

一种基于模型进行预测的方法和装置

技术领域

[01]本公开涉及机器学习技术领域，特别涉及一种基于模型进行预测的方法和装置。

背景技术

5 [02]在人工智能中，特征工程是一个非常重要的环节。特征工程是将机器学习模型的输入数据转换为特征向量的过程，是机器学习中重要的起始步骤。而嵌入（embedding）是特征工程中的一种数据转换方法，机器学习的输入数据可能是一个高维张量，可以通过嵌入层将高维张量映射到低维的连续空间，达到特征提取的目的。通过将高维空间映射到嵌入这种低维空间，使得在大型输入（典型的是高维度稀疏张量）上的机器学习变得更容易。

10 [03]但是，嵌入层的参数量通常非常大，在某些业务场景中，所有嵌入层参数的内存总量可以达到千亿级别（比如 100GB~1TB 内存）。这种情况对实时在线模型的部署和预测构成很大的挑战：单台机器不能提供足够的共享内存去存储整个模型，因此需要购买定制的机器提供服务，增加了设备购置、运行和维护的成本；并且，单台机器的容量被较大的嵌入层占用，也使得模型运行不能充分发挥底层硬件资源的计算能力，造成资源浪费。

发明内容

[04]有鉴于此，本说明书一个或多个实施例提供一种基于模型进行预测的方法和装置，以降低模型运行成本，且更好的发挥资源能力。

20 [05]具体地，本说明书一个或多个实施例是通过如下技术方案实现的：

[06]第一方面，提供一种基于模型进行预测的方法，所述方法由机器学习模型系统执行，所述系统包括：机器学习模型、以及用于对机器学习模型的输入张量进行转换的嵌入模型；所述嵌入模型和机器学习模型分开部署，嵌入模型部署在嵌入运行环境，所述机器学习模型部署在模型运行环境中；所述方法包括：

25 [07]所述模型运行环境接收所述机器学习模型的输入张量；

[08]所述模型运行环境向嵌入运行环境发送查表请求，所述查表请求中携带所述输入张

量，以请求对所述输入张量进行低维转换；

[09]所述模型运行环境接收嵌入运行环境返回的查表结果，所述查表结果由所述嵌入运行环境根据所述输入张量进行嵌入查询和处理得到；

5 [10]所述模型运行环境将所述查表结果输入机器学习模型，并运行所述机器学习模型完成基于模型进行预测。

[11]第二方面，提供一种机器学习模型系统，所述系统包括：嵌入运行环境和模型运行环境，所述嵌入运行环境中部署嵌入模型，所述模型运行环境中部署机器学习模型；

10 [12]所述模型运行环境，用于接收所述机器学习模型的输入张量；并用于向嵌入运行环境发送查表请求，所述查表请求中携带所述输入张量，以请求对所述输入张量进行低维转换；还用于将所述查表结果输入机器学习模型，并运行所述机器学习模型完成基于模型进行预测；

[13]所述嵌入运行环境，用于根据查表请求，对所述输入张量进行嵌入查询和处理得到查表结果，并将所述查表结果反馈给所述模型运行环境。

15 [14]第三方面，提供一种基于模型进行预测的设备，所述设备包括存储器、处理器，以及存储在存储器上并可在处理器上运行的计算机指令，所述处理器执行指令时以实现嵌入运行环境和模型运行环境，所述嵌入运行环境用于运行嵌入模型，所述模型运行环境用于运行机器学习模型；所述处理器执行指令用于实现以下步骤：

[15]所述模型运行环境接收所述机器学习模型的输入张量；

20 [16]所述模型运行环境向嵌入运行环境发送查表请求，所述查表请求中携带所述输入张量，以请求对所述输入张量进行低维转换；

[17]所述模型运行环境接收嵌入运行环境返回的查表结果，所述查表结果由所述嵌入运行环境根据所述输入张量进行嵌入查询和处理得到；

[18]所述模型运行环境将所述查表结果输入机器学习模型，并运行所述机器学习模型完成基于模型进行预测。

25 [19]本说明书一个或多个实施例的基于模型进行预测的方法和装置，由于将模型与嵌入进行了分离部署，当单台机器的内存不足够时，可以使用分离的其他内存存储模型或嵌入，从而降低了模型运行和维护成本；并且，这种分离部署结构，可以使得网络虚拟执行单元的计算资源充分的用于模型运行预测，充分发挥底层硬件资源的计算能力。

附图说明

[20] 为了更清楚地说明本说明书一个或多个实施例或现有技术中的技术方案，下面将对实施例或现有技术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说明书一个或多个实施例中记载的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[21] 图 1 为本说明书一个或多个实施例提供的一种机器学习模型系统；

