(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115147618 A
(43)申请公布日 2022.10.04

(21)申请号 202110277005.9
(22)申请日 2021.03.15
(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
(72)发明人 许雨顺 杨溢 刘静 邱禄瑜
(74)专利代理机构 深圳市深佳知识产权代理事务所(普通合伙) 44285
专利代理师 王仲凯
(51)Int.Cl.
G06V 10/46(2022.01)
G06V 10/774(2022.01)
G06V 10/82(2022.01)
G06V 10/764(2022.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

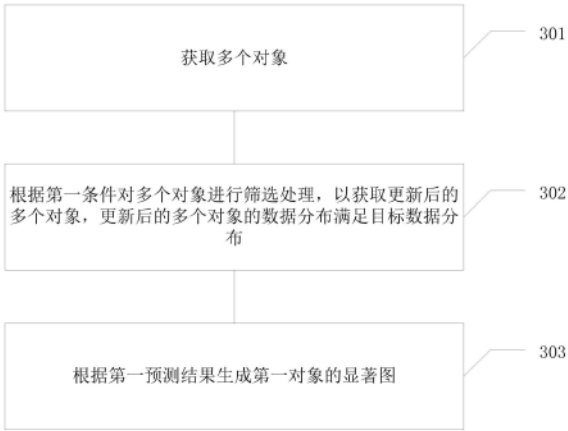
权利要求书5页 说明书28页 附图12页

(54)发明名称

一种生成显著图的方法、异常对象检测的方法以及装置

(57)摘要

本申请实施例涉及人工智能领域，公开了一种生成显著图的方法。方法包括：对第一对象进行扰动处理，以获取多个对象。对多个对象进行筛选，以获取更新后的多个对象。该更新后的多个对象满足目标数据分布，目标模型的训练样本也是满足目标数据分布的。将更新后的多个对象输入至目标模型中，以输出第一预测结果。根据第一预测结果和更新后的多个对象生成第一对象的显著图，以提升显著图的准确性。



CN 115147618 A

1.一种生成显著图的方法,其特征在于,包括:

获取多个对象,所述多个对象是对第一对象进行扰动处理后获取的;

根据第一条件对所述多个对象进行筛选处理,以获取更新后的所述多个对象,所述更新后的多个对象满足目标数据分布,所述目标数据分布是根据训练样本获取的,所述训练样本用于对预设模型进行训练,以得到目标模型;

基于所述更新后的多个对象获取所述目标模型的输入;

根据所述目标模型输出的第一预测结果和所述更新后的多个对象,生成所述第一对象的显著图。

2.根据权利要求1所述的方法,其特征在于,所述第一条件为删除所述多个对象中的目标对象,所述目标对象的特征和所述目标模型的权重向量之间的距离超过预设阈值,所述目标对象的特征是通过所述目标模型对所述目标对象进行特征提取后获取的。

3.根据权利要求2所述的方法,其特征在于,所述目标对象的特征具体是通过第一特征提取层提取的,所述第一特征提取层是所述目标模型中多个特征提取层中的任意一层特征提取层,所述目标对象的特征和所述目标模型的权重向量之间的距离具体是所述目标对象的特征和第二特征提取层的权重向量之间的距离,所述第二特征提取层是所述多个特征提取中的任意一层特征提取层。

4.根据权利要求3所述的方法,其特征在于,所述第一特征提取层和所述第二特征提取层是不同的特征提取层。

5.根据权利要求3或4所述的方法,其特征在于,所述第一特征提取层具体是所述特征提取模型中多个特征提取层中的倒数第二层特征提取层,所述多个特征提取层首尾相连,所述目标对象是所述多个特征提取层中的第一层特征提取层的输入,所述第二特征提取层具体是所述多个特征提取中的最后一层特征提取层。

6.根据权利要求3至5任一项所述的方法,其特征在于,所述目标对象的特征和所述第二特征提取层的目标权重向量之间的距离超过所述预设阈值,所述目标对象的特征和所述第二特征提取层的目标权重向量之间的距离是多个距离中的最大距离,所述多个距离包括所述目标对象的特征和所述第二特征提取层的每个所述权重向量之间的距离。

7.根据权利要求1至6任一项所述的方法,其特征在于,所述目标模型是通过第一损失值更新所述预设模型获取的,所述第一损失值是根据所述训练样本的特征和所述预设模型的权重向量之间的偏差确定的,所述训练样本的特征是所述预设模型对所述训练样本进行特征提取后获取的。

8.根据权利要求7所述的方法,其特征在于,所述目标模型具体是通过所述第一损失值和第二损失值更新所述预设模型后获取的,所述第二损失值是根据目标结果和所述训练样本的真实结果之间的偏差确定的,所述目标结果是根据第二预测结果和预设函数确定的,所述第二预测结果是所述预设模型针对所述训练样本的预测结果,所述预设函数的输入是所述第二预测结果,所述预设函数的输出是所述目标结果,所述预设函数的输出和所述预设函数的输入负相关。

9.根据权利要求1至7任一项所述的方法,其特征在于,所述方法还包括:

设置所述更新后的多个对象的权重为第一权重;

设置剩余的多个对象的权重为第二权重,所述剩余的多个对象是所述多个对象中除了

所述更新后的多个对象之外的对象，所述第一权重大于所述第二权重；

所述基于所述更新后的多个对象获取所述目标模型的输入，包括：

根据所述第一权重和所述更新后的多个对象获取第一结果，所述第一结果为所述目标模型的输入，所述目标模型的输入还包括第二结果，所述第二结果是根据所述第二权重和所述剩余的多个对象获取的。

10.一种异常对象检测的方法，其特征在于，包括：

获取多个对象；

通过特征提取模型对目标对象进行特征提取，以获取所述目标对象的特征，所述目标对象是所述多个对象中的任意一个所述对象，所述特征提取模型是通过第一损失值更新预设模型后获取的，所述第一损失值是根据训练样本的特征和所述预设模型的权重向量之间的偏差确定的，所述训练样本的特征是所述预设模型对所述训练样本进行特征提取后获取的；

获取所述目标对象的特征和所述特征提取模型的权重向量之间的距离，其中，在所述距离超过预设阈值的情况下，所述目标对象为异常对象。

11.根据权利要求10所述的方法，其特征在于，所述通过特征提取模型对目标对象进行特征提取，以获取所述目标对象的特征，包括：

通过第一特征提取层对所述目标对象进行特征提取，以获取所述目标对象的特征，所述第一特征提取层是所述特征提取模型中多个特征提取层中的任意一层特征提取层；

所述获取所述目标对象的特征和所述特征提取模型的权重向量之间的距离，包括：

获取所述目标对象的特征和第二特征提取层的权重向量之间的距离，所述第二特征提取层是所述多个特征提取中的任意一层特征提取层。

12.根据权利要求11所述的方法，其特征在于，所述第一特征提取层和所述第二特征提取层是不同的特征提取层。

13.根据权利要求11或12所述的方法，其特征在于，所述第一特征提取层是所述特征提取模型中多个特征提取层中的倒数第二层特征提取层，所述多个特征提取层首尾相连，所述目标对象是所述多个特征提取层中的第一层特征提取层的输入，所述第二特征提取层是所述多个特征提取中的最后一层特征提取层。

14.根据权利要求11至13任一项所述的方法，其特征在于，所述第二特征提取层的权重向量为多个，所述获取所述目标对象的特征和第二特征提取层的权重向量之间的距离，包括：

获取所述目标对象的特征和所述第二特征提取层的每个所述权重向量之间的距离，其中，在多个所述距离中的最大距离超过所述预设阈值的情况下，所述目标对象为异常样本。

15.根据权利要求10至14任一项所述的方法，其特征在于，所述特征提取模型具体是通过所述第一损失值和第二损失值更新所述预设模型后获取的，所述第二损失值是根据目标结果和所述训练样本的真实结果之间的偏差确定的，所述目标结果是根据第一预测结果和预设函数确定的，所述第一预测结果是所述预设模型针对所述训练样本的预测结果，所述预设函数的输入是所述第一预测结果，所述预设函数的输出是所述目标结果，所述预设函数的输出和所述预设函数的输入负相关。

16.根据权利要求10至15任一项所述的方法，其特征在于，所述多个对象是对第一图像

扰动处理后获取的，所述方法还包括：

从所述多个对象中删除所述目标对象，以获取更新后的所述多个对象，所述更新后的多个对象用于获取所述第一图像的显著图。

17.根据权利要求10至15任一项所述的方法，其特征在于，所述多个对象是对第一图像扰动处理后获取的，所述方法还包括：

若所述距离不超过所述预设阈值，则确定所述目标对象为正常对象；

若所述目标对象为异常对象，则设定所述目标对象的权重为第一权重；

若所述目标对象为正常对象，则设定所述目标对象的权重为第二权重，所述第二权重大于所述第一权重；

根据所述第一权重或者第二权重，对所述目标对象的特征进行处理，以获取处理后的所述目标对象，所述处理后的目标对象用于获取所述第一图像的显著图。

18.一种生成显著图的装置，其特征在于，包括：

获取模块，用于获取多个对象，所述多个对象是对第一对象进行扰动处理后获取的；

筛选模块，用于根据第一条件对所述获取模块获取的所述多个对象进行筛选处理，以获取更新后的所述多个对象，所述更新后的多个对象满足目标数据分布，所述目标数据分布是根据训练样本获取的，所述训练样本用于对预设模型进行训练，以得到目标模型；

生成模块，用于根据所述目标模型输出的第一预测结果和所述更新后的多个对象，生成所述第一对象的显著图，所述目标模型的输入是基于所述更新后的多个对象获取的。

19.根据权利要求18所述的装置，其特征在于，所述第一条件为删除所述多个对象中的目标对象，所述目标对象的特征和所述目标模型的权重向量之间的距离超过预设阈值，所述目标对象的特征是通过所述目标模型对所述目标对象进行特征提取后获取的。

20.根据权利要求19所述的装置，其特征在于，所述目标对象的特征具体是通过第一特征提取层提取的，所述第一特征提取层是所述目标模型中多个特征提取层中的任意一层特征提取层，所述目标对象的特征和所述目标模型的权重向量之间的距离具体是所述目标对象的特征和第二特征提取层的权重向量之间的距离，所述第二特征提取层是所述多个特征提取中的任意一层特征提取层。

21.根据权利要求20所述的装置，其特征在于，所述第一特征提取层和所述第二特征提取层是不同的特征提取层。

22.根据权利要求20或21所述的装置，其特征在于，所述第一特征提取层具体是所述特征提取模型中多个特征提取层中的倒数第二层特征提取层，所述多个特征提取层首尾相连，所述目标对象是所述多个特征提取层中的第一层特征提取层的输入，所述第二特征提取层具体是所述多个特征提取中的最后一层特征提取层。

23.根据权利要求20至22任一项所述的装置，其特征在于，所述目标对象的特征和所述第二特征提取层的目标权重向量之间的距离超过所述预设阈值，所述目标对象的特征和所述第二特征提取层的目标权重向量之间的距离是多个距离中的最大距离，所述多个距离包括所述目标对象的特征和所述第二特征提取层的每个所述权重向量之间的距离。

24.根据权利要求18至23任一项所述的装置，其特征在于，所述目标模型是通过第一损失值更新所述预设模型获取的，所述第一损失值是根据所述训练样本的特征和所述预设模型的权重向量之间的偏差确定的，所述训练样本的特征是所述预设模型对所述训练样本进

行特征提取后获取的。

25.根据权利要求24所述的装置，其特征在于，所述目标模型具体是通过所述第一损失值和第二损失值更新所述预设模型后获取的，所述第二损失值是根据目标结果和所述训练样本的真实结果之间的偏差确定的，所述目标结果是根据第二预测结果和预设函数确定的，所述第二预测结果是所述预设模型针对所述训练样本的预测结果，所述预设函数的输入是所述第二预测结果，所述预设函数的输出是所述目标结果，所述预设函数的输出和所述预设函数的输入负相关。

26.根据权利要求18至25任一项所述的装置，其特征在于，还包括权重模块，

所述权重模块，用于:

设置所述更新后的多个对象的权重为第一权重；

设置剩余的多个对象的权重为第二权重，所述剩余的多个对象是所述多个对象中除了所述更新后的多个对象之外的对象，所述第一权重大于所述第二权重；

所述目标模型的一个输入具体是基于所述更新后的多个对象和所述第一权重获取的，所述目标模型的另一个输入是基于所述剩余的多个对象和所述第二权重获取的。

27.一种异常对象检测的装置，其特征在于，包括：

第一获取模块，用于获取多个对象；

特征提取模块，用于通过特征提取模型对目标对象进行特征提取，以获取所述目标对象的特征，所述目标对象是所述多个对象中的任意一个所述对象，所述特征提取模型是通过第一损失值更新预设模型后获取的，所述第一损失值是根据训练样本的特征和所述预设模型的权重向量之间的偏差确定的，所述训练样本的特征是所述预设模型对所述训练样本进行特征提取后获取的；

第二获取模块，用于获取所述目标对象的特征和所述特征提取模型的权重向量之间的距离，其中，在所述距离超过预设阈值的情况下，所述目标对象为异常对象。

28.根据权利要求27所述的装置，其特征在于，所述特征提取模块，具体用于：

通过第一特征提取层对所述目标对象进行特征提取，以获取所述目标对象的特征，所述第一特征提取层是所述特征提取模型中多个特征提取层中的任意一层特征提取层；

所述获取所述目标对象的特征和所述特征提取模型的权重向量之间的距离，包括：

获取所述目标对象的特征和第二特征提取层的权重向量之间的距离，所述第二特征提取层是所述多个特征提取中的任意一层特征提取层。

29.根据权利要求28所述的装置，其特征在于，所述第一特征提取层和所述第二特征提取层是不同的特征提取层。

30.根据权利要求28或29所述的装置，其特征在于，所述第一特征提取层是所述特征提取模型中多个特征提取层中的倒数第二层特征提取层，所述多个特征提取层首尾相连，所述目标对象是所述多个特征提取层中的第一层特征提取层的输入，所述第二特征提取层是所述多个特征提取中的最后一层特征提取层。

31.根据权利要求28至30任一项所述的装置，其特征在于，所述第二获取模块，具体用于：

获取所述目标对象的特征和所述第二特征提取层的每个所述权重向量之间的距离，其中，在多个所述距离中的最大距离超过所述预设阈值的情况下，所述目标对象为异常样本。

32.根据权利要求27至31任一项所述的装置，其特征在于，所述特征提取模型具体是通过所述第一损失值和第二损失值更新所述预设模型后获取的，所述第二损失值是根据目标结果和所述训练样本的真实结果之间的偏差确定的，所述目标结果是根据第一预测结果和预设函数确定的，所述第一预测结果是所述预设模型针对所述训练样本的预测结果，所述预设函数的输入是所述第一预测结果，所述预设函数的输出是所述目标结果，所述预设函数的输出和所述预设函数的输入负相关。

33.根据权利要求27至32任一项所述的装置，其特征在于，所述多个对象是对第一图像扰动处理后获取的，所述装置还包括生成模块，所述生成模块，用于：

从所述多个扰动后的第一图像中删除所述目标对象，以获取更新后的所述多个对象，所述更新后的多个对像用于获取所述第一图像的显著图。

34.根据权利要求27至32任一项所述的装置，其特征在于，所述多个对象是对第一图像扰动处理后获取的，在所述距离不超过预设阈值的情况下，所述目标对象为正常对象所述装置还包括权重模块和生成模块，

所述权重模块，用于：

若所述目标对象为异常对象，则设定所述目标对象的权重为第一权重；

若所述目标对象为正常对象，则设定所述目标对象的权重为第二权重，所述第二权重大于所述第一权重；

所述生成模块，用于根据所述第一权重或者第二权重，对所述目标对象的特征进行处理，以获取处理后的所述目标对象，所述处理后的目标对象用于获取所述第一图像的显著图。

35.一种生成显著图的装置，其特征在于，包括处理器，所述处理器和存储器耦合，所述存储器存储有程序指令，当所述存储器存储的程序指令被所述处理器执行时实现权利要求1至9中任一项所述的方法。

36.一种异常对象检测的装置，其特征在于，包括处理器，所述处理器和存储器耦合，所述存储器存储有程序指令，当所述存储器存储的程序指令被所述处理器执行时实现权利要求10至17中任一项所述的方法。

37.一种计算机可读存储介质，其特征在于，包括程序，当其在计算机上运行时，使得计算机执行如权利要求1至9中任一项所述的方法，或者，使得计算机执行如权利要求10至17中任一项所述的方法。

38.一种电路系统，其特征在于，所述电路系统包括处理电路，所述处理电路配置为执行如权利要求1至9中任一项所述的方法，或者配置为执行如权利要求10至17中任一项所述的方法。

39.一种计算机程序产品，其特征在于，所述计算机程序产品包括指令，当所述指令由电子设备加载并执行，使电子设备执行权利要求1至9中任一项所述的方法，或者使得电子设备执行权利要求10至17中任一项所述的方法。

40.一种芯片，其特征在于，所述芯片与存储器耦合，用于执行所述存储器中存储的程序，以执行如权利要求1至9任一项所述的方法，或者执行如权利要求10至17中任一项所述的方法。

一种生成显著图的方法、异常对象检测的方法以及装置

技术领域

[0001] 本申请涉及人工智能领域，具体涉及一种生成显著图的方法、异常对象检测的方法以及装置。

背景技术

[0002] 人工智能(artificial intelligence，AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式作出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

[0003] AI最广为人知的缺点是“黑盒”性质，这意味着用户不知道模型如何以及为何会得出一定的输出。例如，当用户将一张猫的图像输入分类模型中，分类模型预测这是汽车时，很难理解为什么会导致它产生这个预测。在过去的几年里，AI在不同领域的许多应用中取得显著成就。但是，随着人们越来越依赖机器学习模型，来自这些模型的决策影响到人类的生活，如何确保AI做出的决定是可信的?这是当前大数据与深度学习为基础的人工智能存在的最大问题：不可解释。换言之，模型缺乏透明性和可解释性，这将严重影响用户对其的信任程度，进而限制模型在现实任务(尤其是风险敏感类任务，如无人驾驶、医疗保健、金融类任务等)中的应用和发展。

[0004] 因为黑盒模型不具备先天的可解释性，所以需要借助显著图(saliency map)获取哪些特征对模型输出的影响最大。因此，如何使显著图能准确反映特征对模型输出的影响亟待解决。

