(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114863243 B
(45)授权公告日 2024.12.17

(21)申请号 202210463318.8

(22)申请日 2022.04.28

(65)同一申请的已公布的文献号
申请公布号 CN 114863243 A

(43)申请公布日 2022.08.05

(73)专利权人 国家电网有限公司大数据中心
地址 100031 北京市西城区宣武门内大街8号

(72)发明人 朱洪斌 刘圣龙 张舸 江伊雯 王迪 周鑫 吕艳丽 夏雨潇 赵涛 王衡

(74)专利代理机构 北京品源专利代理有限公司 11332
专利代理师 高艳红

(51)Int.Cl.
G06V 10/82(2022.01)
G06N 3/0464(2023.01)
G06N 3/082(2023.01)

(56)对比文件
CN 111667068 A,2020.09.15
CN 113204745 A,2021.08.03

审查员 章鹏

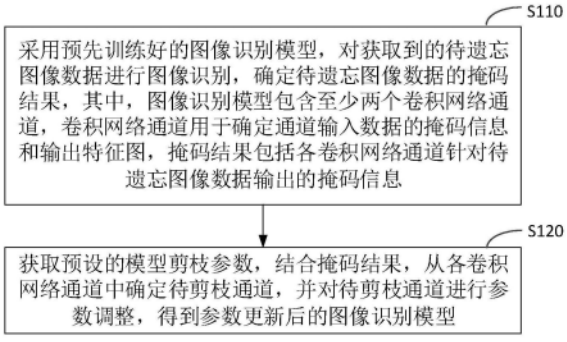
权利要求书2页 说明书10页 附图3页

(54)发明名称

一种模型的数据遗忘方法、装置、设备及存储介质

(57)摘要

本发明公开了一种模型的数据遗忘方法、装置、设备及存储介质。该方法包括:采用预先训练好的图像识别模型,对获取到的待遗忘图像数据进行图像识别,确定所述待遗忘图像数据的掩码结果,其中,所述图像识别模型包含至少两个卷积网络通道,所述卷积网络通道用于确定通道输入数据的掩码信息和输出特征图,所述掩码结果包括各所述卷积网络通道针对所述待遗忘图像数据输出的掩码信息;获取预设的模型剪枝参数,结合所述掩码结果,从各所述卷积网络通道中确定待剪枝通道,并对所述待剪枝通道进行参数调整,得到参数更新后的图像识别模型。本发明在保持模型识别准确度的同时,实现对部分训练数据的完全遗忘,使被删除的训练数据无法恢复,保护用户隐私。



CN 114863243 B

1.一种模型的数据遗忘方法,其特征在于,包括:

采用预先训练好的图像识别模型,对获取到的待遗忘图像数据进行图像识别,确定所述待遗忘图像数据的掩码结果,其中,所述图像识别模型包含至少两个卷积网络通道,所述卷积网络通道用于确定通道输入数据的掩码信息和输出特征图,所述掩码结果包括各所述卷积网络通道针对所述待遗忘图像数据输出的掩码信息;

获取预设的模型剪枝参数,结合所述掩码结果,从各所述卷积网络通道中确定待剪枝通道,并对所述待剪枝通道进行参数调整,得到参数更新后的图像识别模型。

2.根据权利要求1所述的方法,其特征在于,所述获取预设的模型剪枝参数,结合所述掩码结果,从各所述卷积网络通道中确定待剪枝通道,并对所述待剪枝通道进行参数调整,得到参数更新后的图像识别模型,包括:

根据各所述卷积网络通道对应的掩码信息,对各所述卷积网络通道进行遗忘数据相关性排序;

获取预设的模型剪枝比例和模型剪枝权重,根据所述模型剪枝比例,确定待剪枝通道数量,并将遗忘数据相关性排序最高的待剪枝通道数量个卷积网络通道确定为待剪枝通道;

根据所述模型剪枝权重,对所述待剪枝通道中的模型参数进行调整,得到参数更新后的图像识别模型。

3.根据权利要求1所述的方法,其特征在于,在从各所述卷积网络通道中确定待剪枝通道,并对所述待剪枝通道进行参数调整,得到参数更新后的图像识别模型之后,还包括:

根据所述待遗忘图像数据,对参数更新后的图像识别模型进行模型精度测试,得到第一模型测试精度;

当所述第一模型测试精度大于预设的模型遗忘阈值时,则重新根据所述待遗忘图像数据对图像识别模型进行剪枝操作,继续更新模型参数。

4.根据权利要求3所述的方法,其特征在于,在得到第一模型测试精度之后,还包括:

当所述第一模型测试精度小于等于所述模型遗忘阈值时,获取剩余图像数据;

根据所述剩余图像数据,对所述图像识别模型进行模型精度补偿,使所述图像识别模型的第二模型测试精度大于预设的模型测试阈值。

5.根据权利要求1-4中任一项所述的方法,其特征在于,所述图像识别模型的训练过程包括:

对训练图像数据进行图像识别标注,得到标准识别结果;

将所述训练图像数据输入待训练图像识别模型,获得输出的训练识别结果和训练掩码结果,其中,所述待训练图像识别模型包括至少两个待训练卷积网络通道和待训练全连接网络层;

根据所述标准识别结果、所述训练识别结果以及所述训练掩码结果,获得拟合损失函数;

通过所述拟合损失函数对所述待训练图像识别模型进行反向传播,得到所述图像识别模型。

6.根据权利要求5所述的方法,其特征在于,所述将所述训练图像数据输入待训练图像识别模型,获得输出的训练识别结果和训练掩码结果,包括:

将所述训练图像数据输入第一个待训练卷积网络通道,输出对应的训练掩码信息和训练输出特征图;

