

一种模型训练方法及装置

本申请要求于 2021 年 3 月 31 日提交中国专利局、申请号为 202110350815.2、发明名称为“一种模型训练方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5

技术领域

本申请涉及人工智能领域，尤其涉及一种模型训练方法及装置。

背景技术

10 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。

15 基于海量用户和推荐对象的交互行为信息,通过梯度下降等方法,训练推荐模型参数,当模型参数收敛后,可以对用户进行个性化推荐。以个性化推荐系统中的点击率为例,输入特征包括用户特征,商品特征,用户历史行为特征和上下文特征,其中特征又分为连续特征和离散特征。对于离散特征,使用 one-hot 进行稀疏编码,然后通过编码 id 从 embeddingtable 中取出 embedding 向量表达,将原始离散特征 id 进行向量化,然后输入推荐模型中训练学习,进行预测。

20 训练模型规模的大小和数据量对模型在线服务的效果有直接影响。推荐系统中,为了使模型可以更为精准的预测用户兴趣,往往需要使用海量的特征和数据。海量特征带来了对应的海量参数,使模型规模达到数百 GB 甚至 TB,而海量数据使得模型训练所需要算力增加。而且,为了更快速的捕捉用户兴趣,从而为用户提供更好的用户体验和推荐效果,25 深度模型需要频繁和快速更新,这些需求带来棘手的问题:需要基于海量特征和海量数据快速的完成推荐模型训练。

研究人员提出 embedding 层将高维稀疏的 one-hot 向量压缩为低维稠密的向量,也称为 embedding。推荐模型中主要的参数也位于 embedding 层(数百 GB 甚至 TB)。此外,为提升模型预测精度,推荐模型一般会基于乘法网络和 MLP 网络构建特征组合挖掘模块,该模块一般比较复杂,需要比较大的计算量,可用于数据特征之间存在的特征交互和非线性映射。与视觉、自然语言处理(natural language processing, NLP)模型不同,推荐模型有其独有特点。以点击率预测(click through rate, CTR)CTR 模型为例(如图 4),推荐模型包括嵌入层 embeddingLayer 和多层感知机(multi-layer perception, MLP)两个部分。推荐任务的数据包含大量高维稀疏的 id 类特征,一般使用 one-hot 编码,这部分特征直接用于模型训练会导致深度部分参数学习困难。为解决该问题,研究人员提出 embedding 层将高维稀疏的 one-hot 向量压缩为低维稠密的向量,也称为嵌入特征 embedding。

35

在现有的实现中,在进行模型训练时,图形处理器(graphics processing unit, GPU)

-2-

的计算核在完成一批batch的模型训练后，需要将更新后的embedding回传至CPU（例如包括更新后的embedding1），CPU维护有一个embedding表table，其中embeddingtable可以包括部分或全部模型训练所需的embedding，如果接下来一批batch的模型训练同样需要（embedding1），为了保证关注参数的一致性，则需要CPU将更新后的embedding1传递至GPU，进而GPU的计算单元的计算核可以从缓存中获取到更新后的embedding1，并基于更新后的embedding1进行下一批batch的模型训练，因而，GPU需要等待CPU完成将更新后的embedding1传递至GPU的过程，这就会造成时间上的延迟，大大增加了训练的时长。

发明内容

- 10 第一方面，本申请提供了一种模型训练方法，其特征在于，应用于第一处理器，所述第一处理器与第二处理器通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器；所述第一处理器和所述第二处理器为不同类型的处理器，其中，所述第一处理器可以为图形处理器GPU、嵌入式神经网络处理器（neural network processing unit, NPU）或张量处理器（tensor processing unit, TPU），所述第一存储器可以为缓存，
- 15 所述第二处理器可以为中央处理器（central processing unit, CPU）。

所述方法包括：

所述计算核通过所述第二存储器获取第一目标嵌入特征embedding；

- 所述计算核根据所述第一目标embedding，对待训练模型进行训练，以得到更新后的待训练模型以及第二目标embedding，所述第二目标embedding为更新后的所述第一目标embedding；
- 20

本申请实施例中，计算核在获取到当前批batch模型训练所需的embedding（包括第一目标embedding）之后，可以根据当前批batch模型训练所需的embedding（包括第一目标embedding），对所述待训练模型进行训练，以得到更新后的待训练模型以及更新后的embedding（包括第二目标embedding）。

- 25 所述计算核将所述第二目标embedding写入所述第一存储器。

- 可见本申请实施例中，第一处理器在完成第一目标embedding的更新后，将第二目标embedding存储至第一处理器的第一存储器中，在保证第一处理器可以获取到最新更新的目标embedding（第二目标embedding）的前提下，使得第一处理器不需要等待第二处理器将第二目标embedding传递至GPU的过程完成，就可以直接获取到第二目标embedding，并基于第二目标embedding进行模型训练。增加了训练数据的吞吐量，进而提高了模型训练的速度。
- 30

在一种可能的实现中，所述计算核根据所述第一目标embedding，对所述待训练模型进行训练，包括：

- 35 所述计算核根据所述第一目标embedding，对所述待训练模型进行第M批batch的模型训练；所述M为大于1的正整数；

在所述计算核将所述第二目标embedding写入所述第一存储器之后，所述方法还包括：

-3-

所述计算核从所述第一存储器中获取所述第二目标embedding;

所述计算核根据所述第二目标embedding, 对所述更新后的待训练模型进行第M+1批batch的模型训练。

5 若所述第一目标embedding还为接下来一批batch模型训练所需的embedding, 则计算核可以直接从第一存储器中获取更新后的第一目标embedding (也就是第二目标embedding)。若第一存储器中未存储有第M+1批batch的模型训练所需的一部分embedding, 则第二处理器可以将这部分embedding传递至第一存储器。

10 在一种可能的实现中, 在所述计算核根据所述第二目标embedding, 对所述更新后的待训练模型进行第M+1批batch的模型训练之后, 得到第三目标embedding, 所述第三目标embedding为更新后的所述第二目标embedding; 所述计算核可以将所述第三目标embedding写入所述第一存储器。

在一种可能的实现中, 所述计算核通过所述第二存储器获取第一目标embedding, 包括:

15 当所述第一目标embedding为所述第M批batch的模型训练所需的embedding, 且所述第一存储器未存储有所述第一目标embedding, 所述第一存储器从所述第二存储器中获取所述第一目标embedding; 其中, 当所述第一目标embedding为所述第M批batch的模型训练所需的embedding, 且所述第一存储器未存储有所述第一目标embedding时, 第二处理器可以将第一目标embedding从第二存储器中传递至第一存储器, 相当于, 所述第一存储器接收到来自所述第二存储器中的所述第一目标embedding。

所述计算核从所述第一存储器中获取第一目标embedding。

20 例如, 第M批batch的模型训练所需的embedding包括第一目标embedding, 若第二处理器确定第一存储器中未存储有第一目标embedding, 则可以将所述第一目标embedding从第二处理器中的第二存储器写入所述第一存储器。在写入的时间点选择上, 可以将所述计算核对所待训练模型进行第M-1批batch的模型训练的过程中, 所述第二处理器将所述目标embedding从第二存储器写入所述第一存储器。

25 本实施例中, 在前一批batch模型训练的过程中, 第二处理器会将第一存储器中没有存储且下一批batch模型训练过程中所需的embedding传递至第一存储器, 使得计算核可以从第一存储器中获取到下一批batch模型训练过程中所需的全部embedding。

在一种可能的实现中, 所述方法还包括:

30 在所述计算核从所述第一存储器中获取第一目标embedding之前, 所述计算核获取来自所述第二处理器的第一向量以及目标向量, 其中, 所述第一向量用于指示所述第M批batch的模型训练所需的embedding, 所述第一向量中的每个元素用于指示一个embedding, 且不同元素指示不同的embedding, 所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系, 所述每个元素指示的embedding
35 在所述排列关系中出现至少一次; 其中, 排列关系可以为第一向量中的每个元素指示的embedding在第M批batch的模型训练时的使用次数, 以及第一向量中的每个元素指示的embedding在第M批batch的模型训练时之间的先后顺序。

-4-

所述计算核从所述第一存储器中获取第一目标embedding，包括：

所述计算核根据所述第一向量从所述第一存储器中获取所述第M批batch的模型训练时所需的embedding；

5 所述计算核根据所述目标向量以及所述第M批batch的模型训练时所需的embedding，确定在进行所述第M批batch的模型训练时所需的每个embedding的排列关系。

在一种实现中，同一批batch模型训练所需的embedding可以包括相同的embedding，例如，同一批batch模型训练所需的embedding可以为：
D=[embedding1,embedding2,embedding3,embedding3,embedding2,embedding4]，则
10 embedding2和embedding3都是重复出现的，为了降低后续第一处理器的计算核读取embedding的参数量，第二处理器可以对同一批batch模型训练所需的embedding进行去重操作，具体的，所述第二处理器可以获取第一向量，所述第一向量包括的元素为所述第M批batch的模型训练所需的embedding；且第一向量可以指示去重后的模型训练所需的embedding的信息。

示例性的，去重前的第一向量可以为：D=[embedding1,embedding2,embedding3,embedding3,embedding2,embedding4]，对第一向量进行去重操作得到去重后的第一向量U=[embedding1, embedding2, embedding3, embedding4]，并计算得到目标向量Mask=[0, 1, 2, 2, 1, 3]，
15 其中，目标向量的第一个元素为0，表示第一向量中的第一个元素为第二向量中的第一个元素，目标向量的第二个元素为1，表示第一向量中的第二个元素为第二向量中的第二个元素，目标向量的第三个元素为2，表示第一向量中的第三个元素为第二向量中的第三个元素，目标向量的第四个元素为2，表示第一向量中的第四个元素为第二向量中的第三个元素，目标向量的第五个元素为1，表示第一向量中的第五个元素为第二向量中的第二个元素，目标向量的第六个元素为3，表示第一向量中的第六个元素为第二向量中的第四个元素，以此，可以基于第一向量和目标向量还原出去重前的第一向量，其中，去重前的第一向量用于确定
20 在进行所述第M批batch的模型训练时所需的每个embedding的排列关系。