发明内容

[0005] 本申请提供一种生成显著图的方法、异常对象检测的方法以及装置，可以提升显著图的准确度。

[0006] 为解决上述技术问题，本申请实施例提供以下技术方案：

[0007] 第一方面，本申请提供一种生成显著图的方法，方法包括：获取多个对象，多个对象是对第一对象进行扰动处理后获取的。其中，对第一对象进行扰动处理可以包括多种方式，比如可以对第一对象进行加噪处理；或者可以对第一对象进行模糊处理(比如对多个像素的取值在预设范围内进行更改)；或者可以对第一对象进行遮挡处理(比如随机抽样多个像素点进行遮挡)等等。第一对象可以是图像数据、文本数据或者语音数据。根据第一条件对多个对象进行筛选处理，以获取更新后的多个对象，更新后的多个对象满足目标数据分布，目标数据分布是根据训练样本获取的。换言之，更新后的多个对象满足的数据分布和训练样本满足的数据分布是一致的。更新后的多个对象中的每个对象(或者对每个对象进行特征提取后的对象特征)可以通过空间坐标系中的一个坐标进行表示。在空间坐标系下，包括所有对象的坐标的空间可以看做是更新后的多个对象满足的数据分布(以下称为第一数

据分布)。每个训练样本(或者对每个训练进行特征提取后的训练样本特征)也可以通过空间坐标系中的一个坐标进行表示。在空间坐标系下,包括所有训练样本的坐标的空间可以看做该训练样本满足的数据分布。当第一数据分布和第二数据分布之间的偏差在预设范围内时,可以认为第一数据分布和第二数据分布是一致的。在本申请的第一方面提供的方案中,第一数据分布和第二数据分布都是目标数据分布。为了更好的理解,这里结合一个具体的例子进行说明:假设训练样本都是动物类的图像,这些动物类的图像满足目标数据分布,目标数据分布可以用于体现这些动物类图像(或者图像特征)的特点。训练后的模型的输入也应当满足目标数据分布,比如输入也是动物类的图像(模型曾经学习过该输入对象的特征),则训练后的模型可以针对输入进行准确的预测。但是,如果输入是人物类的图像,显然,人物类的图像和动物类的图像具有不同的特点,在空间坐标系下的坐标的分布也应当是不同的,则通过动物类图像训练后的模型,对输入的人物类图像的预测结果的准确率会大大降低。

[0008] 本申请提供的方案发现了扰动后的第一对象可能并不满足目标数据分布,所以对扰动后的第一对象进行筛选处理,获取满足目标数据分布的扰动后的第一对象。将更新后的多个对象输入至目标模型中,以输出第一预测结果。根据第一预测结果和更新后的多个对象生成第一对象的显著图。第一方面提供的方案是对已有的基于扰动的可解释方法进行的改进,通过基于扰动的可解释方法生成的显著图可以获取第一对象的哪些特征对模型输出的影响最大,进而可以对黑盒模型进行解释。

[0009] 本申请提供的方案可以提升输入是图像数据的模型的可解释性,比如提升图像分类模型的可解释性,通过第一方面获取的显著图可以更好的理解输入至图像分类模型的图像的哪部分特征对图像分类模型的输出影响大。本申请提供的方案可以提升输入是语音数据的模型的可解释性,比如提升多轮对话模型的可解释性,通过第一方面获取的显著图可以更好的解释多轮对话模型是基于输入语音的哪部分特征,与用户进行交互。本申请提供的方案可以提升输入是文本数据的模型的可解释性,比如提升机器翻译的可解释性,通过第一方面获取的显著图可以更好的理解输入的文本的哪部分特征对翻译模型的输出影响大。

[0010] 在一种可能的实施方式中,目标模型是通过第一损失值更新预设模型获取的,第一损失值是根据训练样本的特征和预设模型的权重向量之间的偏差确定的,训练样本的特征是预设模型对训练样本进行特征提取后获取的,根据第一条件对多个对象进行筛选处理,以获取更新后的多个对象,包括:若多个对象中包括目标对象,则删除多个对象中的目标对象,以获取更新后的多个对象,目标对象的特征和目标模型的权重向量之间的距离超过预设阈值,目标对象的特征是通过目标模型对目标对象进行特征提取后获取的。在这种实施方式中,随着对预设模型的训练,第一损失值会越来越小,则预设模型的权重向量与训练样本的特征也会越来越接近。训练后的预设模型,即目标模型的权重向量已经可以反映训练样本的数据分布特征。在这种实施方式中,只需要通过正常样本对分类模型进行训练,就可以使分类模型区分正常样本与异常样本,降低了训练分类模型的训练难度以及训练成本,同时还可以提升分类模型的性能。

[0011] 在一种可能的实施方式中,目标对象的特征具体是通过第一特征提取层提取的,第一特征提取层是目标模型中多个特征提取层中的任意一层特征提取层,目标对象的特征

和目标模型的权重向量之间的距离具体是目标对象的特征和第二特征提取层的权重向量之间的距离，第二特征提取层是多个特征提取中的任意一层特征提取层。在这种实施方式中，可以通过任意一层特征提取层提取目标对象的特征与任意一层特征提取层的权重向量之间的距离，表示目标对象的特征和目标模型的权重向量之间的距离。

[0012] 在一种可能的实施方式中，可以根据实际应用的需求设定第一特征提取层和第二特征提取层是相同的特征提取层，还可以根据实际应用的需求设定第一特征提取层和第二特征提取层是不同的特征提取层。

[0013] 在一种可能的实施方式中，第一特征提取层具体是特征提取模型中多个特征提取层中的倒数第二层特征提取层，多个特征提取层首尾相连，目标对象是多个特征提取层中的第一层特征提取层的输入，第二特征提取层具体是多个特征提取中的最后一层特征提取层。在这种实施方式中，给出了一种优选的第一特征提取层和优选的第二特征提取层，增加了方案的多样性。

[0014] 在一种可能的实施方式中，若第二特征提取层的权重向量包括多个，目标对象的特征和第二特征提取层的目标权重向量之间的距离超过预设阈值，目标对象的特征和第二特征提取层的目标权重向量之间的距离是多个距离中的最大距离，多个距离包括目标对象的特征和第二特征提取层的每个权重向量之间的距离。在这种实施方式中，为了更好的识别异常样本，在这种实施方式中，根据多个距离中的最大距离判断目标对象是否为异常样本。

[0015] 在一种可能的实施方式中，目标模型具体是通过第一损失值和第二损失值更新预设模型后获取的，第二损失值是根据目标结果和训练样本的真实结果之间的偏差确定的，目标结果是根据第二预测结果和预设函数确定的，第二预测结果是预设模型针对训练样本的预测结果，预设函数的输入是第二预测结果，预设函数的输出是目标结果，预设函数的输出和预设函数的输入负相关。在这种实施方式中，在模型的训练过程中引入预设函数，该预设函数的输入和输出负相关，根据预设函数的不同输入，选择不同的输出，无需手动调节参数，可以自动根据不同的输入获取不同的输出，避免模型陷入局部最优，减缓模型收敛的速度，使训练后的模型的性能更优。

[0016] 在一种可能的实施方式中，方法还包括：设置更新后的多个对象的权重为第一权重。设置剩余的多个对象的权重为第二权重，剩余的多个对象是多个对象中除了更新后的多个对象之外的对象，第一权重大于第二权重。将第一结果和第二结果输入至目标模型中，以输出第一预测结果，第一结果是根据第一权重和更新后的多个对象确定的，第二结果是根据第二权重和剩余的多个对象确定的。在这种实施方式中，将正常样本设定更大的权重，将异常样本设定更小的权重，在生成显著图的过程中，以正常样本为主，异常样本为辅，削弱异常样本在生成显著图的过程中造成的影响。使生成的显著图能够能较好的反应特征对模型的输出的影响。

[0017] 第二方面，本申请提供一种异常对象检测的方法，包括：获取多个对象。通过特征提取模型对目标对象进行特征提取，以获取目标对象的特征，目标对象是多个对象中的任意一个对象，特征提取模型是通过第一损失值更新预设模型后获取的，第一损失值是根据训练样本的特征和预设模型的权重向量之间的偏差确定的，训练样本的特征是预设模型对训练样本进行特征提取后获取的。获取目标对象的特征和特征提取模型的权重向量之间的

距离。若距离超过预设阈值，则确定目标对象为异常对象。在这种实施方式中，可以通过第一损失值对预设模型进行更新，再基于更新后的预设模型再更新第一损失值，从而再更新预设模型，随着更新次数的增加，预设模型的权重向量与训练样本的特征也会越来越接近。训练后的预设模型，即目标模型的权重向量已经可以反映训练样本的数据分布特征。在这种实施方式中，只需要通过正常样本对分类模型进行训练，就可以使分类模型区分正常样本与异常样本，降低了训练分类模型的训练难度以及训练成本，同时还可以提升分类模型的性能。

[0018] 在一种可能的实施方式中，通过特征提取模型对目标对象进行特征提取，以获取目标对象的特征，包括：通过第一特征提取层对目标对象进行特征提取，以获取目标对象的特征，第一特征提取层是特征提取模型中多个特征提取层中的任意一层特征提取层。获取目标对象的特征和特征提取模型的权重向量之间的距离，包括：获取目标对象的特征和第二特征提取层的权重向量之间的距离，第二特征提取层是多个特征提取中的任意一层特征提取层。在这种实施方式中，可以通过任意一层特征提取层提取目标对象的特征与任意一层特征提取层的权重向量之间的距离，表示目标对象的特征和目标模型的权重向量之间的距离。

[0019] 在一种可能的实施方式中，第一特征提取层和第二特征提取层是同一层特征提取层或者是不同的特征提取层。

[0020] 在一种可能的实施方式中，第一特征提取层是特征提取模型中多个特征提取层中的倒数第二层特征提取层，多个特征提取层首尾相连，目标对象是多个特征提取层中的第一层特征提取层的输入，第二特征提取层是多个特征提取中的最后一层特征提取层。在这种实施方式中，给出了一种优选的第一特征提取层和优选的第二特征提取层，增加了方案的多样性。

[0021] 在一种可能的实施方式中，获取目标对象的特征和第二特征提取层的权重向量之间的距离，包括：若第二特征提取层的权重向量包括多个，则获取目标对象的特征和第二特征提取层的每个权重向量之间的距离。若距离超过预设阈值，则确定目标对象为异常对象，包括：若多个距离中的最大距离超过预设阈值，则确定目标对象为异常样本。在这种实施方式中，为了更好的识别异常样本，在这种实施方式中，根据多个距离中的最大距离判断目标对象是否为异常样本。

[0022] 在一种可能的实施方式中，特征提取模型具体是通过第一损失值和第二损失值更新预设模型后获取的，第二损失值是根据目标结果和训练样本的真实结果之间的偏差确定的，目标结果是根据第一预测结果和预设函数确定的，第一预测结果是预设模型针对训练样本的预测结果，预设函数的输入是第一预测结果，预设函数的输出是目标结果，预设函数的输出和预设函数的输入负相关。在这种实施方式中，在模型的训练过程中引入预设函数，该预设函数的输入和输出负相关，根据预设函数的不同输入，选择不同的输出，无需手动调节参数，可以自动根据不同的输入获取不同的输出，避免模型陷入局部最优，减缓模型收敛的速度，使训练后的模型的性能更优。

[0023] 在一种可能的实施方式中，多个对象是对同一个第一图像进行扰动处理后获取的(多个对象可以通过对同一个图像进行不同的扰动获取)，方法还包括：若距离超过预设阈值，则从多个对象中删除目标对象，以获取更新后的多个对象，更新后的多个对象用于获取

该第一图像的显著图。在这种实施方式中，对多个对象进行筛选处理，保留满足目标数据分布的扰动后的多个对象，利用更新后的多个对象获取第一图像的显著图，提升显著图的准确率。

[0024] 在一种可能的实施方式中，多个对象是对同一个第一图像进行扰动处理后获取的(多个对象可以通过对同一个图像进行不同的扰动获取)，方法还包括：若距离不超过预设阈值，则确定目标对象为正常对象。若目标对象为异常对象，则设定目标对象的权重为第一权重。若目标对象为正常对象，则设定目标对象的权重为第二权重，第二权重大于第一权重。根据第一权重或者第二权重，对目标对象的特征进行处理，以获取处理后的目标对象，处理后的目标对象用于获取第一图像的显著图。在这种实施方式中，将正常样本设定更大的权重，将异常样本设定更小的权重，在生成显著图的过程中，以正常样本为主，异常样本为辅，削弱异常样本在生成显著图的过程中造成的影响。使生成的显著图能够能较好的反应特征对模型的输出的影响。

[0025] 第三方面，本申请提供一种生成显著图的装置，包括：获取模块，用于获取多个对象，多个对象是对第一对象进行扰动处理后获取的。筛选模块，用于根据第一条件对获取模块获取的多个对象进行筛选处理，以获取更新后的多个对象，更新后的多个对象满足目标数据分布，目标数据分布是根据训练样本获取的，训练样本用于对预设模型进行训练，以得到目标模型。预测模块，用于将筛选模块获取的更新后的多个对象输入至目标模型中，以输出第一预测结果。生成模块，用于根据预测模块获取的第一预测结果和更新后的多个对象生成第一对象的显著图。

[0026] 在一种可能的实施方式中，目标模型是通过第一损失值更新预设模型获取的，第一损失值是根据训练样本的特征和预设模型的权重向量之间的偏差确定的，训练样本的特征是预设模型对训练样本进行特征提取后获取的，筛选模块，具体用于：若多个对象中包括目标对象，则删除多个对象中的目标对象，以获取更新后的多个对象，目标对象的特征和目标模型的权重向量之间的距离超过预设阈值，目标对象的特征是通过目标模型对目标对象进行特征提取后获取的。

[0027] 在一种可能的实施方式中，目标对象的特征具体是通过第一特征提取层提取的，第一特征提取层是目标模型中多个特征提取层中的任意一层特征提取层，目标对象的特征和目标模型的权重向量之间的距离具体是目标对象的特征和第二特征提取层的权重向量之间的距离，第二特征提取层是多个特征提取中的任意一层特征提取层。

[0028] 在一种可能的实施方式中，可以根据实际应用的需求设定第一特征提取层和第二特征提取层是相同的特征提取层，还可以根据实际应用的需求设定第一特征提取层和第二特征提取层是不同的特征提取层。

[0029] 在一种可能的实施方式中，第一特征提取层具体是特征提取模型中多个特征提取层中的倒数第二层特征提取层，多个特征提取层首尾相连，目标对象是多个特征提取层中的第一层特征提取层的输入，第二特征提取层具体是多个特征提取中的最后一层特征提取层。

[0030] 在一种可能的实施方式中，若第二特征提取层的权重向量包括多个，目标对象的特征和第二特征提取层的目标权重向量之间的距离超过预设阈值，目标对象的特征和第二特征提取层的目标权重向量之间的距离是多个距离中的最大距离，多个距离包括目标对象

的特征和第二特征提取层的每个权重向量之间的距离。

[0031] 在一种可能的实施方式中，目标模型具体是通过第一损失值和第二损失值更新预设模型后获取的，第二损失值是根据目标结果和训练样本的真实结果之间的偏差确定的，目标结果是根据第二预测结果和预设函数确定的，第二预测结果是预设模型针对训练样本的预测结果，预设函数的输入是第二预测结果，预设函数的输出是目标结果，预设函数的输出和预设函数的输入负相关。

[0032] 在一种可能的实施方式中，还包括权重模块，权重模块，用于:设置更新后的多个对象的权重为第一权重。设置剩余的多个对象的权重为第二权重，剩余的多个对象是多个对象中除了更新后的多个对象之外的对象，第一权重大于第二权重。预测模块，具体用于将第一结果和第二结果输入至目标模型中，以输出第一预测结果，第一结果是根据第一权重和更新后的多个对象确定的，第二结果是根据第二权重和剩余的多个对象确定的。

[0033] 对于本申请第三方面以及各种可能实现方式的具体实现步骤，以及每种可能实现方式所带来的有益效果，均可以参考第一方面中各种可能的实现方式中的描述，此处不再一一赘述。

[0034] 第四方面，本申请提供一种异常对象检测的装置，包括:第一获取模块，用于获取多个对象。特征提取模块，用于通过特征提取模型对目标对象进行特征提取，以获取目标对象的特征，目标对象是多个对象中的任意一个对象，特征提取模型是通过第一损失值更新预设模型后获取的，第一损失值是根据训练样本的特征和预设模型的权重向量之间的偏差确定的，训练样本的特征是预设模型对训练样本进行特征提取后获取的。第二获取模块，用于获取目标对象的特征和特征提取模型的权重向量之间的距离。异常检测模块，用于若距离超过预设阈值，则确定目标对象为异常对象。

[0035] 在一种可能的实施方式中，特征提取模块，具体用于:通过第一特征提取层对目标对象进行特征提取，以获取目标对象的特征，第一特征提取层是特征提取模型中多个特征提取层中的任意一层特征提取层。获取目标对象的特征和特征提取模型的权重向量之间的距离，包括:获取目标对象的特征和第二特征提取层的权重向量之间的距离，第二特征提取层是多个特征提取中的任意一层特征提取层。

[0036] 在一种可能的实施方式中，可以根据实际应用的需求设定第一特征提取层和第二特征提取层是相同的特征提取层，还可以根据实际应用的需求设定第一特征提取层和第二特征提取层是不同的特征提取层。

[0037] 在一种可能的实施方式中，第一特征提取层是特征提取模型中多个特征提取层中的倒数第二层特征提取层，多个特征提取层首尾相连，目标对象是多个特征提取层中的第一层特征提取层的输入，第二特征提取层是多个特征提取中的最后一层特征提取层。

[0038] 在一种可能的实施方式中，第二获取模块，具体用于:若第二特征提取层的权重向量包括多个，则获取目标对象的特征和第二特征提取层的每个权重向量之间的距离。异常检测模块，具体用于若多个距离中的最大距离超过预设阈值，则确定目标对象为异常样本。

[0039] 在一种可能的实施方式中，特征提取模型具体是通过第一损失值和第二损失值更新预设模型后获取的，第二损失值是根据目标结果和训练样本的真实结果之间的偏差确定的，目标结果是根据第一预测结果和预设函数确定的，第一预测结果是预设模型针对训练样本的预测结果，预设函数的输入是第一预测结果，预设函数的输出是目标结果，预设函数