将第一个待训练卷积网络通道输出的训练输出特征图作为输入数据,输入第二个待训练卷积网络通道,输出对应的训练掩码信息和训练输出特征图,以此类推,直至最后一个待训练卷积网络通道输出对应的训练掩码信息和训练输出特征图;

将最后一个待训练卷积网络通道输出的训练输出特征图输入所述待训练全连接网络层,得到训练识别结果;

将各所述待训练卷积网络通道输出的训练掩码信息进行融合,得到训练掩码结果。

7.根据权利要求5所述的方法,其特征在于,所述根据所述标准识别结果、所述训练识别结果以及所述训练掩码结果,获得拟合损失函数,包括:

根据所述标准识别结果和所述训练识别结果,结合预设的交叉熵函数表达式,确定分类损失函数;

根据所述训练掩码结果,结合预设的L1正则化函数表达式,确定正则损失函数;

对所述分类损失函数和所述正则损失函数进行加权融合,得到拟合损失函数。

8.一种模型的数据遗忘装置,其特征在于,包括:

掩码结果确定模块,用于采用预先训练好的图像识别模型,对获取到的待遗忘图像数据进行图像识别,确定所述待遗忘图像数据的掩码结果,其中,所述图像识别模型包含至少两个卷积网络通道,所述卷积网络通道用于确定通道输入数据的掩码信息和输出特征图,所述掩码结果包括各所述卷积网络通道针对所述待遗忘图像数据输出的掩码信息;

模型通道剪枝模块,用于获取预设的模型剪枝参数,结合所述掩码结果,从各所述卷积网络通道中确定待剪枝通道,并对所述待剪枝通道进行参数调整,得到参数更新后的图像识别模型。

9.一种电子设备,其特征在于,所述电子设备包括:

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计算机程序,所述计算机程序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执行权利要求1-7中任一项所述的模型的数据遗忘方法。

10.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质存储有计算机指令,所述计算机指令用于使处理器执行时实现权利要求1-7中任一项所述的模型的数据遗忘方法。

一种模型的数据遗忘方法、装置、设备及存储介质

技术领域

[0001] 本发明涉及机器学习技术领域，尤其涉及一种模型的数据遗忘方法、装置、设备及存储介质。

背景技术

[0002] 随着机器学习技术的不断发展，多数企业正在构建更多的机器学习模型。在实际应用中，想要得到一个足够精确的模型，就需要大量的实际数据去训练对应的神经网络，而公开的数据集往往是难以满足这个要求的。因此，企业通常会收集所需要的用户数据，构建对应的用户数据集，使得训练出的模型能够具有更高的性能。但是对用户而言，在上传自己的数据之后，即使之后再向企业提出删除数据的请求，企业方也往往只是删除用户的原始数据，对于通过用户数据所训练得到的模型，却不会进行调整。

[0003] 近些年的一些研究表明，如果用户的数据曾经被用于训练神经网络模型，那么就可能通过一些攻击手段来获取原始用于训练的数据。例如，成员推理攻击可以通过对模型的输出进行攻击，恢复训练集中的某些图片。而这些攻击方式的存在，使得用户即便已经要求企业删除自己的数据，仍然可能会被第三方通过模型攻击来获取用户的隐私数据。因此，如何有效地在已经训练好的模型中遗忘某些训练数据，对于能否满足用户的被遗忘权至关重要。

发明内容

[0004] 本发明提供了一种模型的数据遗忘方法、装置、设备及存储介质，在保持模型识别准确度的同时，实现对部分训练数据的完全遗忘，从而保护用户隐私。

[0005] 根据本发明的一方面，提供了一种模型的数据遗忘方法，该方法包括：

[0006] 采用预先训练好的图像识别模型，对获取到的待遗忘图像数据进行图像识别，确定所述待遗忘图像数据的掩码结果，其中，所述图像识别模型包含至少两个卷积网络通道，所述卷积网络通道用于确定通道输入数据的掩码信息和输出特征图，所述掩码结果包括各所述卷积网络通道针对所述待遗忘图像数据输出的掩码信息；

[0007] 获取预设的模型剪枝参数，结合所述掩码结果，从各所述卷积网络通道中确定待剪枝通道，并对所述待剪枝通道进行参数调整，得到参数更新后的图像识别模型。

[0008] 根据本发明的另一方面，提供了一种模型的数据遗忘装置，该装置包括：

[0009] 掩码结果确定模块，用于采用预先训练好的图像识别模型，对获取到的待遗忘图像数据进行图像识别，确定所述待遗忘图像数据的掩码结果，其中，所述图像识别模型包含至少两个卷积网络通道，所述卷积网络通道用于确定通道输入数据的掩码信息和输出特征图，所述掩码结果包括各所述卷积网络通道针对所述待遗忘图像数据输出的掩码信息；

[0010] 模型通道剪枝模块，用于获取预设的模型剪枝参数，结合所述掩码结果，从各所述卷积网络通道中确定待剪枝通道，并对所述待剪枝通道进行参数调整，得到参数更新后的图像识别模型。

[0011] 根据本发明的另一方面,提供了一种电子设备,所述电子设备包括:

[0012] 至少一个处理器;以及

[0013] 与所述至少一个处理器通信连接的存储器;其中,

[0014] 所述存储器存储有可被所述至少一个处理器执行的计算机程序,所述计算机程序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执行本发明任一实施例所述的模型的数据遗忘方法。

[0015] 根据本发明的另一方面,提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机指令,所述计算机指令用于使处理器执行时实现本发明任一实施例所述的模型的数据遗忘方法。

[0016] 本发明实施例的技术方案,通过采用预先训练好的图像识别模型,对获取到的待遗忘图像数据进行图像识别,确定待遗忘图像数据的掩码结果,其中,图像识别模型包含至少两个卷积网络通道,卷积网络通道用于确定通道输入数据的掩码信息和输出特征图,掩码结果包括各卷积网络通道针对待遗忘图像数据输出的掩码信息;获取预设的模型剪枝参数,结合掩码结果,从各卷积网络通道中确定待剪枝通道,并对待剪枝通道进行参数调整,得到参数更新后的图像识别模型,本发明在保持模型识别准确度的同时,实现对部分训练数据的完全遗忘,使被删除的训练数据无法恢复,保护用户隐私。

