

数据处理的方法与装置

- 5 本申请要求于2020年03月20日提交中国专利局、申请号为202010202053.7、申请名称为“数据处理的方法与装置”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。

技术领域

- 10 本申请涉及人工智能领域，具体涉及一种数据处理的方法与装置。

背景技术

- 随着互联网技术快速发展，产生信息过载(information overload)的问题。为了解决信息过载的问题，推荐系统(recommender system, RS)应运而生。在推荐系统中，点击率(click-through rate, CTR)预估是非常重要的一个环节，判断一个商品是否进行推荐，需要根据预估的CTR来进行。在进行CTR预估时，除了考虑单个特征外，还需要考虑特征之间的组合。为了表征特征之间的组合，因子分解机(factorization machine, FM)模型被提了出来，FM模型中包括单个特征的所有组合情况的特征组合项。现有技术，通常采用基于FM构建CTR预估模型。
- 15 FM模型中的特征组合项的数量随特征组合的阶数呈指数上升，因此，当特征组合的阶数越来越高，特征组合项的数量是巨大的，导致FM模型的训练的计算量超大。针对这个问题，特征组合筛选(feature interaction selection, FIS)被提出来。由于手工进行特征组合筛选费时费力，业界提出自动特征组合筛选(automatic feature interaction selection, AutoFIS)。
- 20 现有的自动特征组合筛选方案，通过在所有可能的特征组合子集构成的搜索空间中搜索最优的子集，来实现特征组合筛选，搜索过程耗能高，算力消耗大。
- 25 现有的自动特征组合筛选方案，通过在所有可能的特征组合子集构成的搜索空间中搜索最优的子集，来实现特征组合筛选，搜索过程耗能高，算力消耗大。

发明内容

- 30 本申请提供一种数据处理的方法与装置，可以减小特征组合筛选的计算量和算力消耗。

- 第一方面，提供了一种数据处理的方法，该方法包括：为第一模型中的每个特征组合项添加架构参数，得到第二模型，其中，所述第一模型为基于因子分解机FM的模型，所述架构参数用于表征相应特征组合项的重要性；对所述第二模型中的所述架构参数进行优化学习，获得优化后的架构参数；根据所述优化后的架构参数，基于所述第一模型或所述第二模型，获得经过特征组合项删减后的第三模型。
- 35 基于FM的模型表示基于FM原理构建的模型，例如包括如下中任一种模型：FM模型、DeepFM模型、IPNN模型、AFM模型与NFM模型等。

第三模型可以是在第一模型基础上进行特征组合项删减后得到的模型。

或者，第三模型可以是在第二模型基础上进行特征组合项删减后得到的模型。

其中，可以采用多种方式，确定要删减的特征组合项，或者说，确定要保留（即筛选）的特征组合项。

5 可选地，作为一种实现方式，可以将优化后的架构参数中取值小于阈值的架构参数对应的特征组合项删除。

其中，该阈值用于表征是否保留特征组合项的判断准则，例如，若一个特征组合项的优化后的架构参数的取值小于该阈值，则表明，该特征组合项被删除，若一个特征组合项的优化后的架构参数的取值达到该阈值，则表明，该特征组合项被保留（即被筛选）。

10 该阈值可以根据实际应用需求确定。例如，可以通过模型训练的方式，获取该阈值的取值。本申请对该阈值的获取方式不作限定。

可选地，作为另一种实现方式，在一些架构参数在优化结束后取值变为零的情况下，可以直接将优化后取值不为零的架构参数所对应的特征组合项作为保留的特征组合项，得到第三模型。

15 可选地，作为又一种实现方式，在一些架构参数在优化结束后取值变为零的情况下，进一步地，在优化后取值不为零的架构参数所对应的特征组合项中，将取值小于阈值的架构参数对应的特征组合项删除，得到第三模型。

20 在现有的自动特征组合筛选的方案中，将所有可能的特征组合子集当作搜索空间（search space），用离散算法在随机抽取的 n 个候选子集中选出最好的候选子集作为筛选后的特征组合。其中，在评估每个候选子集时，均需要执行一遍训练操作，导致计算量大，消耗算力大。

25 在本申请中，通过在基于 FM 的模型中引入架构参数，通过对架构参数的优化学习，可以实现特征组合项的筛选。也就是说，本申请只需执行架构参数的优化学习，这样的一次模型训练过程，就可以实现特征组合项筛选，而无需现有技术中针对多个候选子集的多遍训练，因此，可以有效减小特征组合筛选的计算量，从而可以节省算力，同时可以提高特征组合筛选的效率。

此外，现有的自动特征组合筛选方案由于计算量大，消耗算力大，导致无法应用于训练时间较长的深度学习模型。

30 在本申请中，只需通过架构参数的优化学习过程，就可实现特征组合筛选，或者说，只需一次端到端的模型训练过程，即可完成特征组合项的筛选，使得特征组合项的筛选（或搜索）的时间可以等价于一次模型训练的时间，因此，可以应用于训练时间较长的深度学习模型。

现有技术的 FM 模型，由于需要枚举所有特征组合情况，导致很难扩展到更高阶。

35 在本申请中，通过在基于 FM 的模型中引入架构参数，并通过架构参数的优化学习，可以实现特征组合筛选，因此，本申请的方案，可以支持基于 FM 的模型中的特征组合项扩展到更高阶。

可以采用多种优化算法（或优化器）对第二模型中的所述架构参数进行优化学习。

结合第一方面，在第一方面的一种可能的实现方式中，所述优化学习使得，所述优化后的架构参数呈现稀疏性。

在本申请中，通过对架构参数的优化学习，使得架构参数呈现稀疏性，有利于后续特

征组合项的删减。

可选地，在所述优化学习使得所述优化后的架构参数呈现稀疏性的实现方式中，所述根据所述优化后的架构参数，基于所述第一模型或所述第二模型，获得经过特征组合项删减后的第三模型，包括：基于所述第一模型或所述第二模型，将所述优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得所述第三模型。

作为一种实现方式，在第一模型中，将所述优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得所述第三模型。

作为另一种实现方式，在第二模型中，将所述优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得所述第三模型。

应理解，在第二模型的基础上经过特征组合项删减，获得第三模型，可以使得第三模型中具有优化后的表征特征组合项的重要性的架构参数，后续通过第三模型的训练，可以进一步学习特征组合项的重要性。

结合第一方面，在第一方面的一种可能的实现方式中，所述优化学习使得，至少一个特征组合项的架构参数的取值在优化结束后等于零。

假设将优化结束后取值为零的架构参数对应的特征组合项视为不重要的特征组合项，则所述优化学习使得，至少一个特征组合项的架构参数的取值在优化结束后等于零，可以视为，可以使得不重要的特征组合项的架构参数在优化结束后等于零。

可选地，所述优化学习，使用 GRDA 优化器对所述第二模型中的所述架构参数进行优化，所述 GRDA 优化器使得至少一个特征组合项的架构参数的取值在优化过程中趋向于零。

在本申请中，通过架构参数的优化学习使得一些架构参数趋向于零，相当于在架构参数的优化过程中就剔除了一些不重要的特征组合项，或者说，通过架构参数的优化学习，既实现了架构参数的优化，又实现了特征组合项的筛选，可以有效提高特征组合筛选的效率，进一步减少计算量和算力消耗。

此外，在架构参数的优化过程中，剔除一些不重要的特征组合项，还可以避免这些不重要的特征组合项所产生的噪声，从而可以使得在架构参数的优化过程中模型就逐渐接近理想模型，这样也可以使得模型中的其他参数（例如，未剔除的特征组合项的架构参数与模型参数）的估计更加准确。

可选地，在所述优化学习使得优化后的架构参数呈现稀疏性，并使得至少一个特征组合项的架构参数的取值在优化结束后等于零的实现方式中，所述根据所述优化后的架构参数，基于所述第一模型或所述第二模型，获得经过特征组合项删减后的第三模型，包括：将优化后的架构参数对应的特征组合项之外的特征组合项删除，获得第三模型。