25 在一种可能的实现中，在所述计算核根据所述第一目标embedding，对所述待训练模型进行训练之前，所述方法还包括：

所述计算核获取来自第三处理器中第三存储器的第三目标embedding，所述第三目标embedding和所述第一目标embedding为所述第M批batch模型训练所需的embedding，所述第一处理器和所述第三处理器为相同类型的处理器；

30 所述计算核根据所述第一目标embedding，对所述待训练模型进行训练，包括：

所述计算核通过所述待训练模型，对所述第一目标embedding和所述第三目标embedding进行处理，以得到目标输出，并基于所述目标输出确定第一梯度；

所述计算核基于所述第一梯度更新所述待训练模型以及所述第一目标embedding。

在包括多个计算节点的场景中，每个计算节点可以负责当前batch模型训练所需的
35 embedding的一部分，且负责一部分embedding的更新，在一种可能的实现中，计算核需要从其他服务器节点的第三处理器的第三存储器中获取第三目标embedding，并基于第一目标embedding和第三目标embedding来进行模型训练过程中的前馈过程，然而在更新时，第一

处理器中的计算核仅需要基于第一梯度来更新第一目标embedding。

在一种可能的实现中，所述基于所述目标输出确定所述第一梯度，包括：基于所述目标输出确定所述第一梯度和第二梯度；

5 所述基于所述目标输出确定所述第一梯度和所述第二梯度之后，所述方法还包括：

将所述第二梯度传递至第三处理器的计算核，以便所述第三处理器的计算核根据所述第二梯度更新所述第三目标embedding。

10 具体的，可以采用第一处理器和第三处理器之间的数据通信（例如allreduce通信）同步embedding的梯度，以使得第三处理器可以得到当前batch模型训练后第一处理器计算得到的梯度，之后第三处理器可以对自己负责的embedding（第三目标embedding）进行更新。

在一种可能的实现中，所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU，所述第一存储器为缓存，所述第二处理器为中央处理器CPU。

15 第二方面，本申请实施例提供了一种模型训练方法，应用于第二处理器，所述第二处理器与第一处理通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器，所述第一处理器和所述第二处理器为不同类型的处理器；所述方法包括：

所述第二处理器获取所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息，所述M为大于1的正整数；

20 其中，第M批batch模型训练所需的embedding信息可以指示第M批batch模型训练所需的embedding，例如该信息可以但不限于为指示embedding的标识信息或者指针信息，这里并不限定。

所述第二处理器根据所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息，将第一目标embedding写入所述第一存储器，
25 所述第一目标embedding为所述第M批batch的模型训练所需的且所述第一存储器未存储的embedding。

在一种可能的实现中，所述将第一目标embedding写入所述第一存储器，包括：

30 在所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中，将所述第一目标embedding写入所述第一存储器。

在一种可能的实现中，在将第一目标embedding写入所述第一存储器之前，所述方法还包括：

35 在所述第一存储器中可用的存储空间低于存储所述第一目标embedding所需的存储空间时，所述第二处理器将所述第一存储器中的至少一个embedding删除，并将所述至少一个embedding写入所述第二存储器，其中，所述至少一个embedding为不属于所述第M批batch的模型训练所需的embedding。

-6-

由于第一存储器中的存储空间是有限的，若所述第二处理器基于所述第一存储器中可用的存储空间低于存储所述第一目标embedding所需的存储空间，则需要将所述第一存储器中存储的至少一个embedding删除，以使得第一存储器中有足够的空间可以存储来自第二处理器传递过来的embedding。

- 5 例如，第M批batch的模型训练所需的embedding包括第一目标embedding、embedding1、embedding2、embedding3，第一存储器中存储有embedding3、embedding4、embedding5，未存储有第一目标embedding、embedding1、embedding2，且第一存储器的存储大小为4，则第一存储器中可用的存储空间(1)低于存储所述第一目标embedding所需的存储空间(3)，
10 则可以将所述第一存储器中存储的2个embedding（包括embedding4、embedding5）删除，以使得第一存储器中有足够的空间可以存储来自第二处理器传递过来的embedding（包括第一目标embedding、embedding1、embedding2）。

为了保证参数一致性，计算核还可以将删除的embedding写入第二存储器，以使得在下次第二处理器将embedding传递至第一处理器的第一存储器时，传递的embedding是最新更新过的embedding，具体的，第二处理器可以将至少一个embedding写入第二存储器中。

- 15 具体的，需要删除的embedding肯定不是当前批batch的模型训练所需的embedding，否则如果删除了当前批batch的模型训练所需的embedding，第二处理器还需要额外的时间将该embedding传递至第一处理器的第一存储器中。

- 此外，还可以选择在针对所述待训练模型的模型训练过程中，所述多个embedding中未被使用时间最长的embedding作为上述需要删除的embedding，在进行针对所述待训练模型的模型训练过程中，所述多个embedding中未被使用时间最长的embedding
20 可以认为是不经常使用的embedding，将这些embedding删除后，之后很长时间模型训练所需的embedding都可能不包括这些embedding，也就是降低了第二处理器传输数据到第一处理器的频率，通过这种选择方式，可以降低从第二处理器到第一处理器的数据传输开销，也就是降低了计算系统通信量，从而进一步提升了系统训练模型的速度。

- 25 在一种可能的实现中，所述方法还包括：

- 在所述将所述第一目标embedding写入所述第一存储器之前，所述第二处理器获取第一向量以及目标向量，其中，所述第一向量用于指示所述第M批batch的模型训练所需的embedding，所述第一向量中的每个元素用于指示一个embedding，且不同元素指示不同的embedding，所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的
30 每个元素指示的embedding的排列关系，所述每个元素指示的embedding在所述排列关系中出现至少一次；

所述第二处理器将所述第一向量以及所述目标向量传递至所述第一处理器。

在一种可能的实现中，所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU，所述第一存储器为缓存，所述第二处理器为中央处理器CPU。

35

第三方面，本申请提供了一种模型训练装置，应用于第一处理器，所述第一处理器与第二处理通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第

-7-

二存储器;所述第一处理器和所述第二处理器为不同类型的处理器,其中所述计算核包括:

获取单元,用于通过所述第二存储器获取第一目标嵌入特征embedding;

模型训练单元,用于根据所述第一目标embedding,对待训练模型进行训练,以得到更新后的待训练模型以及第二目标embedding,所述第二目标embedding为更新后的所述第一

5 目标embedding;

数据写入单元,用于将所述第二目标embedding写入所述第一存储器。

在一种可能的实现中,所述模型训练单元,用于根据所述第一目标embedding,对所述待训练模型进行第M批batch的模型训练,所述M为大于1的正整数;

10 所述获取单元,用于在所述计算核将所述第二目标embedding写入所述第一存储器之后,从所述第一存储器中获取所述第二目标embedding;

所述模型训练单元,还用于:

根据所述第二目标embedding,对所述更新后的待训练模型进行第M+1批batch的模型训练。

15

在一种可能的实现中,所述模型训练单元,还用于:

根据所述第二目标embedding,对所述更新后的待训练模型进行第M+1批batch的模型训练之后,得到第三目标embedding,所述第三目标embedding为更新后的所述第二目标embedding;

20 所述数据写入单元,还用于将所述第三目标embedding写入所述第一存储器。

在一种可能的实现中,当所述第一目标embedding为所述第M批batch的模型训练所需的embedding,且所述第一存储器未存储有所述第一目标embedding,所述第一存储器从所述第二存储器中获取所述第一目标embedding;

25 所述获取单元,具体用于:

从所述第一存储器中获取第一目标embedding。

在一种可能的实现中,所述第一存储器获取来自所述第二存储器中的所述第一目标embedding,包括:

30 在所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中,所述第一存储器从所述第二存储器中获取所述第一目标embedding。

在一种可能的实现中,所述获取单元,还用于在所述计算核从所述第一存储器中获取第一目标embedding之前,获取来自所述第二处理器的第一向量以及目标向量,其中,所述

35 第一向量用于指示所述第M批batch的模型训练所需的embedding,所述第一向量中的每个元素用于指示一个embedding,且不同元素指示不同的embedding,所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关

- 8 -

系，所述每个元素指示的embedding在所述排列关系中出现至少一次；

根据所述第一向量从所述第一存储器中获取所述第M批batch的模型训练时所需的embedding；

5 根据所述目标向量以及所述第M批batch的模型训练时所需的embedding，确定在进行所述第M批batch的模型训练时所需的每个embedding的排列关系。

在一种可能的实现中，所述获取单元，还用于：

10 在所述计算核根据所述第一目标embedding，对所述待训练模型进行训练之前，获取来自第三处理器中第三存储器的第三目标embedding，所述第三目标embedding和所述第一目标embedding为所述第M批batch模型训练所需的embedding，所述第一处理器和所述第三处理器为相同类型的处理器；

所述模型训练单元，具体用于：

通过所述待训练模型，对所述第一目标embedding和所述第三目标embedding进行处理，以得到目标输出，并基于所述目标输出确定第一梯度；

15 基于所述第一梯度更新所述待训练模型以及所述第一目标embedding。

在一种可能的实现中，所述模型训练单元，具体用于：

基于所述目标输出确定所述第一梯度和第二梯度；

20 所述计算核还包括：第二数据传递单元，用于在所述基于所述目标输出确定所述第一梯度和所述第二梯度之后，将所述第二梯度传递至第三处理器的计算核，以便所述第三处理器的计算核根据所述第二梯度更新所述第三目标embedding。

在一种可能的实现中，所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU，所述第一存储器为缓存，所述第二处理器为中央处理器CPU。