[0017] 应当理解,本部分所描述的内容并非旨在标识本发明的实施例的关键或重要特征,也不用于限制本发明的范围。本发明的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0018] 为了更清楚地说明本发明实施例中的技术方案,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他的附图。

[0019] 图1a是根据本发明实施例一提供的一种模型的数据遗忘方法的流程图;

[0020] 图1b是根据本发明实施例一提供的一种模型的数据遗忘方法中卷积网络通道的原理示意图;

[0021] 图1c是根据本发明实施例一提供的一种模型的数据遗忘方法中图像识别模型的原理示意图;

[0022] 图2是根据本发明实施例二提供的一种模型的数据遗忘装置的结构示意图;

[0023] 图3是实现本发明实施例的模型的数据遗忘方法的电子设备的结构示意图。

具体实施方式

[0024] 为了使本技术领域的人员更好地理解本发明方案,下面将结合本发明实施例中的附图,对本发明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发明一部分的实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都应当属于本发明保护的范围。

[0025] 需要说明的是,本发明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,以便这里描述的本发明的实施例能够以除了在这里图示或描述的那些以外的顺序实施。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单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,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单元。

[0026] 实施例一

[0027] 图1a为本发明实施例一提供了一种模型的数据遗忘方法的流程图,本实施例可适用于删除部分训练数据后对模型进行调参的情况,该方法可以由模型的数据遗忘装置来执行,该模型的数据遗忘装置可以采用硬件和/或软件的形式实现,该模型的数据遗忘装置可配置于计算机中。如图1a所示,该方法包括:

[0028] S110、采用预先训练好的图像识别模型,对获取到的待遗忘图像数据进行图像识别,确定待遗忘图像数据的掩码结果,其中,图像识别模型包含至少两个卷积网络通道,卷积网络通道用于确定通道输入数据的掩码信息和输出特征图,掩码结果包括各卷积网络通道针对待遗忘图像数据输出的掩码信息。

[0029] 在本实施例中,图像识别模型由训练图像数据训练得到,而待遗忘图像数据是训练图像数据中的部分数据。待遗忘图像数据可以是模型训练数据中的一条数据,也可以是多条数据组成的数据集合,待遗忘图像数据可以用D_forget表示。

[0030] 具体的,对于获取到的待遗忘图像数据,可以将其作为图像识别模型的输入数据,得到对应的输出,包括模型的识别结果和掩码结果。在使用图像识别模型识别图像时,主要关注模型输出的识别结果,而在对模型进行调整,使其遗忘部分训练数据时,主要关注模型输出的掩码结果。本实施例中的图像识别模型包含至少两个卷积网络通道,每个卷积网络通道都会输出对应的掩码信息,所有掩码信息就组成了掩码结果。掩码结果可以包含各个卷积网络通道与待遗忘图像数据的相关性。

[0031] 本实施例的为了在保持图像识别模型识别准确度的同时,实现对部分训练数据的完全遗忘,满足这个要求,首先需要调整图像识别模型的结构,设计一种有利于数据遗忘的网络模型,并且需要在遗忘用户数据之后还能保持比较高的准确度。利用这种网络结构,在训练时设计特定的优化目标,使得不同的数据所需要通过的计算路径变得稀疏,从而减少相互之间的关联。在进行数据遗忘时,首先统计出需要遗忘的数据所通过的计算路径,之后针对于这些路径进行定制化的剪枝操作,删除与这些数据有关的模型参数,实现对部分数据的遗忘,使得攻击者无法再恢复出被删除的数据。通过这种方式,企业可以在收到用户的数据删除请求后,实现完全的数据遗忘,起到对用户隐私的保护作用。

[0032] 可选的,本实施例所使用的图像识别模型的训练过程可以包括以下步骤:

[0033] A1、对训练图像数据进行图像识别标注,得到标准识别结果。

[0034] 具体的,可以对训练图像数据进行人工标注,得到标准识别结果。

[0035] A2、将训练图像数据输入待训练图像识别模型,获得输出的训练识别结果和训练掩码结果,其中,待训练图像识别模型包括至少两个待训练卷积网络通道和待训练全连接网络层。

[0036] 在本实施例中,待训练图像识别模型基于多通道卷积神经网络添加了掩码生成功

能,从而控制输入数据的计算路径。首先,可以构建基本的卷积模块,每个卷积模块可以由三部分组成,分别为卷积计算单元(Conv),归一化单元(BN)和激活函数单元(Ac)。对于每个卷积模块,会对输入特征图(x)进行计算,得到输出特征图为y=Ac(BN(Conv(x)))。将这些卷积模块连接起来,可以实现基本的卷积神经网络。其次,对每一个卷积模块,在其上添加一个掩码模块。掩码模块首先使用平均池化单元(AvgPool)对输入的数据按通道进行压缩,得到一组通道显著值,之后使用全连接单元(FC)计算得到一组掩码,掩码的长度与卷积层的输出通道数相同,也即m=FC(AvgPool(x))。最后将掩码与卷积层的输出相乘,再输入到归一化层和激活层,就形成了一个待训练卷积网络通道,得到待训练卷积网络通道的输出特征图为y=Ac(BN(Conv(x)·m))。同时,每个掩码模块的计算结果m作为掩码信息输出。

