测试模型中,以获得各个待测试模型的输出结果,然后根据输出结果确定测试结果,进一步地,在该测试结果未满足预定条件的情况下,按照预定规则,基于上述样本生成新样本,并将所生成的新样本加入测试样本集。在该模型测试的方法被循环执行的情况下,在该模型测试的方法被循环执行的情况下,一方面对待检测模型的准确性能和/或测试充分程度进行评估,另一方面将基于原样本生成的、与原样本具有差异化的新样本。如此既能避免人工收集样本造成的人力浪费,又可以对模型进行进一步的测试,以及将扩充后的测试样本集用于训练模型从而提升模型的准确性。

[84]根据另一方面的实施例,还提供一种模型测试的装置。图5示出根据一个实施例的用于模型测试的装置的示意性框图。如图5所示,用于模型测试的装置500包括:获取单元51,配置为从测试样本集中获取样本;测试单元52,配置为将该样本输入模型集包括的多个待测试模型中,以获得各个待测试模型的输出结果,其中,该多个待测试模型包括至少一个神经网络模型;确定单元53,配置为根据上述输出结果确定测试结果,测试结果包括第一测试结果和第二测试结果中的至少一项,第一测试结果包括各个待测试模型的输出结果的准确性,第二测试结果包括上述至少一个神经网络模型的神经元覆盖率;生成单元54,配置为在测试结果未满足预定条件的情况下,按照预定规则,基于该样本生成新样本,并将该新样本加入上述测试样本集。

[85]根据一方面的实施方式,在测试结果包括第一测试结果的情况下,确定单元52还包括:结果确定模块(未示出),配置为对所述多个待测试模型中包括的功能相同的待测试模型,投票确定正确输出结果,其中,票数最高的输出结果为正确输出结果;准确性确定模块(未示出),配置为确定输出结果为正确输出结果的待测试模型输出结果准确。

[86]在一个可能的设计中,模型集中包括有标准模型,确定单元52包括:结果确定模块(未示出),配置为基于标准模型的输出结果,确定与标准模型功能相同的待测试模型的正确输出结果;准确性确定模块(未示出),配置为确定输出结果为正确输出结果的待测试模型输出准确。

[87]根据另一方面的实施方式,在测试结果包括第二测试结果的情况下,确定单元包括:统计模块(未示出),配置为根据上述至少一个神经网络模型中各个神经元的输出是否满足输出条件,统计各个神经网络模型的神经元覆盖数,其中,神经元覆盖数包括,测试过程中满足过输出条件的神经元个数;覆盖率确定模块,配置为基于神经元覆盖数确定各个神经网络模型的神经元覆盖率,其中,神经元覆盖率和以下比值正相关:神经元覆盖数与总神经元个数的比值。

[88]进一步地,在测试结果包括第二测试结果的情况下,预定条件包括,上述至少一个神经网络模型的神经元覆盖率均大于预设覆盖率阈值。

[89]更进一步地,预定条件还可以包括,上述多个待测试模型的输出结果准确度均大于预设准确度阈值。其中,该准确度可以和以下比值正相关:历史测试结果准确的次数与总测试次数的比值。

[90]在一个实施例中,生成模块 54 进一步配置为:获取上述样本中各个特征的设定范围;基于各个特征的设定范围为样本添加随机扰动项,生成新样本。

[91]在一种可能的设计中,装置 500 还包括以下中的一项或多项:

[92]剔除单元(未示出),配置为剔除模型集中测试结果满足预定条件的待测试模型;
10 添加单元(未示出),配置为在模型集中加入新的待测试模型。

[93]通过以上装置,可以不断从测试样本集中获取样本,循环执行模型测试的方法,产生更多差异化的样本,极大提高模型测试的有效性。并且,通过产生的差异化样本,还可以进一步用于训练模型从而提升模型的准确性。

[94]根据另一方面的实施例,还提供一种计算机可读存储介质,其上存储有计算机程序,
15 当所述计算机程序在计算机中执行时,令计算机执行结合图 2 所描述的方法。

[95]根据再一方面的实施例,还提供一种计算设备,包括存储器和处理器,所述存储器中存储有可执行代码,所述处理器执行所述可执行代码时,实现结合图 2 所述的方法。

[96]本领域技术人员应该可以意识到,在上述一个或多个示例中,本发明所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时,可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。
20

[97]以上所述的具体实施方式,对本发明的目的、技术方案和有益效果进行了进一步详细说明,所应理解的是,以上所述仅为本发明的具体实施方式而已,并不用于限定本发明的保护范围,凡在本发明的技术方案的基础之上,所做的任何修改、等同替换、改进等,均应包括在本发明的保护范围之内。
25