的输出和预设函数的输入负相关。

[0040] 在一种可能的实施方式中，多个对象是对同一个第一图像进行扰动处理后获取的(多个对象可以通过对同一个图像进行不同的扰动获取)，异常检测模块，具体用于：若距离超过预设阈值，则从多个扰动后的第一图像中删除目标对象，以获取更新后的多个扰动后的第一图像，更新后的多个扰动后的第一图像用于获取第一图像的显著图。

[0041] 在一种可能的实施方式中，多个对象是对同一个第一图像进行扰动处理后获取的(多个对象可以通过对同一个图像进行不同的扰动获取)，装置还包括权重模块，异常检测模块，还用于：若距离不超过预设阈值，则确定目标对象为正常对象。权重模块，用于：若目标对象为异常对象，则设定目标对象的权重为第一权重。若目标对象为正常对象，则设定目标对象的权重为第二权重，第二权重大于第一权重。根据第一权重或者第二权重，对目标对象的特征进行处理，以获取处理后的目标对象，处理后的目标对象用于获取第一图像的显著图。

[0042] 对于本申请第四方面以及各种可能实现方式的具体实现步骤，以及每种可能实现方式所带来的有益效果，均可以参考第二方面中各种可能的实现方式中的描述，此处不再一一赘述。

[0043] 第五方面，本申请提供一种生成显著图的装置，包括处理器，处理器和存储器耦合，存储器存储有程序指令，当存储器存储的程序指令被处理器执行时实现第一方面或第一方面任意一种可能的实施方式中的方法。

[0044] 对于本申请第五方面以及各种可能实现方式的具体实现步骤，以及每种可能实现方式所带来的有益效果，均可以参考第一方面中各种可能的实现方式中的描述，此处不再一一赘述。

[0045] 第六方面，本申请提供一种异常对象检测的装置，包括处理器，处理器和存储器耦合，存储器存储有程序指令，当存储器存储的程序指令被处理器执行时实现第二方面或第二方面任意一种可能的实施方式中的方法。

[0046] 对于本申请第六方面以及各种可能实现方式的具体实现步骤，以及每种可能实现方式所带来的有益效果，均可以参考第二方面中各种可能的实现方式中的描述，此处不再一一赘述。

[0047] 第七方面，本申请提供一种计算机可读存储介质，包括程序，当其在计算机上运行时，使得计算机执行如第一方面或第一方面任意一种可能的实施方式中的方法，或者，使得计算机执行如第二方面或第二方面任意一种可能的实施方式中的方法。

[0048] 第八方面，本申请提供一种电路系统，电路系统包括处理电路，处理电路配置为执行如第一方面或第一方面任意一种可能的实施方式中的方法，或者配置为执行如第二方面或第二方面任意一种可能的实施方式中的方法。

[0049] 第九方面，本申请提供一种计算机程序产品，计算机程序产品包括指令，当指令由电子设备加载并执行，使电子设备执行第一方面或第一方面任意一种可能的实施方式中的方法，或者使得电子设备执行第二方面或第二方面任意一种可能的实施方式中的方法。

[0050] 第十方面，本申请提供一种芯片，芯片与存储器耦合，用于执行存储器中存储的程序，以执行如第一方面或第一方面任意一种可能的实施方式中的方法，或者执行如第二方面或第二方面任意一种可能的实施方式中的方法。

[0051] 对于本申请第七方面至第十方面以及各种可能实现方式的具体实现步骤，以及每种可能实现方式所带来的有益效果，均可以参考第一方面或第二方面中各种可能的实现方式中的描述，此处不再一一赘述。

附图说明

[0052] 图1为人工智能主体框架的一种结构示意图；

[0053] 图2为一种基于扰动的可解释方法的流程示意图；

[0054] 图3为本申请实施例提供的生成显著图的方法一种流程示意图；

[0055] 图4为本申请实施例提供的生成显著图的方法另一种流程示意图；

[0056] 图5为本申请实施例提供的模型训练方法一种流程示意图；

[0057] 图6为本申请实施例提供的模型训练方法的一种应用场景的示意图；

[0058] 图7为本申请实施例提供的模型训练的方法另一种流程示意图；

[0059] 图8为本申请实施例提供的生成显著图的一种流程示意图；

[0060] 图9为本申请实施例提供的生成显著图的一种流程示意图；

[0061] 图10为本申请实施例提供的方案的一种应用场景的示意图；

[0062] 图11为本申请实施例提供的方案的一种应用场景的示意图；

[0063] 图12为本申请实施例提供的方案的实验效果图；

[0064] 图13为本申请实施例提供的一种系统的架构示意图；

[0065] 图14为本申请实施例提供的一种执行设备的架构示意图；

[0066] 图15为本申请实施例提供的生成显著图的装置的一种结构示意图；

[0067] 图16为本申请实施例提供的异常对象检测装置的一种结构示意图；

[0068] 图17为本申请实施例提供的训练装置的一种结构示意图；

[0069] 图18为本申请实施例提供的执行设备设备的一种结构示意图；

[0070] 图19为本申请实施例提供的芯片的一种结构示意图。

具体实施方式

[0071] 本申请实施例提供了一种生成显著图的方法，对扰动后的第一对象进行筛选处理，经过筛选处理后，更新后的扰动后的第一对象满足目标数据分布，目标数据分布是根据训练样本的数据分布获取的。利用满足目标数据分布的扰动后的第一图像生成显著图，使显著图能够能准确的反应特征对模型的输出的影响，对模型的解释更具有说服力，有利于提升用户对模型的信任程度。

[0072] 下面结合附图，对本申请的实施例进行描述。本领域普通技术人员可知，随着技术的发展和新场景的出现，本申请实施例提供的技术方案对于类似的技术问题，同样适用。

[0073] 首先对人工智能系统总体工作流程进行描述，请参见图1，图1示出的为人工智能主体框架的一种结构示意图，下面从“智能信息链”(水平轴)和“IT价值链”(垂直轴)两个维度对上述人工智能主体框架进行阐述。其中，“智能信息链”反映从数据的获取到处理的一列过程。举例来说，可以是智能信息感知、智能信息表示与形成、智能推理、智能决策、智能执行与输出的一般过程。在这个过程中，数据经历了“数据—信息—知识—智慧”的凝练过程。“IT价值链”从人工智能的底层基础设施、信息(提供和处理技术实现)到系统的产业生

态过程,反映人工智能为信息技术产业带来的价值。

[0074] (1)基础设施

[0075] 基础设施为人工智能系统提供计算能力支持,实现与外部世界的沟通,并通过基础平台实现支撑。通过传感器与外部沟通;计算能力由智能芯片提供,作为示例,该智能芯片包括中央处理器(central processing unit,CPU)、神经网络处理器(neural-network processing unit,NPU)、图形处理器(graphics processing unit,GPU)、专用集成电路(application specific integrated circuit,ASIC)、现场可编程逻辑门阵列(field programmable gate array,FPGA)等硬件加速芯片;基础平台包括分布式计算框架及网络等相关的平台保障和支持,可以包括云存储和计算、互联互通网络等。举例来说,传感器和外部沟通获取数据,这些数据提供给基础平台提供的分布式计算系统中的智能芯片进行计算。

[0076] (2)数据

[0077] 基础设施的上一层的数据指示人工智能领域的数据来源。数据涉及到图形、图像、语音、文本,还涉及到传统设备的物联网数据,包括已有系统的业务数据以及力、位移、液位、温度、湿度等感知数据。

[0078] (3)数据处理

[0079] 数据处理通常包括数据训练,机器学习,深度学习,搜索,推理,决策等方式。

[0080] 其中,机器学习和深度学习可以对数据进行符号化和形式化的智能信息建模、抽取、预处理、训练等。

[0081] 推理是指在计算机或智能系统中,模拟人类的智能推理方式,依据推理控制策略,利用形式化的信息进行机器思维和求解问题的过程,典型的功能是搜索与匹配。

[0082] 决策是指智能信息经过推理后进行决策的过程,通常提供分类、排序、预测等功能。

[0083] (4)通用能力

[0084] 对数据经过上面提到的数据处理后,进一步基于数据处理的结果可以形成一些通用的能力,比如可以是算法或者一个通用系统,例如,图像的分类、图像的个性化管理、电池充电个性化管理、文本分析、计算机视觉的处理、语音识别等等。

[0085] (5)智能产品及行业应用

[0086] 智能产品及行业应用指人工智能系统在各领域的产品和应用,是对人工智能整体解决方案的封装,将智能信息决策产品化、实现落地应用,其应用领域主要包括:智能终端、智能制造、智能交通、智能家居、智能医疗、智能安防、自动驾驶、智慧城市等。

[0087] 本申请实施例可以应用于上述各种领域中需要获取显著图(saliency map)的场景中。显著图是显示每个像素独特性的图像,目标在于将一般图像的表示简化或是改变为更容易分析的样式。举例来说,某个像素在一张图片的标签中具有较高的贡献值,其会在显著图中以较明显的方式被显示出来。

[0088] 为了更好的理解这一应用场景,下面对模型的可解释算法进行介绍。目前的可解释算法主要有两种类型,一种是基于梯度的可解释方法,另一种是基于扰动的可解释方法。其中,基于扰动的可解释方法,解释模块与模型之间是解耦的,更具有通用性,目前该方法的应用更为广泛。

[0089] 如图2所示，为一种基于扰动的可解释方法的流程示意图。如图2所示，以第一对象对图像为例进行说明，基于扰动的可解释方法可以包括如下几个步骤：首先对待处理图片(图像)进行扰动处理，从待处理图片中随机抽样多个像素点进行遮挡(mask)，以获取扰动后的第一图像。其中，对待处理图片进行扰动处理的方式可以有多种方式，比如可以对第一图像进行模糊处理，或者可以对第一图像进行透明化处理，或者还可以对第一图像进行加噪处理等等。然后，将扰动后的第一图像输入到模型中，使模型针对第一图像进行预测，以得到预测结果。其中模型可以是执行任意一种任务的模型，比如该模型可以是用于执行图像分类任务的模型，还可以是用于执行图像分割任务的模型等等，本申请实施例对此并不进行限定。随之，对扰动后的第一图像的特征进行线性加权处理，以得到第一图像的显著图，其中，对应的权重是根据预测结果确定的。以该模型是用于执行分类任务模型为例进行说明，假设扰动区域对应的图像特征包括M1和M2，其中，M1对应的扰动后的图像是I1，M2对应的扰动后的图像是I2，假设分类模型针对I1预测正确的概率为P1，针对I2预测正确的概率为P2，则对预测结果和扰动后的图像特征进行线性加权处理，即I1*P1+I2*P2，该结果用于表示第一图像的显著图。

[0090] 通过显著图，可以很容易观察到哪些特征对模型的输出的影响最大。为了更容易理解，下面以一个具体的例子对如何通过显著图解释模型进行说明，假设第一图像是包括猫的图像，对第一图像进行扰动，比如通过对猫的眼睛区域(M1)进行扰动，获取扰动后的图像I1；通过对猫的嘴巴区域(M2)进行扰动，获取扰动后的图像I2；通过对猫的尾巴区域(M3)进行扰动，获取扰动后的图像I3。假设分类模型针对I1预测I1是猫的概率为0.3，针对I2预测I2是猫的概率为0.6，针对I3预测I3是猫的概率为0.1。可见，当将猫的尾巴区域M3进行扰动后，模型预测第一对象是猫的概率最低，说明猫的尾巴区域M3的特征对模型的输出影响最大。0.3*I1+0.6*I2+0.1*I3表示的结果用于获取显著图。

[0091] 然而，申请人发现目前基于扰动的可解释方法存在很大的缺陷，即对第一对象进行扰动处理后，获取的多个扰动后的第一图像可能和该模型训练时采用的训练样本的数据分布是不一致的，这将会影响模型解释的准确性和可靠性。需要说明的是，可解释人工智能(explainable artificial intelligence，XAI)领域的研究人员并不能容易地意识到这一问题(该问题属于概率领域知识)，导致目前所有的基于扰动的可解释方法都存在这一缺陷，申请人发现了这一个问题，并提出了具体的改进方式。为了更好的展示这一缺陷，下面结合公式1-1进行说明。

[0092] $$p(y|d,x)=\frac{p(y,d|x)}{p(d|x)} \qquad (1\text{-}1)$$

[0093] 其中，y代表预测结果，以分类模型为例，则y表示某一类别。d表示满足目标数据分布的训练样本(将训练样本满足的数据分布称为目标数据分布)。x表示扰动后的第一图像(本申请也简称为扰动样本或者扰动后的图像，他们表示相同的意思)。p(y|d,x)表示x满足目标数据分布，并且x属于y的概率；p(y|d,x)的计算过程可以通过公式1-1进行表示。其中p(y,d|x)表示x满足目标数据分布，并且满足目标数据分布的数据被预测为y的概率，p(d|x)表示x属于目标数据分布的概率。当x属于异常样本时，即当x不满足目标数据分布时，x属于y的概率应当是很小的。这是因为分类模型是通过训练样本对模型进行训练的，训练后的模型只能针对已经学习过的知识对输入做出预测，即训练后的模型的输入应当与训练样本的

数据分布是一致的。所以当输入x不满足目标数据分布时(以下将不满足目标数据分布的x称为异常样本x),异常样本x可能不属于预设的任意一种类别(包括y),即异常样本x属于y的概率很小。但是实际情况是,针对异常样本x,模型将它预测为y的概率会很大。结合上述公式1-1进行说明,当x是异常样本时,p(y,d|x)的值会很小,假设是0.09,p(d|x)的值也会很小,假设是0.1,二者的比值为0.9,即p(y|d,x)的值为0.9,说明模型预测异常样本属于y的概率非常大。以上分析可见,当输入x不满足目标数据分布时,模型有非常高的概率会分类错误。由于显著图的确定与分类结果紧密关联,分类结果的错误将会导致显著图不能准确反映特征对模型的输出的影响,这将严重影响用户对模型的信任程度。

[0094] 针对以上发现的问题,本申请提供一种生成显著图的方法,对扰动后的第一图像进行筛选处理,经过筛选的扰动后的第一图像的数据分布和训练样本的数据分布式是一致的。利用经过筛选的扰动后的第一图像生成显著图,使显著图能够能准确的反应特征对模型的输出的影响,对模型的解释更具有说服力,有利于提升用户对模型的信任程度。下面结合具体的实施例对本申请提供的方案进行介绍。

[0095] 请参阅图3,图3为本申请实施例提供的生成显著图的方法一种流程示意图,方法可以包括:

[0096] 301、获取多个对象。

[0097] 多个对象是对第一对象进行扰动处理后获取的。

[0098] 可以采用多种方式对第一对象进行扰动处理,示例性的,可以对第一对象进行加噪处理;或者可以对第一对象进行模糊处理(比如对多个像素的取值在预设范围内进行更改);或者可以对第一对象进行遮挡处理(比如随机抽样多个像素点进行遮挡)等等。本申请实施例并不对扰动的方式进行限定,扰动的目的是为了使扰动后的第一对象中的部分区域与扰动前的第一对象中的部分区域产生差异。

[0099] 多个对象可以是图像数据、语音数据或者文本数据等等,本申请实施例对此并不进行限定。多个对象是图像数据时,本申请提供的方案可以提升输入是图像数据的模型的可解释性,比如提升图像分类模型的可解释性,更好的理解输入至图像分类模型的图像的哪部分特征对图像分类模型的输出影响大。多个对象是语音数据时,本申请提供的方案可以提升输入是语音数据的模型的可解释性,比如提升多轮对话模型的可解释性,更好的解释多轮对话模型是基于输入语音的哪部分特征,与用户进行交互。多个对象是文本数据时,本申请提供的方案可以提升输入是文本数据的模型的可解释性,比如提升机器翻译的可解释性,更好的理解输入的文本的哪部分特征对翻译模型的输出影响大。

[0100] 在一个优选的实施方式中,多个对象是对医疗图像(比如CT片、B超片、核磁共振片等)进行扰动处理后获取的。在另一个优选的实施方式中,多个对象是对交通图像(比如通过车载相机获取的图像)进行扰动处理后获取的。

[0101] 302、根据第一条件对多个对象进行筛选处理,以获取更新后的多个对象,更新后的多个对象的数据分布满足目标数据分布。

[0102] 目标数据分布是根据训练样本的数据分布获取的,训练样本用于对预设模型进行训练,以得到目标模型。

[0103] 本申请提供的方案发现了对第一对象进行扰动处理后获取的多个对象,可能并不满足目标数据分布,这将会影响模型解释的准确性和可靠性。假设训练样本包括图像1,图

像2,图像3以及图像4,根据这些训练样本获取目标数据分布。训练后的模型的输入也应当是满足目标数据分布的,否则训练后的模型针对输入做出的预测结果的准确率会大大降低。这里通过一个实例进行解释,假设训练样本都是动物类的图像,这些动物类的图像满足目标数据分布,训练后的模型的输入也应当满足目标数据分布,比如输入也是动物类的图像(模型曾经学习过该输入的特征),则训练后的模型可以针对输入进行准确的预测。但是,如果输入是人物类的图像或者是模型未学习过的动物类的图像,则训练后的模型针对输入的预测结果的准确率会大大降低。本申请提供的方案发现了扰动后的第一对象可能并不满足目标数据分布,所以对扰动后的第一对象进行筛选处理,保留满足目标数据分布的扰动后的第一对象。利用这些满足目标数据分布的扰动后的第一对象作为模型的输入,提升模型预测的准确率。

[0104] 本申请提供的方案可以通过多种方式对多个对象进行筛选处理,下面给出几种具体的筛选方式。

[0105] 在一种可能的实施方式中,可以将该多个对象作为分类模型的输入,该分类模型是通过训练样本对初始分类模型进行训练后获取的,训练后的分类模型可以用于识别异常样本。将分类模型的输出作为第一预设函数的输入,若第一预设函数的输出的值大于第一预设阈值,则认为该对象是正常样本,若第一预设函数的输出的值不大于第一预设阈值,则认为该对象是异常样本,第一预设函数的输出和第一预设函数的输入正相关。正常样本的集合即为更新后的多个对象。

