

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 12 月 24 日 (24.12.2020)



(10) 国际公布号
WO 2020/253049 A1

- (51) 国际专利分类号:
G06N 3/08 (2006.01)
- (21) 国际申请号: PCT/CN2019/117724
- (22) 国际申请日: 2019 年 11 月 13 日 (13.11.2019)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201910534815.0 2019年6月20日 (20.06.2019) CN
- (71) 申请人: 平安科技(深圳)有限公司(PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国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, Guangdong 518000 (CN)。
- (72) 发明人: 郑立颖(ZHENG, Liying); 中国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, Guangdong 518000 (CN)。徐亮(XU, Liang); 中国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, Guangdong 518000 (CN)。阮晓雯(RUAN, Xiaowen); 中国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, Guangdong 518000 (CN)。

(74) 代理人: 北京中强智尚知识产权代理有限公司 (BEIJING STRONG WISDOM INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市朝阳区北苑东路 19 号院 2 号楼 2608 室, Beijing 100012 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) Title: DATA PROCESSING METHOD AND DEVICE, STORAGE MEDIUM, AND ELECTRONIC DEVICE

(54) 发明名称: 数据处理方法和装置、及存储介质和电子装置

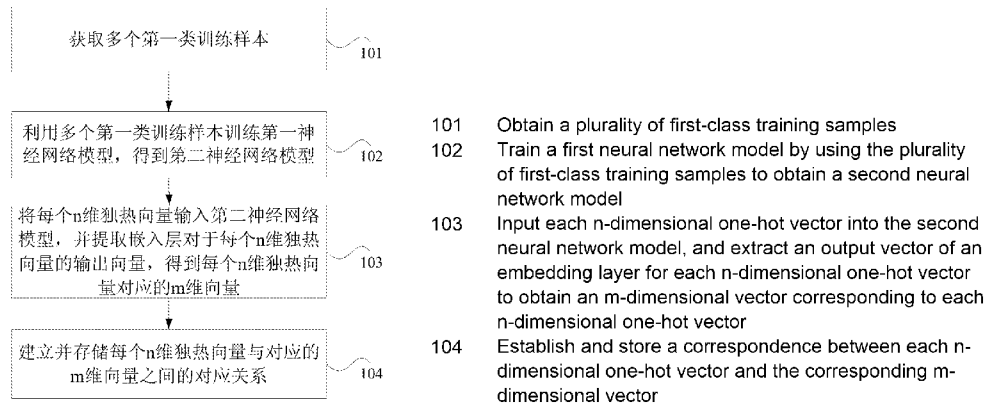


图 1

(57) Abstract: The present application provides a data processing method and device, a storage medium, and an electronic device. The method comprises: obtaining a plurality of first-class training samples, each first-class training sample comprising an n-dimensional one-hot vector and a corresponding first-class training label; training a first neural network model by using the plurality of first-class training samples to obtain a second neural network model, the first layer of the first neural network model being an embedding layer, the embedding layer being used for outputting an m-dimensional vector for the n-dimensional vector, and n being greater than m; inputting each n-dimensional one-hot vector into the second neural network model, and extracting an output vector of the embedding layer for each n-dimensional one-hot vector; and establishing and storing a correspondence between each n-dimensional one-hot vector and the corresponding m-dimensional vector. By means of the present application, the technical problems in the related art of excessive memory occupation and a low training speed in a training process of machine learning caused by identifying different objects with one-hot vectors are solved.

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

- 关于申请人有权申请并被授予专利 (细则
4.17(ii))
- 发明人资格 (细则4.17(iv))

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

(57) 摘要: 本申请提供了一种数据处理方法和装置、及存储介质和电子装置, 其中, 该方法包括: 获取多个第一类训练样本, 每个第一类训练样本包括一个 n 维独热向量与对应的第一类训练标签; 利用多个第一类训练样本训练第一神经网络模型, 得到第二神经网络模型, 第一神经网络模型的第一层为嵌入层, 嵌入层用于针对 n 维的向量输出 m 维的向量, $n > m$; 将每个 n 维独热向量输入第二神经网络模型, 并提取嵌入层对于每个 n 维独热向量的输出向量, 建立并存储每个 n 维独热向量与对应的 m 维向量之间的对应关系。通过本申请, 解决了相关技术中通过独热向量来标识不同对象导致机器学习的训练过程占用内存过多、训练速度较慢的技术问题。

数据处理方法和装置、及存储介质和电子装置

本申请要求与 2019 年 6 月 20 日提交中国专利局、申请号为 2019105348150、申请名称为“数据处理方法和装置、及存储介质和电子装置”的中国专利申请的优先权，其全部内容通过引用结合在申请中。

