(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112465141 A
(43)申请公布日 2021.03.09

(21)申请号 202011501677.5

(22)申请日 2020.12.18

(71)申请人 平安科技(深圳)有限公司
地址 518000 广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼

(72)发明人 成冠举 李葛 曾婵 高鹏

(74)专利代理机构 深圳市沃德知识产权代理事务所(普通合伙) 44347
代理人 高杰 于志光

(51)Int.Cl.
G06N 3/08(2006.01)
G06N 3/04(2006.01)

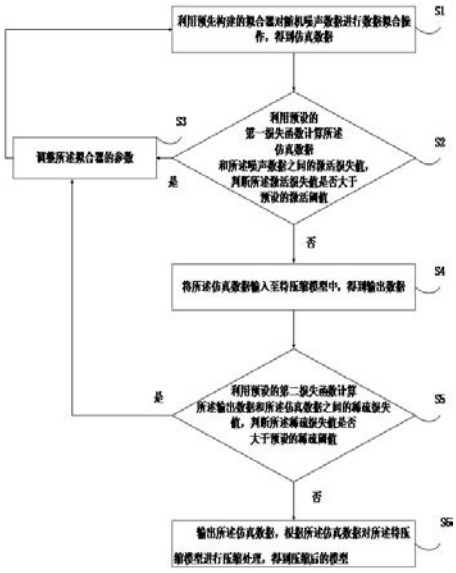
权利要求书2页 说明书9页 附图2页

(54)发明名称

模型压缩方法、装置、电子设备及介质

(57)摘要

本发明涉及数据处理技术，揭露一种模型压缩方法，包括：利用预先构建的拟合器对随机噪声数据进行数据拟合，得到仿真数据并计算仿真数据和噪声数据的激活损失值，在激活损失值大于预设的激活阈值，调整所述拟合器的参数，直到激活损失值小于或等于预设的激活阈值，将仿真数据输入至待压缩模型中得到输出数据；计算输出数据和仿真数据的稀疏损失值，在稀疏损失值大于预设的稀疏阈值，调整拟合器的内部参数，直到稀疏损失值小于或者等于预设的稀疏阈值，输出仿真数据并对待压缩模型进行压缩，得到压缩后的模型。本发明还揭露一种模型压缩装置、电子设备及存储介质。本发明不需要获取训练数据、网络结构和参数等，就能实现模型的压缩。



CN 112465141 A

1.一种模型压缩方法,其特征在于,所述方法包括:

步骤A:利用预先构建的拟合器对随机噪声数据进行数据拟合操作,得到仿真数据;

步骤B:利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值,在所述激活损失值大于预设的激活阈值时,调整所述拟合器的参数并返回上述的步骤A,直到所述激活损失值小于或等于预设的激活阈值时,将所述仿真数据输入至待压缩模型中,得到输出数据;

步骤C:利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值,在所述稀疏损失值大于预设的稀疏阈值时,调整所述拟合器的内部参数并返回上述的步骤A,直到所述稀疏损失值小于或者等于预设的稀疏阈值时,输出所述仿真数据;

步骤D:根据所述仿真数据对所述待压缩模型进行压缩处理,得到压缩后的模型。

2.如权利要求1所述的模型压缩方法,其特征在于,所述利用预先构建的拟合器对随机噪声数据进行数据拟合操作,得到仿真数据,包括:

利用所述拟合器中的长短期记忆网络对所述噪声数据进行预测,得到拟合数据集;

利用激活函数对所述拟合数据集进行压缩,得到压缩数据集;

对所述压缩数据集进行向量化处理,得到仿真数据。

3.如权利要求2所述的模型压缩方法,其特征在于,所述对所述压缩数据集进行向量化处理,得到仿真数据,包括:

利用Word2Vec算法将所述压缩数据集中的压缩数据映射为特征向量;

按照所述特征向量的序列对所述特征向量进行拼接,得到所述仿真数据。

4.如权利要求1所述的模型压缩方法,其特征在于,所述利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值,包括:

利用下述第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值:

$$\mathcal{L}_{act} = -\frac{1}{n}\sum_{m} \| f_T^m \|_1$$

其中,$\mathcal{L}_{act}$为所述激活损失值,n为所述噪声数据的样本数,f_T^m为所述仿真数据中的第m个数据,$||\ ||_1$是L1范数。

5.如权利要求1所述的模型压缩方法,其特征在于,所述利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值,包括:

利用下述第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值:

$$\mathcal{L}_{cls} = \frac{1}{x}\sum_{m} \mathcal{H}_{softmax}(y_T^m,\ t^m)$$

其中,$\mathcal{L}_{cls}$为所述稀疏损失值,x为所述仿真数据的样本数,y_T^m是所述输出数据中的第m个数据,t^m为预设的参数,$\mathcal{H}_{softmax}$为softmax损失函数。

6.如权利要求1至5中任意一项所述的模型压缩方法,其特征在于,所述根据所述仿真数据对所述待压缩模型进行压缩处理,得到压缩后的模型,包括:

将所述仿真数据输入至预设的标准压缩模型中进行向量运算,得到所述标准压缩模型输出的第一特征,将所述仿真数据输入至所述待压缩模型中进行向量运算,得到所述待压缩模型输出的第二特征;

根据所述第一特征和所述第二特征确定所述待压缩模型的损失函数;

根据所述损失函数对所述待压缩模型进行反向传播，得到压缩后的模型。

7.如权利要求6所述的模型压缩方法，其特征在于，所述根据所述第一特征和所述第二特征确定所述待压缩模型的损失函数，包括：

根据所述第一特征和所述第二特征进行求差计算，得到差值函数；

将所述差值函数进行范数转换处理并求其平方，得到损失函数。

8.一种模型压缩装置，其特征在于，所述装置包括：

数据拟合模块，用于利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据；

激活损失模块，用于利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值，在所述激活损失值大于预设的激活阈值时，调整所述拟合器的参数，直到所述激活损失值小于或等于预设的激活阈值时，将所述仿真数据输入至待压缩模型中，得到输出数据；

稀疏损失模块，用于利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值，在所述稀疏损失值大于预设的稀疏阈值时，调整所述拟合器的内部参数，直到所述稀疏损失值小于或者等于预设的稀疏阈值时，输出所述仿真数据；

模型压缩模块，用于根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型。

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执行如权利要求1至7中任一项所述的模型压缩方法。

10.一种计算机可读存储介质，存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至7中任一项所述的模型压缩方法。

模型压缩方法、装置、电子设备及介质

技术领域

[0001] 本发明涉及数据处理领域，尤其涉及一种模型压缩方法、装置、电子设备及计算机可读存储介质。

背景技术

[0002] 大数据时代深度学习模型运用的越来越频繁，为了将深度学习模型应用到移动设备、传感器等小型设备，有时必须将深度学习模型进行压缩裁剪才能部署到小型设备。

[0003] 目前主流的深度学习压缩方法都要基于原始训练数据集、网络结构、参数等进行模型的压缩，如知识蒸馏方法和基于元数据的方法，前者需要大量的原始训练数据，而后者需要模型的网络结构和参数，但由于法律、隐私等原因，训练数据、网络结构和参数通常很难获取到。

发明内容

[0004] 本发明提供一种模型压缩方法、装置、电子设备及计算机可读存储介质，其主要目的在于提供一种不需要获取训练数据、网络结构和参数而进行模型压缩的方案。

[0005] 为实现上述目的，本发明提供的一种模型压缩方法，包括：

利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据；

利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值，在所述激活损失值大于预设的激活阈值时，调整所述拟合器的参数并返回利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据，直到所述激活损失值小于或等于预设的激活阈值时，将所述仿真数据输入至待压缩模型中，得到输出数据；

利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值，在所述稀疏损失值大于预设的稀疏阈值时，调整所述拟合器的内部参数并返回利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据，直到所述稀疏损失值小于或者等于预设的稀疏阈值时，输出所述仿真数据；

根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型。

[0006] 可选地，所述利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据，包括：

利用所述拟合器中的长短期记忆网络对所述噪声数据进行预测，得到拟合数据集；

利用激活函数对所述拟合数据集进行压缩，得到压缩数据集；

对所述压缩数据集进行向量化处理，得到仿真数据。

[0007] 可选地，所述对所述压缩数据集进行向量化处理，得到仿真数据，包括：

利用Word2Vec算法将所述压缩数据集中的压缩数据映射为特征向量；

按照所述特征向量的序列对所述特征向量进行拼接，得到所述仿真数据。

[0008] 可选地，所述利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值，包括：

利用下述第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值：

$$\mathcal{L}_{act} = -\frac{1}{n}\sum_{m} \| f_T^m \|_1$$

其中，$\mathcal{L}_{act}$为所述激活损失值，n为所述噪声数据的样本数，f_T^m为所述仿真数据中的第m个数据，$\| \ \|_1$是L1范数。