[0106] 在一种可能的实施方式中,可以将该多个对象作为重构模型(auto encoder)的输入,通过重构模型对该多个对象进行重构处理,使重构后的该多个对象的数据分布满足目标数据分布。获取每个对象在重构前和重构后之间的偏差,将偏差未超过预设阈值的筛选出来,通过这些偏差未超过预设阈值对应的对象构成更新后的多个对象。需要说明的是,本申请的实施方式中,预设阈值可以设置多个,比如这个实施方式中的预设阈值和上一段介绍的实施方式中的预设阈值可能并不相同,以下对此不再重复解释说明。

[0107] 在一种优选的实施方式中,删除多个对象中的目标对象,以获取更新后的多个对象。其中,目标对象的特征和目标模型的权重向量之间的距离超过预设阈值,目标对象的特征是通过目标模型对目标对象进行特征提取后获取的。目标模型是通过第一损失值更新预设模型获取的,第一损失值是根据训练样本的特征和预设模型的权重向量之间的偏差确定的,训练样本的特征是预设模型对训练样本进行特征提取后获取的。在这种实施方式中,目标模型在迭代训练的过程中,通过第一损失值更新预设模型,使预设模型的权重向量可以更接近训练样本的特征。训练好的预设模型,即目标模型的权重向量可以反映训练样本的分布,因此可以通过比较输入目标模型的对象与目标模型的权重向量之间的距离,来判断输入至目标模型的对象是否满足目标数据分布。这种删除目标对象的实施方式将在下文图5对应的实施例中展开介绍。

[0108] 本申请将多个对象中除了更新后的多个对象中的其他对象称为异常对象或者异常样本。比如在上述步骤302提到的几种实施方式中,将最大概率未超过预设阈值对应的对象称为异常样本;将偏差超过预设阈值对应的对象称为异常样本;将目标对象称为异常样本。本申请也将更新后的多个对象中的每一个对象称为正常样本(或者正常对象)。

[0109] 303、根据第一预测结果生成第一对象的显著图。

[0110] 在一种可能的实施方式中，可以将更新后的多个对象输入至目标模型中，以输出第一预测结果，并根据该第一预测结果生成第一对象的显著图。举例说明，假设第一对象是第一图像1，执行了步骤301后，针对第一图像1，可以获取多个扰动后的第一图像，比如多个扰动后的第一图像包括扰动图像1，扰动图像2，扰动图像3，扰动图像4以及扰动图像5。执行了步骤302后，获取了更新后的多个对象，比如更新后的多个对象包括扰动图像1，扰动图像2以及扰动图像3。即通过步骤302，获取了扰动图像1，扰动图像2以及扰动图像3为正常样本，扰动图像4和扰动图像5为异常样本。在这种实施方式中，可以将扰动图像1，扰动图像2以及扰动图像3作为目标模型的输入，以输出第一预测结果，扰动图像4和扰动图像5从多个对象中被剔除，不再作为目标模型的输入。对扰动图像1，扰动图像2以及扰动图像3进行加权处理，以获取第一图像的显著图，其中，扰动图像1，扰动图像2以及扰动图像3的权重根据第一预测结果确定，具体的，扰动图像1的权重根据目标模型针对扰动图像1的预测结果确定，具体的，扰动图像2的权重根据目标模型针对扰动图像2的预测结果确定，具体的，扰动图像3的权重根据目标模型针对扰动图像3的预测结果确定。在这种实施方式中，只将正常样本作为目标模型的输入，只根据正常样本获取显著图，提升显著图的性能，使生成的显著图能够能准确的反应特征对模型的输出的影响。

[0111] 在一种可能的实施方式中，参照图4，还可以包括步骤3031、设置正常对象的权重为第一权重，设置异常对象的权重为第二权重，第一权重大于第二权重。在这种实施方式中，筛选出正常对象和异常对象后，为正常对象和异常对象设置不同的权重，并且正常对象的权重是大于异常对象的权重的。需要说明的是，所有正常对象的权重可以是相同的，也可以是不同的，所有异常对象的权重可以是相同的，也可以是不同的。同时，任意一个正常对象的权重都大于任意一个异常对象的权重。在这种实施方式中，通过第一权重对正常对象进行处理，获取处理后的正常对象，通过第二权重对异常对象进行处理，获取处理后的异常对象。将处理后的正常对象和处理后的异常对象都作为目标模型的输入，以获取第一预测结果，根据第一预测结果对处理后的正常对象和处理后的异常对象进行加权处理，以获取第一对象的显著图。举例说明，假设第一对象是第一图像1，执行了步骤301后，针对第一图像1，可以获取多个扰动后的第一图像，比如多个扰动后的第一图像包括扰动图像1，扰动图像2，扰动图像3，扰动图像4以及扰动图像5。执行了步骤302后，获取了更新后的多个对象，比如更新后的多个对象包括扰动图像1，扰动图像2以及扰动图像3。即通过步骤302，获取了扰动图像1，扰动图像2以及扰动图像3为正常样本，扰动图像4和扰动图像5为异常样本。在这种实施方式中，为每一个扰动图像设定权重，假设设定正常样本的权重为0.9，异常样本的权重为0.1，即扰动图像1，扰动图像2以及扰动图像3对应的权重为0.9，扰动图像4和扰动图像5对应的权重为0.1，则通过权重对扰动图像进行处理，处理后的扰动图像1可以看做0.9*扰动图像1，处理后的扰动图像2可以看做0.9*扰动图像2，处理后的扰动图像3可以看做0.9*扰动图像3。处理后的扰动图像4可以看做0.1*扰动图像4，处理后的扰动图像5可以看做0.1*扰动图像5。将处理后的扰动图像都作为目标模型的输入，即将0.9*扰动图像1、0.9*扰动图像2、0.9*扰动图像3、0.1*扰动图像4以及0.1*扰动图像5都作为目标模型的输入，以输出第一预测结果。对处理后的扰动图像1，处理后的扰动图像2以及处理后的扰动图像3进行加权处理，以获取第一图像的显著图，其中，处理后的扰动图像1，处理后的扰动图像2以及处理后的扰动图像3的权重根据第一预测结果确定，具体的，处理后的扰动图像1的

权重根据目标模型针对处理后的扰动图像1的预测结果确定，具体的，处理后的扰动图像2的权重根据目标模型针对处理后的扰动图像2的预测结果确定，具体的，处理后的扰动图像3的权重根据目标模型针对处理后的扰动图像3的预测结果确定。在这种实施方式中，将正常样本设定更大的权重，将异常样本设定更小的权重，在生成显著图的过程中，以正常样本为主，异常样本为辅，削弱异常样本在生成显著图的过程中造成的影响。使生成的显著图能够能较好的反应特征对模型的输出的影响。

[0112] 通过以上对图3对应的实施例的介绍可知，对扰动后的第一图像进行筛选处理，经过筛选的扰动后的第一图像的数据分布和训练样本的数据分布式是一致的。利用经过筛选的扰动后的第一图像生成显著图，使显著图能够能准确的反应特征对模型的输出的影响，对模型的解释更具有说服力，有利于提升用户对模型的信任程度。

[0113] 在介绍图3对应的实施例时，在步骤302中提到了一种对异常样本进行筛选的方法。下面将结合图5对这种方式进行具体的说明。需要说明的是，图5对应的实施方式可以和图3对应的实施方式进行结合，图5对应的实施方式也可以作为一个独立的实施方式。当图5对应的实施方式作为一个独立的实施方式时，图5对应的实施方式可以应用于需要进行异常样本检测的应用场景中。下面对图5对应的实施方式可能的应用场景进行介绍：作为一个示例，用户经常会接收到金融类型的短信，其中一些金融类型的短信来自于银行，这些短信对于用户来说是有用的，可能用于通知用户账户余额发生了变动；还有一些金融类型的短信是诈骗短信，如果用户误信了这些短信，比如点了这些诈骗短信中携带的链接，或者联系了诈骗犯，可能为用户带来金钱损失。在这一场景中，存在如何识别诈骗短信的问题，通过本申请提供的方案可以有效的筛选出金融诈骗短信，筛选出的金融炸骗短信不对用户显示，避免用户因为误信诈骗短信而造成的损失。作为另一个示例，通过AI进行医疗图像(比如CT片、B超片、核磁共振片等)的诊断时，需要从大量的医疗图像中识别哪些是正常的医疗图像，哪些是异常的医疗图像，异常的医疗图像说明该医疗图像对应的患者可能患有某种疾病。在这一场景中，通过本申请提供的方案可以有效的提升模型诊断疾病的效率。作为另一个示例，工业制造需要进行故障检测，比如需要采集矿井温度，当矿井温度异常时发出警报，在这一场景中，需要根据采集到的矿井温度，识别矿井的温度是否异常，通过本申请提供的方案可以有效的识别矿井的温度是否异常，并及时发出警报，避免危险的发生。作为另一个示例，希望从大量交易操作样本中发现异常交易操作，从而提前防范欺诈交易。作为另一个示例，希望从网络访问的样本中检测异常访问，从而发现不安全的访问，例如黑客攻击。作为另一个示例，希望从进行各种操作的用户账户中发现异常账户，从而锁定涉嫌进行高风险操作(欺诈交易、刷单等虚假交易、网络攻击)的账户等等。

[0114] 异常样本检测问题的特殊性给分类模型带来极大的挑战。异常样本检测的目的在于区分正常样本与异常样本，然而与传统的分类模型不同，异常样本的出现频率较低，导致难以收集到足够的异常样本来对分类模型进行训练。通过本申请提供的方案，只需要通过正常样本对分类模型进行训练，就可以使分类模型区分正常样本与异常样本，降低了训练分类模型的训练难度以及训练成本，同时还可以提升分类模型的性能。下面对本申请实施例提供的一种模型训练的方法，以及应用训练后的模型进行异常对象检测进行介绍。

[0115] 请参阅图5，图5为本申请实施例提供的模型训练方法一种流程示意图，方法可以包括：

[0116] 501、对训练样本进行特征提取，以获取训练样本的特征。

[0117] 本申请对训练样本的数据类型并不进行限定，可以是图像数据、语音数据或者文字数据等等。训练样本是图像数据时，训练后的模型可以用于对图像数据进行预测，具体的，训练后的模型可以从多个待检测图像中识别异常的图像，比如医疗图像的检测；训练样本是语音数据时，训练后的模型可以用于对以语音数据进行预测，具体的，训练后的模型可以从多个检测语音数据中识别异常的语音数据；再比如，训练样本是文字数据时，则训练后的模型可以用于对文字数据进行预测，以从多个待检测文字数据中识别异常的文字数据，比如金融诈骗类短信的识别。

[0118] 本申请通过预设模型对训练样本进行特征提取，具体的，通过预设模型的特征提取模型(或者说特征提取模块)对训练样本进行特征提取。预设模型可以根据特征提取模型提取的训练样本的特征对训练样本进行预测，根据预测结果和训练样本的实际结果之间的偏差更新预设模型，使更新后的预设模型针对训练样本的预测结果可以更接近实际结果。预设模型根据特征提取模型提取的训练样本的特征可以进行多种预测，比如对训练样本的类别进行预测，对训练样本中目标对象所在区域进行预测等等。

[0119] 特征提取模型包括多个首尾相连的特征提取层，前一层特征提取层的输出作为下一层特征提取层的输入，第一层特征提取层的输入是训练样本。特征提取模型的每一层的工作可以用数学表达式$\vec{y}=a(W\bullet\vec{x}+b)$来描述：从物理层面深度神经网络中的每一层的工作可以理解为通过五种对输入空间(输入向量的集合)的操作，完成输入空间到输出空间的变换(即矩阵的行空间到列空间)，这五种操作包括：1、升维/降维；2、放大/缩小；3、旋转；4、平移；5、“弯曲”。其中1、2、3的操作由$W\bullet\vec{x}$完成，4的操作由+b完成，5的操作则由来实现。这里之所以用“空间”二字来表述是因为被分类的对象并不是单个事物，而是一类事物，空间是指这类事物所有个体的集合。其中，W是权重向量，该向量中的每一个值表示该层神经网络中的一个神经元的权重值。该向量决定着上文所述的输入空间到输出空间的空间变换，即每一层的权重控制着如何变换空间。训练预设模型的目的，也就是最终得到训练好的预设模型的所有层的权重矩阵(由很多层的向量形成的权重矩阵)。因此，预设模型的训练过程本质上就是学习控制空间变换的方式，更具体的就是学习权重矩阵。

[0120] 因为希望预设模型的输出尽可能的接近真正想要预测的值。其中真正想要预测的值与预设模型的训练目标或者说预设模型需要完成的任务相关。比如预设模型用于进行图像分类任务，则预设模型的输出尽可能的接近真实的图像分类结果。为了使预设模型的输出尽可能的接近真正想要预测的值，可以通过比较当前网络的预测值和真正想要的目标值，再根据两者之间的差异情况来更新每一层神经网络的权重向量(当然，在第一次更新之前通常会有初始化的过程，即为深度神经网络中的各层预先配置参数)，比如，如果网络的预测值高了，就调整权重向量让它预测低一些，不断的调整，直到神经网络能够预测出真正想要的目标值。因此，就需要预先定义“如何比较预测值和目标值之间的差异”，这便是损失函数(loss function)或目标函数(objective function)，它们是用于衡量预测值和目标值的差异的重要方程。其中，以损失函数举例，损失函数的输出值(loss)越高表示差异越大，那么预设模型的训练就变成了尽可能缩小这个loss的过程。

[0121] 502、根据训练样本的特征和预设模型的权重向量之间的偏差获取第一损失值。

[0122] 步骤501中介绍到训练的过程是为了尽可能缩小loss，本申请提供的方案的loss中包括第一损失值。随着对预设模型的训练，第一损失值会越来越小，则预设模型的权重向量与训练样本的特征也会越来越接近。训练后的预设模型，即目标模型的权重向量已经可以反映训练样本的数据分布特征。本申请提供的方案可以采用多种方式获取第一损失值，下面对可能的实现方式进行说明。

[0123] 在一个可能的实施方式中，可以通过预设模型的任意一层特征提取层提取的特征作为步骤502中的训练样本的特征，可以通过预设模型的任意一层特征提取层的权重向量作为步骤502中的预设模型的权重向量。比如，通过第一特征提取层对训练样本进行特征提取(以下将通过第一特征提取层对训练样本进行特征提取后获取的特征称为第一特征)，其中第一特征提取层可以是多层特征提取层中的任意一层特征提取层。预设模型的权重向量可以是第二特征提取层的权重向量(以下将第二特征提取层的权重向量称为第一权重向量)，其中，第二特征提取层可以是多层特征提取层中的任意一层特征提取层。可以通过第一特征和第一权重向量之间的偏差获取第一损失值。

[0124] 在一个可能的实施方式中，第一特征提取层和第二特征提取层是同一层特征提取层。

[0125] 在一个可能的实施方式中，第一特征提取层和第二特征提取层是不同的特征提取层。在一个优选的实施方式中，第一特征提取层是特征提取模型中多个特征提取层中的倒数第二层特征提取层，多个特征提取层首尾相连，训练样本是多个特征提取层中的第一层特征提取层的输入，第二特征提取层是多个特征提取中的最后一层特征提取层。

[0126] 在一个可能的实施方式中，预设模型的权重向量可以是对多个特征提取层的权重向量进行加权处理后获取的结果。比如，对全部特征提取层的权重向量进行加权处理，认为该加权处理后的结果为预设模型的权重向量。再比如，可以对第一特征提取层的权重向量和第二特征提取层的权重向量进行加权处理，认为该结果为预设模型的权重向量。

[0127] 在一个可能的实施方式中，每一层特征提取层的权重向量可能包括多个，则通过每一层特权提取层对一个训练样本进行特征提取后，可能获取训练样本的多个特征。则根据每一个特征和每个权重向量之间的偏差可以获取多个第一损失值，可以通过该多个第一损失值更新预设模型，使每一个权重向量都可以更接近每一个特征。举例说明，假设第一特征提取层包括权重向量1，权重向量2以及权重向量3，其中根据权重向量1对输入训练样本进行特征提取，获取特征1；根据权重向量2对输入训练样本进行特征提取，获取特征2；根据权重向量3对输入训练样本进行特征提取，获取特征3。根据权重向量1和特征1之间的偏差可以获取损失值1，根据权重向量2和特征2之间的偏差可以获取损失值2，根据权重向量3和特征3之间的偏差可以获取损失值3。第一损失值包括损失值1，损失值2以及损失值3，通过第一损失值更新预设模型的过程，也是损失1，损失2以及损失3不断缩小的过程，使权重向量1更靠近特征1，权重向量2更靠近特征2，权重向量3更靠近特征3。当目标模型是分类模型时，继续举例说明，假设预设类别包括狗，猫以及兔子。通过权重向量1提取的特征1更关注输入训练样本中属于狗的特征，权重向量2提取的特征2更关注输入训练样本中属于猫的特征，权重向量3提取的特征3更关注输入训练样本中属于兔子的特征。在对预设模型进行训练的过程中，随着损失值1不断的缩小，权重向量1将更能够反映狗的特征，可以将目标模型的权重向量1看做狗这一类别的类别中心，其中类别中心可以用于表示狗这一类别的典型

特征。可以将目标模型的权重向量2看做猫这一类别的类别中心，其中类别中心可以用于表示猫这一类别的典型特征。可以将目标模型的权重向量3看做兔子这一类别的类别中心，其中类别中心可以用于表示兔子这一类别的典型特征。

[0128] 503、根据第一损失值更新预设模型。

[0129] 通过步骤502获取的第一损失值对预设模型进行更新，重复执行步骤501至步骤503直至满足预设的停止条件。其中预设的停止条件可以理解为预设的模型已经收敛，或者迭代训练的次数已经到达预设的次数。

[0130] 通过步骤501至步骤503对预设的模型进行训练，获取训练后的预设模型，即目标模型。可以通过该目标模型识别正常样本和异常样本。

[0131] 由图5对应的实施例可知，通过本申请提供的方案对预设模型进行训练，在训练的过程中只需要利用正常样本作为训练数据，无需利用异常样本。相比于异常样本难以获取的特征，正常样本更容易获取，可以大大减少对模型的训练难度，降低模型训练的成本。在模型的训练过程中，通过第一损失值更新预设模型，随着对预设模型的训练，第一损失值会越来越小，则预设模型的权重向量与训练样本的特征也会越来越接近。训练后的预设模型，即目标模型的权重向量已经可以反映训练样本的数据分布特征。在推理阶段，将对象输入至目标模型中，通过对比目标模型的权重向量和输入对象的特征之间的距离，就可以判断输入对象是否是异常样本。