可选地，在第一模型中，将优化后的架构参数对应的特征组合项之外的特征组合项删除，获得第三模型。即第三模型是在第一模型基础上进行特征组合项删减后得到的。

可选地，将完成架构参数优化后的第二模型作为第三模型。即第三模型是在第二模型基础上进行特征组合项删减后得到的。

应理解，在第二模型的基础上经过特征组合项删减，获得第三模型，可以使得第三模型中具有优化后的表征特征组合项的重要性的架构参数，后续通过第三模型的训练，可以进一步学习特征组合项的重要性。

可选地，在所述优化学习使得优化后的架构参数呈现稀疏性，并使得至少一个特征组合项的架构参数的取值在优化结束后等于零的实现方式中，所述根据所述优化后的架构参数，基于所述第一模型或所述第二模型，获得经过特征组合项删减后的第三模型，包括：将优化后的架构参数对应的特征组合项之外的特征组合项删除，再将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得第三模型。

可选地，第一模型中，将优化后的架构参数对应的特征组合项之外的特征组合项删除，再将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得第三模型。

可选地，在完成架构参数优化后的第二模型中，将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得第三模型。

应理解，在第二模型的基础上经过特征组合项删减，获得第三模型，可以使得第三模型中具有优化后的表征特征组合项的重要性的架构参数，后续通过第三模型的训练，可以进一步学习特征组合项的重要性。

结合第一方面，在第一方面的一种可能的实现方式中，所述方法还包括对所述第二模型中的模型参数进行优化学习，所述优化学习包括，对所述第二模型中的模型参数进行标量化处理。

模型参数表示，第二模型中除了架构参数之外，特征组合项本身的权重参数。或者说，模型参数表示第一模型中原有的参数。

作为一种实现方式，所述优化学习包括，对所述第二模型中的模型参数进行批量归一化（batch normalization，BN）处理。

应理解，通过对特征组合项的模型参数进行标量化处理，可以实现该特征组合项的模型参数与架构参数的去耦合，可以使得架构参数能更准确地体现该特征组合项的重要性，进而可以提高架构参数的优化准确度。

结合第一方面，在第一方面的一种可能的实现方式中，对所述第二模型中的所述架构参数进行优化学习和对所述第二模型中的所述模型参数进行优化学习，包括：使用相同的训练数据，对所述第二模型中的架构参数与模型参数同时进行优化学习，获得优化后的架构参数。

换句话说，在优化过程的每轮训练中，基于同一批训练数据，对架构参数与模型参数同时进行优化学习。

或者说，将第二模型中的架构参数与模型参数视为同一级别的决策变量，对第二模型中的架构参数与模型参数同时进行优化学习，获得优化后的架构参数。

在本申请中，通过对第二模型中的架构参数与模型参数进行一层优化处理，以实现第二模型中的架构参数的优化学习，可以实现架构参数与模型参数的同时优化，从而可以减少第二模型中的架构参数的优化学习过程的耗时，进而有助于提高特征组合项筛选的效率。

结合第一方面，在第一方面的一种可能的实现方式中，所述方法还包括：训练所述第三模型，以得到点击率 CTR 预估模型或转化率 CVR 预估模型。

第二方面，提供了一种数据处理的方法，该方法包括：将目标对象的数据输入点击率 CTR 预估模型或转化率 CVR 预估模型，获得所述目标对象的预估结果；根据所述目标对

象的预估结果，确定所述目标对象的推荐情况。

其中，所述 CTR 预估模型或所述 CVR 预估模型通过第一方面的方法获得。

其中，训练第三模型的步骤包括：使用该目标对象的训练样本，训练第三模型，以得到 CTR 预估模型或 CVR 预估模型。

- 5 可选地，对架构参数进行优化学习的步骤包括：使用所述目标对象的训练样本中相同的训练数据，对所述第二模型中的架构参数与模型参数同时进行优化学习，获得优化后的架构参数。

第三方面，提供一种数据处理的装置，该装置包括如下单元。

- 10 第一处理单元，用于为第一模型中的每个特征组合项添加架构参数，得到第二模型，其中，所述第一模型为基于因子分解机 FM 的模型，所述架构参数用于表征相应特征组合项的重要性。

第二处理单元，用于对所述第二模型中的所述架构参数进行优化学习，获得优化后的架构参数。

- 15 第三处理单元，用于根据所述优化后的架构参数，基于所述第一模型或所述第二模型，获得经过特征组合项删减后的第三模型。

结合第三方面，在第三方面的一种可能的实现方式中，所述第二处理单元对架构参数进行的优化学习使得，所述优化后的架构参数呈现稀疏性。

- 20 结合第三方面，在第三方面的一种可能的实现方式中，所述第三处理单元用于，基于所述第一模型或所述第二模型，将所述优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得所述第三模型。

结合第三方面，在第三方面的一种可能的实现方式中，所述第二处理单元对架构参数进行的优化学习使得，至少一个特征组合项的架构参数的取值在优化结束后等于零。

- 25 结合第三方面，在第三方面的一种可能的实现方式中，所述第三处理单元用于，使用广义正规对偶平均 GRDA 优化器对所述第二模型中的所述架构参数进行优化，所述 GRDA 优化器使得至少一个特征组合项的架构参数的取值在优化过程中趋向于零。

结合第三方面，在第三方面的一种可能的实现方式中，所述第二处理单元还用于，对所述第二模型中的模型参数进行优化学习，所述优化学习包括，对所述第二模型中的模型参数进行标量化处理。

- 30 结合第三方面，在第三方面的一种可能的实现方式中，所述第二处理单元用于，对所述第二模型中的模型参数进行批量归一化 BN 处理。

结合第三方面，在第三方面的一种可能的实现方式中，第二处理单元用于，使用相同的训练数据，对所述第二模型中的架构参数与模型参数同时进行优化学习，获得优化后的架构参数。

- 35 结合第三方面，在第三方面的一种可能的实现方式中，所述装置还包括：训练单元用于，训练所述第三模型，以得到 CTR 预估模型或 CVR 预估模型。

第四方面，提供一种数据处理的装置，该装置包括如下单元。

第一处理单元用于，将目标对象的数据输入点击率 CTR 预估模型或转化率 CVR 预估模型，获得所述目标对象的预估结果。

第一处理单元用于，根据所述目标对象的预估结果，确定所述目标对象的推荐情况。

其中，所述 CTR 预估模型或所述 CVR 预估模型通过第一方面的方法获得。

其中，训练第三模型的步骤包括：使用该目标对象的训练样本，训练第三模型，以得到 CTR 预估模型或 CVR 预估模型。

5 可选地，对架构参数进行优化学习的步骤包括：使用所述目标对象的训练样本中相同的训练数据，对所述第二模型中的架构参数与模型参数同时进行优化学习，获得优化后的架构参数。

第五方面，提供一种图像处理的装置，该装置包括：存储器，用于存储程序；处理器，用于执行存储器存储的程序，当存储器存储的程序被执行时，处理器用于执行上述第一方面或第二方面中的方法。

10 第六方面，提供一种计算机可读介质，该计算机可读介质存储用于设备执行的程序代码，该程序代码包括用于执行上述第一方面或第二方面中的方法。

第七方面，提供一种包含指令的计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机执行上述第一方面或第二方面中的方法。

15 第八方面，提供一种芯片，所述芯片包括处理器与数据接口，所述处理器通过所述数据接口读取存储器上存储的指令，执行上述第一方面或第二方面中的方法。

可选地，作为一种实现方式，所述芯片还可以包括存储器，所述存储器中存储有指令，所述处理器用于执行所述存储器上存储的指令，当所述指令被执行时，所述处理器用于执行上述第一方面或第二方面中的方法。

20 第九方面，提供一种电子设备，该电子设备包括上述第三方面、第四方面、第五方面或第六方面提供的装置。

基于上文描述可知，在本申请中通过在基于 FM 的模型中引入架构参数，并通过架构参数的优化学习，可以实现特征组合项的筛选。也就是说，本申请通过架构参数的优化学习可以实现特征组合项的筛选，而无需现有技术中针对多个候选子集的多遍训练，因此，可以有效减小特征组合筛选的计算量，从而可以节省算力，同时可以提高特征组合筛选的效率。