[0009] 可选地，所述利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值，包括：

利用下述第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值：

$$\mathcal{L}_{cls} = \frac{1}{x}\sum_{m} \mathcal{H}_{softmax}(y_T^m,\ t^m)$$

其中，$\mathcal{L}_{cls}$为所述稀疏损失值，x为所述仿真数据的样本数，y_T^m是所述输出数据中的第m个数据，t^m为预设的参数，$\mathcal{H}_{softmax}$为softmax损失函数。

[0010] 可选地，所述根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型，包括：

将所述仿真数据输入至预设的标准压缩模型中进行向量运算，得到所述标准压缩模型输出的第一特征，将所述仿真数据输入至所述待压缩模型中进行向量运算，得到所述待压缩模型输出的第二特征；

根据所述第一特征和所述第二特征确定所述待压缩模型的损失函数；

根据所述损失函数对所述待压缩模型进行反向传播，得到压缩后的模型。

[0011] 可选地，所述根据所述第一特征和所述第二特征确定所述待压缩模型的损失函数，包括：

根据所述第一特征和所述第二特征进行求差计算，得到差值函数；

将所述差值函数进行范数转换处理并求其平方，得到损失函数。

[0012] 为了解决上述问题，本发明还提供一种模型压缩装置，所述装置包括：

数据拟合模块，用于利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据；

激活损失模块，用于利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值，在所述激活损失值大于预设的激活阈值时，调整所述拟合器的参数，直到所述激活损失值小于或等于预设的激活阈值时，将所述仿真数据输入至待压缩模型中，得到输出数据；

稀疏损失模块，用于利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值，在所述稀疏损失值大于预设的稀疏阈值时，调整所述拟合器的内部参数，直到所述稀疏损失值小于或者等于预设的稀疏阈值时，输出所述仿真数据；

模型压缩模块，用于根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型。

[0013] 为了解决上述问题，本发明还提供一种电子设备，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执行上述的模型压缩方法。

[0014] 为了解决上述问题,本发明还提供一种计算机可读存储介质,存储有计算机程序,所述计算机程序被处理器执行时实现上述的模型压缩方法。

[0015] 本发明实施例利用预先构建的拟合器对随机噪声数据进行数据拟合操作,得到仿真数据;利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值,在所述激活损失值大于预设的激活阈值时,调整所述拟合器的参数,直到所述激活损失值小于或等于预设的激活阈值时,将所述仿真数据输入至待压缩模型中,得到输出数据;利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值,在所述稀疏损失值大于预设的稀疏阈值时,调整所述拟合器的内部参数,直到所述稀疏损失值小于或者等于预设的稀疏阈值时,输出所述仿真数据;利用两个损失函数对拟合器仿真出的数据进行验证,得到最接近噪声数据的仿真数据,根据所述仿真数据对所述待压缩模型进行压缩处理,得到压缩后的模型。因此本发明提出的模型压缩方法、装置及计算机可读存储介质,可以解决需要获取很难获取到的训练数据、网络结构和参数进行模型压缩的问题。

附图说明

[0016] 图1为本发明一实施例提供的模型压缩方法的流程示意图;

图2为本发明一实施例提供的模型压缩装置的模块示意图;

图3为本发明一实施例提供的实现模型压缩方法的电子设备的内部结构示意图;

本发明目的的实现、功能特点及优点将结合实施例,参照附图做进一步说明。

具体实施方式

[0017] 应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。

[0018] 本申请实施例提供一种模型压缩方法。所述模型压缩方法的执行主体包括但不限于服务端、终端等能够被配置为执行本申请实施例提供的该方法的电子设备中的至少一种。换言之,所述模型压缩方法可以由安装在终端设备或服务端设备的软件或硬件来执行,所述软件可以是区块链平台。所述服务端包括但不限于:单台服务器、服务器集群、云端服务器或云端服务器集群等。

[0019] 参照图1所示,为本发明实施例提供的一种模型压缩方法的流程示意图。在本实施例中,所述模型压缩方法包括:

S1、利用预先构建的拟合器对随机噪声数据进行数据拟合操作,得到仿真数据。

[0020] 本发明实施例中,所述随机噪声数据是从高斯分布中采样得到的随机高斯噪音。所述拟合器是将噪声数据不断进行线性拟合处理,生成逼近于真实数据的仿真数据。

[0021] 具体地,所述利用预先构建的拟合器对随机噪声数据进行数据拟合操作,得到仿真数据,包括:

利用所述拟合器中的长短期记忆网络对所述噪声数据进行预测,得到拟合数据集;

利用激活函数对所述拟合数据集进行压缩,得到压缩数据集;

对所述压缩数据集进行向量化处理,得到仿真数据。

[0022] 其中,所述长短期记忆网络可以训练所述随机噪音从高斯分布到拟合分布的映射,同时为了防止过拟合的发生,所述长短期记忆网络的每一层神经网络会增加dropout机

制。所述激活函数可以是tanh函数，利用所述tanh函数将所述拟合数据集中的数据压缩到-1到1之间，以便后续进行向量化操作。

[0023] 进一步地，所述对所述压缩数据集进行向量化处理，得到仿真数据，包括：

利用Word2Vec算法将所述压缩数据集中的压缩数据映射为特征向量；

按照所述特征向量的序列对所述特征向量进行拼接，得到所述仿真数据。

[0024] 其中，所述Word2Vec算法可以将数据映射为统一维度的向量，所述Word2Vec算法适用于在对于一个序列的数据且序列局部数据间存在着很强的关联的情况，可以用来对数据进行更泛化的分析。

[0025] 详细地，利用预先构建的拟合器对随机噪声数据进行数据拟合操作，可以得到一个与所述随机噪声数据接近的仿真数据，用于代替所述随机噪声数据进行后续的模型压缩。

[0026] S2、利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值。

[0027] 本发明实施例中，所述第一损失函数：

$$\mathcal{L}_{act} = -\frac{1}{n}\sum_{m} \| f_T^m \|_1$$

其中，$\mathcal{L}_{act}$为所述激活损失值，n为所述噪声数据的样本数，f_T^m为所述仿真数据中的第m个数据，$\| \ \|_1$是L1范数。L1范数主要是为了获得稀疏性，加上负号是为了尽量不稀疏，让f_T^m尽可能多的被激活。

[0028] 在所述激活损失值大于预设的激活阈值时，本发明实施例调整所述拟合器的参数并返回上述的S1，重新利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据。

[0029] 优选地，所述拟合器的参数可以是拟合器的权重、梯度等。

[0030] 在所述激活损失值小于或等于预设的激活阈值时，执行S3、将所述仿真数据输入至待压缩模型中，得到输出数据。

[0031] 其中，所述第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值，将所述激活损失值和预设的激活阈值进行比较，进而调整所述拟合器的参数，直至所述仿真数据和所述噪声数据之间的激活损失值收敛，此时调整得到的拟合器符合标准，无需再调整其参数。

[0032] S4、利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值。

[0033] 本发明实施例中，所述第二损失函数可以是

$$\mathcal{L}_{cls} = \frac{1}{x}\sum_{m} \mathcal{H}_{softmax}(y_T^m,\ t^m)$$

其中，$\mathcal{L}_{cls}$为所述稀疏损失值，x为所述仿真数据的样本数，y_T^m是所述输出数据中的第m个数据，t^m为预设的参数，$\mathcal{H}_{softmax}$为softmax损失函数。

[0034] 在所述稀疏损失值大于预设的稀疏阈值时，本发明实施例调整所述拟合器的内部参数并返回上述的S1，利用预先构建的拟合器重新对随机噪声数据进行数据拟合操作，得到仿真数据。

[0035] 在所述稀疏损失值小于或者等于预设的稀疏阈值时，执行S5，输出所述仿真数据，并根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型。

[0036] 本发明实施例中，所述根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型，包括：

将所述仿真数据输入至预设的标准压缩模型中进行向量运算，得到所述标准压缩模型输出的第一特征，将所述仿真数据输入至所述待压缩模型中进行向量运算，得到所述待压缩模型输出的第二特征；

根据所述第一特征和所述第二特征确定所述待压缩模型的损失函数；

根据所述损失函数对所述待压缩模型进行反向传播，得到压缩后的模型。

[0037] 具体地，所述根据所述第一特征和所述第二特征确定所述待压缩模型的损失函数，包括：

根据所述第一特征和所述第二特征进行求差计算，得到差值函数；

将所述差值函数进行范数转换处理并求其平方，得到损失函数。

[0038] 如图2所示，是本发明模型压缩装置的模块示意图。

[0039] 本发明所述模型压缩装置100可以安装于电子设备中。根据实现的功能，所述模型压缩装置100可以包括数据拟合模块101、激活损失模块102、稀疏损失模块103、模型压缩模块104。本发明所述模块也可以称之为单元，是指一种能够被电子设备处理器所执行，并且能够完成固定功能的一系列计算机程序段，其存储在电子设备的存储器中。