[0132] 在一种可能的实施方式中，获取多个对象；通过目标模型对目标对象进行特征提取，以获取目标对象的特征，目标对象是多个对象中的任意一个对象。这里的目标模型是图5对应的实施例中描述的目标模型。获取目标对象的特征和目标模型的权重向量之间的距离。若距离超过预设阈值，则确定目标对象为异常对象。

[0133] 在一种可能的实施方式中，通过第一特征提取层对目标对象进行特征提取，以获取目标对象的特征，第一特征提取层是目标模型中多个特征提取层中的任意一层特征提取层。获取目标对象的特征和第二特征提取层的权重向量之间的距离，第二特征提取层是多个特征提取中的任意一层特征提取层。

[0134] 在一种可能的实施方式中，第一特征提取层和第二特征提取层是同一层特征提取层或者是不同的特征提取层。

[0135] 在一种可能的实施方式中，第一特征提取层是目标模型中多个特征提取层中的倒数第二层特征提取层，多个特征提取层首尾相连，目标对象是多个特征提取层中的第一层特征提取层的输入，第二特征提取层是多个特征提取中的最后一层特征提取层。

[0136] 在一种可能的实施方式中，若第二特征提取层的权重向量包括多个，则获取目标对象的特征和第二特征提取层的每个权重向量之间的距离；若多个距离中的最大距离超过预设阈值，则确定目标对象为异常样本。上文已经介绍了一个权重向量可以看做一个类别的类别中心。参阅图6，假设一共有10个权重向量分别是W1至W10。根据该权重向量W1至W10可以获取目标数据分布空间，使目标数据分布空间包括权重向量W1至W10。具体的，根据该权重向量W1至W10获取的目标数据分布空间范围的大小，可以根据实际情况进行确定，本申请实施例对此并不进行限定。图6中的方框用于表示目标数据分布空间，理论上，每个训练样本都包含于该目标数据分布空间内。在筛选异常样本的时候，可以获取目标对象的特征和第二特征提取层的每个权重向量之间的距离，并根据多个距离中的最大距离判断目标对

象是否为异常样本。继续参阅图6，假设目标对象的特征为f(x)，超过预设阈值，分别获取f(x)和W1至W10中每一个权重向量之间的距离，从图6中可以看出，f(x)和W1之间的距离最远，则根据f(x)和W1之间的距离，则确定目标对象为异常样本。之所以通过多个距离中的最大距离判断目标对象是否为异常样本，是为了有更好的识别效果。比如，继续参阅图6，若根据f(x)和W10之间的距离判断目标对象是否为异常样本，由于f(x)和W10之间的距离很近(假设f(x)和w10之间的距离是f(x)和W1至W10中每一个权重向量之间的距离的最小值)，可能这一距离并没有超过预设阈值，则认为f(x)不是异常样本。但是如果f(x)不在目标数据分布空间内(如图6所示，f(x)不在方框内部)，应当属于异常样本。所以为了更好的识别异常样本，在这种实施方式中，根据多个距离中的最大距离判断目标对象是否为异常样本。

[0137] 在一个优选的实施方式中，还可以进一步优化训练过程，使训练后的预设模型，即目标模型的性能更佳。下面结合一个具体的实施方式对此进行说明。

[0138] 请参阅图7，图7为本申请实施例提供的模型训练的方法另一种流程示意图，方法可以包括：

[0139] 701、对训练样本进行特征提取，以获取训练样本的特征。

[0140] 702、根据训练样本的特征和预设模型的权重向量之间的偏差获取第一损失值。

[0141] 步骤701和步骤702可以参照图5对应的实施例总的步骤501和步骤502进行理解，这里不再重复赘述。

[0142] 703、根据目标结果和训练样本的真实结果之间的偏差获取第二损失值。

[0143] 目标结果是根据第一预测结果和预设函数确定的，第一预测结果是预设模型针对训练样本的预测结果，预设函数的输入是第一预测结果，预设函数的输出是目标结果，预设函数的输出和预设函数的输入负相关。

[0144] 在图3对应的实施方式中，对模型的训练过程进行了介绍，训练的过程是使损失值不断减小的过程，使更新后的预设模型针对训练样本的预测结果可以更接近实际结果。假设经过少数几轮训练后，预测模型针对训练样本的预测结果就接近了实际结果，则训练就停止了。理论上，模型训练的轮次越多，越有利于提升训练后的模型的性能，模型在每一轮的训练中，不断的学习，提升预测的准确度。所以为了减缓模型收敛的速度，使模型可以进行更多轮次的训练，还可以对每一次训练获取的预测结果进行处理，拉近概率最大的预测结果和实际结果之间的距离，相比于没有对每一次训练获取的预测结果进行处理，经过处理后，每一次训练获取的损失值将会增大。举例说明，假设在一次训练中，预设模型针对训练样本1的预测结果为训练样本1属于猫的概率为0.9，训练样本1属于老虎的概率为0.1。将这次训练的预测结果作为预设函数的输入，得到目标结果1和目标结果2，由于预设函数的输出和预设函数的输入负相关，则目标结果1可能是训练样本1属于猫的概率为0.6，训练样本1属于老虎的概率为0.4。相比于未处理前的预测结果(0.9，0.1)，处理后的预测结果(0.6，0.4)中的各个概率之间比较均衡，有利于减缓模型收敛的速度。第二损失值包括根据目标结果1和训练样本1的真实类别之间的偏差获取的损失值，以及根据目标结果2和训练样本1的真实类别之间的偏差获取的损失值。

[0145] 704、根据第一损失值和第二损失值更新预设模型。

[0146] 通过步骤702和步骤703获取的第一损失值和第二损失值对预设模型进行更新，重复执行步骤701至步骤704直至满足预设的停止条件。其中预设的停止条件可以理解为预设

的模型已经收敛。

[0147] 由图7对应的实施例可知，在图5对应的实施的基础上，在模型的训练过程中引入预设函数，该预设函数的输入和输出负相关，根据预设函数的不同输入，选择不同的输出，无需手动调节参数，可以自动根据不同的输入获取不同的输出(也称为获取不同的温度temperature)，避免模型陷入局部最优，减缓模型收敛的速度，使训练后的模型的性能更优。

[0148] 在一个优选的实施方式中，可以通过图5对应的实施例提供的方法获取的目标模型对图3对应的实施例中的多个扰动后的样本进行筛选，识别多个扰动后的样本中的异常样本。或者可以通过图7对应的实施例提供的方法获取的目标模型对图3对应的实施例中的多个扰动后的样本进行筛选，识别多个扰动后的样本中的异常样本。将异常样本从多个扰动后的样本中进行删除，通过删除了异常样本的多个扰动后的样本获取第一对象的显著图。为了更好的展示这一方案，下面结合两个典型的流程图对此进行说明，应当理解下述两个典型的流程图只是提供了两种可能的实施例结合的方式，除此之外，还可能有其他的结合方式。比如下述图8对应的实施例中，以第一对象是第一图像进行的说明，第一对象当然还可以是其他类型的数据，具体参照图3对应的实施例进行理解，这里不再重复赘述。

[0149] 请参阅图8，图8为本申请实施例提供的生成显著图的一种流程示意图。获取第一图像，对第一图像进行扰动处理，以获取扰动后的第一图像。在图8所示的流程中，从第一图像中随机选择多组像素进行遮挡，根据每一组选取遮挡的像素点，可以获取一个扰动后的第一图像。通过异常样本筛选模块对获取到的扰动后的第一图像进行异常样本的筛选，将扰动后的第一图像中的异常样本筛选出来并进行删除，将删除了异常样本的扰动后的第一图像作为分类模型的输入。根据分类模型的输出对删除了异常样本的扰动后的第一图像进行加权处理，以获取加权结果，将获取到的加权结果输入至XAI解释器，以获取第一图像的显著图。请参阅图9，对筛选异常样本的过程进行进一步的说明。将扰动后的第一图像输入至目标模型，通过目标模型对每个扰动后的第一图像进行特征提取，获取目标模型中多个特征提取层的倒数第二层特征提取层提取的特征以及最后一层特征提取层之间的距离。若包括多个距离，则判断多个距离中的最大距离是否超过预设阈值，若超过预设阈值，则该扰动后的第一图像是异常样本，将其删除，若没有超过预设阈值，则该扰动后的第一图像是正常样本。将正常样本作为分类模型的输入。根据分类模型的输出对正常样本进行加权处理，以获取加权结果，将获取到的加权结果输入至XAI解释器，以获取第一图像的显著图。此外，对预设模型进行训练获取目标模型的过程中，利用了第一损失值和第二损失值，并引入预设函数，延缓模型的收敛速度，以提升目标模型的性能。

[0150] 通过本申请实施例提供的方案，可以提升模型解释的准确性，下面以本申请提供的方案应用于几个典型的场景为例，对本申请提供的方案进行介绍。

[0151] 参阅图10，本申请提供的方案可以应用于医疗图片检测的应用场景中。比如针对同一身体区域获取多个医疗图像(比如肺部CT片)。该目标模型的训练样本是正常的医疗图像，正常的医疗图像是指该医疗图像的来源对象(人物)是健康的，未患有疾病的。将该多个医疗图像输入至目标模型中，可以筛选出异常样本，这里的异常样本是非正常的医疗图像，非正常的医疗图像是指该医疗图像的来源对象(人物)可能是患有疾病的。在另一种可能的实施方式中，对每个医疗图像进行扰动处理，对经过扰动处理后的医疗图像进行筛选处理，

以将扰动处理后的医疗图像中的异常样本进行删除。该目标模型的训练样本是非正常的医疗图像。将经过扰动处理后的医疗图像输入至目标模型中，可以筛选出异常样本，这里的异常样本是不满足训练样本的数据分布的样本。将扰动处理后的医疗图像中的正常样本输入目标模型中，根据目标模型的预测结果获取每个医疗图像的显著图。使医疗图像的显著图可以更好的解释。

[0152] 参阅图11，本申请提供的方案可以应用于智能驾驶的应用场景中。比如获取交通图像(比如通过车载相机获取的图像)。对每个交通图像进行扰动处理，对经过扰动处理后的交通图像进行筛选处理，以将扰动处理后的交通图像中的异常样本进行删除。将扰动处理后的交通图像中的正常样本输入目标模型中，根据目标模型的预测结果获取每个交通图像的显著图。使交通图像的显著图能更好的展示交通图像的哪部分特征对目标模型的输出影响更大，提高模型的可解释性。

[0153] 为了更直观地理解本方案所带来的有益效果，以下结合数据对本申请实施例带来的有益效果进行说明。

[0154] 在一种测试实验中，测试数据集为ImageNet数据集，通过在ImageNet数据集上，根据模型ResNet50进行图像分类的测试实验，并对分类的结果进行解释，具体的通过显著图的方式对分类的结果进行解释。测试结果通过图12进行展示。如图12所示，通过对三种典型的原始图像进行图像分类，一种原始图像中包括多个目标(多条鱼、多朵花)，另一种原始图像远距离场景(街道场景)，还有一种原始图像中是单一的目标，且该目标在原始图像中占据的区域比较大。衡量可解释方法准确性的指标包括定位性(localization)。基于显著图的解释应当显著高亮图片中与标签相关的像素，localization这一指标利用目标检测数据集，对于同一图片同一个标签，通过显著图的高亮部分和真值(ground truth)的重合度来衡量显著图的定位能力，定位能力越高解释性方法越准确。本申请提供的方案由于对扰动后的图像进行了筛选，通过正常样本获取的显著图，相比于已有的方案，可以大幅度提示localization这一指标。具体的，在多目标场景中，本申请更能准确高亮出多个目标，而已有方案在高亮单个目标之外，其他光亮部分不能准确捕捉到额外目标；街道场景中，本申请高亮出了两个红绿灯，而已有方案只高亮了一个；在单一大目标场景中，本申请的高亮部分和目标轮廓也更加接近。本申请提供的方案在localization指标上有显著提高——多目标场景提升63.1%，街道场景提升58.3%。

[0155] 在另一种测试实验中，选用DenseNet-BC和ResNet50分别作为目标模型。参阅表1：当选用DenseNet-BC作为目标模型时，将Cifar10数据集作为正常样本，TinyImageNet数据集作为异常样本，进行推理时，本申请提供的方案相比于已有的方案，接收者操作特征曲线下方的面积大小(area under receiver operating characteristic curve，AUROC)指标提升2.9%，真正率达到95%时对应的真负率(true negative rate at 95%True Positive rate，TNR@TPR95)指标提升26.2%。当选用ResNet50作为目标模型时，将ImageNet数据集作为正常样本据，将Gaussian Noise数据集作为异常样本进行推理时，AUROC指标提升9.2%，TNR@TPR95指标提升98.3%。其中AUROC指标和TNR@TPR95指标越高，表示检测异常样本的准确率越高。

[0156]

DenseNet-BC			ResNet50		
	其他方案	本方案		其他方案	本方案
AUROC	95.48%	98.23%	AUROC	91.53%	99.9964%
TNR@TPR95	72.33%	91.26%	TNR@TPR95	50.44%	100

[0157] 表1

[0158] 为了便于更好的理解本方案，先结合图13对本申请实施例提供的一种系统进行介绍，请参阅图13，图13为本申请实施例提供的一种系统的架构图，在图13中，系统200包括执行设备210、训练设备220、数据库230和数据存储系统240。

[0159] 在训练阶段，数据库230中存储有训练数据集合，数据库230具体可以表现为任意形式的存储介质，不限定为传统意义上的数据库。训练数据集合中可以有多个训练样本。本申请对训练样本的数据类型并不进行限定，比如训练样本可以是图像数据，或者训练样本可以是语音数据，或者训练样本可以是文字数据。需要说明的是，通常情况训练数据集中包括的训练样本的数据类型是相同的。训练设备220生成预设模型，并利用数据库中的训练数据集合对预设模型进行迭代训练，得到成熟的预设模型(即目标模型)。上文对应的实施例中已经对如何对预设模型进行训练以获取目标模型进行了详细的介绍，这里不再重复赘述。

[0160] 在推理阶段，执行设备210可以调用数据存储系统240中的数据、代码等，也可以将数据、指令等存入数据存储系统240中。数据存储系统240可以配置于执行设备210中，也可以为执行设备210外部的存储器。执行设备210可以调用成熟的预设模型提取第一对象的特征，并根据提取出的第一对象的特征执行分类任务，根据预测结果获取显著图(参照上文关于获取显著图的相关实施例进行理解)，或者根据提取出的第一对象的特征进行异常样本的检测(参照上文关于异常样本检测的相关实施例进行理解)。

[0161] 本申请的一些实施例中，例如图3中，“用户”可以直接与执行设备210进行交互，也即执行设备210与客户设备集成于同一设备中。作为示例，在一些应用场景中，执行设备210可以表现为终端设备，比如手机、摄像头、智能家居等等。则在推理阶段，用户可以通过执行设备210输入第一对象，比如用户通过摄像头进行拍照，摄像头获取的图像作为成熟的预设模型的输入。在另一些应用场景中，执行设备210具体可以表现为配置有显示屏的执行设备，则在推理阶段，执行设备210在完成一个任务(或者多个任务)之后，可以向用户展示预设模型的输出结果。比如执行设备210执行了图像分类任务之后，向用户展示图像分类的结果。执行设备210还可以表现为其它形态，此处不一一进行列举，但图3仅是本发明实施例提供的架构示意图，图中所示设备、器件、模块等之间的位置关系不构成任何限制。

[0162] 在本申请的另一些实施例中，执行设备210和客户设备可以为分别独立的设备，执行设备210配置有输入/输出接口，与客户设备进行数据交互，“用户”可以通过客户设备的输入/输出接口向执行设备210输入至少一个任务，执行设备210通过输入/输出接口将处理结果返回给客户设备。

[0163] 以上对本申请实施例提供的一种显著图的获取方法、模型的训练方法以及异常对象检测的方法进行了介绍，通过本申请提供的方案可以提升显著图的性能，还可以降低异常检测模型的训练难度，提升异常对象检测的准确度。

[0164] 可以理解的是，为了实现上述功能，下面还提供用于实施上述方案的相关设备。这

些相关设备包含了执行各个功能相应的硬件结构和/或软件模块。本领域技术人员应该很容易意识到,结合本文中所公开的实施例描述的各示例的模块及算法步骤,本申请能够以硬件或硬件和计算机软件的结合形式来实现。某个功能究竟以硬件还是计算机软件驱动硬件的方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0165] 参阅图14,为本申请实施例提供的一种执行设备的架构示意图。如图14所示,执行设备包含了执行各个功能相应的硬件结构和软件模块。其中,该执行设备用于接收输入数据,可解释工具包根据输入数据输出显著图。可解释工具包根据输入数据输出显著图的过程中,需要调用AI框架,其中,AI框架中部署有上述实施例中描述的目标模型。计算过程需要硬件结构的支持,该执行设备中还包括处理器,用于运行软件中的代码。其中处理器可以是中央处理器(central processing units,CPU)、网络处理器(neural-network processing unit,NPU)、图形处理器(graphics processing unit,GPU)、数字信号处理器(digital signal processor,DSP)、专用集成电路(application specific integrated circuit,ASIC)或现场可编程逻辑门阵列(field programmable gate array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者也可以是任何常规的处理器等。

[0166] 具体参阅图15,图15为本申请实施例提供的生成显著图的装置的一种结构示意图。该生成显著图的装置可以包括获取模块1501,筛选模块1502,预测模块1503,生成模块1504。

[0167] 在一种可能的实施方式中,该生成显著图的装置包括:获取模块1501,用于获取多个对象,多个对象是对第一对象进行扰动处理后获取的。筛选模块1502,用于根据第一条件对获取模块1501获取的多个对象进行筛选处理,以获取更新后的多个对象,更新后的多个对象满足目标数据分布,目标数据分布是根据训练样本获取的,训练样本用于对预设模型进行训练,以得到目标模型。预测模块1503,用于将筛选模块1502获取的更新后的多个对象输入至目标模型中,以输出第一预测结果。生成模块1504,用于根据预测模块1503获取的第一预测结果和更新后的多个对象生成第一对象的显著图。