25

第四方面，本申请提供了一种模型训练装置，应用于第二处理器，所述第二处理器与第一处理通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器，所述第一处理器和所述第二处理器为不同类型的处理器；所述装置包括：

30 获取单元，用于获取所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息，所述M为大于1的正整数；

数据写入单元，用于根据所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息，将第一目标embedding写入所述第一存储器，所述第一目标embedding为所述第M批batch的模型训练所需的且所述第一存储器未存储的embedding。

35

在一种可能的实现中，所述数据写入单元，用于在所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中，将所述第一目标embedding写入所述第一存储器。

在一种可能的实现中，所述数据写入单元，还用于在将第一目标embedding写入所述第一存储器之前，在所述第一存储器中可用的存储空间低于存储所述第一目标embedding所需的存储空间时，将所述第一存储器中的至少一个embedding删除，并将所述至少一个embedding写入所述第二存储器，其中，所述至少一个embedding为不属于所述第M批batch的模型训练所需的embedding。

在一种可能的实现中，所述获取单元，还用于在所述将所述第一目标embedding写入所述第一存储器之前，获取第一向量以及目标向量，其中，所述第一向量用于指示所述第M批batch的模型训练所需的embedding，所述第一向量中的每个元素用于指示一个embedding，且不同元素指示不同的embedding，所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系，所述每个元素指示的embedding在所述排列关系中出现至少一次；

所述数据写入单元，还用于将所述第一向量以及所述目标向量传递至所述第一处理器。

在一种可能的实现中，所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU，所述第一存储器为缓存，所述第二处理器为中央处理器CPU。

第五方面，本申请实施例提供了一种模型训练装置，可以包括存储器、处理器以及总线系统，其中，存储器用于存储程序，处理器用于执行存储器中的程序，以执行如上述第一方面、第二方面及其任一可选的方法。

第六方面，本申请实施例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中存储有计算机程序，当其在计算机上运行时，使得计算机执行上述第一方面、第二方面及其任一可选的方法。

第七方面，本申请提供了一种计算机程序产品，包括代码，当代码在计算机上运行时，使得计算机执行上述第一方面、第二方面及其任一可选的方法。

第八方面，本申请提供了一种芯片系统，该芯片系统包括处理器，用于支持执行设备或训练设备实现上述方面中所涉及的功能，例如，发送或处理上述方法中所涉及的数据；或，信息。在一种可能的设计中，所述芯片系统还包括存储器，所述存储器，用于保存执行设备或训练设备必要的程序指令和数据。该芯片系统，可以由芯片构成，也可以包括芯片和其他分立器件。

本申请实施例提供了一种模型训练方法，应用于第一处理器，所述第一处理器与第二处理器通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器；所述第一处理器和所述第二处理器为不同类型的处理器，所述方法包括：所述计算核通过所述第二存储器获取第一目标嵌入特征embedding；所述计算核根据所述第一目标embedding，对待训练模型进行训练，以得到更新后的待训练模型以及第二目标embedding，所述第二目标embedding为更新后的所述第一目标embedding；所述计算核将所述第二目标

embedding写入所述第一存储器。通过上述方式，第一处理器在完成embedding的更新后，将更新后的embedding存储至第一处理器的第一存储器中，在保证第一处理器可以获取到最新更新的embedding的前提下，使得第一处理器不需要等待第二处理器将第二目标embedding传递至GPU的过程完成，就可以直接获取到更新后的embedding，并基于更新后的embedding进行下一轮模型训练，本申请可以处理器在进行模型训练时的训练数据的处理速度。

附图说明

- 图 1 为人工智能主体框架的一种结构示意图；
图 2 为一种应用架构示意；
图 3 为一种应用架构示意；
图 4 为一种模型的结构示意；
图 5 为本申请实施例提供的一种模型训练方法的实施例示意；
图 6 为一种模型训练方法的示意；
图 7 至图 12 为本申请实施例中的模型训练方法的实施例示意；
图 13 至图 16 为本申请实施例的效果示意；
图 17 和图 18a 为本申请实施例的软件模块示意；
图 18b 为本申请实施例中的模型训练方法的实施例示意；
图 19 和图 20 为本申请实施例提供的一种模型训练装置的示意；
图 21 是本申请实施例提供的训练设备一种结构示意图；
图 22 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

- 下面结合本发明实施例中的附图对本发明实施例进行描述。本发明的实施方式部分使用的术语仅用于对本发明的具体实施例进行解释，而非旨在限定本发明。

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。

- 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术语在适当情况下可以互换，这仅仅是描述本申请的实施例中对相同属性的对象在描述时所采用的区分方式。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，以便包含一系列单元的过程、方法、系统、产品或设备不必限于那些单元，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单元。

- 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题框架进行阐述。其中，“智能信息链”反映从数据的获取到处理的一系列过程。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。

(1) 基础设施

- 5 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（CPU、NPU、GPU、ASIC、FPGA 等硬件加速芯片）提供；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。

10 (2) 数据

基础设施的上一层的数据用于表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。

(3) 数据处理

- 15 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。

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。

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。

- 20 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。

(4) 通用能力

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，翻译，文本的分析，计算机视觉的处理，语音识别，图像的识别等等。

25 (5) 智能产品及行业应用

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、智能交通、智能医疗、自动驾驶、智慧城市等。

接下来描述本申请实施例的应用场景。

- 30 本申请实施例可以应用于计算系统，其中计算系统可以是服务器或者终端设备，其中终端设备可以是可移动设备、计算机等等，计算系统可以包括第一处理器和第二处理器，其中，第一处理器可以为图形处理器 GPU、嵌入式神经网络处理器 NPU 或张量处理器 TPU 等等，第二处理器可以为中央处理器 CPU。

- 35 以计算系统为服务器，第一处理器为 GPU，第二处理器为 CPU 为例，参照图 2，图 2 为本申请实施例提供的一种服务器的架构示意，如图 2 所示，服务器可以包括 CPU 以及多个和 CPU 互联的 GPU，其中，CPU 可以包括内存（后续实施例中可以称之为第二存储器），CPU 可以通过总线来读写内存，GPU 可以包括计算核以及缓存（或者称之为显存存储器）

(后续实施例中可以称之为第一存储器)。CPU 和 GPU 之间可以通过 PCIe 等总线传输/交换数据, GPU 与 GPU 之间通过 PCIe Switch/NVSwitch 互联。

参照图 3, 图 3 为本申请实施例提供的一种服务器集群的架构示意, 如图 3 所示, 服务器集群可以包括多个互联的服务器节点, 其中, 服务器节点是一个虚拟概念, 包括一个多核多线程的 CPU 节点以及若干个 GPU 节点。多个服务器节点之间可以使用以太网/InfiniBand 等网络实现互联。

本申请实施例应用的计算系统可以为服务器或者多个服务器节点, 其中嵌入特征 embedding (或者称之为 embedding 参数) 存储于 CPU 节点的内存之中, GPU 节点的缓存可用于缓存部分 embedding。CPU 节点的硬盘用于存储训练样本, CPU 节点的计算核或线程主要用于数据读取、embedding 在 CPU 与 GPU 之间搬运。GPU 节点的计算核用于神经网络的前向、反向计算, GPU 节点之间通过 PCIe 等方式进行数据交换/通信。

由于本申请实施例涉及大量神经网络的应用, 为了便于理解, 下面先对本申请实施例涉及的相关术语及神经网络等相关概念进行介绍。

(1) 神经网络

神经网络可以是由神经元组成的, 神经元可以是指以 x_s 和截距 1 为输入的运算单元, 该运算单元的输出可以为:

其中, $s=1, 2, \dots, n$, n 为大于 1 的自然数, W_s 为 x_s 的权重, b 为神经单元的偏置。 f 为神经单元的激活函数(activation functions), 用于将非线性特性引入神经网络中, 来将神经元中的输入信号转换为输出信号。该激活函数的输出信号可以作为下一层卷积层的输入, 激活函数可以是 sigmoid 函数。神经网络是将多个上述单一的神经单元联结在一起形成的网络, 即一个神经单元的输出可以是另一个神经单元的输入。每个神经单元的输入可以与前一层的局部接受域相连, 来提取局部接受域的特征, 局部接受域可以是由若干个神经单元组成的区域。

(2) 嵌入层(embedding 层)

嵌入层可以称为输入嵌入(input embedding)层。当前输入可以为文本输入, 例如可以为一段文本, 也可以为一个句子。文本可以为中文文本, 也可以为英文文本, 还可以为其他语言文本。嵌入层在获取当前输入后, 可以对该当前输入中各个词进行嵌入处理, 可得到各个词的特征向量。在一些实施例中, 所述嵌入层包括输入嵌入层和位置编码(positional encoding)层。在输入嵌入层, 可以对当前输入中的各个词进行词嵌入处理, 从而得到各个词的词嵌入向量。在位置编码层, 可以获取各个词在该当前输入中的位置, 进而对各个词的位置生成位置向量。在一些示例中, 各个词的位置可以为各个词在该当前输入中的绝对位置。以当前输入为“几号应还花呗”为例, 其中的“几”的位置可以表示为第一位, “号”的位置可以表示为第二位, ……。在一些示例中, 各个词的位置可以为各个词之间的相对位置。仍以当前输入为“几号应还款”为例, 其中的“几”的位置可以表示为“号”之前, “号”的位置可以表示为“几”之后、“应”之前, ……。当得到当前输入中各个词的词嵌入向量和位置向量时, 可以将各个词的位置向量和对应的词嵌入向量进行组合, 得到各个

词特征向量，即得到该当前输入对应的多个特征向量。多个特征向量可以表示为具有预设维度的嵌入向量。可以设定该多个特征向量中的特征向量个数为 M ，预设维度为 H 维，则该多个特征向量可以表示为 $M \times H$ 的嵌入向量。