[0037] 在待训练卷积网络通道中,卷积计算单元通过卷积核参数对输入的特征图进行卷积计算,并将输入通道数变换到输出通道数,归一化单元对卷积计算的结果进行归一化处理,之后通过激活单元对特征进行非线性处理。另一部分是掩码模块,主要包括平均池化单元和全连接单元,对输入的特征图计算得到输出通道的显著值。以输入特征图为原始图片为例,输入特征图的尺寸为3*32*32,卷积计算单元使用3*3大小的卷积核,输入通道为3,输出通道为64,则卷积计算单元的参数尺寸为64*3*3*3,通过卷积计算单元后,特征图的尺寸为64*32*32;另一方面,掩码模块通过池化单元和全连接单元计算得到输出通道的显著值,尺寸为64*1,其中每个数据归一化到[0,1]之间,用于表示输出通道的重要性,之后将特征图与通道显著值相乘,得到增加了掩码的特征图,尺寸仍然为64*32*32;最后通过归一化和激活单元得到最终输出的特征图,尺寸为64*32*32。同时,将掩码模块的输出也一并添加到结果中,留待后续计算正则损失使用。

[0038] 另外,在使用待训练图像识别模型之前,可以对模型进行初始化,本发明中所使用的初始化方式可以为高斯初始化,即对模型中的每个参数,都随机地从高斯分布中进行采样,并作为初始值。

[0039] 进一步的,A2的具体实现步骤可以是:

[0040] A21、将训练图像数据输入第一个待训练卷积网络通道,输出对应的训练掩码信息和训练输出特征图。

[0041] A22、将第一个待训练卷积网络通道输出的训练输出特征图作为输入数据,输入第二个待训练卷积网络通道,输出对应的训练掩码信息和训练输出特征图,以此类推,直至最后一个待训练卷积网络通道输出对应的训练掩码信息和训练输出特征图。

[0042] A23、将最后一个待训练卷积网络通道输出的训练输出特征图输入待训练全连接网络层,得到训练识别结果。

[0043] A24、将各待训练卷积网络通道输出的训练掩码信息进行融合,得到训练掩码结果。

[0044] 示例性的,图1b是根据本发明实施例一提供的一种模型的数据遗忘方法中卷积网络通道的原理示意图。如图1b所示,将输入特征图输入待训练卷积网络通道,经过计算可以输出掩码信息和输出特征图。

[0045] 本实施例使用添加了掩码模块的卷积模块构建完整的神经网络,将多个卷积模块及掩码模块级联起来,其中每个卷积模块对应一个掩码模块,形成一个卷积网络通道。第一个卷积网络通道的输入为原始的图像数据,之后每个卷积网络通道的输入为上一级卷积网

络通道的输出特征图,在最后一个卷积网络通道之后添加一个全连接单元,用于输出最终识别结果。同时,在训练过程中,将每个卷积网络通道的掩码信息也作为辅助结果输出。

[0046] 示例性的,图1c是根据本发明实施例一提供的一种模型的数据遗忘方法中图像识别模型的原理示意图。如图1c所示,将训练图像数据作为输入图像输入搭建好的神经网络模型,输出的识别结果和掩码结果即为训练图像数据对应的训练识别结果和训练掩码结果。

[0047] A3、根据标准识别结果、训练识别结果以及训练掩码结果,获得拟合损失函数。

[0048] 具体的,由于标准识别结果为人为标注真实的识别结果,而训练识别结果为模型在训练过程中计算得到的,因此标准识别结果和训练识别结果必然会存在一定误差,可以根据标准识别结果、训练识别结果,结合模型辅助生成的训练掩码结果,计算得到拟合损失函数。

[0049] 进一步的,A3的具体实现步骤可以是:

[0050] A31、根据标准识别结果和训练识别结果,结合预设的交叉熵函数表达式,确定分类损失函数。

[0051] A32、根据训练掩码结果,结合预设的L1正则化函数表达式,确定正则损失函数。

[0052] A33、对分类损失函数和正则损失函数进行加权融合,得到拟合损失函数。

[0053] 在本实施例中,可以预先设计模型的训练损失函数。模型的训练损失可以主要包含两个部分,一个是分类误差损失函数,主要使用交叉熵函数进行计算,使得模型的预测结果尽可能拟合实际训练的数据,具体的形式可以为$L_1=-\frac{1}{N}\sum_i\sum_{c=1}^{M}y_{ic}\log p_{ic}$,其中,N可以表示样本数目,M可以表示类别数目,$y_{ic}$可以表示样本的的真实类别是否为c,$p_{ic}$可以表示预测样本属于类别c的概率。另一个是掩码结果的损失函数,可以使用L1正则化函数进行计算,使得掩码的输出结果尽可能变得稀疏,也就是激活更少的卷积计算通道,使得每个样本都只使用更少的通道就能够完成计算,从而使不同样本之间所通过的相同通道数目减少,也即减少了不同样本之间的耦合性,使得模型在进行遗忘部分数据时能够减少对其他数据的精度影响,具体的形式可以为$L_2=||w||_1=\Sigma_i|w_i|$,其中w可以表示各个掩码模块的输出结果,即各待训练卷积网络通道输出的训练掩码信息。在得到分类损失函数和正则损失函数后,可以将两者相加,得到模型的总训练损失,即拟合损失函数$L=L_1+L_2$。在计算拟合损失函数时,也可以根据实际需求调整分类损失函数和正则损失函数的权重。

[0054] A4、通过拟合损失函数对待训练图像识别模型进行反向传播,得到图像识别模型。

[0055] 进一步的,A4的具体实现步骤可以是:对拟合损失函数求导,确定各待训练卷积网络通道的参数梯度,采用梯度更新方法更新各待训练卷积网络通道的参数,得到图像识别模型。

[0056] 具体的,在得到拟合损失函数后,可以对拟合损失函数求导,得到各个通道参数的梯度,并使用梯度更新算法更新模型的参数,完成一组训练过程,直到全部数据集都被训练完毕,或者模型的模型测试精度大于预设的模型测试阈值,完成模型训练,保存模型的参数,之后可以将模型部署到实际的应用中。