25 此外，采用本申请提供的方案，可以支持基于 FM 的模型中的特征组合项扩展到更高阶。

附图说明

30 图 1 是 FM 模型的架构示意图。

图 2 是 FM 模型的训练示意图。

图 3 是本申请实施例提供的数据处理的方法的示意性流程图。

图 4 是本申请实施例中的 FM 模型的架构示意图。

图 5 是本申请实施例提供的数据处理的方法的另一示意性流程图。

35 图 6 是本申请实施例提供的数据处理的方法的又一示意性流程图。

图 7 是本申请实施例提供的数据处理的装置的示意性框图。

图 8 是本申请实施例提供的数据处理的装置的另一示意性框图。

图 9 是本申请实施例提供的数据处理的装置的又一示意性框图。

图 10 是本申请实施例提供的一种芯片硬件结构示意图。

具体实施方式

下面将结合附图，对本申请中的技术方案进行描述。

随着当今技术的飞速发展，数据量与日俱增，为了解决信息过载(information overload)的问题，推荐系统(recommender system, RS)被提了出来。推荐系统通过将用户的历史行为或者用户的兴趣偏好或者用户的人口统计学特征送给推荐算法，然后推荐系统运用推荐算法来产生用户可能感兴趣的项目列表。

在推荐系统中，点击率(click-through rate, CTR)预估（或者还包括转化率（conversion rate, CVR）预估）是非常重要的一个环节。判断一个商品的是否进行推荐需要根据预估的点击率来进行。在进行 CTR 预估时，除了考虑单特征外，往往还要考虑特征之间的组合。特征组合对于推荐排序非常重要，因子分解机（factorization machine, FM）可以体现特征之间的组合。因子分解机可以简称为 FM 模型。

根据特征组合项的最大阶数，FM 模型可以称为*阶 FM 模型。例如，特征组合项的阶数最大为二阶的 FM 模型可以称为二阶 FM 模型，特征组合项的阶数最大为三阶的 FM 模型可以称为三阶 FM 模型。

特征组合项的阶数表征特征组合项是几个特征的交互项。例如，2 个特征的交互项可以称为二阶特征组合项，3 个特征的交互项可以称为三阶特征组合项。

作为一个示例，二阶 FM 模型如下列公式（1）所示。

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^m w_i x_i + \sum_{i=1}^m \sum_{j=i+1}^m \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j \quad (1)$$

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle := \sum_{f=1}^k v_{i,f} \cdot v_{j,f}$$

$\mathbf{x}$ 表示特征向量， x_i 表示第 i 个特征， x_j 表示第 j 个特征。 m 表示特征维度，也可以称为特征的域（field）。 w_0 表示全局偏置， $w_0 \in \mathbb{R}$ 。 w_i 表示第 i 个特征的强度， $\mathbf{w} \in \mathbb{R}^m$ 。 $\mathbf{v}_i$ 表示第 i 个特征 x_i 的辅助向量， $\mathbf{v}_j$ 表示第 j 个特征 x_j 的辅助向量。 k 为 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 的维度。二维矩阵 $\mathbf{v} \in \mathbb{R}^{m \times k}$ 。

其中， $x_i x_j$ 表示第 i 个特征 x_i 与第 j 个特征 x_j 的组合。

$\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ 表示 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 的内积，表示第 i 个特征 x_i 与第 j 个特征 x_j 之间的相互作用。 $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ 也可以理解为特征组合项 $x_i x_j$ 的权重参数，例如， $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ 可以记为 w_{ij} 。

在本文中，将 $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ 记为特征组合项 $x_i x_j$ 的权重参数。

可选地，公式（1）还可以表达为下列公式（2）：

$$l_{fm} = \langle \mathbf{w}, \mathbf{x} \rangle + \sum_{i=1}^m \sum_{j>i}^m \langle \mathbf{e}_i, \mathbf{e}_j \rangle \quad (2)$$

其中，公式（2）中的 $\langle \mathbf{e}_i, \mathbf{e}_j \rangle$ 表示公式（1）中的 $\langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$ ，公式（2）中的 $\langle \mathbf{w}, \mathbf{x} \rangle$ 表示公式（1）中的 $w_0 + \sum_{i=1}^m w_i x_i$ 。

作为另一个示例，三阶 FM 模型如下列公式（3）所示。

$$l_{jm}^{3rd} = \langle w, x \rangle + \sum_{i=1}^m \sum_{j>i}^m \langle e_i, e_j \rangle + \sum_{i=1}^m \sum_{j>i}^m \sum_{t>j}^m \langle e_i, e_j, e_t \rangle \quad (3)。$$

FM 模型中包括单个特征的所有组合情况的特征组合项。例如，在公式(1)或公式(2)所示的二阶 FM 模型中，包括单个特征的所有二阶特征组合情况的组合特征项。又例如，在公式(3)所示的三阶 FM 模型中，包括单个特征的所有二阶特征组合情况的组合特征项，以及所有三阶特征组合情况的组合特征项。

例如，业界将获取特征 x_i 的辅助向量 $\mathbf{v}_i$ 的操作称为嵌入 (embedding)，将根据特征 x_i 及其辅助向量 $\mathbf{v}_i$ 构造特征交互项的操作称为交互 (interaction)。图 1 为 FM 模型的架构示意图。如图 1 所示，FM 模型可以视为神经网络模型，包括输入 (input) 层、嵌入 (embedding) 层、交互 (interaction) 层与输出 (output) 层。输入层用于生成特征，域 1、域、...、域 m 表示特征维度。嵌入层用于产生特征的辅助向量。交互层用于根据特征及其辅助向量生成特征组合项。输出层用于输出 FM 模型的输出。

现有技术，通常基于 FM 进行 CTR 预估或 CVR 预估。

当前技术中，基于 FM 的模型包括：FM 模型、DeepFM 模型、IPNN 模型、AFM 模型与 NFM 模型等。

作为示例而非限定，建立 FM 模型的流程如图 2 所示。

S210，枚举特征组合项进入 FM 模型。

例如，使用公式(1)或公式(3)构建 FM 模型。

S220，训练 FM 模型直至收敛，即获得可以投入应用的 FM 模型。

完成 FM 模型的训练后，可以使用训练好的 FM 模型进行上线推理，如图 2 所示的步骤 S230。

如前文描述，FM 模型中包括单个特征的所有组合情况的特征组合项。因此，FM 模型的训练的计算量与耗时都是巨大的。

此外，从公式(1)与公式(3)可以看出，FM 模型中的特征组合项的数量随特征维度以及特征组合的阶数的增加而剧增。

例如，在公式(1)中，当特征维度 m 增大，特征组合项的数量呈指数增大。再例如，从二阶 FM 模型到三阶 FM 模型，随着特征组合的阶数的增大，FM 模型中特征组合项的数量的增加是巨大的。

因此，特征维度以及特征组合的阶数的增大为 FM 模型的推理时延与计算量带来巨大负担，因此，限制了 FM 模型能够容纳的最大特征维度以及特征组合的阶数。例如，当前技术的 FM 模型很难扩展到更高阶。

针对这个问题，特征组合筛选 (feature interaction selection, FIS) 被提出来。

一些现有技术采用手工筛选的方式进行特征组合筛选。筛选好的特征组合可能需要花费很多工程师若干年的探索。这种手工筛选的方式不但消耗巨大的人力，而且可能会漏掉重要的特征组合项。

针对手工筛选的弊端，业界提出自动特征组合筛选 (automatic feature interaction selection, AutoFIS) 的方案。相对于手工筛选，自动特征组合筛选可以在短时间内筛选到有价值的特征组合。