技术领域

本申请涉及数据处理领域，具体而言，涉及一种数据处理方法和装置、及存储介质和电子装置。

背景技术

在机器学习问题中，需要对每个对象分配唯一对应的 ID，通常是通过不同的向量来标识不同的对象。发明人发现目前对对象分配唯一对应的 ID 的方法主要有两种：label encoder（标签编码）以及 one-hot encoder（独热编码）。label encoder 的具体做法是为每一个对象分配一个数值，例如，对于红、黄、绿三种颜色，分别对应的分配数值 1、2、3 表示，但是，label encoder 会引入数值大小信息，例如，1+3 平均为 2，但是，红和绿的平均不是黄色，因此，label encoder 并不适用于一般的场景。one-hot encoder 的具体做法为，用 n 维独热向量来分别标识 n 个对象，向量中的每个元素为 0 或 1，对于一个具体的对象，其对应的独热向量仅有一个对应位置的元素为 1，其余的元素均为 0，例如，对于红、黄、绿三种颜色，分别用向量【0，0，1】、【0，1，0】、【1，0，0】表示。但是，one-hot encoder 会增加向量维度，如果有成百上千个对象，需要采用成百上千个维度的向量来标识，在运算时，会消耗大量的内存，使得运行速度变慢。

针对相关技术中存在的上述问题，目前尚未发现有效的解决方案。

发明内容

本申请实施例提供了一种数据处理方法和装置、及存储介质和电子装置，以至少解决现有技术中的通过独热向量来标识不同对象导致机器学习的训练过程占用内存过多、训练速度较慢的技术问题。

根据本申请的一个实施例，提供了一种数据处理方法，包括：获取多个第一类训练样本，其中，每个第一类训练样本包括一个 n 维独热向量与对应的第一类训练标签；利用多个第一类训练样本训练第一神经网络模型，得到第二神经网络模型，其中，第一神经网络模型的第一层为嵌入层，嵌入层用于针对 n 维的向量输出 m 维的向量， $n > m$ ；将每个 n 维独热向量输入第二神经网络模型，并提取嵌入层对于每个 n 维独热向量的输出向量，得到每个 n 维独热向量对应的 m 维向量；建立并存储每个 n 维独热向量与对应的 m 维向量之间的对应关系。

根据本申请的另一个实施例，提供了一种数据处理装置，该装置包括：第一获取模块，用于获取多个第一类训练样本，其中，每个第一类训练样本包括一个 n 维独热向量与对应的第一类训练标签；第一训练模块，用于利用多个第一类训练样本训练第一神经网络模型，得到第二神经网络

络模型，其中，第一神经网络模型的第一层为嵌入层，嵌入层用于针对 n 维的向量输出 m 维的向量， $n > m$ ；第一执行模块，用于将每个 n 维独热向量输入第二神经网络模型，并提取嵌入层对于每个 n 维独热向量的输出向量，得到每个 n 维独热向量对应的 m 维向量；第二执行模块，用于建立并存储每个 n 维独热向量与对应的 m 维向量之间的对应关系。

根据本申请的又一个实施例，还提供了一种电子装置，包括存储器和处理器，所述存储器中存储有计算机可读指令，所述处理器被设置为运行所述计算机可读指令以执行上述任一项方法实施例中的步骤。

通过本申请，通过在任意需要通过不同的 n 维独热向量来标识不同对象来训练一个神经网络模型的情况下，将神经网络模型的第一层嵌入层设计为一个可以将 n 维向量输出为 m 维向量的嵌入层，并利用一些标注好训练标签的 n 维独热向量对该神经网络模型进行训练，使得该神经网络模型的嵌入层可以更准确地用 m 维向量来表达 n 维独热向量的信息，从而可以利用训练后的神经网络模型来对任意 n 维独热向量降维至 m 维，减少在训练过程中占用的内存，提高神经网络模型的训练速度，缩短神经网络模型的训练时间，解决了相关技术中通过独热向量来标识不同对象导致机器学习的训练过程占用内存过多、训练速度较慢的技术问题。

附图说明

此处所说明的附图用来提供对本申请的进一步理解，构成本申请的一部分，本申请的示意性实施例及其说明用于解释本申请，并不构成对本申请的不当限定。在附图中：

图 1 是根据本申请实施例的数据处理方法的流程图；

图 2 是根据本申请实施例的数据处理装置的示意图；

图 3 是本申请实施例的一种电子装置的硬件结构框图。

具体实施方式

为了使本技术领域的人员更好地理解本申请方案，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分的实施例，而不是全部的实施例，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。基于本申请中的实施例，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，都应当属于本申请保护的范围。

需要说明的是，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，以便这里描述的本申请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。