[0040] 在本实施例中，关于各模块/单元的功能如下：

所述数据拟合模块101，用于利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据；

所述激活损失模块102，用于利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值，在所述激活损失值大于预设的激活阈值时，调整所述拟合器的参数，直到所述激活损失值小于或等于预设的激活阈值时，将所述仿真数据输入至待压缩模型中，得到输出数据；

所述稀疏损失模块103，用于利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值，在所述稀疏损失值大于预设的稀疏阈值时，调整所述拟合器的内部参数，直到所述稀疏损失值小于或者等于预设的稀疏阈值时，输出所述仿真数据；

所述模型压缩模块104，用于根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型。

[0041] 详细地，所述模型压缩装置100中的各模块由电子设备的处理器所执行时，可以实现一种模型压缩方法，所述模型压缩方法的具体实施步骤如下：

步骤一、所述数据拟合模块101利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据。

[0042] 本发明实施例中，所述随机噪声数据是从高斯分布中采样得到的随机高斯噪音。所述拟合器是将噪声数据不断进行线性拟合处理，生成逼近于真实数据的仿真数据。

[0043] 具体地，所述数据拟合模块101利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据，包括：

利用所述拟合器中的长短期记忆网络对所述噪声数据进行预测，得到拟合数据集；

利用激活函数对所述拟合数据集进行压缩，得到压缩数据集；

对所述压缩数据集进行向量化处理，得到仿真数据。

[0044] 其中，所述长短期记忆网络可以训练所述随机噪音从高斯分布到拟合分布的映射，同时为了防止过拟合的发生，所述长短期记忆网络的每一层神经网络会增加dropout机制。所述激活函数可以是tanh函数，利用所述tanh函数将所述拟合数据集中的数据压缩到-1到1之间，以便后续进行向量化操作。

[0045] 进一步地，所述对所述压缩数据集进行向量化处理，得到仿真数据，包括：

利用Word2Vec算法将所述压缩数据集中的压缩数据映射为特征向量；

按照所述特征向量的序列对所述特征向量进行拼接，得到所述仿真数据。

[0046] 步骤二、所述激活损失模块102利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值。

[0047] 本发明实施例中，所述第一损失函数：

$$\mathcal{L}_{act} = -\frac{1}{n}\sum_{m} \| f_T^m \|_1$$

其中，$\mathcal{L}_{act}$为所述激活损失值，n为所述噪声数据的样本数，f_T^m为所述仿真数据中的第m个数据，$||\ ||_1$是L1范数。L1范数主要是为了获得稀疏性，加上负号是为了尽量不稀疏，让f_T^m尽可能多的被激活。

[0048] 在所述激活损失值大于预设的激活阈值时，本发明实施例调整所述拟合器的参数并返回上述的步骤一，重新利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据。

[0049] 优选地，所述拟合器的参数可以是拟合器的权重、梯度等。

[0050] 在所述激活损失值小于或等于预设的激活阈值时，执行步骤三、将所述仿真数据输入至待压缩模型中，得到输出数据。

[0051] 步骤四、所述稀疏损失模块103利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值。

[0052] 本发明实施例中，所述第二损失函数可以是

$$\mathcal{L}_{cls} = \frac{1}{x}\sum_{m} \mathcal{H}_{softmax}(y_T^m,\ t^m)$$

其中，$\mathcal{L}_{cls}$为所述稀疏损失值，x为所述仿真数据的样本数，y_T^m是所述输出数据中的第m个数据，t^m为预设的参数，$\mathcal{H}_{softmax}$为softmax损失函数。

[0053] 在所述稀疏损失值大于预设的稀疏阈值时，本发明实施例调整所述拟合器的内部参数并返回上述的步骤一，利用预先构建的拟合器重新对随机噪声数据进行数据拟合操作，得到仿真数据。

[0054] 在所述稀疏损失值小于或者等于预设的稀疏阈值时，执行步骤五，输出所述仿真数据，并根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型。