[22] 图 2 为本说明书一个或多个实施例提供的一种基于模型进行预测的方法；

[23] 图 3 为本说明书一个或多个实施例提供的一种机器学习模型系统。

具体实施方式

10 [24] 为了使本技术领域的人员更好地理解本说明书一个或多个实施例中的技术方案，下面将结合本说明书一个或多个实施例中的附图，对本说明书一个或多个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说明书一部分实施例，而不是全部的实施例。基于本说明书一个或多个实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都应当属于本公开保护的范围。

15 [25] 嵌入 (embedding) 是特征工程中常用的一种数据转换方法，可以用于将大型稀疏向量映射到一个保留语义关系的低维空间。嵌入的实现是通过训练嵌入模型，并通过该嵌入模型进行嵌入查询，可以将机器学习模型的输入张量 (如，稀疏张量) 作为索引查询嵌入模型，从而将原始的高维张量映射到低维空间。

[26] 以输入张量是 one-hot encoding 方法表示为例，One-hot encoding 方法将每个类别表示成一个很长的向量，这个向量的维度是类型的总个数，其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的类型。这种情况下，如果采用 one-hot encoding，生成的特征维度往往非常大。例如，在自然语言领域，one-hot encoding 可以把每个词表示为一个很长的向量。对于这样的高维向量，人工智能领域一种比较流行的方法就是将这些特征连接到嵌入 (embedding) 层，从而映射到低维空间，达到特征提取的目的。

25 [27] 由上述可以看到，嵌入与机器学习模型的关系是，嵌入可以用于对机器学习模型的输入张量进行高维空间向低维空间的转换操作，并将转换后的结果再输入进机器学习模型。通过嵌入的高维空间向低维空间的转换，使得在大型输入 (典型的是高维度稀疏张

量)上的机器学习变得更加容易。同时,为一个机器学习模型学习到的嵌入模型,也可以被其他机器学习模型复用;嵌入模型可以单独训练,也可以作为机器学习模型的一部分进行学习。例如,该嵌入模型可以是训练得到的嵌入表。

5 [28]本说明书至少一个实施例提供了一种机器学习模型系统,如图1所示,该系统可以包括:嵌入运行环境11和模型运行环境12。其中,嵌入运行环境11中部署了嵌入模型,具体可以是部署有嵌入表。模型运行环境12中可以部署机器学习模型,例如可以部署深度神经网络模型。

10 [29]其中,上述的嵌入运行环境11或者模型运行环境12可以是物理执行单元,或者也可以是虚拟执行单元。例如,物理执行单元可以是服务器等物理设备,虚拟执行单元可以是虚拟化容器等。可以将嵌入模型部署在嵌入运行环境11,嵌入运行环境11还可以执行嵌入查找和处理等;可以将机器学习模型部署在模型运行环境12,以由模型运行环境12运行该模型进行预测。

15 [30]图1所示的机器学习模型系统是一种将嵌入和机器学习模型进行分离部署的系统,例如,上述的嵌入运行环境11和模型运行环境12可以是虚拟化容器,具体的,可以运行有容器实例,例如,可以使用docker作为容器。还可以采用VMware虚拟机,Amazon Firecracker等作为虚拟执行单元。

20 [31]在分离部署的系统架构中,输入张量可以是输入到模型运行环境12,再由模型运行环境12向嵌入运行环境11发送查表请求。嵌入运行环境11可以进行嵌入查找,例如,以输入张量为索引查询嵌入运行环境11中部署的嵌入表,将高维张量映射到低维空间,并将查表结果返回给模型运行环境12。模型运行环境12可以根据查表结果继续运行模型。

25 [32]为了实现模型运行环境12和嵌入运行环境11之间的通信交互,两者之间可以进行网络互连,以实现上述的查表请求的发送,或者查表结果的反馈。例如,当嵌入运行环境11和模型运行环境12是虚拟化容器进行物理部署时,模型运行环境12以及其所需查表的嵌入运行环境11可以部署在同一个物理节点上,或者也可以部署在不同的物理节点上。当两者位于同一物理节点时,模型运行环境12发送查表请求时可以是本地发送。当两者位于不同的物理节点时,模型运行环境12发送查表请求可以是远程发送。

[33]图2是本说明书至少一个实施例提供的基于模型进行预测的方法的流程,该方法基于图1所示部署的机器学习模型系统,当机器学习模型系统按照图1所示部署后,可以

执行该基于模型进行预测的方法，将模型应用于预测。本例子以模型运行环境和嵌入运行环境是虚拟化容器为例进行说明，此时模型运行环境可以称为网络虚拟执行单元，嵌入运行环境可以称为嵌入虚拟执行单元。并且，在下面的描述中，嵌入模型以嵌入表为例进行说明。