实施例 1

本实施例提供了一种数据处理方法，可以运行中移动终端、手持终端或类似的运算设备之中。运行在不同的运算设备仅是方案在执行主体上的差异，本领域人员可预见在不同运算设备中运行能够产生相同的技术效果。

本实施例提供的数据处理方法，在任意需要通过不同的 n 维独热向量来标识不同对象来训练一个神经网络模型的情况下，将神经网络模型的第一层嵌入层设计为一个可以将 n 维向量输出为 m 维向量的嵌入层，并利用一些标注好训练标签的 n 维独热向量对该神经网络模型进行训练，使得该神经网络模型的嵌入层可以更准确地用 m 维向量来表达 n 维独热向量的信息，从而可以利用训练后的神经网络模型来对任意 n 维独热向量降维至 m 维，减少在训练过程中占用的内存，提高神经网络模型的训练速度，缩短神经网络模型的训练时间，解决了相关技术中通过独热向量来标识不同对象导致机器学习的训练过程占用内存过多、训练速度较慢的技术问题。

如图 1 所示，本实施例提供的数据处理方法包括如下步骤：

步骤 101，获取多个第一类训练样本，其中，每个第一类训练样本包括一个 n 维独热向量与对应的第一类训练标签；每个第一类训练样本是一个对象， n 维独热向量用于通过向量标识该对象以使机器能够进行识别和处理，对应的第一类训练标签也是一个向量，用于标识该对象的一种特征或属性，该特征或属性是神经网络模型的训练目标，也即，神经网络模型用于在训练之后针对每一个用 n 维独热向量标识的对象来输出一个对应的第一类训练标签。例如，第 i 个第一类训练样本 A_i 为 $[x_{i1}, x_{i2}, \dots, x_{in}, y_i]$ ，其中， $x_{i1} \sim x_{in}$ 为 n 维独热向量 N_i 的 n 个元素，在 $x_{i1} \sim x_{in}$ 中，仅有一个元素为 1，其余元素均为 0， y_i 为与 n 维独热向量 N_i 对应的第一类训练标签，用于表示 n 维独热向量 N_i 所表示的对象的类别。

可选的，在获取多个第一类训练样本时，首先，获取互不相同的 n 个 n 维独热向量，每个 n 维独热向量为一个对象的编码，例如，用向量 $[0, \dots, 0, 1, 0, \dots]$ 表示北京；其次，获取每个 n 维独热向量对应的第一类训练标签，第一类训练标签用于训练神经网络模型实现第一类分类目标，例如，第一类训练标签可以是城市的规模，对应于北京的第一类训练标签为 1，用于表示北京是一线城市；最后，组合每个 n 维独热向量与对应的第一类训练标签，得到 n 个第一类训练样本，利用第一类训练样本对模型进行训练之后，得到的模型用于对输入的城市进行分类，确定输入城市的规模。

其中，每个 n 维独热向量用于表示一个对象，可选的，可以先获取 n 个对象的标识，并将 n 个对象的标识通过互不相同的 n 个 n 维独热向量一一对应的表示，以得到每个 n 维独热向量与对应的对象标识之间的对应关系，并存储。例如，对于红色，用向量 $[0, 0, 1]$ 表示，将“红色”与向量 $[0, 0, 1]$ 的对应关系存储起来。利用多个第一类训练样本能够将第一神经网络模型训练为第二神经网络模型，以使得第二神经网络模型具有与第一类训练样本所对应的分类功能。例如，为了得到一个在输入城市名称之后，能够输出城市规模（分为一线、二线、三线等）的神经网络模型，获取 n 个第一类训练样本，以其中一个城市 C 为例，城市 C 对应的第一类训练标签为 2，用于表示城市 C 为二线城市，城市 C 的名称所对应的 n 维独热向量为 $[x_{i1}, x_{i2}, \dots, 0, \dots, x_{in}]$ ，将城市 C 的 n 维独热向量和第一类训练标签组合起来，得到对应于城市 C 的第一类训练样本

为 $[x_{i1}, x_{i2}, \dots, 0, \dots, x_{in}, 2]$ 。

步骤 102, 利用多个第一类训练样本训练第一神经网络模型, 得到第二神经网络模型, 其中, 第一神经网络模型的第一层为嵌入层, 嵌入层用于针对 n 维的向量输出 m 维的向量, 其中, $n > m$;