[0055] 本发明实施例中，所述根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型，包括：

将所述仿真数据输入至预设的标准压缩模型中进行向量运算，得到所述标准压缩模型输出的第一特征，将所述仿真数据输入至所述待压缩模型中进行向量运算，得到所述待压

缩模型输出的第二特征；

根据所述第一特征和所述第二特征确定所述待压缩模型的损失函数；

根据所述损失函数对所述待压缩模型进行反向传播，得到压缩后的模型。

[0056] 具体地，所述根据所述第一特征和所述第二特征确定所述待压缩模型的损失函数，包括：

根据所述第一特征和所述第二特征进行求差计算，得到差值函数；

将所述差值函数进行范数转换处理并求其平方，得到损失函数。

[0057] 如图3所示，是本发明实现模型压缩方法的电子设备的结构示意图。

[0058] 所述电子设备1可以包括处理器10、存储器11和总线，还可以包括存储在所述存储器11中并可在所述处理器10上运行的计算机程序，如模型压缩程序12。

[0059] 其中，所述存储器11至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、移动硬盘、多媒体卡、卡型存储器(例如：SD或DX存储器等)、磁性存储器、磁盘、光盘等。所述存储器11在一些实施例中可以是电子设备1的内部存储单元，例如该电子设备1的移动硬盘。所述存储器11在另一些实施例中也可以是电子设备1的外部存储设备，例如电子设备1上配备的插接式移动硬盘、智能存储卡(Smart Media Card，SMC)、安全数字(Secure Digital，SD)卡、闪存卡(Flash Card)等。进一步地，所述存储器11还可以既包括电子设备1的内部存储单元也包括外部存储设备。所述存储器11不仅可以用于存储安装于电子设备1的应用软件及各类数据，例如模型压缩程序12的代码等，还可以用于暂时地存储已经输出或者将要输出的数据。

[0060] 所述处理器10在一些实施例中可以由集成电路组成，例如可以由单个封装的集成电路所组成，也可以是由多个相同功能或不同功能封装的集成电路所组成，包括一个或者多个中央处理器(Central Processing unit，CPU)、微处理器、数字处理芯片、图形处理器及各种控制芯片的组合等。所述处理器10是所述电子设备的控制核心(Control Unit)，利用各种接口和线路连接整个电子设备的各个部件，通过运行或执行存储在所述存储器11内的程序或者模块(例如执行模型压缩程序等)，以及调用存储在所述存储器11内的数据，以执行电子设备1的各种功能和处理数据。

[0061] 所述总线可以是外设部件互连标准(peripheral component interconnect，简称PCI)总线或扩展工业标准结构(extended industry standard architecture，简称EISA)总线等。该总线可以分为地址总线、数据总线、控制总线等。所述总线被设置为实现所述存储器11以及至少一个处理器10等之间的连接通信。

[0062] 图3仅示出了具有部件的电子设备，本领域技术人员可以理解的是，图3示出的结构并不构成对所述电子设备1的限定，可以包括比图示更少或者更多的部件，或者组合某些部件，或者不同的部件布置。

[0063] 例如，尽管未示出，所述电子设备1还可以包括给各个部件供电的电源(比如电池)，优选地，电源可以通过电源管理装置与所述至少一个处理器10逻辑相连，从而通过电源管理装置实现充电管理、放电管理、以及功耗管理等功能。电源还可以包括一个或一个以上的直流或交流电源、再充电装置、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。所述电子设备1还可以包括多种传感器、蓝牙模块、Wi-Fi模块等，在此不再赘述。

[0064] 进一步地，所述电子设备1还可以包括网络接口，可选地，所述网络接口可以包括有线接口和/或无线接口(如WI-FI接口、蓝牙接口等)，通常用于在该电子设备1与其他电子设备之间建立通信连接。

[0065] 可选地，该电子设备1还可以包括用户接口，用户接口可以是显示器(Display)、输入单元(比如键盘(Keyboard))，可选地，用户接口还可以是标准的有线接口、无线接口。可选地，在一些实施例中，显示器可以是LED显示器、液晶显示器、触控式液晶显示器以及OLED(Organic Light-Emitting Diode，有机发光二极管)触摸器等。其中，显示器也可以适当的称为显示屏或显示单元，用于显示在电子设备1中处理的信息以及用于显示可视化的用户界面。

[0066] 应该了解，所述实施例仅为说明之用，在专利申请范围上并不受此结构的限制。

[0067] 所述电子设备1中的所述存储器11存储的模型压缩程序12是多个指令的组合，在所述处理器10中运行时，可以实现：

利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据；

利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值，在所述激活损失值大于预设的激活阈值时，调整所述拟合器的参数并返回利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据，直到所述激活损失值小于或等于预设的激活阈值时，将所述仿真数据输入至待压缩模型中，得到输出数据；