(3) 深度学习推荐系统

5 基于海量用户和推荐对象的交互行为信息，通过梯度下降等方法，训练推荐模型参数，当模型参数收敛后，可以对用户进行个性化推荐。以个性化推荐系统中的点击率为例，输入特征包括用户特征，商品特征，用户历史行为特征和上下文特征，其中特征又分为连续特征和离散特征。对于离散特征，使用 one-hot 进行稀疏编码，然后通过编码 id 从 embeddingtable 中取出 embedding 向量表达，将原始离散特征 id 进行向量化，然后输入推荐模型中训练学习，进行预测。

10 在深度学习推荐系统的训练过程中，训练模型规模的大小和数据量对模型在线服务的效果有直接影响。推荐系统中，为了使模型可以更为精准的预测用户兴趣，往往需要使用海量的特征和数据。海量特征带来了对应的海量参数，使模型规模达到数百 GB 甚至 TB，而海量数据使得模型训练所需要算力增加。而且，为了更快速的捕捉用户兴趣，从而为用户提供更好的用户体验和推荐效果，深度模型需要频繁和快速更新，这些需求带来棘手的问题：需要基于海量特征和海量数据快速的完成推荐模型训练。

15 与视觉、自然语言处理 (natural language processing, NLP) 模型不同，推荐模型有其独有特点。以点击率预测 (click through rate, CTR) CTR 模型为例 (如图 4)，推荐模型包括嵌入层 embeddingLayer 和多层感知机 (multi-layer perception, MLP) 两个部分。推荐任务的数据包含大量高维稀疏的 id 类特征，一般使用 one-hot 编码，这部分特征直接用于模型训练会导致深度部分参数学习困难。为解决该问题，研究人员提出 embedding 层将高维稀疏的 one-hot 向量压缩为低维稠密的向量，也称为嵌入特征 embedding。推荐模型中主要的参数也位于 embedding 层 (数百 GB 甚至 TB)。此外，为提升模型预测精度，推荐模型一般会基于乘法网络和 MLP 网络构建特征组合挖掘模块，该模块一般比较复杂，需要比较大的计算量，可用于数据特征之间存在的特征交互和非线性映射。

30 参照图5，图5为本申请实施例提供的一种模型训练方法，所述第一处理器与第二处理器通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器；所述第一处理器和所述第二处理器为不同类型的处理器，其中，所述第一处理器可以为GPU、NPU或TPU，所述第一存储器可以为缓存，所述第二处理器可以为CPU。

如图5所示，本申请实施例提供的一种模型训练方法包括：

501、所述计算核通过所述第二存储器获取第一目标嵌入特征embedding。

本申请实施例中，计算核在进行模型训练时，可以获取到待训练模型，其中待训练模型可以包括嵌入层以及其他网络层，以推荐模型为例，待训练模型可以包括MLP层。

35 为了进行待训练模型的训练，计算核可以从第一处理器的第一存储器中获取到当前批batch所需的embedding (包括第一目标embedding)。为了方便描述，接下来以第一处理器为GPU，第二处理器为CPU为例进行说明。

在现有的实现中，计算核在完成一批batch的模型训练后，需要将更新后的embedding回传至CPU（例如包括更新后的embedding1），CPU维护有一个embedding表table，其中embeddingtable可以包括部分或全部模型训练所需的embedding，如果接下来一批batch的模型训练同样需要（embedding1），为了保证关注参数的一致性，则需要CPU将更新后的embedding1传递至GPU，进而GPU的计算核可以从缓存中获取到更新后的embedding1，并基于更新后的embedding1进行下一批batch的模型训练，因而，GPU需要等待CPU完成将更新后的embedding1传递至GPU的过程，这就会造成时间上的延迟，大大增加了训练的时长。

具体的，可以参照图6，在时刻1，CPU可以确定下一批batch的模型训练所需的embedding为数据1，2，3，在时刻2，CPU可以将数据1，2，3传递至GPU，在时刻3，GPU的计算核可以基于数据1，2，3完成模型训练，并将更新后的数据1，2，3传递至CPU，同时在时刻3，CPU可以确定下一批batch的模型训练所需的embedding为数据2，3，4，在时刻4，CPU可以将数据2，3，4传递至GPU，在时刻5，GPU的计算核可以基于数据2，3，4完成模型训练，并将更新后的数据2，3，4传递至CPU，同时在时刻5，CPU可以确定下一批batch的模型训练所需的embedding为数据1，3，4，在时刻6，CPU可以将数据1，3，4传递至GPU。在时刻2、4、6，GPU需要等待CPU完成将embedding传递至GPU的过程，大大增加了训练的时长。

本申请实施例提供的模型训练方法可以解决上述问题。

具体的，参照图9，在一种实现中，若当前要进行的模型训练为第M批batch的模型训练，则所述第二处理器可以判断第一存储器中是否存储有第M批batch的模型训练所需的embedding，若第二处理器确定第一存储器中未存储有所述第M批batch的模型训练所需的embedding中的部分或全部，则可以从第二存储器中获取上述第一存储器中未存储且模型训练所需的embedding，并将第一存储器中未存储且模型训练所需的embedding传递至第一存储器。

例如，第M批batch的模型训练所需的embedding包括第一目标embedding，若第二处理器确定第一存储器中未存储有第一目标embedding，则第二处理器可以将所述第一目标embedding从第二存储器写入所述缓存。

在写入的时间点选择上，可以将所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中，所述第二处理器将所述第一目标embedding从所述第二存储器写入所述第一存储器。

由于第一存储器中的存储空间是有限的，若所述第二处理器基于所述第一存储器中可用的存储空间低于存储所述目标embedding所需的存储空间，则需要将所述第一存储器中存储的至少一个embedding删除，以使得第一存储器中有足够的空间可以存储来自第二处理器传递过来的embedding。

例如，第M批batch的模型训练所需的embedding包括第一目标embedding、embedding1、embedding2、embedding3，第一存储器中存储有embedding3、embedding4、embedding5，未存储有第一目标embedding、embedding1、embedding2，且第一存储器的存储大小为4，则第一存储器中可用的存储空间(1)低于存储所述第一目标embedding所需的存储空间(3)，

则可以将所述第一存储器中存储的2个embedding（包括embedding4、embedding5）删除，以使得第一存储器中有足够的空间可以存储来自第二处理器传递过来的embedding（包括第一目标embedding、embedding1、embedding2）。

5 参照图10，为了保证参数一致性，计算核还可以将删除的embedding写入第二存储器，以使得在下一第二次处理器将embedding传递至第一处理器的第一存储器时，传递的embedding是最新更新过的embedding，具体的，第二处理器可以将至少一个embedding写入第二存储器中。

接下来描述，第二处理器如何确定第一存储器中应该删除哪些embedding。

10 本申请实施例中，可以根据预设条件，从所述第一存储器中存储的多个embedding中获取所述至少一个embedding，其中，所述预设条件包括如下条件的至少一种：在进行针对于所述待训练模型的模型训练过程中，所述多个embedding中未被使用时间最长的embedding；以及，不属于所述第M批batch的模型训练所需的embedding。

15 具体的，需要删除的embedding肯定不是当前批batch的模型训练所需的embedding，否则如果删除了当前批batch的模型训练所需的embedding，CPU还需要额外的时间将该embedding传递至第一处理器的第一存储器中。

20 此外，在进行针对于所述待训练模型的模型训练过程中，所述多个embedding中未被使用时间最长的embedding可以认为是不经常使用的embedding，将这些embedding删除后，之后很长时间模型训练所需的embedding都可能不包括这些embedding，也就是降低了CPU传输数据到第一处理器的频率，通过这种选择方式，可以降低从CPU到第一处理器的数据传输开销，也就是降低了计算系统通信量，从而进一步提升了系统训练模型的速度。

在进行模型训练的过程中，为了使得第一处理器的计算核可以知道当前批batch的模型训练所需的embedding，第二处理器你需要将指示当前批batch的模型训练所需的embedding的信息传递至计算核，进而计算核可以基于上述信息从第一存储器中读取当前批batch的模型训练所需的embedding以及embedding之间的排列关系。

25 具体的，在所述计算核从所述第一存储器中获取第一目标embedding之前，所述计算核可以获取来自所述第二处理器的第一向量以及目标向量，其中，所述第一向量用于指示所述第M批batch的模型训练所需的embedding，所述第一向量中的每个元素用于指示一个embedding，且不同元素指示不同的embedding，所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系，所述每个元素指示的embedding在所述排列关系中出现至少一次；其中，排列关系可以为第一向量中的每个元素指示的embedding在第M批batch的模型训练时的使用次数，以及第一向量中的每个元素指示的embedding在第M批batch的模型训练时之间的先后顺序。

30 在一种实现中，同一批batch模型训练所需的embedding可以包括相同的embedding，例如，同一批batch模型训练所需的embedding可以为：
35 $D=[\text{embedding1}, \text{embedding2}, \text{embedding3}, \text{embedding3}, \text{embedding2}, \text{embedding4}]$ ，则embedding2和embedding3都是重复出现的，为了降低后续第一处理器的计算核读取embedding的参数量，第二处理器可以对同一批batch模型训练所需的embedding进行去重操

- 16 -

作，具体的，所述第二处理器可以获取第一向量，所述第一向量包括的元素为所述第M批batch的模型训练所需的embedding；且第一向量可以指示去重后的模型训练所需的embedding的信息。

示例性的，去重前的第一向量可以为：D=[embedding1,embedding2,embedding3,embedding3,embedding2,embedding4]，对第一向量进行去重操作得到去重后的第一向量U=[embedding1, embedding2, embedding3, embedding4]，并计算得到目标向量Mask=[0, 1, 2, 2, 1, 3]，其中，目标向量的第一个元素为0，表示第一向量中的第一个元素为第二向量中的第一个元素，目标向量的第二个元素为1，表示第一向量中的第二个元素为第二向量中的第二个元素，目标向量的第三个元素为2，表示第一向量中的第三个元素为第二向量中的第三个元素，目标向量的第四个元素为2，表示第一向量中的第四个元素为第二向量中的第三个元素，目标向量的第五个元素为1，表示第一向量中的第五个元素为第二向量中的第二个元素，目标向量的第六个元素为3，表示第一向量中的第六个元素为第二向量中的第四个元素，以此，可以基于第一向量和目标向量还原出去重前的第一向量，其中，去重前的第一向量用于确定出在进行所述第M批batch的模型训练时所需的每个embedding的的排列关系。