[0168] 在一种可能的实施方式中,目标模型是通过第一损失值更新预设模型获取的,第一损失值是根据训练样本的特征和预设模型的权重向量之间的偏差确定的,训练样本的特征是预设模型对训练样本进行特征提取后获取的,筛选模块1502,具体用于:若多个对象中包括目标对象,则删除多个对象中的目标对象,以获取更新后的多个对象,目标对象的特征和目标模型的权重向量之间的距离超过预设阈值,目标对象的特征是通过目标模型对目标对象进行特征提取后获取的。

[0169] 在一种可能的实施方式中,目标对象的特征具体是通过第一特征提取层提取的,第一特征提取层是目标模型中多个特征提取层中的任意一层特征提取层,目标对象的特征和目标模型的权重向量之间的距离具体是目标对象的特征和第二特征提取层的权重向量之间的距离,第二特征提取层是多个特征提取中的任意一层特征提取层。

[0170] 在一种可能的实施方式中,第一特征提取层和第二特征提取层是同一层特征提取层或者是不同的特征提取层。

[0171] 在一种可能的实施方式中，第一特征提取层具体是特征提取模型中多个特征提取层中的倒数第二层特征提取层，多个特征提取层首尾相连，目标对象是多个特征提取层中的第一层特征提取层的输入，第二特征提取层具体是多个特征提取中的最后一层特征提取层。

[0172] 在一种可能的实施方式中，若第二特征提取层的权重向量包括多个，目标对象的特征和第二特征提取层的目标权重向量之间的距离超过预设阈值，目标对象的特征和第二特征提取层的目标权重向量之间的距离是多个距离中的最大距离，多个距离包括目标对象的特征和第二特征提取层的每个权重向量之间的距离。

[0173] 在一种可能的实施方式中，目标模型具体是通过第一损失值和第二损失值更新预设模型后获取的，第二损失值是根据目标结果和训练样本的真实结果之间的偏差确定的，目标结果是根据第二预测结果和预设函数确定的，第二预测结果是预设模型针对训练样本的预测结果，预设函数的输入是第二预测结果，预设函数的输出是目标结果，预设函数的输出和预设函数的输入负相关。

[0174] 在一种可能的实施方式中，该生成显著图的装置还包括权重模块1505，权重模块1505，用于:设置更新后的多个对象的权重为第一权重。设置剩余的多个对象的权重为第二权重，剩余的多个对象是多个对象中除了更新后的多个对象之外的对象，第一权重大于第二权重。预测模块1503，具体用于将第一结果和第二结果输入至目标模型中，以输出第一预测结果，第一结果是根据第一权重和更新后的多个对象确定的，第二结果是根据第二权重和剩余的多个对象确定的。

[0175] 需要说明的是，图15中所示的生成显著图的装置中各模块之间的信息交互、执行过程等内容，与本申请中图3至图9对应的各个方法实施例基于同一构思，具体内容可参见本申请前述所示的方法实施例中的叙述，此处不再赘述。

[0176] 具体参阅图16，图16为本申请实施例提供的异常对象检测装置的一种结构示意图。该异常对象检测装置可以包括第一获取模块1601，特征提取模块1602，第二获取模块1603，异常检测模块1604。

[0177] 在一种可能的实施方式中，本申请提供一种异常对象检测的装置，包括:第一获取模块1601，用于获取多个对象。特征提取模块1602，用于通过特征提取模型对目标对象进行特征提取，以获取目标对象的特征，目标对象是多个对象中的任意一个对象，特征提取模型是通过第一损失值更新预设模型后获取的，第一损失值是根据训练样本的特征和预设模型的权重向量之间的偏差确定的，训练样本的特征是预设模型对训练样本进行特征提取后获取的。第二获取模块1603，用于获取目标对象的特征和特征提取模型的权重向量之间的距离。异常检测模块1604，用于若距离超过预设阈值，则确定目标对象为异常对象。

[0178] 在一种可能的实施方式中，特征提取模块1602，具体用于:通过第一特征提取层对目标对象进行特征提取，以获取目标对象的特征，第一特征提取层是特征提取模型中多个特征提取层中的任意一层特征提取层。获取目标对象的特征和特征提取模型的权重向量之间的距离，包括:获取目标对象的特征和第二特征提取层的权重向量之间的距离，第二特征提取层是多个特征提取中的任意一层特征提取层。

[0179] 在一种可能的实施方式中，第一特征提取层和第二特征提取层是同一层特征提取层或者是不同的特征提取层。

[0180] 在一种可能的实施方式中，第一特征提取层是特征提取模型中多个特征提取层中的倒数第二层特征提取层，多个特征提取层首尾相连，目标对象是多个特征提取层中的第一层特征提取层的输入，第二特征提取层是多个特征提取中的最后一层特征提取层。

[0181] 在一种可能的实施方式中，第二获取模块1603，具体用于：若第二特征提取层的权重向量包括多个，则获取目标对象的特征和第二特征提取层的每个权重向量之间的距离。异常检测模块1604，具体用于若多个距离中的最大距离超过预设阈值，则确定目标对象为异常样本。

[0182] 在一种可能的实施方式中，特征提取模型具体是通过第一损失值和第二损失值更新预设模型后获取的，第二损失值是根据目标结果和训练样本的真实结果之间的偏差确定的，目标结果是根据第一预测结果和预设函数确定的，第一预测结果是预设模型针对训练样本的预测结果，预设函数的输入是第一预测结果，预设函数的输出是目标结果，预设函数的输出和预设函数的输入负相关。

[0183] 在一种可能的实施方式中，多个对象是多个扰动后的第一图像，多个扰动后的第一图像是对第一图像进行扰动处理后获取的，异常检测模块1604，具体用于：若距离超过预设阈值，则从多个扰动后的第一图像中删除目标对象，以获取更新后的多个扰动后的第一图像，更新后的多个扰动后的第一图像用于获取第一图像的显著图。

[0184] 在一种可能的实施方式中，多个对象是多个扰动后的第一图像，多个扰动后的第一图像是对第一图像进行扰动处理后获取的，装置还包括权重模块1605，异常检测模块1604，还用于：若距离不超过预设阈值，则确定目标对象为正常对象。权重模块1605，用于：若目标对象为异常对象，则设定目标对象的权重为第一权重。若目标对象为正常对象，则设定目标对象的权重为第二权重，第二权重大于第一权重。根据第一权重或者第二权重，对目标对象的特征进行处理，以获取处理后的目标对象，处理后的目标对象用于获取第一图像的显著图。

[0185] 需要说明的是，图16中所示的异常对象检测装置中各模块之间的信息交互、执行过程等内容，与本申请中图3至图9对应的各个方法实施例基于同一构思，具体内容可参见本申请前述所示的方法实施例中的叙述，此处不再赘述。

[0186] 本申请实施例还提供一种训练装置，请参阅图17，图17为本申请实施例提供的训练装置的一种结构示意图。训练装置1700上可以部署有图5、图6中所描述的预设模型。具体的，训练装置1700可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上中央处理器(central processing units，CPU)1722(例如，一个或一个以上处理器)和存储器1732，一个或一个以上存储应用程序1742或数据1744的存储介质1730(例如一个或一个以上海量存储设备)。其中，存储器1732和存储介质1730可以是短暂存储或持久存储。在一个实施例中，存储器1732为随机存储存储器(random access memory，RAM)，可以与中央处理器1722直接交换数据，用于加载数据1744和应用程序1742和/或操作系统1741以供中央处理器1722直接运行与运用，通常作为操作系统或其他正在运行中的程序的临时数据存储媒介。存储在存储介质1730的程序可以包括一个或一个以上模块(图示没标出)，每个模块可以包括对训练装置中的一系列指令操作。更进一步地，中央处理器1722可以设置为与存储介质1730通信，在训练装置1700上执行存储介质1730中的一系列指令操作。

[0187] 训练装置1700还可以包括一个或一个以上电源1726，一个或一个以上有线或无线

网络接口1750，一个或一个以上输入输出接口1758，和/或，一个或一个以上操作系统1741，例如Windows ServerTM，Mac OS XTM，UnixTM，LinuxTM，FreeBSDTM等等。

[0188] 需要说明的是，中央处理器1722还用于执行图5、图6中预设模型执行的其他步骤，对于中央处理器1722执行图5、图6对应实施例中的预设模型执行的步骤的具体实现方式以及带来的有益效果，均可以参考图5、图6对应的各个方法实施例中的叙述，此处不再一一赘述。

[0189] 本申请实施例还提供一种执行设备，请参阅图18，图18为本申请实施例提供的执行设备设备的一种结构示意图。执行设备1800上可以部署有图3至图9中所描述的目标模型，用于执行图3至图9中的生成显著图对应步骤或者异常样本检测的步骤。具体的，执行设备1800包括：接收器1801、发射器1802、处理器1803和存储器1804(其中执行设备1800中的处理器1803的数量可以为一个或多个，图18中以一个处理器为例)，其中，处理器1803可以包括应用处理器18031和通信处理器18032。在本申请的一些实施例中，接收器1801、发射器1802、处理器1803和存储器1804可通过总线或其它方式连接。

[0190] 存储器1804可以包括只读存储器和随机存取存储器，并向处理器1803提供指令和数据。存储器1804的一部分还可以包括非易失性随机存取存储器(non-volatile random access memory，NVRAM)。存储器1804存储有处理器和操作指令、可执行模块或者数据结构，或者它们的子集，或者它们的扩展集，其中，操作指令可包括各种操作指令，用于实现各种操作。

[0191] 处理器1803控制执行设备的操作。具体的应用中，执行设备的各个组件通过总线系统耦合在一起，其中总线系统除包括数据总线之外，还可以包括电源总线、控制总线和状态信号总线等。但是为了清楚说明起见，在图中将各种总线都称为总线系统。

[0192] 上述本申请实施例揭示的方法可以应用于处理器1803中，或者由处理器1803实现。处理器1803可以是一种集成电路芯片，具有信号的处理能力。在实现过程中，上述方法的各步骤可以通过处理器1803中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器1803可以是通用处理器、数字信号处理器(digital signal processing，DSP)、微处理器或微控制器，还可进一步包括专用集成电路(application specific integrated circuit，ASIC)、现场可编程门阵列(field-programmable gate array，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。该处理器1803可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。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。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成，或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器，闪存、只读存储器，可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器1804，处理器1803读取存储器1804中的信息，结合其硬件完成上述方法的步骤。

[0193] 接收器1801可用于接收输入的数字或字符信息，以及产生与执行设备的相关设置以及功能控制有关的信号输入。发射器1802可用于通过接口输出数字或字符信息；发射器1802还可用于通过上述接口向磁盘组发送指令，以修改磁盘组中的数据；发射器1802还可以包括显示屏等显示设备。

[0194] 在一种情况下，本申请实施例中，应用处理器18031用于执行有图3至图9中对应的

实施例中描述的目标模型执行的方法。

[0195] 对于应用处理器18031执行图3至图9对应实施例中目标模型的功能的具体实现方式以及带来的有益效果，均可以参考图3至图9对应的各个方法实施例中的叙述，此处不再一一赘述。

[0196] 应当理解，上述仅为本申请实施例提供的一个例子，并且，车辆可具有比示出的部件更多或更少的部件，可以组合两个或更多个部件，或者可具有部件的不同配置实现。

[0197] 在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。

[0198] 本申请实施例提供的执行设备和训练设备具体可以为芯片，以使芯片执行上述图3至图9所示实施例描述的方法。可选地，所述存储单元为所述芯片内的存储单元，如寄存器、缓存等，所述存储单元还可以是所述无线接入设备端内的位于所述芯片外部的存储单元，如只读存储器(read-only memory，ROM)或可存储静态信息和指令的其他类型的静态存储设备，随机存取存储器(random access memory，RAM)等。

[0199] 具体的，请参阅图19，图19为本申请实施例提供的芯片的一种结构示意图，所述芯片可以表现为神经网络处理器NPU 190，NPU 190作为协处理器挂载到主CPU(Host CPU)上，由Host CPU分配任务。NPU的核心部分为运算电路1903，通过控制器1904控制运算电路1903提取存储器中的矩阵数据并进行乘法运算。

[0200] 在一些实现中，运算电路1903内部包括多个处理单元(Process Engine，PE)。在一些实现中，运算电路1903是二维脉动阵列。运算电路1903还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中，运算电路1903是通用的矩阵处理器。

[0201] 举例来说，假设有输入矩阵A，权重矩阵B，输出矩阵C。运算电路从权重存储器1902中取矩阵B相应的数据，并缓存在运算电路中每一个PE上。运算电路从输入存储器1901中取矩阵A数据与矩阵B进行矩阵运算，得到的矩阵的部分结果或最终结果，保存在累加器(accumulator)1908中。

[0202] 统一存储器1906用于存放输入数据以及输出数据。权重数据直接通过存储单元访问控制器(Direct Memory Access Controller，DMAC)1905被搬运到权重存储器1902中。输入数据也通过DMAC被搬运到统一存储器1906中。

[0203] 总线接口单元1910(Bus Interface Unit，简称BIU)，用于取指存储器1909从外部存储器获取指令，还用于存储单元访问控制器1905从外部存储器获取输入矩阵A或者权重矩阵B的原数据。

[0204] DMAC主要用于将外部存储器DDR中的输入数据搬运到统一存储器1906或将权重数据搬运到权重存储器1902中或将输入数据数据搬运到输入存储器1901中。

[0205] 向量计算单元1907包括多个运算处理单元，在需要的情况下，对运算电路的输出做进一步处理，如向量乘，向量加，指数运算，对数运算，大小比较等等。主要用于神经网络中非卷积/全连接层网络计算，如Batch Normalization(批归一化)，像素级求和，对特征平面进行上采样等。

[0206] 在一些实现中，向量计算单元1907能将经处理的输出的向量存储到统一存储器1906。例如，向量计算单元1907可以将线性函数和/或非线性函数应用到运算电路1903的输

出，例如对卷积层提取的特征平面进行线性插值，再例如累加值的向量，用以生成激活值。在一些实现中，向量计算单元1907生成归一化的值、像素级求和的值，或二者均有。在一些实现中，处理过的输出的向量能够用作到运算电路1903的激活输入，例如用于在神经网络中的后续层中的使用。

[0207] 控制器1904连接的取指存储器(instruction fetch buffer)1909，用于存储控制器1904使用的指令；统一存储器1906，输入存储器1901，权重存储器1902以及取指存储器1909均为On-Chip存储器。外部存储器私有于该NPU硬件架构。

[0208] 其中，循环神经网络中各层的运算可以由运算电路1903或向量计算单元1907执行。

[0209] 其中，上述任一处提到的处理器，可以是一个通用中央处理器，微处理器，ASIC，或一个或多个用于控制上述第一方面方法的程序执行的集成电路。

[0210] 本申请实施例提供还提供一种芯片，该芯片包括：处理单元和通信单元，所述处理单元例如可以是处理器，所述通信单元例如可以是输入/输出接口、管脚或电路等。该处理单元可执行存储单元存储的计算机执行指令，以使芯片执行上述图3至图9中所描述的方法。可选地，所述存储单元为所述芯片内的存储单元，如寄存器、缓存等，所述存储单元还可以是所述无线接入设备端内的位于所述芯片外部的存储单元，如只读存储器(read-only memory，ROM)或可存储静态信息和指令的其他类型的静态存储设备，随机存取存储器(random access memory，RAM)等。具体地，前述的处理单元或者处理器可以是中央处理器(central processing unit，CPU)、网络处理器(neural-network processing unit，NPU)、图形处理器(graphics processing unit，GPU)、数字信号处理器(digital signal processor，DSP)、专用集成电路(application specific integrated circuit，ASIC)或现场可编程逻辑门阵列(field programmable gate array，FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者也可以是任何常规的处理器等。

[0211] 本申请实施例中还提供一种计算机可读存储介质，该计算机可读存储介质中存储有用于训练模型的程序，当其在计算机上运行时，使得计算机执行上述图3至图9中所描述的方法。

[0212] 本申请实施例中还提供一种包括计算机程序产品，当其在计算机上运行时，使得计算机执行如前述图3至图9所示实施例描述的方法中的步骤。所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线(例如同轴电缆、光纤、数字用户线(DSL))或无线(例如红外、无线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存储的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。所述可用介质可以是磁性介质，(例如，软盘、硬盘、磁带)、光介质(例如，DVD)、或者半导体介质(例如固态硬盘Solid State Disk(SSD))等。

[0213] 本申请实施例中还提供一种电路系统，所述电路系统包括处理电路，所述处理电路配置为执行如前述图3至图9所示实施例描述的方法中的步骤。

[0214] 通过以上的实施方式的描述，所属领域的技术人员可以清楚地了解到本申请可借助纯软件或软件加必需的通用硬件的方式来实现，当然也可以通过专用硬件包括专用集成电路、专用CLU、专用存储器、专用元器件等来实现。一般情况下，凡由计算机程序完成的功能都可以很容易地用相应的硬件来实现，而且，用来实现同一功能的具体硬件结构也可以是多种多样的，例如模拟电路、数字电路或专用电路等。但是，对本申请而言更多情况下软件程序实现是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在可读取的存储介质中，如计算机的软盘、U盘、移动硬盘、ROM、RAM、磁碟或者光盘等，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，训练设备，或者网络设备等)执行本申请各个实施例所述的方法。此外，该计算机软件产品也可以控件、驱动程序、独立或可下载软件对象等形式体现。

[0215] 本申请的说明书和权利要求书及上述附图中的术语“第一”，“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的实施例能够以除了在这里图示或描述的内容以外的顺序实施。本申请中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A和/或B，可以表示：单独存在A，同时存在A和B，单独存在B这三种情况，另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或模块的过程，方法，系统，产品或设备不必限于清楚地列出的那些步骤或模块，而是可包括没有清楚地列出的或对于这些过程，方法，产品或设备固有的其它步骤或模块。在本申请中出现的对步骤进行的命名或者编号，并不意味着必须按照命名或者编号所指示的时间/逻辑先后顺序执行方法流程中的步骤，已经命名或者编号的流程步骤可以根据要实现的技术目的变更执行次序，只要能达到相同或者相类似的技术效果即可。本申请中所出现的模块的划分，是一种逻辑上的划分，实际应用中实现时可以有另外的划分方式，例如多个模块可以结合成或集成在另一个系统中，或一些特征可以忽略，或不执行，另外，所显示的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端口，模块之间的间接耦合或通信连接可以是电性或其他类似的形式，本申请中均不作限定。并且，作为分离部件说明的模块或子模块可以是也可以不是物理上的分离，可以是也可以不是物理模块，或者可以分布到多个电路模块中，可以根据实际的需要选择其中的部分或全部模块来实现本申请方案的目的。