在得到多个第一类训练样本之后, 利用多个第一类训练样本训练第一神经网络模型, 训练目标为使第一神经网络模型对于输入的 n 维独热向量能够得到对应的第一类训练标签。需要说明的是, 本申请为了减少训练过程中占用的内存, 提高训练速度, 缩短训练时间, 将第一神经网络模型的输入层 (第一层) 设计为一个输入向量维度为 n 、输出向量维度为 m 的嵌入层, 其中 $m < n$ 。嵌入层的输入向量 (也即第一神经网络模型的输入向量) 的维度被预先配置为 n , 嵌入层的输出向量维度被配置为 m , 从而, 嵌入层用于接收输入的 n 维独热向量, 并输出对应的 m 维向量。相应的, 嵌入层之后的其它层可以设计为常规的神经网络模型的模式, 例如, 第一神经网络模型的第二层为全连接层、输出层为归一化层, 输出层用于输出分类目标, 其中, 与嵌入层相连的层的输入向量维度为 m , 第一神经网络模型的输出层的输出向量维度为第一类训练标签的维度。

步骤 103, 将每个 n 维独热向量输入第二神经网络模型, 并提取嵌入层对于每个 n 维独热向量的输出向量, 得到每个 n 维独热向量对应的 m 维向量;

在利用多个第一类训练样本训练第一神经网络模型之后, 得到的模型为第二神经网络模型, 也即, 训练后的神经网络模型, 从而, 利用第二神经网络模型对每个 n 维独热向量进行降维, 具体的做法为, 将每个 n 维独热向量输入第二神经网络模型, 第二神经网络模型的嵌入层输出的 m 维向量则为对应的降维向量, 其中, m 为预先配置参数。需要采用训练后的神经网络模型来对 n 维独热向量进行降维的原因是, 需要通过第一类训练本来对第一神经网络模型进行训练, 使得该神经网络模型的嵌入层输出的 m 维向量更准确的表达 n 维独热向量所标识的对象。举例而言, 该步骤在一种应用场景中, 如果存在 100 个待分配标记向量的城市, 如果采用独热向量, 需要用 100 维的独热向量表示, 在使用本申请实施例提供的方法之后, 能够将维度降为指定的维度 m , 例如, 用 10 维向量标识, 并且, 降维后的向量能够包含其它特征, 例如, 如果第一类训练标签用于表示城市的规模, 则降维后的向量携带有城市规模的特征信息。

步骤 104, 建立并存储每个 n 维独热向量与对应的 m 维向量之间的对应关系。在得到每个 n 维独热向量与对应的 m 维向量之间的对应关系之后, 存储起来, 以用于作为训练其它模型的编码。

可选的, 在建立并存储每个 n 维独热向量与对应的 m 维向量之间的对应关系之后, 如果需要训练另一种分类模型, 该方法还包括如下步骤:

步骤 201, 获取输入的每个 m 维向量对应的第二类训练标签。

步骤 202, 基于每个 m 维向量与对应的第二类训练标签, 生成多个第二类训练样本;

步骤 203, 利用多个第二类样本训练第三神经网络模型, 得到第四神经网络模型, 也即, 得到了另一种分类模型。其中, 第二类训练标签为需要训练的另一种分类模型的训练目标, 第二类训练标签通过一个指定维度的向量来表示, 第四神经网络模型是所需的另一种分类模型, 也即, 第四神经网络模型用于针对每一个输入的 m 维向量输出指定维度的第二类训练标签。

可选的，第一神经网络模型的参数 m 可以是由第三神经网络模型的输入向量决定的。在利用多个第一类训练样本训练第一神经网络模型之前，获取输入的用于配置第三神经网络模型的输入向量维度的参数，并根据参数配置第一神经网络模型的嵌入层的输出维度。

需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。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一个存储介质（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可以是手机，计算机，服务器，或者网络设备等）执行本申请各个实施例所述的方法。

实施例 2

在本实施例中还提供了一种数据处理装置，该装置用于实现上述实施例 1 及其优选实施方式，对于本实施例中未详述的术语或实现方式，可参见实施例 1 中的相关说明，已经进行过说明的不再赘述。如以下所使用的术语“模块”，是可以实现预定功能的软件和/或硬件的组合。尽管以下实施例所描述的装置较佳地以软件来实现，但是硬件，或者软件和硬件的组合的实现也是可以被构想的。

图 2 是根据本申请实施例的数据处理装置的示意图，如图 2 所示，该装置包括：第一获取模块 10，第一训练模块 20，第一执行模块 30，第二执行模块 40。

其中，第一获取模块用于获取多个第一类训练样本，其中，每个第一类训练样本包括一个 n 维独热向量与对应的第一类训练标签；第一训练模块用于利用多个第一类训练样本训练第一神经网络模型，得到第二神经网络模型，其中，第一神经网络模型的第一层为嵌入层，嵌入层用于针对 n 维的向量输出 m 维的向量， $n > m$ ；第一执行模块用于将每个 n 维独热向量输入第二神经网络模型，并提取嵌入层对于每个 n 维独热向量的输出向量，得到每个 n 维独热向量对应的 m 维向量；第二执行模块用于建立并存储每个 n 维独热向量与对应的 m 维向量之间的对应关系。