在一种实现中，第一处理器可以根据所述第一向量从所述第一存储器中获取所述第M批batch的模型训练时所需的embedding，之后再根据所述目标向量以及所述第M批batch的模型训练时所需的embedding，确定在进行所述第M批batch的模型训练时所需的每个embedding的的排列关系。

502、所述计算核根据所述第一目标embedding，对待训练模型进行训练，以得到更新后的待训练模型以及第二目标embedding，所述第二目标embedding为更新后的所述第一目标embedding。

本申请实施例中，计算核在获取到当前批batch模型训练所需的embedding（包括第一目标embedding）之后，可以根据当前批batch模型训练所需的embedding（包括第一目标embedding），对所述待训练模型进行训练，以得到更新后的待训练模型以及更新后的embedding（包括第二目标embedding）。

在一种实现中，计算核可以基于当前批batch模型训练所需的embedding（包括第一目标embedding）以及待训练模型进行前向和反向的计算，以得到待训练模型对应的梯度以及当前批batch模型训练所使用的各个embedding对应的梯度，以待训练模型包括MLP为例，待训练模型的网络梯度可以包括神经网络梯度(MLP_G)；之后可以基于待训练模型的网络梯度以及当前批batch模型训练所需的embedding的梯度来更新待训练模型的参数以及更新embedding。

在一种实现中，若计算系统还可以包括和第一处理器相同处理器类型的第三处理器，第一处理器可以完成当前批batch模型训练所需的embedding的一部分的梯度的计算，在得到embedding梯度后，可以通过第一处理器之间的通信来同步embedding的梯度。

具体的，所述计算核获取来自第三处理器中第三存储器的第三目标embedding，所述第三目标embedding和所述第一目标embedding为所述第M批batch模型训练所需的embedding，

所述第一处理器和所述第三处理器为相同类型的处理器；所述计算核通过所述待训练模型，对所述第一目标embedding和所述第三目标embedding进行处理，以得到目标输出，并基于所述目标输出确定第一梯度和第二梯度；计算核可以基于所述第一梯度更新所述待训练模型以及所述第一目标embedding，并将所述第二梯度传递至第三处理器的计算核，以便所述第三处理器的计算核根据所述第二梯度更新所述第三目标embedding。

在包括多个计算节点的场景中，每个计算节点可以负责当前batch模型训练所需的embedding的一部分，且负责一部分embedding的更新，在一种可能的实现中，计算核需要从其他服务器节点的第三处理器的第三存储器中获取第三目标embedding，并基于第一目标embedding和第三目标embedding来进行模型训练过程中的前馈过程，然而在更新时，第一处理器中的计算核仅需要基于第一梯度来更新第一目标embedding。

503、所述计算核将所述第二目标embedding写入所述第一存储器。

参照图7，本申请实施例中，计算核可以将更新后的待训练模型以及更新后的embedding（包括第二目标embedding）写入所述第一存储器。

若所述第一目标embedding还为接下来一批batch模型训练所需的embedding，则计算核可以直接从第一存储器中获取更新后的第一目标embedding（也就是第二目标embedding）。若第一存储器中未存储有第M+1批batch的模型训练所需的一部分embedding，则第二处理器可以将这部分embedding传递至第一存储器。

参照图8，具体的，第一目标embedding为第M批batch以及第M+1批batch模型训练所需的embedding，所述M为大于1的正整数，所述计算核可以根据所述第一目标embedding，对所述待训练模型进行第M批batch的模型训练，在进行第M+1批batch的模型训练时，所述计算核从所述第一存储器中获取所述第二目标embedding，并根据所述第二目标embedding，对所述更新后的待训练模型进行第M+1批batch的模型训练，以得到第三目标embedding，之后所述计算核可以将所述第三目标embedding写入所述第一存储器。

以第一处理器为GPU、第二处理器为CPU为例，如图11所示，假设GPU的第一存储器的大小为6，在模型训练的开始阶段，在时刻1，CPU可以确定下一批batch的模型训练所需的embedding为embedding1、embedding2以及embedding3（图11中示出的为数据1，2，3）；

在时刻2，CPU可以将数据1，2，3传输至GPU的第一存储器；

在时刻3，GPU的计算核可以基于数据1，2，3完成模型训练，CPU可以确定下一批batch的模型训练所需的embedding为数据2，3，4，由于第一存储器中未存储有数据4且第一存储器可用的存储空间足够，则CPU可以将数据4传递至GPU的第一存储器；

在时刻4，GPU的计算核可以基于数据2，3，4完成模型训练，CPU可以确定下一批batch的模型训练所需的embedding为数据1，3，5，由于第一存储器中未存储有数据5且第一存储器可用的存储空间足够，则CPU可以将数据5传递至GPU的第一存储器；

在时刻5，GPU的计算核可以基于数据1，3，5完成模型训练，CPU可以确定下一批batch的模型训练所需的embedding为数据1，6，7，由于第一存储器中未存储有数据6，7，且第一存储器可用的存储空间不足够（缺少一个），则CPU可以将数据4删除并回传至CPU的第二存储器，则CPU可以将数据6，7传递至GPU的第一存储器；

在时刻6, GPU的计算核可以基于数据1, 6, 7完成模型训练。

可见本申请实施例中, 第一处理器在完成第一目标embedding的更新后, 将第二目标embedding存储至第一处理器的第一存储器中, 在保证第一处理器可以获取到最新更新的目标embedding (第二目标embedding) 的前提下, 使得第一处理器不需要等待第二处理器将第二目标embedding传递至GPU的过程完成, 就可以直接获取到第二目标embedding, 并基于第二目标embedding进行模型训练。增加了训练数据的吞吐量, 进而提高了模型训练的速度。

此外, 在前一批batch模型训练的过程中, 第二处理器会将第一存储器中没有存储且下一批batch模型训练过程中所需的embedding传递至第一存储器, 使得第二处理器的计算核可以从第一存储器中获取到下一批batch模型训练过程中所需的全部embedding。

参照图12, 在包括多个计算节点的场景中, 每个计算节点可以负责当前batch模型训练所需的embedding的一部分, 且负责一部分embedding的更新, 在一种可能的实现中, 计算核需要从其他服务器节点的第三处理器的第三存储器中获取第三目标embedding, 并基于第一目标embedding和第三目标embedding来进行模型训练过程中的前馈过程, 然而在更新时, 第一处理器中的计算核仅需要基于第一梯度来更新第一目标embedding。

具体的, 在模型训练的过程中, 所述计算核可以通过所述待训练模型, 对所述第一目标embedding和所述第三目标embedding进行处理, 以得到目标输出, 并基于所述目标输出确定第一梯度和第二梯度, 所述计算核可以基于所述第一梯度更新所述待训练模型以及所述第一目标embedding, 并将所述第二梯度传递至第三处理器的计算核, 以便所述第三处理器的计算核根据所述第二梯度更新所述第三目标embedding。

在一种实现中, 可以采用第一处理器和第三处理器之间的数据通信 (例如allreduce通信) 同步embedding的梯度, 以使得第三处理器可以得到当前batch模型训练后第一处理器计算得到的第二梯度, 之后第三处理器可以基于第二梯度对自己负责的embedding (第三目标embedding) 进行更新。

示例性的, 每个服务器节点为3个特征, 当前batch模型训练所需的embedding为[1,2,3,3,2,4], 应理解, 这里的1可以表示embedding1, 2可以表示embedding2, 3可以表示embedding3, 4可以表示embedding4, 前三个属于第一个服务器节点, 后三个为第二个服务器节点; 每个服务器节点都读入数据[1,2,3,3,2,4], CPU可以将[0,1,2]传输至GPU节点1, [2,1,3]传输至GPU节点2;

此时, GPU节点1的第一存储器中存储有2和4的embedding, GPU节点2的第一存储器中存储有1和3的embedding, 采用GPU之间的数据通信 (例如Allreduce通信) 交换两个GPU节点的embedding数据, 每个GPU都可以得到E=[1,2,3,4]。

具体的, 可以采用GPU之间的数据通信 (例如Allreduce通信) 同步embedding的梯度, 以使得每个GPU可以得到当前batch所有GPU节点计算得到的梯度, 之后每一个GPU节点可以对自己负责的embedding进行更新, 如GPU节点1负责特征2和4的参数更新。

接下来基于实现结果, 对本申请实施例提供的模型训练方法进行验证。

具体的, 可以使用 TB 级的 Criteo 数据集进行验证。通过特征筛选, 分别构建 10GB

embeddingTable (33M 个特征 id) 的数据和 100GB embeddingTable 数据 (330M 个特征 id)。测试的模型是 DeepFM 模型。实验硬件环境为每一台服务器有两颗 2 18 核 36 线程的 CPU，有 8 张 32G 显存的 V100 卡，GPU 之间通过 PCIe 互联，服务器之间通过 100Gb RDMA 相连，每个服务器有接近 1T 内存。

5 如图 13 所述，针对于吞吐量实验，本申请实施例取得了更高的训练吞吐量，在 10GB 模型上分别是 HugeCTR 和参数服务器 (PS) 的 4.7 倍和 1.7 倍，在 100GB 模型上分别是 HugeCTR 和参数服务器的 7.4 和 1.4 倍。

如图 14 所示，在加入去重操作后，GPU-CPU 和 GPU-GPU 的通信量大大减少。

10 如图 15 所示，相比传统 Pre-fetch 和参数服务器 (Host)，可以发现本申请实施例的训练速度最快。图 16 比较了这三种技术的模型收敛曲线，可以发现本申请实施例 (对应于 Cache 曲线) 的收敛速度最快。