当前技术中，提出一种自动特征组合筛选的方案。该方案将所有可能的特征组合子集

当作搜索空间 (search space)，用离散算法在随机抽取的 n 个候选子集中选出最好的候选子集作为筛选后的特征组合。其中，在评估每个候选子集时，均需要执行一遍训练操作，因此，该方案的计算量较大，所需消耗的算力较大。此外，在评估每个候选子集时，若完整训练整个模型，可以提高评价精确度，但是会导致搜索能耗 (search cost) 巨大；若用小批量 (mini-batches) 训练做近似，则会导致评价不精确。此外，在该方案中，随着特征组合的阶数的增大，搜索空间呈指数级别增长，会加剧搜索过程能耗。

因此，现有的自动特征组合筛选的方案，计算量大，算力消耗大，搜索过程耗能高。

针对上述问题，本申请提出一种自动特征组合筛选的方案，相对于现有技术，可以降低自动特征组合筛选的算力消耗，提高自动特征组合筛选的效率。

图 3 为本申请实施例提供的数据处理的方法 300 的示意性流程图，该方法 300 包括如下步骤 S310、S320 与 S330。

S310，为第一模型中的每个特征组合项添加架构参数，得到第二模型。

其中，第一模型为基于因子分解机 (factorization machine, FM) 的模型，也就是说，第一模型中包括单个特征的所有组合情况的特征组合项，或者说，第一模型中枚举了所有组合情况的特征组合项。

例如，第一模型可以为如下基于 FM 的模型中的任一种：FM 模型、DeepFM 模型、IPNN 模型、AFM 模型与 NFM 模型。

作为示例，第一模型为如公式 (1) 或公式 (2) 所示的二阶 FM 模型，或者第一模型为如公式 (3) 所示的三阶 FM 模型。

本申请是要实现特征组合项的筛选，第一模型可以视为是待进行特征组合项删减的模型。

为第一模型中的每个特征组合项添加架构参数，指的是，为第一模型中的每个特征组合项添加一个系数，本申请中，将该系数称为架构参数。该架构参数用于表征相应特征组合项的重要性。为第一模型中的每个特征组合项添加架构参数后得到的模型记为第二模型。

作为一个示例，假设第一模型为如公式 (1) 所示的二阶 FM 模型，则第二模型可以如下列公式 (4) 所示。

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^m w_i x_i + \sum_{i=1}^m \sum_{j=i+1}^m \alpha_{(i,j)} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j \quad (4)$$

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle := \sum_{f=1}^k v_{i,f} \cdot v_{j,f}$$

$\mathbf{x}$ 表示特征向量， x_i 表示第 i 个特征， x_j 表示第 j 个特征。 m 表示特征维度，也可以称为特征的域 (field)。 w_0 表示全局偏置， $w_0 \in \mathbb{R}$ 。 w_i 表示第 i 个特征的强度， $\mathbf{w} \in \mathbb{R}^m$ 。 $\mathbf{v}_i$ 表示第 i 个特征 x_i 的辅助向量， $\mathbf{v}_j$ 表示第 j 个特征 x_j 的辅助向量。 k 为 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 的维度。二维矩阵 $\mathbf{v} \in \mathbb{R}^{m \times k}$ 。

其中， $x_i x_j$ 表示第 i 个特征 x_i 与第 j 个特征 x_j 的组合。

$\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ 表示特征组合项 $x_i x_j$ 的权重参数， $\alpha_{(i,j)}$ 表示特征组合项 $x_i x_j$ 的架构参数。

$\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ 表示 $\mathbf{v}_i$ 与 $\mathbf{v}_j$ 的内积，表示第 i 个特征 x_i 与第 j 个特征 x_j 之间的相互作用。

$\langle v_i, v_j \rangle$ 也可以理解为特征组合项 $x_i x_j$ 的权重参数, 例如, $\langle v_i, v_j \rangle$ 可以记为 w_{ij} 。

假设第一模型表达为公式 (2) 所示的二阶 FM 模型, 则第二模型可以表达为下列公式 (5)。

$$l_{autoFIS} = \langle w, x \rangle + \sum_{i=1}^m \sum_{j>i}^m \alpha_{(i,j)} \langle e_i, e_j \rangle \quad (5)$$

5 其中, $\alpha_{(i,j)}$ 表示特征组合项的架构参数。

作为另一个示例, 假设第一模型为如公式 (3) 所示的三阶 FM 模型, 则第二模型可以如下列公式 (6) 所示。

$$l_{autoFIS} = \langle w, x \rangle + \sum_{i=1}^m \sum_{j>i}^m \alpha_{(i,j)} \langle e_i, e_j \rangle + \sum_{i=1}^m \sum_{j>i}^m \sum_{t>j}^m \alpha_{(i,j,t)} \langle e_i, e_j, e_t \rangle \quad (6)$$

其中, $\alpha_{(i,j)}$ 与 $\alpha_{(i,j,t)}$ 表示特征组合项的架构参数。

10 为了便于理解与描述, 在本文中做如下约定, 将第一模型中特征组合项原有的权重参数 (如公式 (4) 中的 $\langle v_i, v_j \rangle$) 称为模型参数。

也就是说, 在第二模型中, 每个特征组合项具有两类系数参数: 模型参数与架构参数。

图 4 为本申请实施例提供的特征组合项筛选的示意图。图 4 中的嵌入层与交互层的含义与图 1 中的嵌入层与交互层的含义相同。如图 4 所示, 在本申请实施例中, 为交互层中的每个特征组合项添加架构参数 $\alpha_{(i,j)}$ (如图 4 中所示的 $\alpha_{(1,2)}$ 、 $\alpha_{(1,m)}$ 、 $\alpha_{(m-1,m)}$)。可以
15 将图 4 中交互层视为第一模型, 为特征组合项添加了架构参数 $\alpha_{(i,j)}$ 后的交互层可以视为第二模型。

S320, 对第二模型中的架构参数进行优化学习, 获得优化后的架构参数。

例如, 使用训练数据, 对第二模型中的架构参数进行优化学习, 获得优化后的架构参
20 数。

例如, 优化后的架构参数可以视为是得第二模型中架构参数的最优值 α^* 。

在本申请实施例中, 架构参数表征相应特征组合项的重要性, 因此, 对架构参数的优化学习, 相当于是, 学习每个特征组合项的重要性或者每个特征组合项对模型预测的贡献度。也就是说, 优化后的架构参数表征了学习到的特征组合项的重要性。

25 换句话说, 在本申请实施例中, 每个特征组合项的贡献 (或者重要性) 可以采用端到端的方式由架构参数来学习。

S330, 根据优化后的架构参数, 基于第一模型或第二模型, 获得经过特征组合项删减后的第三模型。

第三模型可以是在第一模型基础上进行特征组合项删减后得到的模型。

30 或者, 第三模型可以是在第二模型基础上进行特征组合项删减后得到的模型。

其中, 可以采用多种方式, 确定要删减的特征组合项, 或者说, 确定要保留 (即筛选) 的特征组合项。

可选地, 作为一种实现方式, 可以将优化后的架构参数中取值小于阈值的架构参数对应的特征组合项删除。

35 其中, 该阈值用于表征是否保留特征组合项的判断准则, 例如, 若一个特征组合项的优化后的架构参数的取值小于该阈值, 则表明, 该特征组合项被删除, 若一个特征组合项的优化后的架构参数的取值达到该阈值, 则表明, 该特征组合项被保留 (即被筛选)。

该阈值可以根据实际应用需求确定。例如，可以通过模型训练的方式，获取该阈值的取值。本申请对该阈值的获取方式不作限定。

继续参见图 4，假设优化后的架构参数 $\alpha_{(1,2)}$ 小于阈值，则可以将架构参数 $\alpha_{(1,2)}$ 对应的特征组合项删除，假设优化后的架构参数 $\alpha_{(1,m)}$ 达到阈值，则可以将架构参数 $\alpha_{(1,m)}$ 对应的特征组合项保留。根据优化后的架构参数进行特征组合项删减之后，获得下一层模型，如图 4 所示。图 3 实施例中的第三模型例如为图 4 中所示的下一层模型。