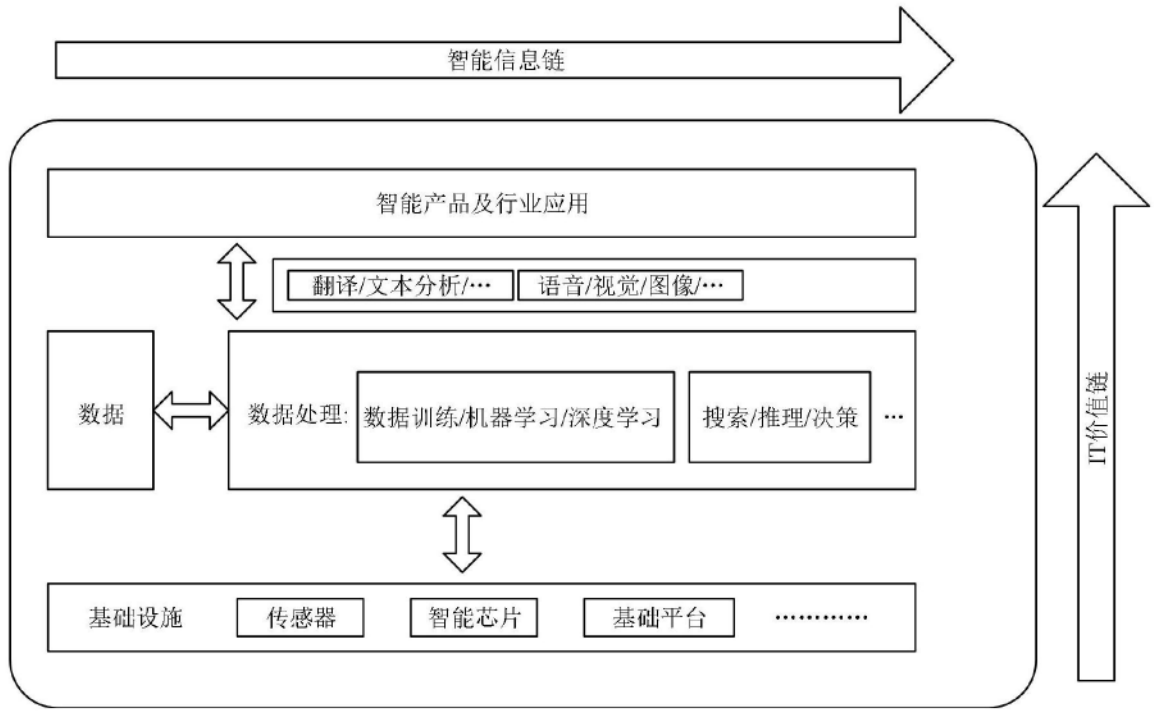

智能信息链
智能产品及行业应用
翻译/文本分析/…
语音/视觉/图像/…
数据
数据处理:
数据训练/机器学习/深度学习
搜索/推理/决策
…
基础设施
传感器
智能芯片
基础平台
……
IT价值链

图1

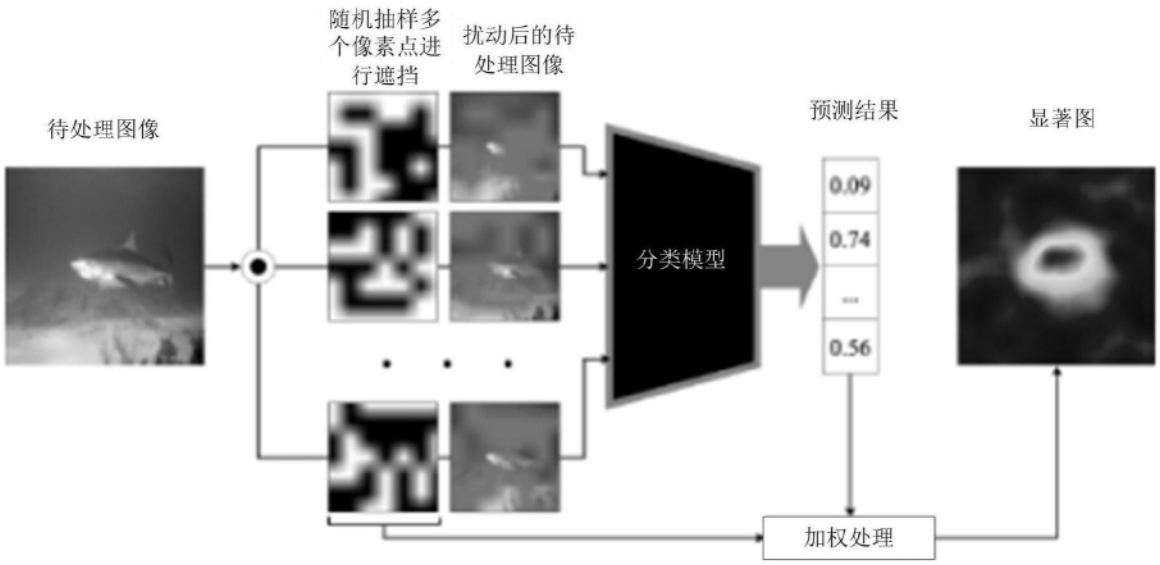

随机抽样多个像素点进行遮挡
扰动后的待处理图像
待处理图像
预测结果
显著图
分类模型
0.09
0.74
0.56
加权处理

图2


获取多个对象
301
根据第一条件对多个对象进行筛选处理，以获取更新后的多个对象，更新后的多个对象的数据分布满足目标数据分布
302
根据第一预测结果生成第一对象的显著图
303

图3


获取多个对象
301
根据第一预设条件对多个对象进行筛选处理，以获取更新后的多个对象，更新后的多个对象的数据分布满足目标数据分布
302
设置正常对象的权重为第一权重，设置异常对象的权重为第二权重，第一权重大于第二权重
3031
根据第一预测结果生成第一对象的显著图
303

图4


对训练样本进行特征提取，以获取训练样本的特征
501
根据训练样本的特征和预设模型的权重向量之间的偏差获取第一损失值
502
根据第一损失值更新预设模型
503

图5

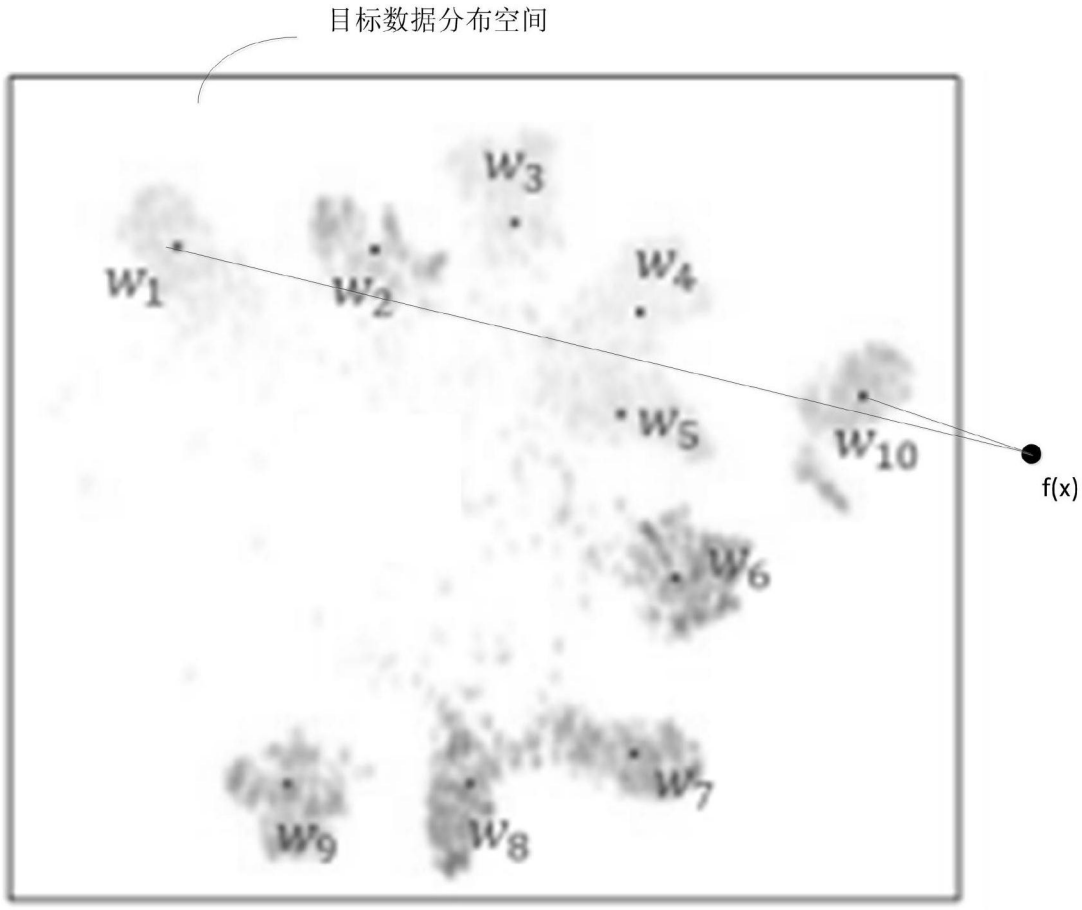


图6


对训练样本进行特征提取，以获取训练样本的特征
701
根据训练样本的特征和预设模型的权重向量之间的偏差获取第一损失值
702
根据目标结果和训练样本的真实结果之间的偏差获取第二损失值
703
根据第一损失值和第二损失值更新预设模型
704

图7

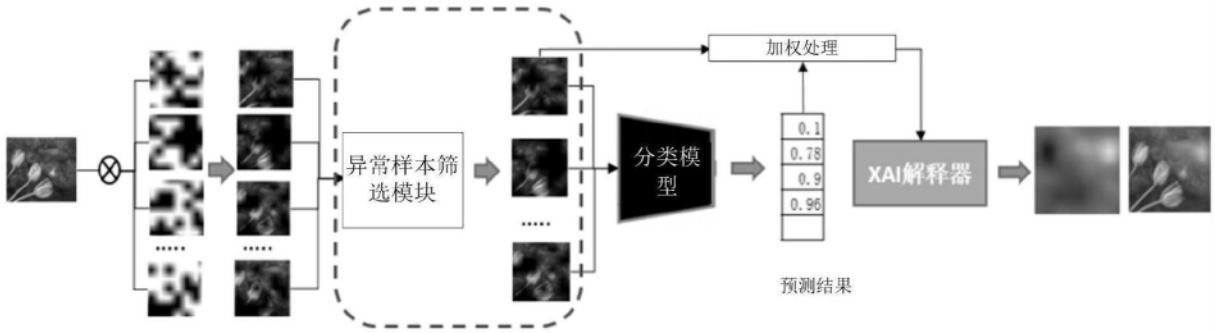

加权处理
异常样本筛选模块
分类模型
0.1
0.78
0.9
0.96
XAI解释器
预测结果

图8

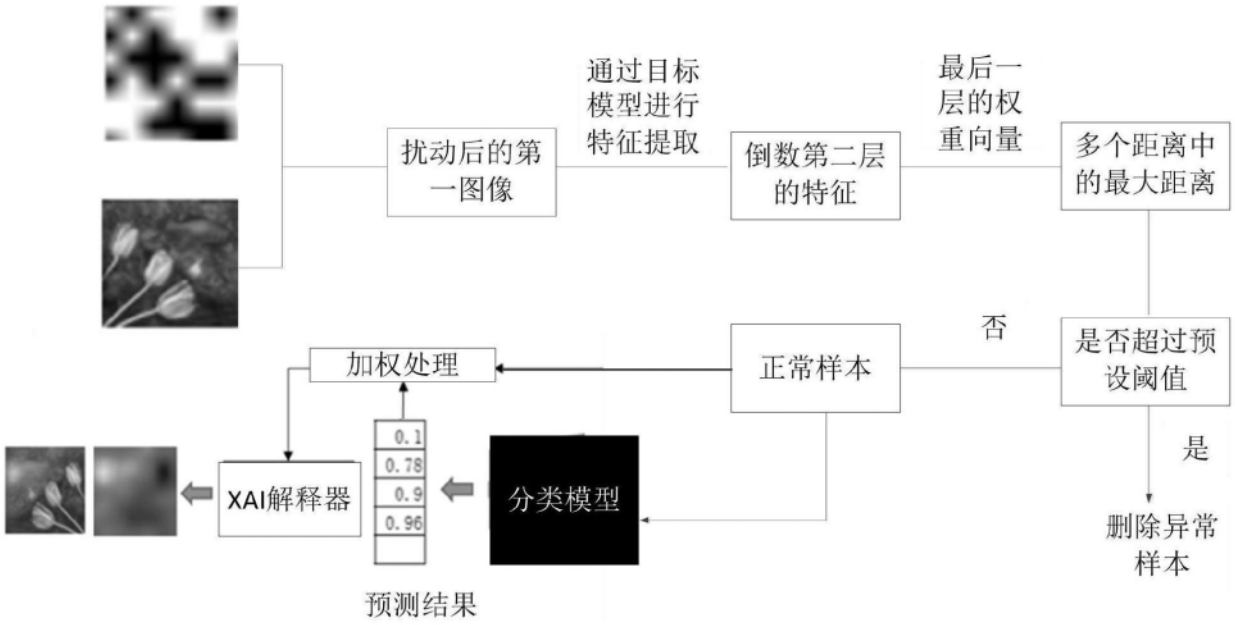
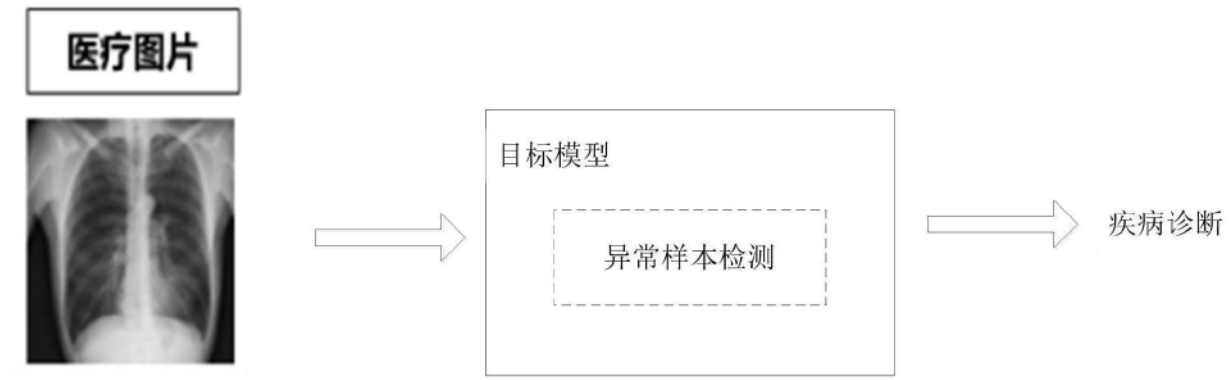

扰动后的第一图像
通过目标模型进行特征提取
倒数第二层的特征
最后一层的权重向量
多个距离中的最大距离
是否超过预设阈值
否
是
正常样本
删除异常样本
加权处理
XAI解释器
分类模型
0.1
0.78
0.9
0.96
预测结果