接下来从软件模块的角度，描述本申请实施例中的模型训练方法：

15 参照图17和图18a，在CPU节点的硬盘或其他存储位置存放训练数据 (embedding)，在 CPU 的内存存储 embeddingtable；数据读取模块可以用于训练样本的读取；稀疏数据编码压缩模块用于对当前 batch 的训练样本进行去重操作 (或者称之为压缩编码)，输出目标向量 Mask 和第二向量；参数预取管理模块可以维护缓存 cache，保障在 CPU 维护 embedding 表，因此可支持超大规模；待训练模型在每一个 GPU 节点都有一个备份；GPU 可以包括 embeddingCache (缓存)，每个 GPU 有一个 Cache 来缓存一部分 embedding；embedding 拉取模块，用于获取第二向量；稀疏数据解码模块，用于基于目标向量 Mask 和第二向量，得到前向计算所需的 embedding (第一向量)，训练模块，用于进行前向和反向梯度计算，参数更新；集合通信模块，用于同步待训练模型，同步 embedding 的梯度信息。

20 本申请实施例提供了一种模型训练方法，应用于第一处理器，所述第一处理器与第二处理器通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器；所述第一处理器和所述第二处理器为不同类型的处理器，所述方法包括：所述计算核通过所述第二存储器获取第一目标嵌入特征 embedding；所述计算核根据所述第一目标 embedding，对待训练模型进行训练，以得到更新后的待训练模型以及第二目标 embedding，所述第二目标 embedding 为更新后的所述第一目标 embedding；所述计算核将所述第二目标 embedding 写入所述第一存储器。通过上述方式，第一处理器在完成 embedding 的更新后，将更新后的 embedding 存储至第一处理器的第一存储器中，在保证第一处理器可以获取到最新更新的 embedding 的前提下，使得第一处理器不需要等待第二处理器将第二目标 embedding 传递至 GPU 的过程完成，就可以直接获取到更新后的 embedding，并基于更新后的 embedding 进行下一轮模型训练，本申请可以处理器在进行模型训练时的训练数据的处理速度。

35 以上以第一处理器为执行主体描述了本申请实施例中的模型训练方法，接下来以第二处理器为执行主体描述本申请实施例中的模型训练方法。

参照图18b，图18b为本申请实施例提供的一种模型训练方法的流程示意，如图18b所示，

本申请实施例提供的一种模型训练方法，应用于第二处理器，所述第二处理器与第一处理通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器，所述第一处理器和所述第二处理器为不同类型的处理器；所述方法包括：

5 1801、所述第二处理器获取所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息，所述M为大于1的正整数；

其中，第M批batch模型训练所需的embedding信息可以指示第M批batch模型训练所需的embedding，例如该信息可以但不限于为指示embedding的标识信息。

10 1802、所述第二处理器根据所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息，将第一目标embedding写入所述第一存储器，所述第一目标embedding为所述第M批batch的模型训练所需的且所述第一存储器未存储的embedding。

15 若当前要进行的模型训练为第M批batch的模型训练，则所述第二处理器可以判断第一存储器中是否存储有第M批batch的模型训练所需的embedding，若第二处理器确定第一存储器中未存储有所述第M批batch的模型训练所需的embedding中的部分或全部，则可以从第二存储中获取上述第一存储器中未存储且模型训练所需的embedding，并将第一存储器中未存储且模型训练所需的embedding传递至第一存储器。

例如，第M批batch的模型训练所需的embedding包括第一目标embedding，若第二处理器确定第一存储器中未存储有第一目标embedding，则第二处理器可以将所述第一目标embedding从第二存储器写入所述缓存。

20 在写入的时间点选择上，可以将所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中，所述第二处理器将所述第一目标embedding从所述第二存储器写入所述第一存储器。

25 在一种可能的实现中，在将第一目标embedding写入所述第一存储器之前，在所述第一存储器中可用的存储空间低于存储所述第一目标embedding所需的存储空间时，所述第二处理器将所述第一存储器中的至少一个embedding删除，并将所述至少一个embedding写入所述第二存储器，其中，所述至少一个embedding为不属于所述第M批batch的模型训练所需的embedding。

30 由于第一存储器中的存储空间是有限的，若所述第二处理器基于所述第一存储器中可用的存储空间低于存储所述第一目标embedding所需的存储空间，则需要将所述第一存储器中存储的至少一个embedding删除，以使得第一存储器中有足够的空间可以存储来自第二处理器传递过来的embedding。

35 例如，第M批batch的模型训练所需的embedding包括第一目标embedding、embedding1、embedding2、embedding3，第一存储器中存储有embedding3、embedding4、embedding5，未存储有第一目标embedding、embedding1、embedding2，且第一存储器的存储大小为4，则第一存储器中可用的存储空间(1)低于存储所述第一目标embedding所需的存储空间(3)，则可以将所述第一存储器中存储的2个embedding（包括embedding4、embedding5）删除，以使得第一存储器中有足够的空间可以存储来自第二处理器传递过来的embedding（包括第

一目标embedding、embedding1、embedding2)。

为了保证参数一致性，计算核还可以将删除的embedding写入第二存储器，以使得在下次第二处理器将embedding传递至第一处理器的第一存储器时，传递的embedding是最新更新过的embedding，具体的，第二处理器可以将至少一个embedding写入第二存储器中。

5 具体的，需要删除的embedding肯定不是当前批batch的模型训练所需的embedding，否则如果删除了当前批batch的模型训练所需的embedding，第二处理器还需要额外的时间将该embedding传递至第一处理器的第一存储器中。

此外，还可以选择在进行针对于所述待训练模型的模型训练过程中，所述多个embedding中未被使用时间最长的embedding作为上述需要删除的embedding，在进行针对于
10 所述待训练模型的模型训练过程中，所述多个embedding中未被使用时间最长的embedding可以认为是不经常使用的embedding，将这些embedding删除后，之后很长时间模型训练所需的embedding都可能不包括这些embedding，也就是降低了第二处理器传输数据到第一处理器的频率，通过这种选择方式，可以降低从第二处理器到第一处理器的数据传输开销，也就是降低了计算系统通信量，从而进一步提升了系统训练模型的速度。

15 在一种可能的实现中，在所述将所述第一目标embedding写入所述第一存储器之前，所述第二处理器可以获取第一向量以及目标向量，其中，所述第一向量用于指示所述第M批batch的模型训练所需的embedding，所述第一向量中的每个元素用于指示一个embedding，且不同元素指示不同的embedding，所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系，所述每个元素指示的
20 embedding在所述排列关系中出现至少一次，之后所述第二处理器可以将所述第一向量以及所述目标向量传递至所述第一处理器。

关于第一向量以及目标向量的具体描述可以参照图5对应的实施例中的相关描述，这里不再赘述。

25 在一种可能的实现中，所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU，所述第一存储器为缓存，所述第二处理器为中央处理器CPU。

参照图19，图19为本申请实施例提供的一种模型训练装置，应用于第一处理器，所述第一处理器与第二处理通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器；所述第一处理器和所述第二处理器为不同类型的处理器，其中所述计算核1900包括：

30 获取单元1901，用于通过所述第二存储器获取第一目标嵌入特征embedding；

关于获取单元1901的具体描述可以参照步骤501的描述，这里不再赘述。

模型训练单元1902，用于根据所述第一目标embedding，对待训练模型进行训练，以得到更新后的待训练模型以及第二目标embedding，所述第二目标embedding为更新后的所述第一目标embedding；

35 关于模型训练单元1902的具体描述可以参照步骤502的描述，这里不再赘述。

数据写入单元1903，用于将所述第二目标embedding写入所述第一存储器。

关于数据写入单元1903的具体描述可以参照步骤503的描述，这里不再赘述。

- 22 -

在一种可能的实现中，所述模型训练单元，还用于：

根据所述第二目标embedding，对所述更新后的待训练模型进行第M+1批batch的模型训练之后，得到第三目标embedding，所述第三目标embedding为更新后的所述第二目标embedding；

5 所述数据写入单元，还用于将所述第三目标embedding写入所述第一存储器。

在一种可能的实现中，当所述第一目标embedding为所述第M批batch的模型训练所需的embedding，且所述第一存储器未存储有所述第一目标embedding，所述第一存储器从所述第二存储器中获取所述第一目标embedding；

所述获取单元，具体用于：

10 从所述第一存储器中获取第一目标embedding。

在一种可能的实现中，所述第一存储器获取来自所述第二存储器中的所述第一目标embedding，包括：

在所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中，所述第一存储器从所述第二存储器中获取所述第一目标embedding。

15 在一种可能的实现中，所述获取单元，还用于在所述计算核对所述第一存储器中获取第一目标embedding之前，获取来自所述第二处理器的第一向量以及目标向量，其中，所述第一向量用于指示所述第M批batch的模型训练所需的embedding，所述第一向量中的每个元素用于指示一个embedding，且不同元素指示不同的embedding，所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系，所述每个元素指示的embedding在所述排列关系中出现至少一次；

20 根据所述第一向量从所述第一存储器中获取所述第M批batch的模型训练时所需的embedding；

根据所述目标向量以及所述第M批batch的模型训练时所需的embedding，确定在进行所述第M批batch的模型训练时所需的每个embedding的排列关系。

25 在一种可能的实现中，所述获取单元，还用于：

在所述计算核根据所述第一目标embedding，对所述待训练模型进行训练之前，获取来自第三处理器中第三存储器的第三目标embedding，所述第三目标embedding和所述第一目标embedding为所述第M批batch模型训练所需的embedding，所述第一处理器和所述第三处理器为相同类型的处理器；