可选的，获取模块包括：第一获取单元，用于获取互不相同的 n 个 n 维独热向量；第二获取单元，用于获取每个 n 维独热向量对应的第一类训练标签；组合单元，用于组合每个 n 维独热向量与对应的第一类训练标签，得到 n 个第一类训练样本。

可选的，第一获取单元包括：第三获取单元，用于获取 n 个对象的标识；标识单元，用于将 n 个对象的标识通过互不相同的 n 个 n 维独热向量一一对应的表示。

可选的，该装置还包括：第二获取模块，用于在建立并存储每个 n 维独热向量与对应的 m 维向量之间的对应关系之后，获取输入的每个 m 维向量对应的第二类训练标签；生成模块，用于基于每个 m 维向量与对应的第二类训练标签，生成多个第二类训练样本；第二训练模块，用于利用

多个第二类样本训练第三神经网络模型，得到第四神经网络模型。

可选的，该装置还包括：第三获取模块，用于在利用多个第一类训练样本训练第一神经网络模型之前，获取输入的用于配置第三神经网络模型的输入向量维度的参数；第一配置模块，用于根据参数配置第一神经网络模型的嵌入层的输出维度。

可选的，该装置还包括：第四获取模块，用于在利用多个第一类训练样本训练第一神经网络模型之前，获取输入的用于配置第一神经网络模型的嵌入层的输出维度的参数；第二配置模块，用于根据参数配置第一神经网络模型的嵌入层的输出维度。

可选的，第一神经网络模型的第二层为全连接层、输出层为归一化层。

需要说明的是，上述各个模块是可以通过软件或硬件来实现的，对于后者，可以通过以下方式实现，但不限于此：上述模块均位于同一处理器中；或者，上述各个模块以任意组合的形式分别位于不同的处理器中。

显然，本领域的技术人员应该明白，上述的本申请各模块或各步骤可以用通用的计算装置来实现，它们可以集中在单个的计算装置上，或者分布在多个计算装置所组成的网络上，可选地，它们可以用计算装置可执行的程序代码来实现，从而，可以将它们存储在存储装置中由计算装置来执行，并且在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤，或者将它们分别制作成各个集成电路模块，或者将它们中的多个模块或步骤制作成单个集成电路模块来实现。这样，本申请不限制于任何特定的硬件和软件结合。

实施例 3

本申请的实施例还提供了一种非易失性可读存储介质，该存储介质中存储有计算机可读指令，其中，该计算机可读指令被设置为运行时执行上述任一项方法实施例中的步骤。可选地，在本实施例中，上述存储介质可以包括但不限于：U 盘、只读存储器（Read-Only Memory，简称为 ROM）、随机存取存储器（Random Access Memory，简称为 RAM）、移动硬盘、磁碟或者光盘等各种可以存储计算机可读指令的介质。

实施例 4

本申请的实施例还提供了一种电子装置，包括存储器和处理器，该存储器中存储有计算机可读指令，该处理器被设置为运行计算机可读指令以执行上述任一项方法实施例中的步骤。

可选地，上述电子装置还可以包括传输设备以及输入输出设备，其中，该传输设备和上述处理器连接，该输入输出设备和上述处理器连接。图 3 是本申请实施例的一种电子装置的硬件结构框图。如图 3 所示，电子装置可以包括一个或多个（图 3 中仅示出一个）处理器 302（处理器 302 可以包括但不限于微处理器 MCU 或可编程逻辑器件 FPGA 等的处理装置）和用于存储数据的存储器 304，可选地，上述电子装置还可以包括用于通信功能的传输设备 306 以及输入输出设备 308。本领域普通技术人员可以理解，图 3 所示的结构仅为示意，其并不对上述电子装置的结构造成限定。例如，电子装置还可包括比图 3 中所示更多或者更少的组件，或者具有与图 3 所示不同的配置。存储器 304 可用于存储计算机可读指令，例如，应用程序的软件程序以及模块，如本申请实

实施例中的图像的识别方法对应的计算机可读指令，处理器 302 通过运行存储在存储器 304 内的计算机可读指令，从而执行各种功能应用以及数据处理，即实现上述的方法。存储器 304 可包括高速随机存储器，还可包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器 304 可进一步包括相对于处理器 302 远程设置的存储器，这些远程存储器可以通过网络连接至电子装置。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。传输装置 306 用于经由一个网络接收或者发送数据。上述的网络具体实例可包括电子装置的通信供应商提供的无线网络。在一个实例中，传输装置 306 包括一个网络适配器（Network Interface Controller，简称为 NIC），其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个实例中，传输装置 306 可以为射频（Radio Frequency，简称为 RF）模块，其用于通过无线方式与互联网进行通讯。