[0057] S120、获取预设的模型剪枝参数,结合掩码结果,从各卷积网络通道中确定待剪枝通道,并对待剪枝通道进行参数调整,得到参数更新后的图像识别模型。

[0058] 在本实施例中,对于训练完成的模型,如果需要遗忘一部分训练集中的数据,就需要进行模型的遗忘过程。模型的遗忘主要依赖于掩码模块,通过计算待遗忘图像数据的掩码结果,可以得到识别待遗忘图像数据时所激活的卷积通道,之后可以通过剪枝去除这些通道的数据,从而完成对数据的遗忘。

[0059] 可选的,S120可以通过以下步骤具体实现:

[0060] S1201、根据各卷积网络通道对应的掩码信息,对各卷积网络通道进行遗忘数据相关性排序。

[0061] S1202、获取预设的模型剪枝比例和模型剪枝权重,根据模型剪枝比例,确定待剪枝通道数量,并将遗忘数据相关性排序最高的待剪枝通道数量个卷积网络通道确定为待剪枝通道。

[0062] S1203、根据模型剪枝权重,对待剪枝通道中的模型参数进行调整,得到参数更新后的图像识别模型。

[0063] 实际应用中,当待遗忘图像数据包含多条图像数据时,可以将每条图像数据对应的掩码结果按照位置进行累加,最终得到整个遗忘数据集在模型的卷积模块中计算得到的掩码累加值,此数据可体现模型参数与待遗忘图像数据之间的相关性,累加值越大,说明对应的卷积网络通道参数与输入数据的相关性越强,因此需要优先处理相关性较大的通道参数,以达到数据遗忘的目的。可以将掩码累加值按照大小进行排序,根据预先设置好的模型剪枝比例P,确定相关性较高的掩码,并通过掩码模块与卷积模块的输出通道关系对应到相关的卷积网络通道,将这些卷积网络通道标记为待剪枝通道。根据预先设置的模型剪枝权重W,对待剪枝通道的模型参数进行修改,本实施例可以对待剪枝通道中的每个模型参数分别乘以(1-W),模型剪枝比例W越大,则剪枝后模型参数的变化就越大,模型中残留的信息就越少,从而达到数据遗忘的目的。

[0064] 本发明实施例的技术方案,通过采用预先训练好的图像识别模型,对获取到的待遗忘图像数据进行图像识别,确定待遗忘图像数据的掩码结果,其中,图像识别模型包含至少两个卷积网络通道,卷积网络通道用于确定通道输入数据的掩码信息和输出特征图,掩码结果包括各卷积网络通道针对待遗忘图像数据输出的掩码信息;获取预设的模型剪枝参数,结合掩码结果,从各卷积网络通道中确定待剪枝通道,并对待剪枝通道进行参数调整,得到参数更新后的图像识别模型,本发明实施例在保持模型识别准确度的同时,实现对部分训练数据的完全遗忘,使被删除的训练数据无法恢复,保护用户隐私。

[0065] 在上述方案的基础上,本实施例提供的模型的数据遗忘方法还可以包括以下步骤:

[0066] S130、根据待遗忘图像数据,对参数更新后的图像识别模型进行模型精度测试,得到第一模型测试精度。

[0067] 具体的,在对模型进行剪枝操作后,可以使用待遗忘图像数据对模型的精度进行测试,得到第一模型测试精度,如果第一模型测试精度大于预设的模型遗忘阈值时,可以进行S140;否则进行S150。

[0068] S140、当第一模型测试精度大于预设的模型遗忘阈值时,则重新根据待遗忘图像数据对图像识别模型进行剪枝操作,继续更新模型参数。

[0069] 具体的,当第一模型测试精度大于预设的模型遗忘阈值时,可以认为剪枝效果不

符合数据遗忘标准,可以返回S120继续重复剪枝。

[0070] S150、当第一模型测试精度小于等于模型遗忘阈值时,获取剩余图像数据;根据剩余图像数据,对图像识别模型进行模型精度补偿,使图像识别模型的第二模型测试精度大于预设的模型测试阈值。

[0071] 具体的,当第一模型测试精度小于等于模型遗忘阈值时,可以认为剪枝效果符合数据遗忘标准,则停止剪枝操作,进行使用剩余图像数据,对模型的精度进行补偿训练,用于弥补剪枝导致的精度下降。

[0072] 在本实施例中,剩余图像数据可以理解为模型训练数据中除待遗忘图像数据以外的其他数据,可以表示为D_retain。可以使用剩余图像数据,通过重复步骤A1～A4,使得模型参数在剩余图像数据上进行微调,使模型的测试精度恢复到数据遗忘前的状态,完成数据遗忘和精度补偿,保存模型的参数,之后可以将模型重新部署到实际的应用中。

[0073] 本发明实施例提供的模型的数据遗忘方法,可以在不影响深度学习模型性能的情况下,实现对一部分训练数据的遗忘算法,同时保持对剩余图像数据的预测精度。本发明实施例通过控制输入数据的计算路径,减少了不同类型数据的计算单元的重叠部分,使得在进行数据遗忘时,通过剪枝操作去除所遗忘数据的计算单元,并且减少对其余数据集的影响,并通过精度补偿过程弥补剪枝所带来的精度损失,从而使模型保持高可用性。

[0074] 实施例二

[0075] 图2为本发明实施例二提供的一种模型的数据遗忘装置的结构示意图。如图2所示,该装置包括:

[0076] 掩码结果确定模块210,用于采用预先训练好的图像识别模型,对获取到的待遗忘图像数据进行图像识别,确定所述待遗忘图像数据的掩码结果,其中,所述图像识别模型包含至少两个卷积网络通道,所述卷积网络通道用于确定通道输入数据的掩码信息和输出特征图,所述掩码结果包括各所述卷积网络通道针对所述待遗忘图像数据输出的掩码信息。