30 所述模型训练单元，具体用于：

通过所述待训练模型，对所述第一目标embedding和所述第三目标embedding进行处理，以得到目标输出，并基于所述目标输出确定第一梯度；

基于所述第一梯度更新所述待训练模型以及所述第一目标embedding。

在一种可能的实现中，所述模型训练单元，具体用于：

35 基于所述目标输出确定所述第一梯度和第二梯度；

所述计算核还包括：第二数据传递单元，用于在所述基于所述目标输出确定所述第一梯度和所述第二梯度之后，将所述第二梯度传递至第三处理器的计算核，以便所述第三处

理器的计算核根据所述第二梯度更新所述第三目标embedding。

在一种可能的实现中,所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU,所述第一存储器为缓存,所述第二处理器为中央处理器CPU。

5 参照图20,本申请还提供了一种模型训练装置,应用于第二处理器,所述第二处理器与第一处理通信连接,所述第一处理器包括计算核以及第一存储器,所述第二处理器包括第二存储器,所述第一处理器和所述第二处理器为不同类型的处理器;所述装置2000包括:

获取单元2001,用于获取所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息,所述M为大于1的正整数;

10 关于获取单元2001的具体描述可以参照步骤1801的描述,这里不再赘述。

数据写入单元2002,用于根据所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息,将第一目标embedding写入所述第一存储器,所述第一目标embedding为所述第M批batch的模型训练所需的且所述第一存储器未存储的embedding。

15 关于数据写入单元2002的具体描述可以参照步骤1802的描述,这里不再赘述。

在一种可能的实现中,所述数据写入单元,用于在所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中,将所述第一目标embedding写入所述第一存储器。

20 在一种可能的实现中,所述数据写入单元,还用于在将第一目标embedding写入所述第一存储器之前,在所述第一存储器中可用的存储空间低于存储所述第一目标embedding所需的存储空间时,将所述第一存储器中的至少一个embedding删除,并将所述至少一个embedding写入所述第二存储器,其中,所述至少一个embedding为不属于所述第M批batch的模型训练所需的embedding。

25 在一种可能的实现中,所述获取单元,还用于在所述将所述第一目标embedding写入所述第一存储器之前,获取第一向量以及目标向量,其中,所述第一向量用于指示所述第M批batch的模型训练所需的embedding,所述第一向量中的每个元素用于指示一个embedding,且不同元素指示不同的embedding,所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系,所述每个元素指示的embedding在所述排列关系中出现至少一次;

所述数据写入单元,还用于将所述第一向量以及所述目标向量传递至所述第一处理器。

30 在一种可能的实现中,所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU,所述第一存储器为缓存,所述第二处理器为中央处理器CPU。

35 本申请实施例还提供了一种训练设备,请参阅图 21,图 21 是本申请实施例提供的训练设备一种结构示意图,具体的,训练设备 2100 由一个或多个服务器实现,训练设备 2100 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处理器(central processing units, CPU) 2121(例如,一个或一个以上处理器)和存储器 2132,一个或一个以上存储应用程序 2142 或数据 2144 的存储介质 2130(例如一个或一个以上海量存储设备)。其中,存储器 2132 和存储介质 2130 可以是短暂存储或持久存储。存储在存储介质 2130 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设备中的一系列指令操作。更进一步地，处理器 2121 可以设置为与存储介质 2130 通信，在训练设备 2100 上执行存储介质 2130 中的一系列指令操作。

其中，处理器可以为本申请实施例中的第一处理器或第二处理器。

5 训练设备 2100 还可以包括一个或一个以上电源 2126，一个或一个以上有线或无线网络接口 2150，一个或一个以上输入输出接口 2158；或，一个或一个以上操作系统 2141，例如 Windows ServerTM，Mac OS XTM，UnixTM，LinuxTM，FreeBSDTM 等等。

具体的，训练设备可以执行图 5 以及图 18b 对应实施例中的模型训练方法。

10 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执行设备所执行的步骤，或者，使得计算机执行如前述训练设备所执行的步骤。

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进行信号处理的程序，当其在计算机上运行时，使得计算机执行如前述执行设备所执行的步骤，或者，使得计算机执行如前述训练设备所执行的步骤。

15 本申请实施例提供的执行设备、训练设备或终端设备具体可以为芯片，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，以使执行设备内的芯片执行上述实施例描述的数据处理方法，或者，以使训练设备内的芯片执行上述实施例描述的数据处理方法。可选地，所述存储单元为所述芯片内的存储单元，如寄存器、
20 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。

具体的，请参阅图 22，图 22 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 2200，NPU 2200 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 2203，通过控制器 2204 控制运算电路 2203 提取存储器中的向量数据并进行乘法运算。
25

其中，NPU 2200 可以为本申请实施例中的第一处理器。

在一些实现中，运算电路 2203 内部包括多个处理单元（Process Engine，PE）。在一些实现中，运算电路 2203 是二维脉动阵列。运算电路 2203 还可以是一维脉动阵列或者能够
30 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 2203 是通用的向量处理器。

举例来说，假设有输入向量 A，权重向量 B，输出向量 C。运算电路从权重存储器 2202 中取向量 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 2201 中取向量 A 数据与向量 B 进行向量运算，得到的向量的部分结果或最终结果，保存在累加器（accumulator）2208 中。
35

统一存储器 2206 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）2205，DMAC 被搬运到权重存储器 2202

中。输入数据也通过 DMAC 被搬运到统一存储器 2206 中。

BIU 为 Bus Interface Unit 即，总线接口单元 2210，用于 AXI 总线与 DMAC 和取指存储器 (Instruction Fetch Buffer, IFB) 2209 的交互。

总线接口单元 2210 (Bus Interface Unit, 简称 BIU)，用于取指存储器 2209 从外部存储器获取指令，还用于存储单元访问控制器 2205 从外部存储器获取输入向量 A 或者权重向量 B 的原数据。

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 2206 或将权重数据搬运到权重存储器 2202 中或将输入数据数据搬运到输入存储器 2201 中。

向量计算单元 2207 包括多个运算处理单元，在需要的情况下，对运算电路 2203 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。

在一些实现中，向量计算单元 2207 能将经处理的输出的向量存储到统一存储器 2206。例如，向量计算单元 2207 可以将线性函数；或，非线性函数应用到运算电路 2203 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 2207 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 2203 的激活输入，例如用于在神经网络中的后续层中的使用。

控制器 2204 连接的取指存储器 (instruction fetch buffer) 2209，用于存储控制器 2204 使用的指令；

统一存储器 2206，输入存储器 2201，权重存储器 2202 以及取指存储器 2209 均为 On-Chip 存储器。外部存储器私有于该 NPU 硬件架构。

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程序执行的集成电路。

另外需说明的是，以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元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。另外，本申请提供的装置实施例附图中，模块之间的连接关系表示它们之间具有通信连接，具体可以实现为一条或多条通信总线或信号线。

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 CPU、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。

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。

5 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。

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、训练设备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、训练设备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训练设备、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。

10

15

权 利 要 求

1.一种模型训练方法，其特征在于，应用于第一处理器，所述第一处理器与第二处理器通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器；所述第一处理器和所述第二处理器为不同类型的处理器，所述方法包括：

5 所述计算核通过所述第二存储器获取第一目标嵌入特征embedding；
 所述计算核根据所述第一目标embedding，对待训练模型进行训练，以得到更新后的待训练模型以及第二目标embedding，所述第二目标embedding为更新后的所述第一目标embedding；

 所述计算核将所述第二目标embedding写入所述第一存储器。

10

2.根据权利要求1所述的方法，其特征在于，所述计算核根据所述第一目标embedding，对所述待训练模型进行训练，包括：

 所述计算核根据所述第一目标embedding，对所述待训练模型进行第M批batch的模型训练，所述M为大于1的正整数；

15 在所述计算核将所述第二目标embedding写入所述第一存储器之后，所述方法还包括：
 所述计算核从所述第一存储器中获取所述第二目标embedding；

 所述计算核根据所述第二目标embedding，对所述更新后的待训练模型进行第M+1批batch的模型训练。

20

3.根据权利要求2所述的方法，其特征在于，所述方法还包括：

 所述计算核根据所述第二目标embedding，对所述更新后的待训练模型进行第M+1批batch的模型训练之后，得到第三目标embedding，所述第三目标embedding为更新后的所述第二目标embedding；

 所述计算核将所述第三目标embedding写入所述第一存储器。

25

4.根据权利要求1至3任一所述的方法，其特征在于，所述计算核通过所述第二存储器获取第一目标embedding，包括：

 当所述第一目标embedding为所述第M批batch的模型训练所需的embedding，且所述第一存储器未存储有所述第一目标embedding，所述第一存储器从所述第二存储器中获取所述
30 第一目标embedding；

 所述计算核从所述第一存储器中获取第一目标embedding。

5.根据权利要求4所述的方法，其特征在于，所述第一存储器获取来自所述第二存储器中的所述第一目标embedding，包括：

35 在所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中，所述第一存储器从所述第二存储器中获取所述第一目标embedding。

6.根据权利要求4或5所述的方法，其特征在于，所述方法还包括：

在所述计算核从所述第一存储器中获取第一目标embedding之前，所述计算核获取来自所述第二处理器的第一向量以及目标向量，其中，所述第一向量用于指示所述第M批batch的模型训练所需的embedding，所述第一向量中的每个元素用于指示一个embedding，且不同元素指示不同的embedding，所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系，所述每个元素指示的embedding在所述排列关系中出现至少一次；

所述计算核从所述第一存储器中获取第一目标embedding，包括：

所述计算核根据所述第一向量从所述第一存储器中获取所述第M批batch的模型训练时所需的embedding；

所述计算核根据所述目标向量以及所述第M批batch的模型训练时所需的embedding，确定在进行所述第M批batch的模型训练时所需的每个embedding的排列关系。

7.根据权利要求1至6任一所述的方法，其特征在于，在所述计算核根据所述第一目标embedding，对所述待训练模型进行训练之前，所述方法还包括：

所述计算核获取来自第三处理器中第三存储器的第三目标embedding，所述第三目标embedding和所述第一目标embedding为所述第M批batch模型训练所需的embedding，所述第一处理器和所述第三处理器为相同类型的处理器；