图9


医疗图片
目标模型
异常样本检测
疾病诊断


医疗图片
扰动
目标模型
异常样本检测
异常样本删除
可解释算法
显著图

图10


扰动
目标模型
异常样本检测
异常样本删除
可解释算法
显著图

图11

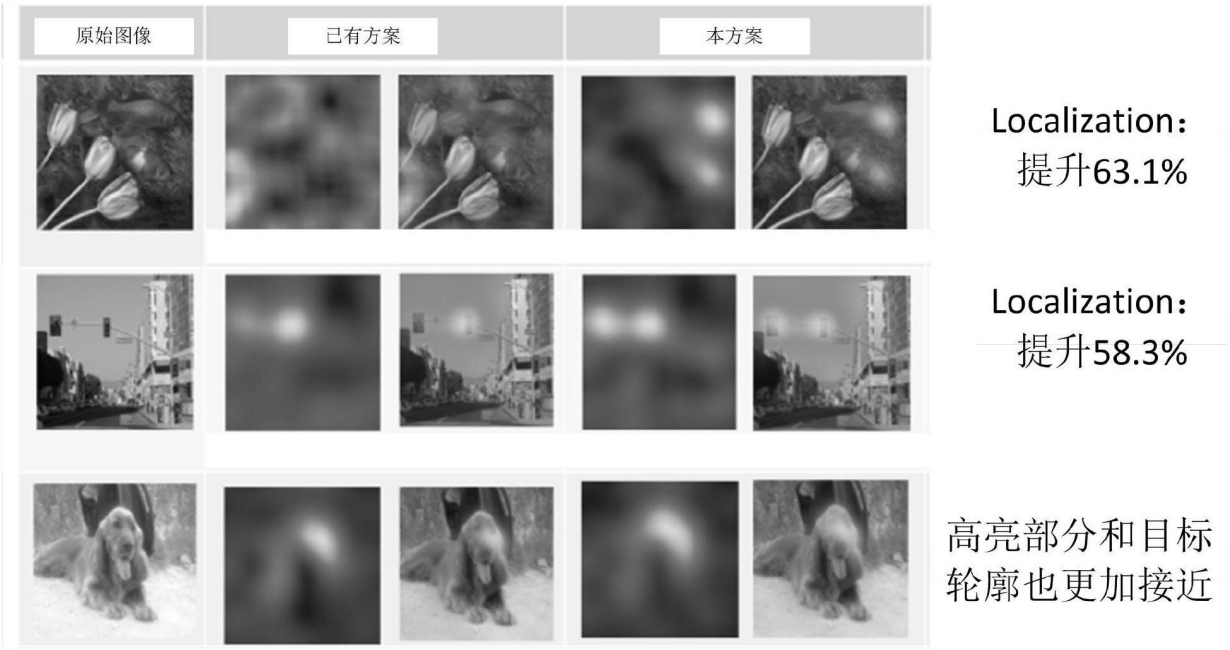

原始图像
已有方案
本方案
Localization：
提升63.1%
Localization：
提升58.3%
高亮部分和目标
轮廓也更加接近

图12

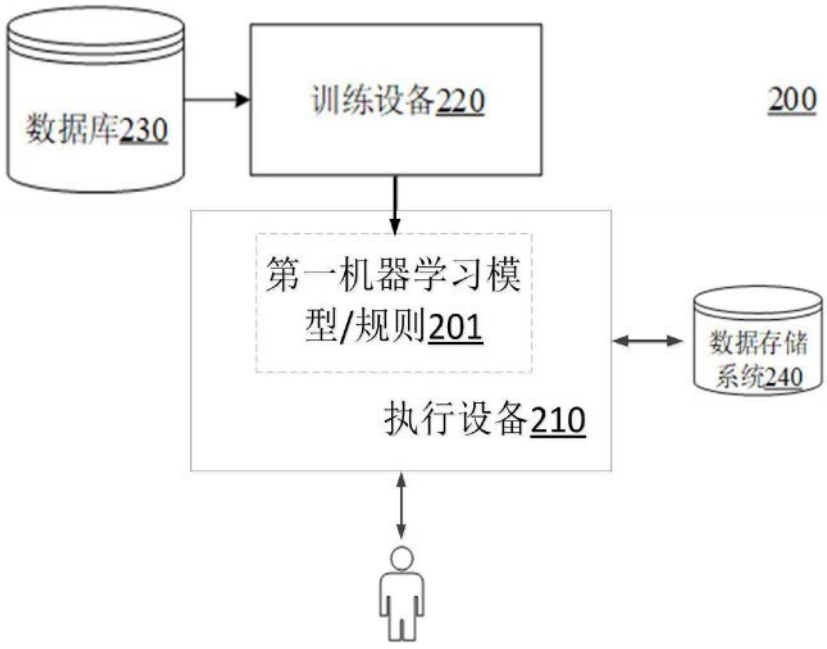

数据库230
训练设备220
200
第一机器学习模
型/规则201
数据存储
系统240
执行设备210

图13

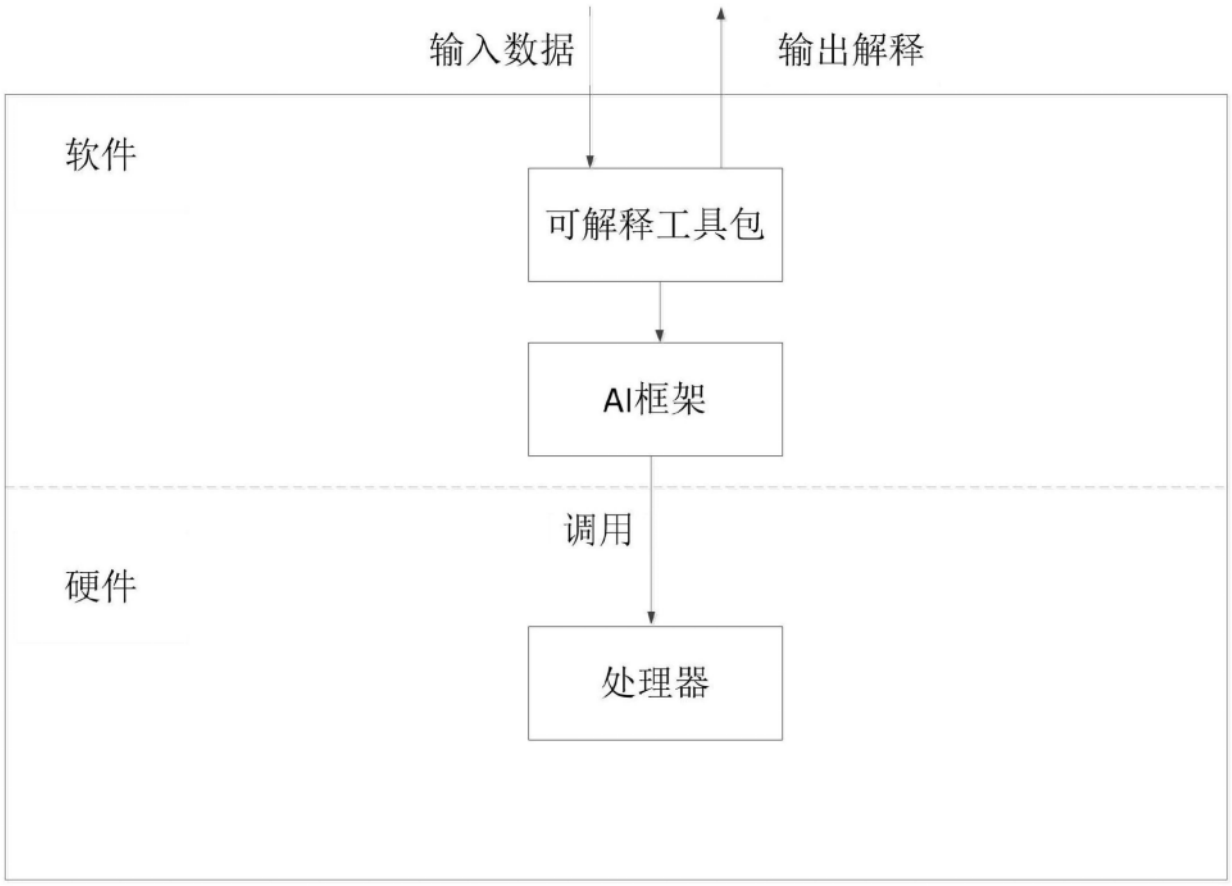

输入数据
输出解释
软件
可解释工具包
AI框架
调用
硬件
处理器

图14

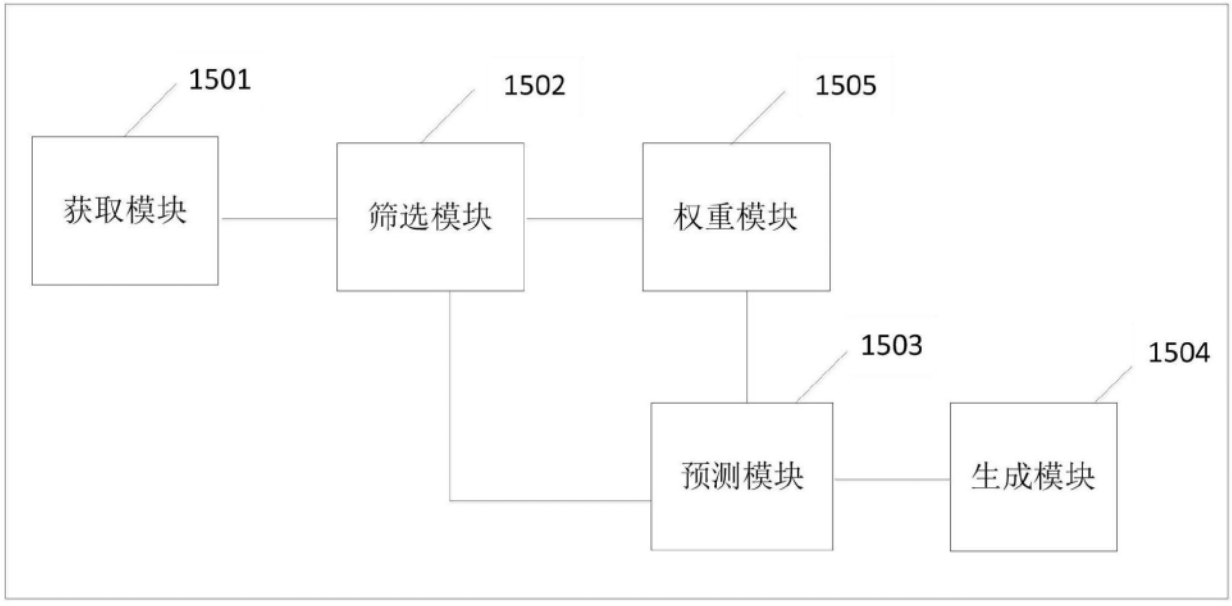

1501
1502
1505
获取模块
筛选模块
权重模块
1503
1504
预测模块
生成模块

图15

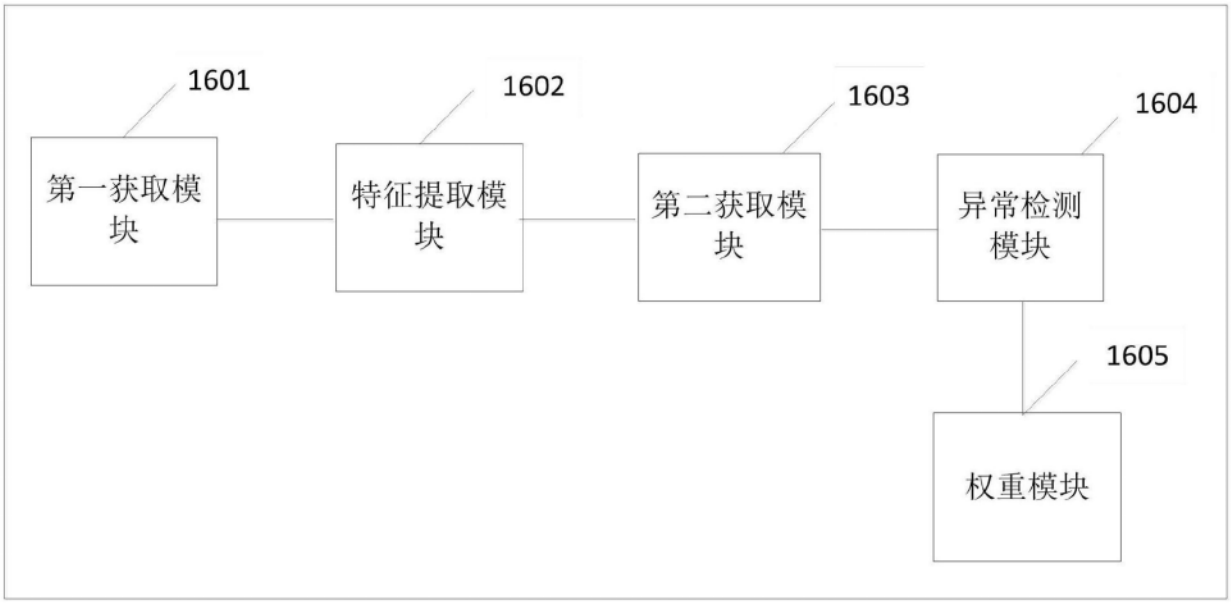

1601
第一获取模块
1602
特征提取模块
1603
第二获取模块
1604
异常检测模块
1605
权重模块

图16

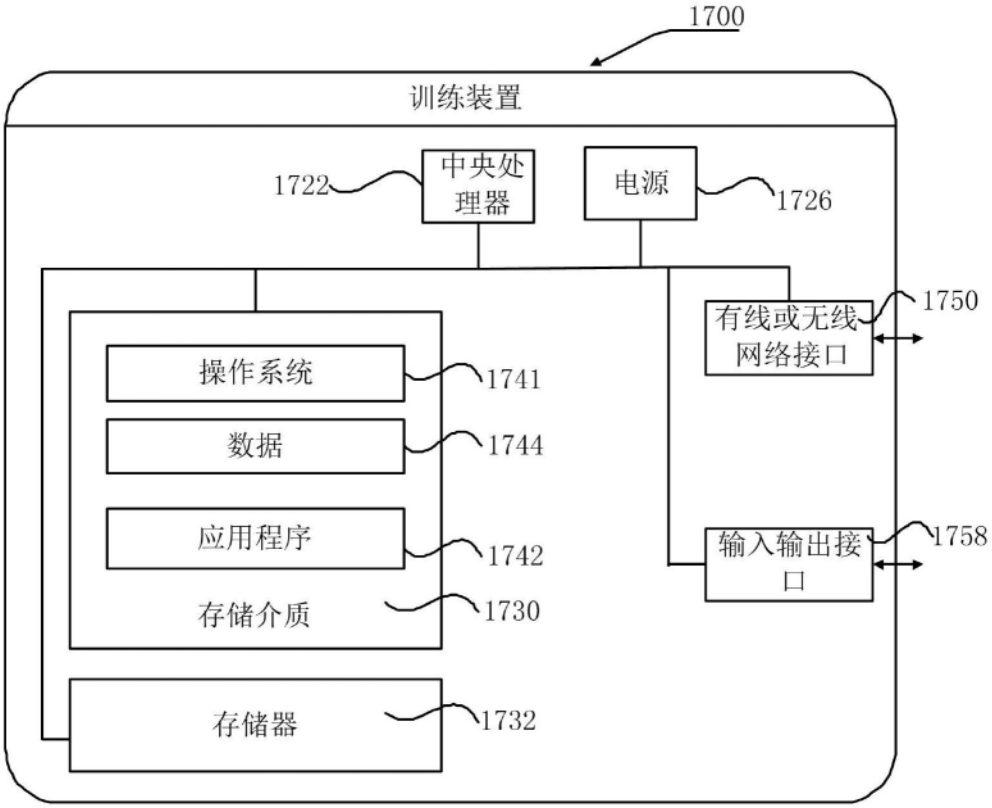

1700
训练装置
1722
中央处理器
电源
1726
1750
有线或无线网络接口
操作系统
1741
数据
1744
应用程序
1742
存储介质
1730
输入输出接口
1758
存储器
1732

图17

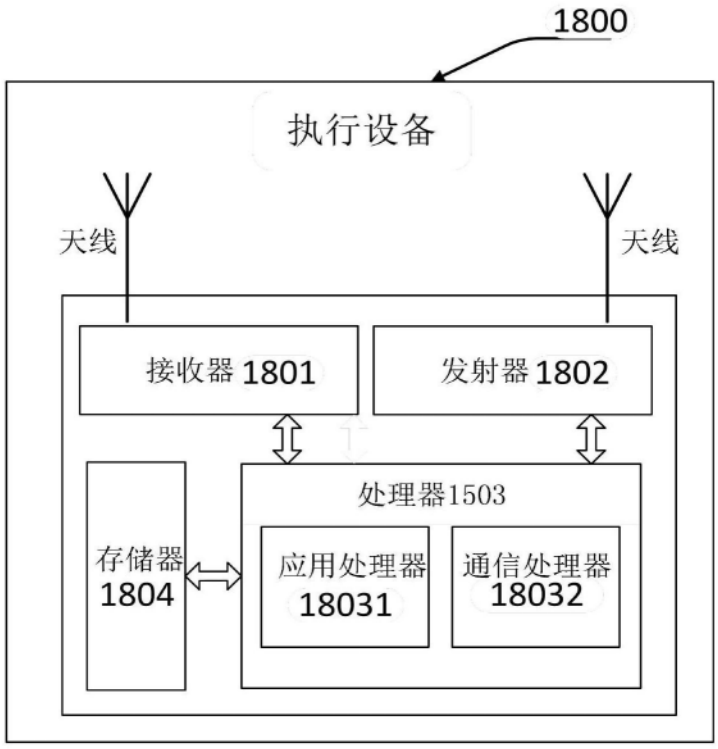

1800
执行设备
天线
天线
接收器 1801
发射器 1802
处理器1503
存储器
1804
应用处理器
18031
通信处理器
18032

图18

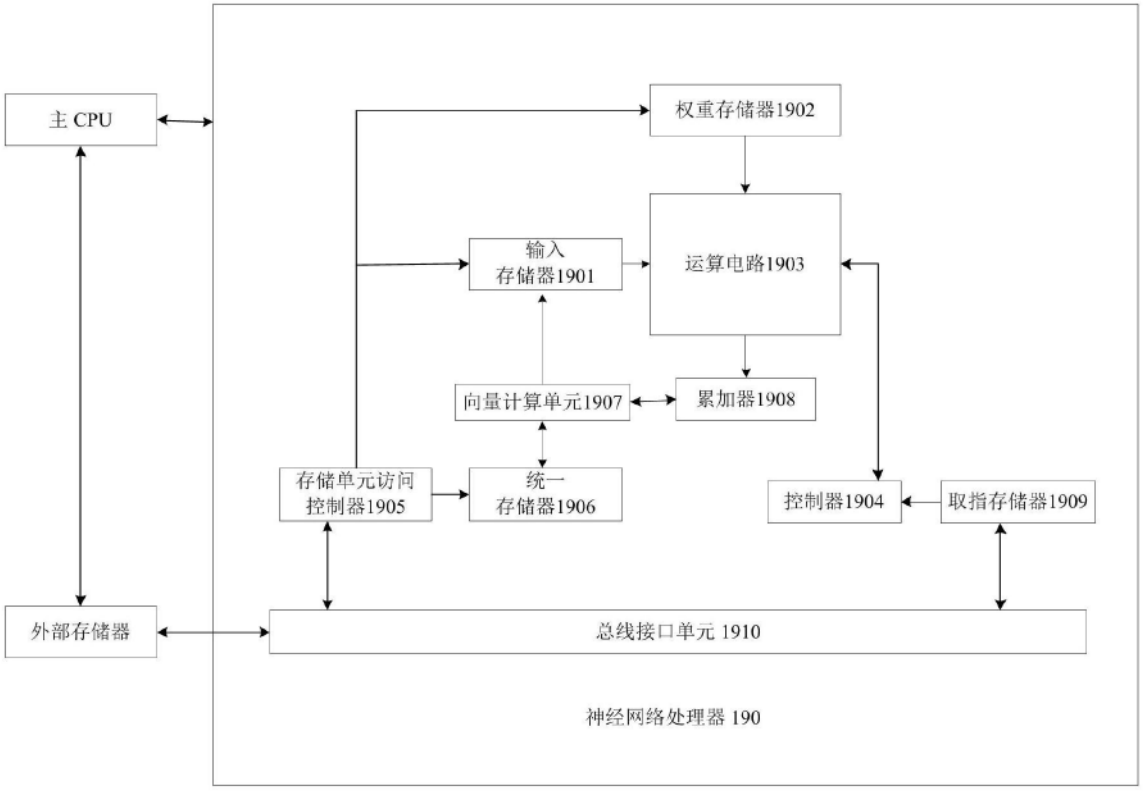

主 CPU
权重存储器1902
输入
存储器1901
运算电路1903
向量计算单元1907
累加器1908
存储单元访问
控制器1905
统一
存储器1906
控制器1904
取指存储器1909
外部存储器
总线接口单元 1910
神经网络处理器 190

图19