作为示例而非限定，如图 4 所示，可以将根据优化后的架构参数决定相应的特征组合项是否删除的操作，记为选择门。

需要说明的是，图 4 仅为示例而非限定。

可选地，作为另一种实现方式，在一些架构参数在优化结束后取值变为零的情况下，可以直接将优化后取值不为零的架构参数所对应的特征组合项作为保留的特征组合项，得到第三模型。

可选地，作为又一种实现方式，在一些架构参数在优化结束后取值变为零的情况下，进一步地，在优化后取值不为零的架构参数所对应的特征组合项中，将取值小于阈值的架构参数对应的特征组合项删除，得到第三模型。

本文中的表述“经过特征组合项的删减后的模型”可以替换为“经过特征组合项的筛选后的模型”。

如前文描述，在现有的自动特征组合筛选的方案中，将所有可能的特征组合子集当作搜索空间（search space），用离散算法在随机抽取的 n 个候选子集中选出最好的候选子集作为筛选后的特征组合。其中，在评估每个候选子集时，均需要执行一遍训练操作，导致计算量大，消耗算力大。

在本申请实施例中，通过在基于 FM 的模型中引入架构参数，通过对架构参数的优化学习，可以实现特征组合项的筛选。也就是说，本申请只需执行架构参数的优化学习，这样的一次模型训练过程，就可以实现特征组合项筛选，而无需现有技术中针对多个候选子集的多遍训练，因此，可以有效减小特征组合筛选的计算量，从而可以节省算力，同时可以提高特征组合筛选的效率。

此外，在现有的自动特征组合筛选方案中，是通过从搜索空间（search space）中搜索候选子集，来实现特征组合筛选。可以理解到，现有技术将特征组合筛选作为离散问题来解决的，即搜索离散的特征组合候选集。

在本申请实施例中，是通过对引入基于 FM 的模型中的架构参数的优化学习，来实现特征组合筛选的。可以理解到，本申请实施例是将现有的搜索离散的特征组合候选集的问题连续化，或者说，将特征组合筛选作为一个连续问题来解决。例如，本申请提供的自动特征组合筛选方案可以表述为，基于连续搜索空间的组合特征搜索方案。换句话说，在本申请实施例中，在基于 FM 的模型中引入架构参数的这一操作，可以视为，是对特征组合项自动筛选这一问题进行的连续建模。

此外，现有的自动特征组合筛选方案由于计算量大，消耗算力大，导致无法应用于训练时间较长的深度学习模型。

在本申请实施例中，只需通过架构参数的优化学习过程，就可实现特征组合筛选，或者说，只需一次端到端的模型训练过程，即可完成特征组合项的筛选，使得特征组合项的

筛选（或搜索）的时间可以等价于一次模型训练的时间，因此，可以应用于训练时间较长的深度学习模型。

如前文描述，现有技术的 FM 模型，由于需要枚举所有特征组合情况，导致很难扩展到更高阶。

- 5 在本申请实施例中，通过在基于 FM 的模型中引入架构参数，并通过架构参数的优化学习，可以实现特征组合筛选，因此，本申请实施例提供的方案，可以支持基于 FM 的模型中的特征组合项扩展到更高阶。

例如，使用本申请实施例提供的方案构造的 FM 模型可以扩展到三阶或更高阶。

- 10 再例如，使用本申请实施例提供的方案构造的 DeepFM 模型可以扩展到三阶或更高阶。

- 15 在本申请实施例中，通过在传统的基于 FM 的模型中引入架构参数，并通过对架构参数的优化学习，可以实现特征组合筛选。或者，换句话说，本申请实施例通过构建一种包括架构参数的基于 FM 的模型，通过对架构参数进行优化学习，可以实现特征组合筛选。其中，构建这种包括架构参数的基于 FM 的模型的方法为，在传统基于 FM 的模型中的每个特征组合项前面添加架构参数。

如图 3 所示，该方法 300 还可以包括步骤 S340。

S340，训练所述第三模型。

- 20 步骤 S340 也可以理解为，重新进行模型训练。可以理解到，通过步骤 S310、步骤 S320 与步骤 S330，实现了特征组合项的删减。在步骤 S340 中，对经过特征组合项的删减后的模型，重新进行训练。

在步骤 S340，可以直接对第三模型进行训练；或者，也可以是在第三模型上增加 L1 正则项，和/或 L2 正则项之后，再进行训练。

例如，可以根据应用需求，确定训练第三模型的训练目的。

- 25 例如，假设要获得点击率(click-through rate, CTR)预估模型，则以获得 CTR 预估模型为训练目标，训练第三模型，以得到 CTR 预估模型。

再例如，例如，假设要获得转化率（conversion rate, CVR）预估模型，则以获得 CVR 预估模型为训练目标，训练第三模型，以得到 CVR 预估模型。

在步骤 S320 中，例如，可以采用多种优化算法（或优化器）对第二模型中的架构参数进行优化学习。

- 30 第一种优化算法。

可选地，在步骤 S320 中，对架构参数的优化学习，使得优化后的架构参数呈现稀疏性。

- 35 例如，在步骤 S320 中，采用套锁算子（least absolute shrinkage and selection operator, Lasso）正则化（Regularization）对第二模型中的架构参数进行优化学习。Lasso 也可以称为最小绝对值收敛和选择算子。

以第二模型为公式（5）为例，在步骤 S320 中，可以采用如下公式（7）对第二模型中的架构参数进行优化学习：

$$L_{search} = L_{\alpha, w}(y, \hat{y}_M) + \lambda \sum_{i, j > i} |\alpha_{(i, j)}| \quad (7)$$

其中, $L_{\alpha,w}(y, \hat{y}_M)$ 表示损失函数, y 表示模型观测值, $\hat{y}_M$ 表示模型预测值。 λ 表示常数, 可以根据具体需求赋值。

应理解, 公式 (7) 表示架构参数优化的约束条件。

通过使得优化后的架构参数呈现稀疏性, 有利于后续特征组合项的删减。

- 5 可选地, 作为一种实施例, 在步骤 S320 中, 对架构参数的优化学习, 使得优化后的架构参数呈现稀疏性; 在步骤 S330 中, 基于第一模型或第二模型, 将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减, 获得第三模型。

其中, 该阈值用于表征是否保留特征组合项的判断准则, 例如, 若一个特征组合项的优化后的架构参数的取值小于该阈值, 则表明, 该特征组合项被删除, 若一个特征组合项的优化后的架构参数的取值达到该阈值, 则表明, 该特征组合项被保留 (即被筛选)。

该阈值可以根据实际应用需求确定。例如, 可以通过模型训练的方式, 获取该阈值的取值。本申请对该阈值的获取方式不作限定。

在本申请实施例中, 通过对架构参数的优化学习, 使得架构参数呈现稀疏性, 有利于特征组合项的筛选。

- 15 应理解, 第二模型中的架构参数表征相应特征组合性的重要性或贡献度, 若优化后的某个架构参数的取值小于阈值, 例如接近于零, 则表明该架构参数所对应的特征组合项不重要或者贡献度很小, 将这样的特征组合项删除 (或称为剔除或剪掉), 不但可以去除其引入的噪声, 还可以减小能耗, 也可以提高模型的推理速度。

因此, 将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减, 是合理地的特征组合筛选操作。

第二种优化算法。

可选地, 在步骤 S320 中, 对架构参数的优化学习, 使得优化后的架构参数呈现稀疏性, 并且使得至少一个特征组合项的架构参数的取值在优化结束后等于零。

- 25 假设将优化结束后取值为零的架构参数对应的特征组合项视为不重要的特征组合项, 则步骤 S320 中对架构参数的优化学习, 可以视为, 可以使得不重要的特征组合项的架构参数在优化结束后等于零。

换句话说, 对架构参数的优化学习使得, 至少一个特征组合项的架构参数的取值在优化过程中趋向于零。

- 30 例如, 在步骤 S320 中, 使用广义正规对偶平均 (generalized regularized dual averaging, GRDA) 优化器对第二模型中的架构参数进行优化。GRDA 优化器不仅可以实现架构参数的稀疏性, 还可以使得至少一个特征组合项的架构参数的取值在优化过程中逐渐趋向于零。