[0077] 模型通道剪枝模块220,用于获取预设的模型剪枝参数,结合所述掩码结果,从各所述卷积网络通道中确定待剪枝通道,并对所述待剪枝通道进行参数调整,得到参数更新后的图像识别模型。

[0078] 可选的,模型通道剪枝模块220包括:

[0079] 数据相关性排序单元,用于根据各所述卷积网络通道对应的掩码信息,对各所述卷积网络通道进行遗忘数据相关性排序;

[0080] 待剪枝通道确定单元,用于获取预设的模型剪枝比例和模型剪枝权重,根据所述模型剪枝比例,确定待剪枝通道数量,并将遗忘数据相关性排序最高的待剪枝通道数量个卷积网络通道确定为待剪枝通道;

[0081] 通道参数剪枝单元,用于根据所述模型剪枝权重,对所述待剪枝通道中的模型参数进行调整,得到参数更新后的图像识别模型。

[0082] 可选的,所述装置还包括第一模型精度测试模块,用于:

[0083] 在从各所述卷积网络通道中确定待剪枝通道,并对所述待剪枝通道进行参数调整,得到参数更新后的图像识别模型之后,根据所述待遗忘图像数据,对参数更新后的图像识别模型进行模型精度测试,得到第一模型测试精度;

[0084] 当所述第一模型测试精度大于预设的模型遗忘阈值时,则重新根据所述待遗忘图

像数据对图像识别模型进行剪枝操作,继续更新模型参数。

[0085] 可选的,所述装置还包括第二模型精度测试模块,用于:

[0086] 在得到第一模型测试精度之后,当所述第一模型测试精度小于等于所述模型遗忘阈值时,获取剩余图像数据;

[0087] 根据所述剩余图像数据,对所述图像识别模型进行模型精度补偿,使所述图像识别模型的第二模型测试精度大于预设的模型测试阈值。

[0088] 可选的,所述图像识别模型的训练过程包括:

[0089] 对训练图像数据进行图像识别标注,得到标准识别结果;

[0090] 将所述训练图像数据输入待训练图像识别模型,获得输出的训练识别结果和训练掩码结果,其中,所述待训练图像识别模型包括至少两个待训练卷积网络通道和待训练全连接网络层;

[0091] 根据所述标准识别结果、所述训练识别结果以及所述训练掩码结果,获得拟合损失函数;

[0092] 通过所述拟合损失函数对所述待训练图像识别模型进行反向传播,得到所述图像识别模型。

[0093] 可选的,所述将所述训练图像数据输入待训练图像识别模型,获得输出的训练识别结果和训练掩码结果,包括:

[0094] 将所述训练图像数据输入第一个待训练卷积网络通道,输出对应的训练掩码信息和训练输出特征图;

[0095] 将第一个待训练卷积网络通道输出的训练输出特征图作为输入数据,输入第二个待训练卷积网络通道,输出对应的训练掩码信息和训练输出特征图,以此类推,直至最后一个待训练卷积网络通道输出对应的训练掩码信息和训练输出特征图;

[0096] 将最后一个待训练卷积网络通道输出的训练输出特征图输入所述待训练全连接网络层,得到训练识别结果;

[0097] 将各所述待训练卷积网络通道输出的训练掩码信息进行融合,得到训练掩码结果。

[0098] 可选的,所述根据所述标准识别结果、所述训练识别结果以及所述训练掩码结果,获得拟合损失函数,包括:

[0099] 根据所述标准识别结果和所述训练识别结果,结合预设的交叉熵函数表达式,确定分类损失函数;

[0100] 根据所述训练掩码结果,结合预设的L1正则化函数表达式,确定正则损失函数;

[0101] 对所述分类损失函数和所述正则损失函数进行加权融合,得到拟合损失函数。

[0102] 本发明实施例所提供的模型的数据遗忘装置可执行本发明任意实施例所提供的模型的数据遗忘方法,具备执行方法相应的功能模块和有益效果。

[0103] 实施例三

[0104] 图3示出了可以用来实施本发明的实施例的电子设备10的结构示意图。电子设备旨在表示各种形式的数字计算机,诸如,膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置,诸如,个人数字处理、蜂窝电话、智能电话、可穿戴设备(如头盔、眼镜、手表

等)和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例,并且不意在限制本文中描述的和/或者要求的本发明的实现。

[0105] 如图3所示,电子设备10包括至少一个处理器11,以及与至少一个处理器11通信连接的存储器,如只读存储器(ROM)12、随机访问存储器(RAM)13等,其中,存储器存储有可被至少一个处理器执行的计算机程序,处理器11可以根据存储在只读存储器(ROM)12中的计算机程序或者从存储单元18加载到随机访问存储器(RAM)13中的计算机程序,来执行各种适当的动作和处理。在RAM 13中,还可存储电子设备10操作所需的各种程序和数据。处理器11、ROM 12以及RAM 13通过总线14彼此相连。输入/输出(I/O)接口15也连接至总线14。

[0106] 电子设备10中的多个部件连接至I/O接口15,包括:输入单元16,例如键盘、鼠标等;输出单元17,例如各种类型的显示器、扬声器等;存储单元18,例如磁盘、光盘等;以及通信单元19,例如网卡、调制解调器、无线通信收发机等。通信单元19允许电子设备10通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0107] 处理器11可以是各种具有处理和计算能力的通用和/或专用处理组件。处理器11的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的处理器、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。处理器11执行上文所描述的各个方法和处理,例如模型的数据遗忘方法。