利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值，在所述稀疏损失值大于预设的稀疏阈值时，调整所述拟合器的内部参数并返回利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据，直到所述稀疏损失值小于或者等于预设的稀疏阈值时，输出所述仿真数据；

根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型。

[0068] 进一步地，所述电子设备1集成的模块/单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读存储介质中。所述计算机可读存储介质可以是易失性的，也可以是非易失性的，例如，所述计算机可读介质可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)。

[0069] 本发明还提供一种计算机可读存储介质，所述可读存储介质，所述可读存储介质存储有计算机程序，所述计算机程序在被电子设备的处理器所执行时，可以实现：

利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据；

利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值，在所述激活损失值大于预设的激活阈值时，调整所述拟合器的参数并返回利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据，直到所述激活损失值小于或等于预设的激活阈值时，将所述仿真数据输入至待压缩模型中，得到输出数据；

利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值，在所述稀疏损失值大于预设的稀疏阈值时，调整所述拟合器的内部参数并返回利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据，直到所述稀疏损失值小于或者等于预设的稀疏阈值时，输出所述仿真数据；

根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型。

[0070] 进一步地，所述计算机可用存储介质可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序等；存储数据区可存储根据区块链节点的使用所创建的数据等。

[0071] 在本发明所提供的几个实施例中，应该理解到，所揭露的设备，装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式。

[0072] 所述作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0073] 另外，在本发明各个实施例中的各功能模块可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用硬件加软件功能模块的形式实现。

[0074] 对于本领域技术人员而言，显然本发明不限于上述示范性实施例的细节，而且在不背离本发明的精神或基本特征的情况下，能够以其他的具体形式实现本发明。

[0075] 因此，无论从哪一点来看，均应将实施例看作是示范性的，而且是非限制性的，本发明的范围由所附权利要求而不是上述说明限定，因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图表记视为限制所涉及的权利要求。

[0076] 此外，显然“包括”一词不排除其他单元或步骤，单数不排除复数。系统权利要求中陈述的多个单元或装置也可以由一个单元或装置通过软件或者硬件来实现。第二等词语用来表示名称，而并不表示任何特定的顺序。

[0077] 最后应说明的是，以上实施例仅用以说明本发明的技术方案而非限制，尽管参照较佳实施例对本发明进行了详细说明，本领域的普通技术人员应当理解，可以对本发明的技术方案进行修改或等同替换，而不脱离本发明技术方案的精神和范围。