作为示例, 在步骤 S320 中, 采用如下公式 (8) 对第二模型中的架构参数进行优化学习:

$$35 \quad \alpha_{i+1} = \arg \min_{\alpha} \left\{ \alpha^T \left(-\alpha_0 + \gamma \sum_{i=0}^i \nabla L(\alpha_i; y_{i+1}) + g(t, \gamma) \|\alpha\|_1 + \frac{1}{2} \|\alpha\|_2^2 \right) \right\} \quad (8)$$

其中, γ 表示学习率。 y_{i+1} 表示模型观测值。 $g(t, \gamma) = c\gamma^{1/2}(t\gamma)^\mu$, c 与 μ 表示可调整的超参数, 调整 c 与 μ 的目标是在模型准确度与架构参数 α 的稀疏性之间找到平衡。

应理解, 公式 (8) 表示架构参数优化的约束条件。

还应理解，在本实施例中，在步骤 S320 中，完成架构参数优化后的第二模型已经为经过特征组合项筛选的模型。

在本申请实施例中，通过架构参数的优化学习使得一些架构参数趋向于零，相当于在架构参数的优化过程中就剔除了一些不重要的特征组合项，或者说，通过架构参数的优化学习，既实现了架构参数的优化，又实现了特征组合项的筛选，可以有效提高特征组合筛选的效率，进一步减少计算量和算力消耗。

此外，在架构参数的优化过程中，剔除一些不重要的特征组合项，还可以避免这些不重要的特征组合项所产生的噪声，从而可以使得在架构参数的优化过程中模型就逐渐接近理想模型，这样也可以使得模型中的其他参数（例如，未剔除的特征组合项的架构参数与模型参数）的估计更加准确。

可选地，作为一种实施例，在步骤 S320 中，对架构参数的优化学习，使得优化后的架构参数呈现稀疏性，并且使得至少一个特征组合项的架构参数的取值在优化结束后等于零；在步骤 S330 中，可以通过如下多种方式，获得第三模型。

方式（1）

在步骤 S330 中，可以直接将优化后的架构参数所对应的特征组合项作为筛选的特征组合项，基于所筛选的特征组合项获得第三模型。

例如，在第一模型中，将优化后的架构参数所对应的特征组合项作为筛选的特征组合项，将其余特征组合项删除，获得第三模型。

再例如，直接经完成架构参数优化的第二模型作为第三模型。

方式（2）

在步骤 S330 中，在优化后的架构参数所对应的特征组合项中，将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得第三模型。

该阈值可以根据实际应用需求确定。例如，可以通过模型训练的方式，获取该阈值的取值。本申请对该阈值的获取方式不作限定。

例如，在第一模型中，将优化后的架构参数对应的特征组合项之外的特征组合项删除，再将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得第三模型。

再例如，在完成架构参数优化后的第二模型中，将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得第三模型。

在本申请实施例中，通过架构参数的优化学习使得一些架构参数趋向于零，相当于在架构参数的优化过程中就剔除了一些不重要的特征组合项，或者说，通过架构参数的优化学习，既实现了架构参数的优化，又实现了特征组合项的筛选，可以有效提高特征组合筛选的效率，进一步减少计算量和算力消耗。

基于上述对步骤 S320 的描述可知，步骤 S330 中，获得经过特征组合项筛选后的第三模型的实现方式，可以根据步骤 S320 中对架构参数的优化方式而确定。下面描述在如下两种情况下，获得第三模型的实现方式。

第一种情况，在步骤 S320 中，对架构参数的优化学习，使得优化后的架构参数呈现稀疏性。

在步骤 S330 中，将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合

项删减, 获取第三模型。该阈值如参见上文描述, 这里不再赘述。

作为示例而非限定, 将架构参数的优化学习完成 (即优选训练收敛) 后得到优化后的架构参数记为架构参数的最优值 α^* 。根据最优值 α^* , 可以确定保留或删除哪些特征组合项。例如, 若一个特征交互项的架构参数的最优值 $\alpha^{*(i,j)}$ 达到阈值, 则该特征交互项应该

5 保留; 若一个特征交互项的架构参数的最优值 $\alpha^{*(i,j)}$ 小于阈值, 则该特征交互项应该删掉。

作为示例, 在第二模型中, 对于每个特征交互项, 设置一个表征该特征交互项是否保留在模型中的开关项 $\psi^{(i,j)}$, 第二模型可以如下公式(9)所示:

$$\hat{y}(\mathbf{x}) := w_0 + \sum_{i=1}^m w_i x_i + \sum_{i=1}^m \sum_{j=i+1}^m \alpha_{(i,j)} \psi_{(i,j)} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j \quad (9)$$

其中, 开关项 $\psi^{(i,j)}$ 的取值可以通过如下公式(10)表示:

$$\psi_{(i,j)} = \begin{cases} 1 & |\alpha^{*(i,j)}| \geq thr \\ 0 & |\alpha^{*(i,j)}| < thr \end{cases} \quad (10)$$

其中, thr 表示阈值。

将开关项 $\psi^{(i,j)}$ 为 0 的特征组合项从第二模型中删掉, 即可得到经过特征组合项筛选后的第三模型。

在本实施例中, 开关项 $\psi^{(i,j)}$ 的设置可以视为是否保留特征组合项的判断标准。

15 可选地, 第三模型可以是基于第一模型经过特征组合项删减后得到的模型。

例如, 将开关项 $\psi^{(i,j)}$ 为 0 的特征组合项从第一模型中删掉, 得到经过特征组合项筛选后的第三模型。

可选地, 第三模型可以是基于第二模型经过特征组合项删减后得到的模型。

20 例如, 将开关项 $\psi^{(i,j)}$ 为 0 的特征组合项从第二模型中删掉, 得到经过特征组合项筛选后的第三模型。

应理解, 在本实施例中, 第三模型中具有优化后的表征特征组合项的重要性的架构参数, 后续通过第三模型的训练, 可以进一步学习特征组合项的重要性。

第二种情况, 在步骤 S320 中, 对架构参数的优化学习, 使得优化后的架构参数呈现稀疏性, 并且使得至少一个特征组合项的架构参数的取值在优化结束后等于零。

25 可选地, 在步骤 S330 中, 将优化后的架构参数对应的特征组合项之外的特征组合项删除, 获得第三模型。

作为一个示例, 在步骤 S330 中, 在第一模型中, 将优化后的架构参数对应的特征组合项之外的特征组合项删除, 获得第三模型。即第三模型是在第一模型基础上进行特征组合项删减后得到的。

30 作为另一个示例, 在步骤 S330 中, 将完成架构参数优化后的第二模型作为第三模型。即第三模型是在第二模型基础上进行特征组合项删减后得到的。

可选地, 在步骤 S330 中, 将优化后的架构参数对应的特征组合项之外的特征组合项删除, 再将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减, 获得

第三模型。

作为一个示例，在步骤 S330 中，在第一模型中，将优化后的架构参数对应的特征组合项之外的特征组合项删除，再将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得第三模型。即第三模型是在第一模型基础上进行特征组合项删减后得到的。

作为另一个示例，在步骤 S330 中，在完成架构参数优化后的第二模型中，将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减，获得第三模型。即第三模型是在第二模型基础上进行特征组合项删减后得到的。

应理解，在第三模型是在第二模型基础上进行特征组合项删减后得到的实施例中，第三模型中具有优化后的表征特征组合项的重要性的架构参数，后续通第三模型的训练，可以进一步学习特征组合项的重要性。

从公式 (4)、公式 (5) 或公式 (6) 可以理解到，第二模型中包括两类参数：架构参数与模型参数。模型参数表示，第二模型中除了架构参数之外，特征组合项本身的权重参数。例如，在第二模型如公式 (4) 的例子中， $\alpha_{(i,j)}$ 表示特征组合项的架构参数为， $\langle \mathbf{v}_i, \mathbf{v}_j \rangle$ 表示特征组合项的模型参数。例如，在第二模型如公式 (5) 的例子中， $\alpha_{(i,j)}$ 表示特征组合项的架构参数为， $\langle e_i, e_j \rangle$ 可以表示特征组合项的模型参数。