[0108] 在一些实施例中,模型的数据遗忘方法可被实现为计算机程序,其被有形地包含于计算机可读存储介质,例如存储单元18。在一些实施例中,计算机程序的部分或者全部可以经由ROM 12和/或通信单元19而被载入和/或安装到电子设备10上。当计算机程序加载到RAM 13并由处理器11执行时,可以执行上文描述的模型的数据遗忘方法的一个或多个步骤。备选地,在其他实施例中,处理器11可以通过其他任何适当的方式(例如,借助于固件)而被配置为执行模型的数据遗忘方法。

[0109] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0110] 用于实施本发明的方法的计算机程序可以采用一个或多个编程语言的任何组合来编写。这些计算机程序可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器,使得计算机程序当由处理器执行时使流程图和/或框图中所规定的功能/操作被实施。计算机程序可以完全在机器上执行、部分地在机器上执行,作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0111] 在本发明的上下文中,计算机可读存储介质可以是有形的介质,其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的计算机程序。计算机可读存储介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、

或半导体系统、装置或设备，或者上述内容的任何合适组合。备选地，计算机可读存储介质可以是机器可读信号介质。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0112] 为了提供与用户的交互，可以在电子设备上实施此处描述的系统和技术，该电子设备具有：用于向用户显示信息的显示装置(例如，CRT(阴极射线管)或者LCD(液晶显示器)监视器)；以及键盘和指向装置(例如，鼠标或者轨迹球)，用户可以通过该键盘和该指向装置来将输入提供给电子设备。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈(例如，视觉反馈、听觉反馈、或者触觉反馈)；并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0113] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如，作为数据服务器)、或者包括中间件部件的计算系统(例如，应用服务器)、或者包括前端部件的计算系统(例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如，通信网络)来将系统的部件相互连接。通信网络的示例包括：局域网(LAN)、广域网(WAN)、区块链网络和互联网。

[0114] 计算系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器，又称为云计算服务器或云主机，是云计算服务体系中的一项主机产品，以解决了传统物理主机与VPS服务中，存在的管理难度大，业务扩展性弱的缺陷。

[0115] 应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本发明中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本发明的技术方案所期望的结果，本文在此不进行限制。

[0116] 上述具体实施方式，并不构成对本发明保护范围的限制。本领域技术人员应该明白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本发明的精神和原则之内所作的修改、等同替换和改进等，均应包含在本发明保护范围之内。

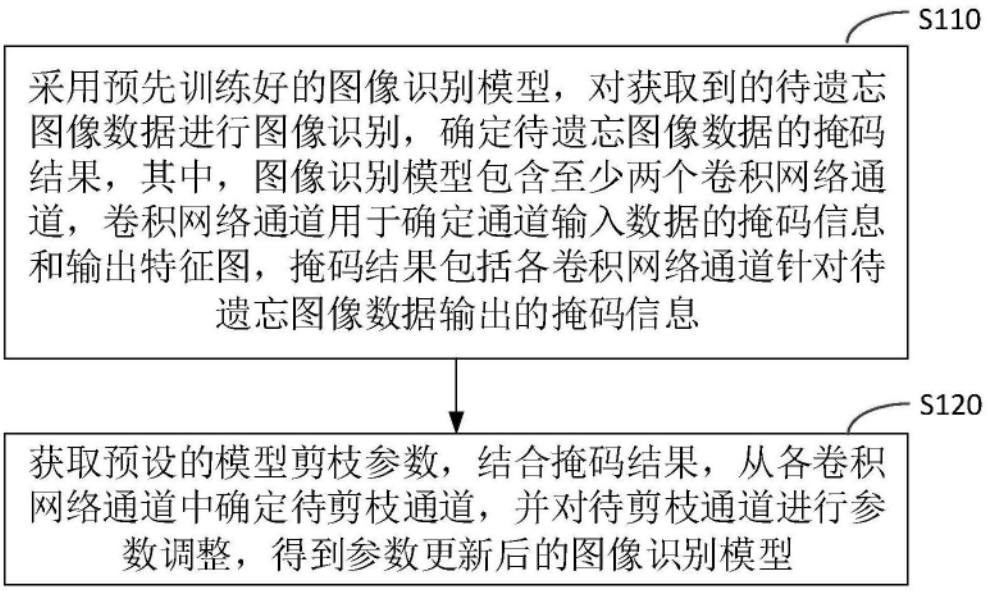

S110
采用预先训练好的图像识别模型，对获取到的待遗忘图像数据进行图像识别，确定待遗忘图像数据的掩码结果，其中，图像识别模型包含至少两个卷积网络通道，卷积网络通道用于确定通道输入数据的掩码信息和输出特征图，掩码结果包括各卷积网络通道针对待遗忘图像数据输出的掩码信息
S120
获取预设的模型剪枝参数，结合掩码结果，从各卷积网络通道中确定待剪枝通道，并对待剪枝通道进行参数调整，得到参数更新后的图像识别模型