以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权利要求

1.一种数据处理方法，所述方法包括：

获取多个第一类训练样本，其中，每个所述第一类训练样本包括一个 n 维独热向量与对应的第一类训练标签；

利用所述多个第一类训练样本训练第一神经网络模型，得到第二神经网络模型，其中，所述第一神经网络模型的第一层为嵌入层，所述嵌入层用于针对 n 维的向量输出 m 维的向量， $n > m$ ；

将每个所述 n 维独热向量输入所述第二神经网络模型，并提取所述嵌入层对于每个所述 n 维独热向量的输出向量，得到每个所述 n 维独热向量对应的 m 维向量；

建立并存储每个所述 n 维独热向量与对应的 m 维向量之间的对应关系。

2.根据权利要求1所述的方法，所述获取多个第一类训练样本，包括：获取互不相同的 n 个 n 维独热向量；获取每个所述 n 维独热向量对应的第一类训练标签；组合每个所述 n 维独热向量与对应的第一类训练标签，得到 n 个第一类训练样本。

3.根据权利要求2所述的方法，所述获取互不相同的 n 个 n 维独热向量，包括：获取 n 个对象的标识；将所述 n 个对象的标识通过所述互不相同的 n 个 n 维独热向量一一对应的表示。

4.根据权利要求1所述的方法，在建立并存储每个所述 n 维独热向量与对应的 m 维向量之间的对应关系之后，所述方法还包括：获取输入的每个所述 m 维向量对应的第二类训练标签；基于每个所述 m 维向量与对应的第二类训练标签，生成多个第二类训练样本；利用所述多个第二类样本训练第三神经网络模型，得到第四神经网络模型。

5.根据权利要求4所述的方法，在利用所述多个第一类训练样本训练第一神经网络模型之前，所述方法还包括：获取输入的用于配置所述第三神经网络模型的输入向量维度的参数；根据所述参数配置所述第一神经网络模型的嵌入层的输出维度。

6.根据权利要求1所述的方法，在利用所述多个第一类训练样本训练第一神经网络模型之前，所述方法还包括：获取输入的用于配置所述第一神经网络模型的嵌入层的输出维度的参数；根据所述参数配置所述第一神经网络模型的嵌入层的输出维度。

7.根据权利要求1所述的方法，所述第一神经网络模型的第二层为全连接层、输出层为归一化层。

8.一种信息处理装置，所述装置包括：

第一获取模块，用于获取多个第一类训练样本，其中，每个所述第一类训练样本包括一个 n 维独热向量与对应的第一类训练标签；

第一训练模块，用于利用所述多个第一类训练样本训练第一神经网络模型，得到第二神经网络模型，其中，所述第一神经网络模型的第一层为嵌入层，所述嵌入层用于针对 n 维的向量输出 m 维的向量， $n > m$ ；

第一执行模块，用于将每个所述 n 维独热向量输入所述第二神经网络模型，并提取所述嵌入层对于每个所述 n 维独热向量的输出向量，得到每个所述 n 维独热向量对应的 m 维向量；

第二执行模块，用于建立并存储每个所述 n 维独热向量与对应的 m 维向量之间的对应关系。

9.根据权利要求 8 所述的装置，所述获取模块包括：第一获取单元，用于获取互不相同的 n 个 n 维独热向量；第二获取单元，用于获取每个所述 n 维独热向量对应的第一类训练标签；组合单元，用于组合每个所述 n 维独热向量与对应的第一类训练标签，得到 n 个第一类训练样本。

10.根据权利要求 9 所述的装置，所述第一获取单元包括：第三获取单元，用于获取 n 个对象的标识；标识单元，用于将所述 n 个对象的标识通过所述互不相同的 n 个 n 维独热向量一一对应的表示。

11.根据权利要求 8 所述的装置，所述装置还包括：第二获取模块，用于在建立并存储每个所述 n 维独热向量与对应的 m 维向量之间的对应关系之后，获取输入的每个所述 m 维向量对应的第二类训练标签；生成模块，用于基于每个所述 m 维向量与对应的第二类训练标签，生成多个第二类训练样本；第二训练模块，用于利用所述多个第二类样本训练第三神经网络模型，得到第四神经网络模型。

12.根据权利要求 11 所述的装置，所述装置还包括：第三获取模块，用于在利用所述多个第一类训练样本训练第一神经网络模型之前，获取输入的用于配置所述第三神经网络模型的输入向量维度的参数；第一配置模块，用于根据所述参数配置所述第一神经网络模型的嵌入层的输出维度。