可以理解到，在对架构参数的优化学习过程中，不仅涉及架构参数的训练，还涉及模型参数的训练。换句话说，在步骤 S320 中对第二模型中的架构参数进行优化学习的过程中，伴随着对第二模型中的模型参数的优化学习。

例如，在图 3 所示实施例中，该方法 300 还包括：对第二模型中的模型参数的优化学习，该优化学习包括对模型参数的标量化处理。

在模型参数的优化学习过程中的每轮训练中，均对第二模型中的模型参数进行标量化处理。

例如，通过对第二模型中的模型参数进行批量归一化 (batch normalization, BN)，实现对第二模型中的模型参数进行标量化处理。

作为示例，以第二模型如公式 (5) 所示为例，可以采用如下公式 (11) 对第二模型中的模型参数进行标量化处理：

$$\langle e_i, e_j \rangle_{BN} = \frac{\langle e_i, e_j \rangle_B - \mu_B(\langle e_i, e_j \rangle_B)}{\sqrt{\sigma_B^2(\langle e_i, e_j \rangle_B) + \theta}} \quad (11)$$

其中， $\langle e_i, e_j \rangle_{BN}$ 表示 $\langle e_i, e_j \rangle$ 的批量归一化。

$\langle e_i, e_j \rangle_B$ 表示 $\langle e_i, e_j \rangle$ 的小批量 (min batch) 数据。

$\mu_B(\langle e_i, e_j \rangle_B)$ 表示 $\langle e_i, e_j \rangle$ 的小批量 (min batch) 数据的平均值。

$\sigma_B^2(\langle e_i, e_j \rangle_B)$ 表示 $\langle e_i, e_j \rangle$ 的小批量 (min batch) 数据的方差。

θ 表示扰动。

继续参见图 4，图 4 中所示的批量归一化 (BN)，表示对第二模型中的模型参数进行

BN 处理。

通过对特征组合项的模型参数进行标量化处理,可以实现该特征组合项的模型参数与架构参数的去耦合,从而可以使得架构参数能更准确地体现该特征组合项的重要性,进而可以提高架构参数的优化准确度。原因如下。

- 5 应理解, e_i 是在模型训练过程中不断更新变化的。 e_i 与 e_j 做内积后, 即 $\langle e_i, e_j \rangle$, 其规格 (scale) 也是不断更新变换的。假设, $\alpha_{(i,j)} \langle e_i, e_j \rangle$ 可以通过 $\left(\frac{\alpha_{(i,j)}}{\eta} \right) (\eta \cdot \langle e_i, e_j \rangle)$ 获得,

其中, 第一项 $\left(\frac{\alpha_{(i,j)}}{\eta} \right)$ 与第二项 $(\eta \cdot \langle e_i, e_j \rangle)$ 具有耦合。如果第二项 $(\eta \cdot \langle e_i, e_j \rangle)$ 不作标量化处理, 则第一项 $\left(\frac{\alpha_{(i,j)}}{\eta} \right)$ 不能绝对代表第二项的重要性, 这样会为系统带来很大的不稳定性。

- 10 在本申请实施例中, 通过对特征组合项的模型参数进行标量化处理, 可以避免出现 $\alpha_{(i,j)} \langle e_i, e_j \rangle$ 可以通过 $\left(\frac{\alpha_{(i,j)}}{\eta} \right) (\eta \cdot \langle e_i, e_j \rangle)$ 获得的现象, 即可以实现该特征组合项的模型参数与架构参数的去耦合。

通过特征组合项的模型参数与架构参数之间的去耦合, 可以使得架构参数能更准确地体现该特征组合项的重要性, 进而可以提高架构参数的优化准确度。

- 15 换句话说, 通过对特征组合项的模型参数进行标量化处理, 可以实现该特征组合项的模型参数与架构参数的去耦合, 可以避免特征组合项的模型参数与架构参数的耦合效应对系统带来较大不稳定性。

- 如上文描述, 第二模型中包括两类参数: 架构参数与模型参数。在对架构参数的优化学习过程中, 不仅涉及架构参数的训练, 还涉及模型参数的训练。换句话说, 在步骤 S320 中对第二模型中的架构参数进行优化学习的过程中, 伴随着对第二模型中的模型参数的优化学习。

为了便于理解与描述, 在下文描述中, 将第二模型中的架构参数记为 α , 将第二模型中的模型参数记为 w (对应上文公式 (4) 中的 $\langle v_i, v_j \rangle$)。

- 25 可选地, 在图 3 所示实施例中, 对第二模型中架构参数 α 的优化处理和对第二模型中模型参数 w 的优化处理, 包括: 对第二模型中的架构参数 α 与模型参数 w 进行两层 (two-level) 优化处理。

换句话说, 在步骤 S320 中, 通过对第二模型中的架构参数 α 与模型参数 w 进行两层 (two-level) 优化处理, 获得优化后的架构参数 α^* 。

- 30 在本实施例中, 将第二模型中的架构参数 α 作为模型超参数进行优化学习, 将第二模型中的模型参数 w 作为模型的参数进行优化学习。也就是说, 将架构参数 α 作为高级决策变量, 将模型参数 w 作为低级决策变量。对于高级决策变量 α 的任一个取值, 都对应一个不同的模型。

可选地，对于高级决策变量 α 的任一个取值所对应的一个模型，评估这个模型时，通过完整训练这个模型来获得最优的模型参数 w_{optimal} 。换句话说，每评估一个架构参数 α 的候选值，完整训练一遍该候选值对应的模型。

5 可选地，对于高级决策变量 α 的任一个取值所对应的一个模型，评估这个模型时，用一个小批量（mini-batches）数据更新该模型一步得到的 w_{t+1} 代替最优的模型参数 w_{optimal} 。

可选地，在图3所示实施例中，对第二模型中架构参数 α 的优化处理和对第二模型中模型参数 w 的优化处理，包括：使用相同的训练数据，对第二模型中的架构参数 α 与模型参数 w 同时进行优化学习。

10 换句话说，在步骤 S320 中，通过使用相同的训练数据，对第二模型中的架构参数 α 与模型参数 w 同时进行优化学习，获得优化后的架构参数 α^* 。

在实施例中，在优化过程的每轮训练中，基于同一批训练数据，对架构参数 α 与模型参数 w 同时进行优化学习。或者说，将第二模型中的架构参数与模型参数视为同一级别的决策变量，对第二模型中的架构参数 α 与模型参数 w 同时进行优化学习，获得优化后的架构参数 α^* 。

15 在本实施例中，对第二模型中的架构参数 α 与模型参数 w 的优化处理，可以称为一层（one-level）优化处理。

例如，第二模型中的架构参数 α 与模型参数 w 在随机梯度下降(stochastic gradient descent, SGD)优化中自由探索其可行域直到收敛。

20 作为示例，采用如下公式(12)实现第二模型中架构参数 α 与模型参数 w 的优化学习：

$$\begin{aligned}\alpha_t &= \alpha_{t-1} - \eta_t \cdot \partial_{\alpha} L_{\text{train}}(w_{t-1}, \alpha_{t-1}) \\ w_t &= w_{t-1} - \delta_t \cdot \partial_w L_{\text{train}}(w_{t-1}, \alpha_{t-1})\end{aligned}\quad (12)$$

其中， α_t 表示进行第 t 步优化后的架构参数。 α_{t-1} 表示进行第 $t-1$ 步优化后的架构参数。 w_t 表示进行第 t 步优化后的模型参数。 w_{t-1} 表示进行第 $t-1$ 步优化后的模型参数。 η_t 表示进行第 t 步优化时架构参数的优化学习率。 δ_t 表示进行第 t 步优化时模型参数的学习率。 $L_{\text{train}}(w_{t-1}, \alpha_{t-1})$ 表示测试集上损失函数在第 t 步优化时的损失函数值。 $\partial_{\alpha} L_{\text{train}}(w_{t-1}, \alpha_{t-1})$ 表示测试集上损失函数在第 t 步优化时对架构参数 α 的梯度。 $\partial_w L_{\text{train}}(w_{t-1}, \alpha_{t-1})$ 表示测试集上损失函数在第 t 步优化时对模型参数 w 的梯度。