图1a

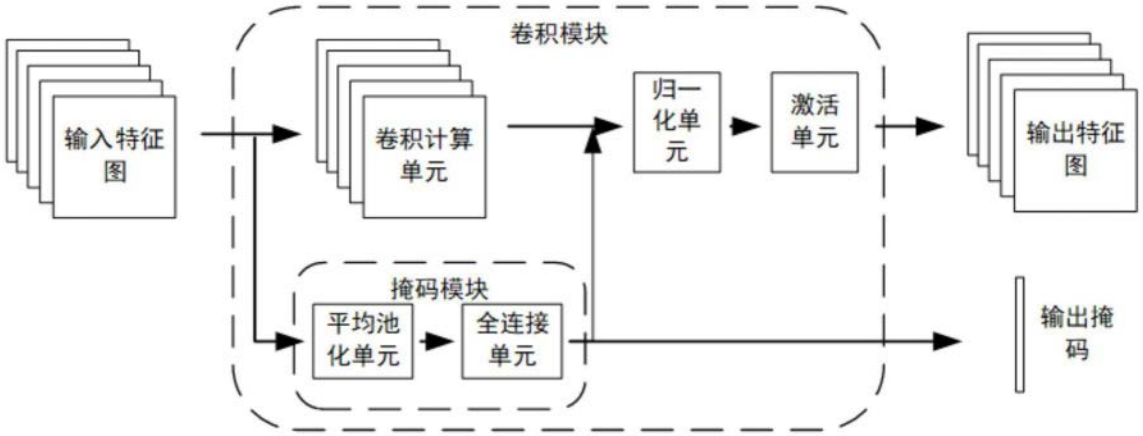

卷积模块
输入特征图
卷积计算单元
归一化单元
激活单元
输出特征图
掩码模块
平均池化单元
全连接单元
输出掩码

图1b

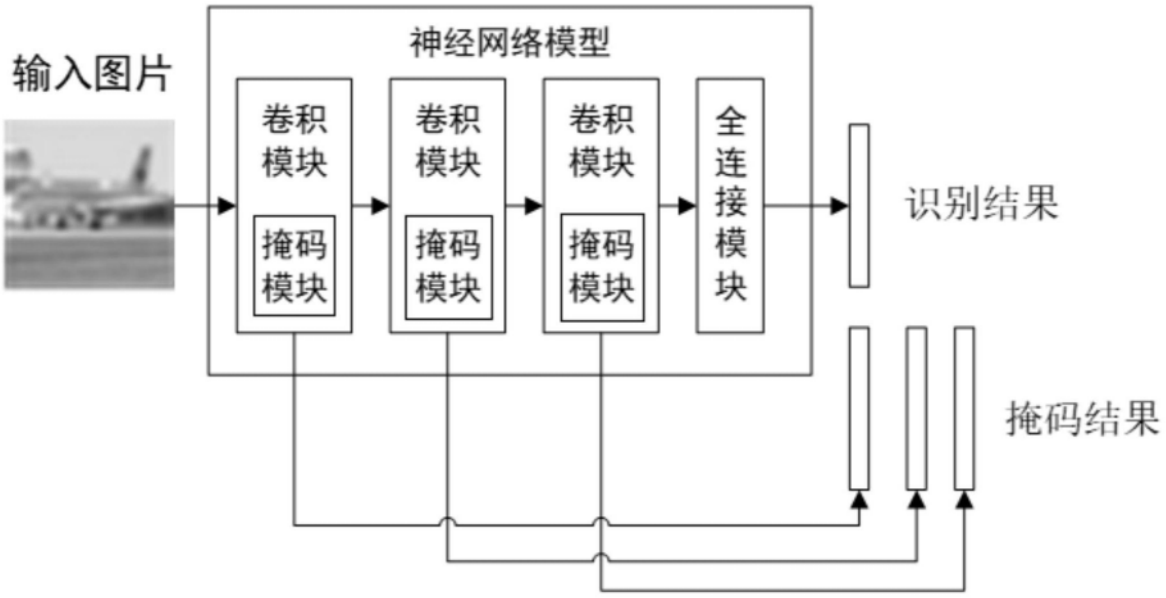

输入图片
神经网络模型
卷积模块
掩码模块
卷积模块
掩码模块
卷积模块
掩码模块
全连接模块
识别结果
掩码结果

图1c

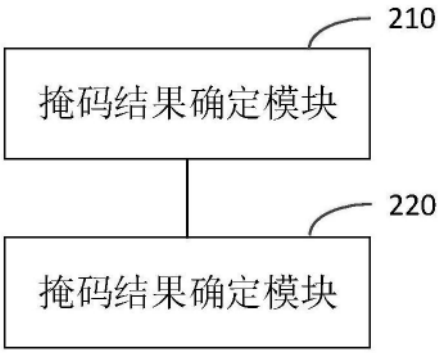

210
掩码结果确定模块
220
掩码结果确定模块

图2

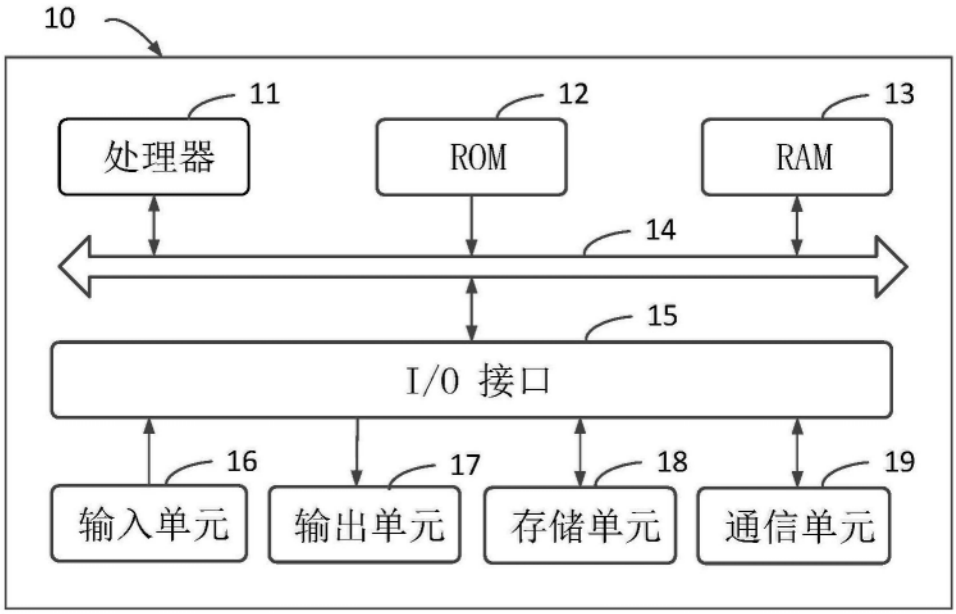


图3