5 [34]如图 2 所示，该方法可以包括：

[35]在步骤 200 中，网络虚拟执行单元接收所述机器学习模型的输入张量。

[36]例如，机器学习模型的输入张量可以被网络虚拟执行单元接收，该输入张量可以是一个高维张量，需要进行低维转换处理。

10 [37]在步骤 202 中，网络虚拟执行单元向嵌入虚拟执行单元发送查表请求，所述查表请求中携带所述输入张量，以请求对所述输入张量进行低维转换。

[38]例如，网络虚拟执行单元可以向嵌入虚拟执行单元发送查表请求，并将步骤 200 中接收到的输入张量发送给嵌入虚拟执行单元，请求对该输入张量进行低维转换处理。

[39]在步骤 204 中，嵌入虚拟执行单元根据所述输入张量进行嵌入表查询和处理，得到查表结果。

15 [40]例如，假设输入张量是 $v=(0,0,0,1,1)$ 。

[41]嵌入虚拟执行单元存储的嵌入表（Embedding 表）可以如下：

表 1: Emb 表

索引	向量
0	(1,2,3,4)
1	(5,6,7,8)
2	(9,10,11,12)
3	(13,14,15,16)
4	(17,18,19,20)

[42]根据输入张量 v 查找嵌入表 Emb，可以得到两个向量：

20 $V1=(13,14,15,16)$ ；

$V2=(17,18,19,20)$ ；

[43]然后，嵌入虚拟执行单元可以再对上述两个向量 $V1$ 和 $V2$ 进行求和或者平均等处理。

例如, $V1+V2$, 或者 $(V1+V2)/2$ 。

[44] 嵌入虚拟执行单元执行上述的求和或者平均的处理后, 将求和结果或者平均结果作为查找结果返回给网络虚拟执行单元。上述例子中, 是以求和或平均为例进行处理, 还可以执行其他处理; 并且, 由嵌入虚拟执行单元执行上述求和或平均等处理, 再返回给网络虚拟执行单元, 可以减少带宽需求。

[45] 在步骤 206 中, 嵌入虚拟执行单元将查表结果返回给网络虚拟执行单元。

[46] 在步骤 208 中, 网络虚拟执行单元将所述查表结果输入机器学习模型, 并运行所述机器学习模型完成基于模型进行预测。

[47] 例如, 网络虚拟执行单元在接收到查表结果后, 可以将该查表结果作为机器学习模型的输入, 并继续运行机器学习模型完成基于模型进行预测。

[48] 本例子的基于模型进行预测的方法, 由于将模型与嵌入进行了分离部署, 当单台机器的内存不足够时, 可以使用分离的其他内存存储模型或嵌入, 从而降低了模型运行和维护成本; 并且, 这种分离部署结构, 可以使得网络虚拟执行单元的计算资源充分的用于模型运行预测, 充分发挥底层硬件资源的计算能力。

[49] 此外, 机器学习模型系统中可以包括至少一个嵌入虚拟执行单元、以及至少一个所述网络虚拟执行单元。每个嵌入虚拟执行单元运行一个嵌入模型, 每个网络虚拟执行单元运行一个机器学习模型。

[50] 请参见图 3, 图 3 以虚拟执行单元是容器为例, 嵌入虚拟执行单元是嵌入容器, 网络虚拟执行单元是网络容器, 并且以网络容器中运行部署有深度神经网络为例, 那么网络容器可以称为深度网络容器。

[51] 图 3 所示的机器学习模型系统中包括多个深度网络容器和多个嵌入容器, 例如, 嵌入容器 31、嵌入容器 32 等, 还包括深度网络容器 33、深度网络容器 34 和深度网络容器 35 等。其中, 可以将同一个嵌入模型的多个拷贝部署在多个容器中, 以可扩展性的支持大量并发的查询访问, 比如, 嵌入容器 32 中部署的嵌入模型可以是嵌入容器 31 中部署的嵌入模型的拷贝。

[52] 嵌入容器和深度网络容器可以运行在同一台物理机器上, 也可以广义上分布在集群的任何角落。不同的网络虚拟执行单元可以配置不同的硬件资源, 该硬件资源可以适应所述网络虚拟执行单元上机器学习模型的运行需求。例如, 所述硬件资源可以包括如下至少一项: 中央处理器 CPU, 或者硬件加速器。其中的硬件加速器可以包括但不限于:

图形处理器 GPU (Graphics Processing Unit), 或者现场可编程门阵列 FPGA (Field - Programmable Gate Array), 或者为专门目的而设计的集成电路 ASIC (Application Specific Integrated Circuit) 芯片。

5 [53] 可以根据深度神经网络的计算复杂度, 业务延迟要求等各种约束分配计算资源, 尽量充分利用底层硬件的计算能力。比如, 在延迟要求苛刻的业务场景中, 可以部署 GPU, FPGA 或者 AI 芯片等硬件加速器到深度神经网络容器中, 以替代多个 CPU 核, 在规定时间内完成相对比较复杂的深度网络模型推理。而对于计算复杂度较低的深度神经网络模型, 单个 CPU 核或许就可以满足要求。在这种情况下, 对应的一个深度网络容器只需要配置一个 CPU 核就够用了。

10 [54] 在图 3 所示的系统中, 嵌入模型被部署在一个具有足够内存的嵌入容器中, 可以通过本地或者远程连接向其他需要嵌入表查询的容器提供服务。例如, 深度网络容器负责运行深度神经网络进行模型推理, 例如, 深度网络容器 33 向嵌入容器 31 发送查表请求是本地发送, 深度网络容器 33 和嵌入容器 31 位于同一节点。深度网络容器收到用户的输入张量后 (可能是密集的, 也可能是稀疏的), 向嵌入容器发出查表请求, 将收到的
15 查表结果输入到深度神经网络中完成模型推理预测。

[55] 此外, 本说明书至少一个实施例中的机器学习模型, 包括但不限于: 深度神经网络模型, 例如 Wide & Deep 模型和 DeepFM 模型。其中, 深度神经网络模型可以是在一个输入层和一个输出层之间添加至少一个隐藏层的机器学习神经网络。

20 [56] 上述实施例阐明的装置或模块, 具体可以由计算机芯片或实体实现, 或者由具有某种功能的产品来实现。一种典型的实现设备为计算机, 计算机的具体形式可以是个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件收发设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任意几种设备的组合。

25 [57] 为了描述的方便, 描述以上装置时以功能分为各种模块分别描述。当然, 在实施本说明书一个或多个实施例时可以把各模块的功能在同一个或多个软件和/或硬件中实现。

[58] 上述图中所示流程中的各个步骤, 其执行顺序不限制于流程图中的顺序。此外, 各个步骤的描述, 可以实现为软件、硬件或者其结合的形式, 例如, 本领域技术人员可以将其实现为软件代码的形式, 可以为能够实现所述步骤对应的逻辑功能的计算机可执行指令。当其以软件的方式实现时, 所述的可执行指令可以存储在存储器中, 并被设备中

的处理器执行。

[59]例如，对应于上述方法，本说明书一个或多个实施例同时提供一种基于模型进行预测设备，所述设备包括存储器、处理器，以及存储在存储器上并可在处理器上运行的计算机指令，所述处理器执行指令时以实现嵌入运行环境和模型运行环境，所述嵌入运行环境用于运行嵌入，所述模型运行环境用于运行机器学习模型；所述处理器执行指令用于实现以下步骤：

[60]所述模型运行环境接收所述机器学习模型的输入张量；

[61]所述模型运行环境向嵌入运行环境发送查表请求，所述查表请求中携带所述输入张量，以请求对所述输入张量进行低维转换；

10 [62]所述模型运行环境接收嵌入运行环境返回的查表结果，所述查表结果由所述嵌入运行环境根据所述输入张量进行嵌入查询和处理得到；

[63]所述模型运行环境将所述查表结果输入机器学习模型，并运行所述机器学习模型完成基于模型进行预测。

15 [64]还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

20 [65]本领域技术人员应明白，本说明书一个或多个实施例可提供为方法、系统或计算机程序产品。因此，本说明书一个或多个实施例可采用完全硬件实施例、完全软件实施例或结合软件和硬件方面的实施例的形式。而且，本说明书一个或多个实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、CD-ROM、光学存储器等）上实施的计算机程序产品的形式。

25 [66]本说明书一个或多个实施例可以在由计算机执行的计算机可执行指令的一般上下文中描述，例如程序模块。一般地，程序模块包括执行特定任务或实现特定抽象数据类型的例程、程序、对象、组件、数据结构等等。也可以在分布式计算环境中实践本说明书一个或多个实施例，在这些分布式计算环境中，由通过通信网络而被连接的远程处理设备来执行任务。在分布式计算环境中，程序模块可以位于包括存储设备在内的本地和远程计算机存储介质中。

[67]本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于数据处理设备实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