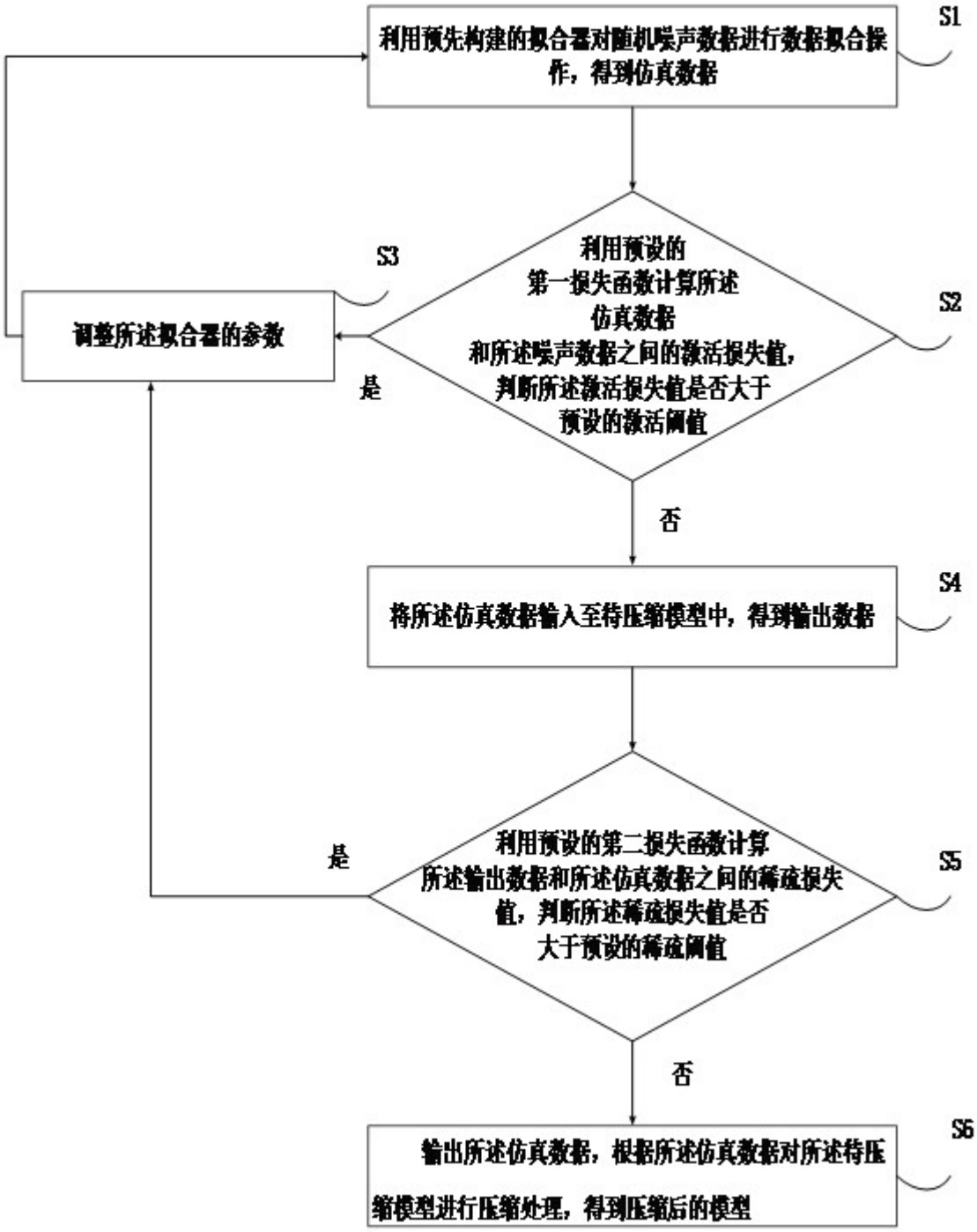

利用预先构建的拟合器对随机噪声数据进行数据拟合操作，得到仿真数据
S1
利用预设的第一损失函数计算所述仿真数据和所述噪声数据之间的激活损失值，判断所述激活损失值是否大于预设的激活阈值
S2
是
S3
调整所述拟合器的参数
否
将所述仿真数据输入至待压缩模型中，得到输出数据
S4
利用预设的第二损失函数计算所述输出数据和所述仿真数据之间的稀疏损失值，判断所述稀疏损失值是否大于预设的稀疏阈值
S5
是
否
输出所述仿真数据，根据所述仿真数据对所述待压缩模型进行压缩处理，得到压缩后的模型
S6

图1


模型压缩装置 100
数据拟合模块 101
激活损失模块 102
稀疏损失模块 103
模型压缩模块 104

图2

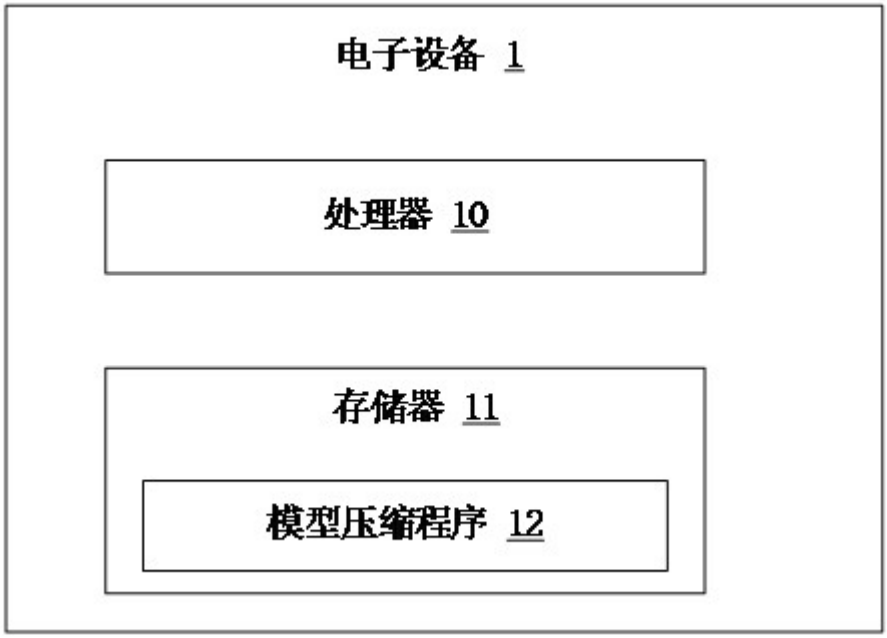

电子设备 1
处理器 10
存储器 11
模型压缩程序 12

图3