13.根据权利要求 8 所述的装置，所述装置还包括：第四获取模块，用于在利用所述多个第一类训练样本训练第一神经网络模型之前，获取输入的用于配置所述第一神经网络模型的嵌入层的输出维度的参数；第二配置模块，用于根据所述参数配置所述第一神经网络模型的嵌入层的输出维度。

14.根据权利要求 8 所述的装置，所述第一神经网络模型的第二层为全连接层、输出层为归一化层。

15.一种非易失性可读存储介质，所述存储介质中存储有计算机可读指令，其中，所述计算机可读指令被设置为运行时执行数据处理方法，包括：获取多个第一类训练样本，其中，每个所述第一类训练样本包括一个 n 维独热向量与对应的第一类训练标签；利用所述多个第一类训练样本训练第一神经网络模型，得到第二神经网络模型，其中，所述第一神经网络模型的第一层为嵌入层，所述嵌入层用于针对 n 维的向量输出 m 维的向量， $n > m$ ；将每个所述 n 维独热向量输入所述第二神经网络模型，并提取所述嵌入层对于每个所述 n 维独热向量的输出向量，得到每个所述 n 维独热向量对应的 m 维向量；建立并存储每个所述 n 维独热向量与对应的 m 维向量之间的对应关系。

16.根据权利要求 15 所述的存储介质，所述计算机可读指令被设置为运行时执行所述获取多个第一类训练样本，包括：获取互不相同的 n 个 n 维独热向量；获取每个所述 n 维独热向量对应的第一类训练标签；组合每个所述 n 维独热向量与对应的第一类训练标签，得到 n 个第一类训练样

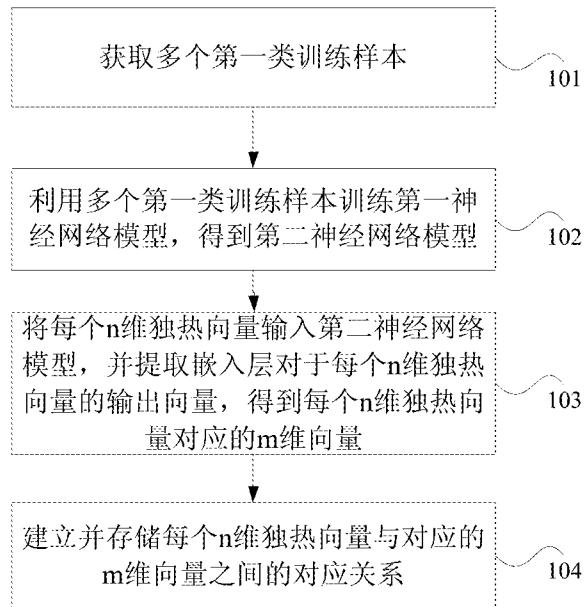
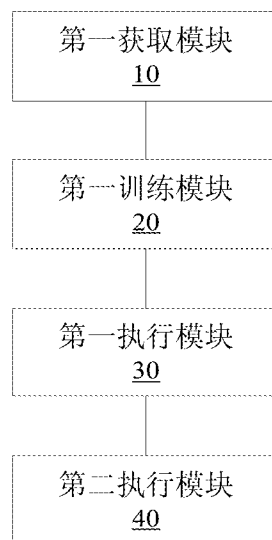
本。

17.根据权利要求 16 所述的存储介质,所述计算机可读指令被设置为运行时执行所述获取互不相同的 n 个 n 维独热向量,包括:获取 n 个对象的标识;将所述 n 个对象的标识通过所述互不相同的 n 个 n 维独热向量一一对应的表示。

18.一种电子装置,包括存储器和处理器,所述存储器中存储有计算机可读指令,所述处理器被设置为运行所述计算机可读指令以执行数据处理方法,包括:获取多个第一类训练样本,其中,每个所述第一类训练样本包括一个 n 维独热向量与对应的第一类训练标签;利用所述多个第一类训练样本训练第一神经网络模型,得到第二神经网络模型,其中,所述第一神经网络模型的第一层为嵌入层,所述嵌入层用于针对 n 维的向量输出 m 维的向量, $n > m$;将每个所述 n 维独热向量输入所述第二神经网络模型,并提取所述嵌入层对于每个所述 n 维独热向量的输出向量,得到每个所述 n 维独热向量对应的 m 维向量;建立并存储每个所述 n 维独热向量与对应的 m 维向量之间的对应关系。