- 5 [68]上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。
- 10 [69]以上所述仅为本说明书一个或多个实施例的较佳实施例而已，并不用以限制本说明书一个或多个实施例，凡在本说明书一个或多个实施例的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本说明书一个或多个实施例保护的范围之内。

权利要求书

- 1、一种基于模型进行预测的方法，所述方法由机器学习模型系统执行，所述系统包括：机器学习模型、以及用于对机器学习模型的输入张量进行转换的嵌入模型；所述嵌入模型和机器学习模型分开部署，所述嵌入模型部署在嵌入运行环境，所述机器学习模型部署在模型运行环境中；
- 5 所述方法包括：
- 所述模型运行环境接收所述机器学习模型的输入张量；
- 所述模型运行环境向嵌入运行环境发送查表请求，所述查表请求中携带所述输入张量，以请求对所述输入张量进行低维转换；
- 10 所述模型运行环境接收嵌入运行环境返回的查表结果，所述查表结果由所述嵌入运行环境根据所述输入张量进行嵌入查询和处理得到；
- 所述模型运行环境将所述查表结果输入机器学习模型，并运行所述机器学习模型完成基于模型进行预测。
- 2、根据权利要求 1 所述的方法，
- 15 所述嵌入运行环境是物理执行单元或者虚拟执行单元。
- 3、根据权利要求 1 所述的方法，所述机器学习模型系统中包括：至少一个嵌入运行环境、以及至少一个所述模型运行环境；每个嵌入运行环境运行一个嵌入模型，每个模型运行环境运行一个机器学习模型；
- 所述模型运行环境向嵌入运行环境发送查表请求，包括：
- 20 所述模型运行环境向本地的嵌入运行环境发送查表请求，所述嵌入运行环境与模型运行环境位于同一个物理节点上；
- 或者，所述模型运行环境向远程的嵌入运行环境发送查表请求，所述嵌入运行环境与模型运行环境位于不同的物理节点上。
- 4、根据权利要求 3 所述的方法，不同的模型运行环境配置不同的硬件资源，所述
- 25 硬件资源适应所述模型运行环境上机器学习模型的运行需求。
- 5、根据权利要求 4 所述的方法，所述硬件资源包括如下至少一项：
- 中央处理器 CPU，或者硬件加速器；
- 所述硬件加速器包括如下任一项：图形处理器 GPU，或者现场可编程门阵列 FPGA，或者为专门目的而设计的集成电路 ASIC 芯片。
- 30 6、根据权利要求 1 所述的方法，所述机器学习模型，包括如下任一项：
- 深度神经网络模型，或者 Wide & Deep 模型，或者 DeepFM 模型。

7、一种机器学习模型系统，所述系统包括：嵌入运行环境和模型运行环境，所述嵌入运行环境中部署嵌入模型，所述模型运行环境中部署机器学习模型；

所述模型运行环境，用于接收所述机器学习模型的输入张量；并用于向嵌入运行环境发送查表请求，所述查表请求中携带所述输入张量，以请求对所述输入张量进行低维转换；还用于将所述查表结果输入机器学习模型，并运行所述机器学习模型完成基于模型进行预测；

所述嵌入运行环境，用于根据查表请求，对所述输入张量进行嵌入查询和处理得到查表结果，并将所述查表结果反馈给所述模型运行环境。

8、根据权利要求7所述的系统，
10 所述嵌入运行环境和模型运行环境，是物理执行单元或者虚拟执行单元。

9、根据权利要求7所述的系统，

所述机器学习模型系统中包括：至少一个嵌入运行环境、以及至少一个所述模型运行环境；每个嵌入运行环境运行至少一个嵌入模型，每个模型运行环境运行至少一个机器学习模型。

15 10、根据权利要求7所述的系统，

不同的模型运行环境配置不同的硬件资源，所述硬件资源适应所述模型运行环境上机器学习模型的运行需求。

11、一种基于模型进行预测的设备，所述设备包括存储器、处理器，以及存储在存储器上并可在处理器上运行的计算机指令，所述处理器执行指令时以实现嵌入运行环境和模型运行环境，所述嵌入运行环境用于运行嵌入模型，所述模型运行环境用于运行机器学习模型；所述处理器执行指令用于实现以下步骤：

所述模型运行环境接收所述机器学习模型的输入张量；

所述模型运行环境向嵌入运行环境发送查表请求，所述查表请求中携带所述输入张量，以请求对所述输入张量进行低维转换；

25 所述模型运行环境接收嵌入运行环境返回的查表结果，所述查表结果由所述嵌入运行环境根据所述输入张量进行嵌入查询和处理得到；

所述模型运行环境将所述查表结果输入机器学习模型，并运行所述机器学习模型完成基于模型进行预测。