所述计算核根据所述第一目标embedding，对所述待训练模型进行训练，包括：

所述计算核通过所述待训练模型，对所述第一目标embedding和所述第三目标embedding进行处理，以得到目标输出，并基于所述目标输出确定第一梯度；

所述计算核基于所述第一梯度更新所述待训练模型以及所述第一目标embedding。

8.根据权利要求7所述的方法，其特征在于，所述基于所述目标输出确定所述第一梯度，包括：基于所述目标输出确定所述第一梯度和第二梯度；

所述基于所述目标输出确定所述第一梯度和所述第二梯度之后，所述方法还包括：

将所述第二梯度传递至第三处理器的计算核，以便所述第三处理器的计算核根据所述第二梯度更新所述第三目标embedding。

9.根据权利要求1至8任一所述的方法，其特征在于，所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU，所述第一存储器为缓存，所述第二处理器为中央处理器CPU。

10.一种模型训练方法，其特征在于，应用于第二处理器，所述第二处理器与第一处理通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器，所述第一处理器和所述第二处理器为不同类型的处理器；所述方法包括：

所述第二处理器获取所述计算核在进行第M批batch模型训练所需的embedding信息以

及所述第一存储器中存储的embedding信息, 所述M为大于1的正整数;

所述第二处理器根据所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息, 将第一目标embedding写入所述第一存储器, 所述第一目标embedding为所述第M批batch的模型训练所需的且所述第一存储器未存储的embedding。

11.根据权利要求10所述的方法, 其特征在于, 所述将第一目标embedding写入所述第一存储器, 包括:

10 在所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中, 将所述第一目标embedding写入所述第一存储器。

12.根据权利要求10或11所述的方法, 其特征在于, 在将第一目标embedding写入所述第一存储器之前, 所述方法还包括:

15 在所述第一存储器中可用的存储空间低于存储所述第一目标embedding所需的存储空间时, 所述第二处理器将所述第一存储器中的至少一个embedding删除, 并将所述至少一个embedding写入所述第二存储器, 其中, 所述至少一个embedding为不属于所述第M批batch的模型训练所需的embedding。

13.根据权利要求10至12任一所述的方法, 其特征在于, 所述方法还包括:

20 在所述将所述第一目标embedding写入所述第一存储器之前, 所述第二处理器获取第一向量以及目标向量, 其中, 所述第一向量用于指示所述第M批batch的模型训练所需的embedding, 所述第一向量中的每个元素用于指示一个embedding, 且不同元素指示不同的embedding, 所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系, 所述每个元素指示的embedding在所述排列关系中出现至少一次;

所述第二处理器将所述第一向量以及所述目标向量传递至所述第一处理器。

14.根据权利要求10至13任一所述的方法, 其特征在于, 所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU, 所述第一存储器为缓存, 所述第二处理器为中央处理器CPU。

15.一种模型训练装置, 其特征在于, 应用于第一处理器, 所述第一处理器与第二处理通信连接, 所述第一处理器包括计算核以及第一存储器, 所述第二处理器包括第二存储器; 所述第一处理器和所述第二处理器为不同类型的处理器, 其中所述计算核包括:

35 获取单元, 用于通过所述第二存储器获取第一目标嵌入特征embedding;

模型训练单元, 用于根据所述第一目标embedding, 对待训练模型进行训练, 以得到更新后的待训练模型以及第二目标embedding, 所述第二目标embedding为更新后的所述第一

目标embedding;

数据写入单元, 用于将所述第二目标embedding写入所述第一存储器。

- 5 16.根据权利要求15所述的装置, 其特征在于, 所述模型训练单元, 用于根据所述第一目标embedding, 对所述待训练模型进行第M批batch的模型训练, 所述M为大于1的正整数;
所述获取单元, 用于在所述计算核将所述第二目标embedding写入所述第一存储器之后, 从所述第一存储器中获取所述第二目标embedding;

所述模型训练单元, 还用于:

- 10 根据所述第二目标embedding, 对所述更新后的待训练模型进行第M+1批batch的模型训练。

17.根据权利要求16所述的装置, 其特征在于, 所述模型训练单元, 还用于:

- 15 根据所述第二目标embedding, 对所述更新后的待训练模型进行第M+1批batch的模型训练之后, 得到第三目标embedding, 所述第三目标embedding为更新后的所述第二目标embedding;

所述数据写入单元, 还用于将所述第三目标embedding写入所述第一存储器。

- 20 18.根据权利要求15至17任一所述的装置, 其特征在于, 当所述第一目标embedding为所述第M批batch的模型训练所需的embedding, 且所述第一存储器未存储有所述第一目标embedding, 所述第一存储器从所述第二存储器中获取所述第一目标embedding;

所述获取单元, 具体用于:

从所述第一存储器中获取第一目标embedding。

- 25 19.根据权利要求18所述的装置, 其特征在于, 所述第一存储器获取来自所述第二存储器中的所述第一目标embedding, 包括:

在所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中, 所述第一存储器从所述第二存储器中获取所述第一目标embedding。

- 30 20.根据权利要求18或19所述的装置, 其特征在于, 所述获取单元, 还用于在所述计算核从所述第一存储器中获取第一目标embedding之前, 获取来自所述第二处理器的第一向量以及目标向量, 其中, 所述第一向量用于指示所述第M批batch的模型训练所需的embedding, 所述第一向量中的每个元素用于指示一个embedding, 且不同元素指示不同的embedding, 所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系, 所述每个元素指示的embedding在所述排列关系中出现至少一次;

根据所述第一向量从所述第一存储器中获取所述第M批batch的模型训练时所需的embedding;

根据所述目标向量以及所述第M批batch的模型训练时所需的embedding，确定在进行所述第M批batch的模型训练时所需的每个embedding的的排列关系。

21.根据权利要求15至20任一所述的装置，其特征在于，所述获取单元，还用于：

- 5 在所述计算核根据所述第一目标embedding，对所述待训练模型进行训练之前，获取来自第三处理器中第三存储器的第三目标embedding，所述第三目标embedding和所述第一目标embedding为所述第M批batch模型训练所需的embedding，所述第一处理器和所述第三处理器为相同类型的处理器；

所述模型训练单元，具体用于：

- 10 通过所述待训练模型，对所述第一目标embedding和所述第三目标embedding进行处理，以得到目标输出，并基于所述目标输出确定第一梯度；
 基于所述第一梯度更新所述待训练模型以及所述第一目标embedding。

22.根据权利要求21所述的装置，其特征在于，所述模型训练单元，具体用于：

- 15 基于所述目标输出确定所述第一梯度和第二梯度；
 所述计算核还包括：第二数据传递单元，用于在所述基于所述目标输出确定所述第一梯度和所述第二梯度之后，将所述第二梯度传递至第三处理器的计算核，以便所述第三处理器的计算核根据所述第二梯度更新所述第三目标embedding。

- 20 23.根据权利要求15至22任一所述的装置，其特征在于，所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU，所述第一存储器为缓存，所述第二处理器为中央处理器CPU。

- 25 24.一种模型训练装置，其特征在于，应用于第二处理器，所述第二处理器与第一处理通信连接，所述第一处理器包括计算核以及第一存储器，所述第二处理器包括第二存储器，所述第一处理器和所述第二处理器为不同类型的处理器；所述装置包括：

获取单元，用于获取所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息，所述M为大于1的正整数；

- 30 数据写入单元，用于根据所述计算核在进行第M批batch模型训练所需的embedding信息以及所述第一存储器中存储的embedding信息，将第一目标embedding写入所述第一存储器，所述第一目标embedding为所述第M批batch的模型训练所需的且所述第一存储器未存储的embedding。

- 35 25.根据权利要求24所述的装置，其特征在于，所述数据写入单元，用于在所述计算核对所述待训练模型进行第M-1批batch的模型训练的过程中，将所述第一目标embedding写入所述第一存储器。

26.根据权利要求24或25所述的装置，其特征在于，所述数据写入单元，还用于在将第

— 32 —

一目标embedding写入所述第一存储器之前，在所述第一存储器中可用的存储空间低于存储所述第一目标embedding所需的存储空间时，将所述第一存储器中的至少一个embedding删除，并将所述至少一个embedding写入所述第二存储器，其中，所述至少一个embedding为不属于所述第M批batch的模型训练所需的embedding。

5

27.根据权利要求24至26任一所述的装置，其特征在于，所述获取单元，还用于在所述将所述第一目标embedding写入所述第一存储器之前，获取第一向量以及目标向量，其中，所述第一向量用于指示所述第M批batch的模型训练所需的embedding，所述第一向量中的每个元素用于指示一个embedding，且不同元素指示不同的embedding，所述目标向量用于指示在进行所述第M批batch的模型训练时所述第一向量中的每个元素指示的embedding的排列关系，所述每个元素指示的embedding在所述排列关系中出现至少一次；

10

所述数据写入单元，还用于将所述第一向量以及所述目标向量传递至所述第一处理器。

28.根据权利要求24至27任一所述的装置，其特征在于，所述第一处理器为图形处理器GPU、嵌入式神经网络处理器NPU或张量处理器TPU，所述第一存储器为缓存，所述第二处理器为中央处理器CPU。

15

29.一种模型训练装置，其特征在于，所述装置包括存储器和处理器；所述存储器存储有代码，所述处理器被配置为获取所述代码，并执行如权利要求1至14任一所述的方法。

20

30.一种计算机存储介质，其特征在于，所述计算机存储介质存储有一个或多个指令，所述指令在由一个或多个计算机执行时使得所述一个或多个计算机实施权利要求1至14任一所述的方法。

25

31.一种计算机程序产品，其特征在于，所述计算机程序产品包括代码，当所述代码被执行时，用于实现权利要求1至14任一项所述的方法的步骤。