19.根据权利要求 18 所述的电子装置,所述处理器被设置为运行所述计算机可读指令以执行所述获取多个第一类训练样本,包括:获取互不相同的 n 个 n 维独热向量;获取每个所述 n 维独热向量对应的第一类训练标签;组合每个所述 n 维独热向量与对应的第一类训练标签,得到 n 个第一类训练样本。

20.根据权利要求 19 所述的电子装置,所述处理器被设置为运行所述计算机可读指令以执行所述获取互不相同的 n 个 n 维独热向量,包括:获取 n 个对象的标识;将所述 n 个对象的标识通过所述互不相同的 n 个 n 维独热向量一一对应的表示。

**图 1****图 2**

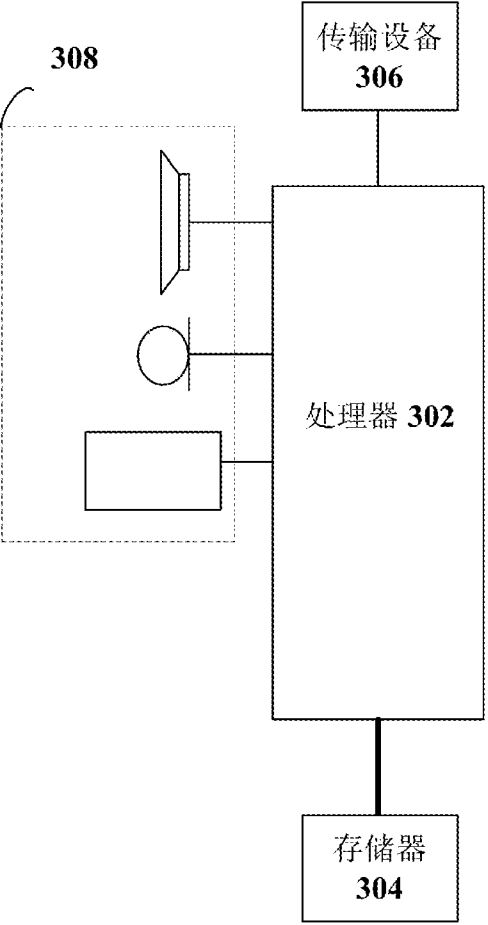


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2019/117724

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N; G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC, IEEE, GOOGLE: 神经网络, 训练, 嵌入, 独热, 降维, 低维, 维度, 样本, 用例, 实例, neural network?, embed+, one 1w hot, onehot, dimension?, reduc+, low+, sample

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110414678 A (PING'AN TECHNOLOGY (SHENZHEN) CO., LTD.) 05 November 2019 (2019-11-05) claims 1-10	1-20
X	CN 108665064 A (ALIBABA GROUP HOLDING LIMITED) 16 October 2018 (2018-10-16) description, paragraphs [0004]-[0028]	1-20
A	US 2018089152 A1 (DIGITAL GENIUS LIMITED) 29 March 2018 (2018-03-29) entire document	1-20
A	GUO, Cheng et al. "Entity Embeddings of Categorical Variables" "arXiv:1604.06737v1 [cs.LG]", 22 April 2016 (2016-04-22), entire document	1-20



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

28 February 2020

Date of mailing of the international search report

19 March 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2019/117724

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	110414678	A	05 November 2019	None	
CN	108665064	A	16 October 2018	None	
US	2018089152	A1	29 March 2018	None	

国际检索报告

国际申请号

PCT/CN2019/117724

A. 主题的分类

G06N 3/08 (2006.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

G06N; G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

CNPAT, CNKI, WPI, EPODOC, IEEE, GOOGLE: 神经网络, 训练, 嵌入, 独热, 降维, 低维, 维度, 样本, 用例, 实例, neural network?, embed+, one 1w hot, onehot, dimension?, reduc+, low+, sample

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 110414678 A (平安科技深圳有限公司) 2019年 11月 5日 (2019 - 11 - 05) 权利要求1-10	1-20
X	CN 108665064 A (阿里巴巴集团控股有限公司) 2018年 10月 16日 (2018 - 10 - 16) 说明书第[0004]-[0028]段	1-20
A	US 2018089152 A1 (DIGITAL GENIUS LIMITED) 2018年 3月 29日 (2018 - 03 - 29) 全文	1-20
A	GUO, Cheng 等. "Entity Embeddings of Categorical Variables" "arXiv:1604.06737v1 [cs.LG]", 2016年 4月 22日 (2016 - 04 - 22), 全文	1-20

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2020年 2月 28日

国际检索报告邮寄日期

2020年 3月 19日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

张枫

电话号码 86-(10)-53961628

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2019/117724

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	110414678	A	2019年 11月 5日	无	
CN	108665064	A	2018年 10月 16日	无	
US	2018089152	A1	2018年 3月 29日	无	