25 在本实施例中，通过对第二模型中的架构参数与模型参数进行一层优化处理，以实现第二模型中的架构参数的优化学习，可以实现架构参数与模型参数的同时优化，从而可以减小第二模型中的架构参数的优化学习过程的耗时，进而有助于提高特征组合项筛选的效率。

完成步骤 S330 后，即完成了特征组合项的筛选，第三模型为完成特征组合项筛选后的模型。

在步骤 S340 中，训练第三模型。

35 其中，可以对第三模型进行训练，或者，也可以在第三模型上增加 L1 正则项，和/

或 L2 正则项之后, 再进行训练。

可以根据应用需求, 确定训练第三模型的训练目的。

例如, 假设要获得点击率(click-through rate, CTR)预估模型, 则以获得 CTR 预估模型为训练目标, 训练第三模型, 以得到 CTR 预估模型。

- 5 再例如, 例如, 假设要获得转化率(conversion rate, CVR)预估模型, 则以获得 CVR 预估模型为训练目标, 训练第三模型, 以得到 CVR 预估模型。

可选地, 第三模型是基于第一模型进行特征组合项删减后得到的模型。详见上文中对步骤 S330 的描述, 这里不再赘述。

- 10 可选地, 第三模型是基于第二模型进行特征组合项删减后得到的模型。详见上文中对步骤 S330 的描述, 这里不再赘述。

应理解, 在完成特征组合项删减(或筛选)后, 在模型中保留架构参数, 对模型进行训练, 可以进一步学习特征组合项的重要性。

- 15 基于上文描述可知, 在本申请实施例, 通过在基于 FM 的模型中引入架构参数, 并通过架构参数的优化学习, 可以实现特征组合项的筛选。也就是说, 本申请通过架构参数的优化学习可以实现特征组合项的筛选, 而无需现有技术中针对多个候选子集的多遍训练, 因此, 可以有效减小特征组合筛选的计算量, 从而可以节省算力, 同时可以提高特征组合筛选的效率。

此外, 采用本申请实施例提供的方案, 可以支持基于 FM 的模型中的特征组合项扩展到更高阶。

- 20 图 5 为本申请实施例提供的自动特征组合筛选的方法 500 的另一示意性流程图。

首先获取训练数据。

例如, 假设特征维度为 m , 则针对 m 个域的特征, 获取训练数据。

S510, 枚举特征组合项进入基于 FM 的模型。

- 25 基于 FM 的模型的可以是如上述公式(1)或公式(2)所示的 FM 模型, 也可以是下列基于 FM 的模型中的任一种模型: DeepFM 模型、IPNN 模型、AFM 模型与 NFM 模型。

枚举特征组合项进入基于 FM 的模型, 指的是, 基于 m 个特征的所有组合情况构造基于 FM 的模型的特征组合项。

应理解, 构造特征组合项时, 会涉及到 m 个特征的辅助向量。

- 30 例如, 可以采用图 1 或图 3 所示的嵌入(embedding)层获取 m 个特征的辅助向量。通过嵌入(embedding)层获取 m 个特征的辅助向量的技术为现有技术, 本文不作赘述。

S520, 在基于 FM 的模型中引入架构参数。具体来说, 为基于 FM 的模型中的每个特征组合项添加一个系数参数, 这个系数参数称为架构参数。

步骤 S520 对应于上文实施例中的步骤 S310, 具体描述详见上文。

- 35 图 5 所示实施例中的基于 FM 的模型对应图 3 所示实施例中的第一模型, 图 5 所示实施例中基于 FM 的模型在添加架构参数后得到的模型对应图 3 所示实施例中的第二模型。

S530, 对架构参数进行优化学习, 直至收敛, 获得优化后的架构参数。

步骤 S530 对应于上文实施例中的步骤 S320, 具体描述详见上文。

S540, 根据优化后的架构参数, 进行特征组合项的删减, 获得经过特征组合项删减后的模型。

步骤 S540 对应于上文实施例中的步骤 S330，具体描述详见上文。

图 5 所示实施例中经过特征组合项删减后的模型对应图 3 所示实施例中的第三模型。

S550，训练经过特征组合项删减后的模型，直至收敛，获得 CTR 预估模型。

步骤 S550 对应于上文实施例中的步骤 S340，具体描述详见上文。

5 训练得到 CTR 预估模型后，可以对该 CTR 预估模型进行上线推理。

例如，将目标对象的数据输入 CTR 预估模型，CTR 预估模型的输出为该目标对象的点击率，根据该点击率，可以判断是否推荐该目标对象。

本申请实施例提供的自动特征组合筛选的方案可以适用于任意基于因子分解机（factorization machine, FM）的模型，例如，FM 模型、DeepFM 模型、IPNN 模型、AFM
10 模型与 NFM 模型等。

作为一个示例，本申请实施例提供的自动特征组合筛选方案可以应用于现有的 FM 模型。

例如，在现有的 FM 模型中引入架构参数，通过架构参数的优化学习，获得各个特征组合项的重要性，然后基于各个特征组合项的重要性实现特征组合的筛选，最终得到经过
15 特征组合筛选后的 FM 模型。

应理解，通过将本申请应用于 FM 模型，可以高效实现 FM 模型的特征组合项筛选，从而可以支持 FM 模型的特征组合项扩展到更高阶。

作为另一个示例，本申请实施例提供的自动特征组合筛选方案可以应用于现有的 deep-FM 模型。

20 例如，在现有的 DeepFM 模型中引入架构参数，通过架构参数的优化学习，获得各个特征组合项的重要性，然后基于各个特征组合项的重要性实现特征组合的筛选，最终得到经过特征组合筛选后的 DeepFM 模型。

应理解，通过将本申请应用于 DeepFM 模型，可以高效实现 DeepFM 模型的特征组合项筛选。

25 如图 6 所示，本申请实施例还提供一种数据处理的方法 600。该方法 600 包括如下步骤 S610 与 S620。

S610，将目标对象的数据输入 CTR 预估模型或 CVR 预估模型，获得目标对象的预估结果。

例如，该目标对象为商品。

30 S620，根据目标对象的预估结果，确定目标对象的推荐情况。

其中，CTR 预估模型或 CVR 预估模型通过如下上文实施例提供的方法 300 获得，即该 CTR 预估模型或 CVR 预估模型通过上文实施例中的步骤 S310 至步骤 S340 获得。详见上文描述，这里不再赘述。

其中，在步骤 S340 中，使用该目标对象的训练样本，训练第三模型，以得到 CTR 预
35 估模型或 CVR 预估模型。

可选地，在步骤 S320 中，使用目标对象的训练样本中相同的训练数据，对第二模型中的架构参数与模型参数同时进行优化学习，获得优化后的架构参数。

或者说，将第二模型中的架构参数与模型参数视为同一级别的决策变量，使用目标对象的训练样本，对第二模型中的架构参数与模型参数同时进行优化学习，获得优化后的架

构参数。

仿真测试：上线 AB 测试 CTR 预测精度显著提升，推理能耗明显降低。

5 作为示例，仿真实验表明，将本申请提供的特征组合筛选方案应用于推荐系统的 DeepFM 模型，上线进行 A/B 测试，游戏下载率可以提升 20%，CTR 预估精确率可以相对提升 20.3%，转化率（conversion rate, CVR）可以相对提升 20.1%，因此，可以有效加快模型的推理速度。

10 作为一个示例，在公开数据集 Avazu 上使用本申请提供的方案获得 FM 模型与 DeepFM 模型，利用本申请的方案获得的 FM 模型与 DeepFM 模型的性能与业界其他的模型的性能的比较结果如表 1 与表 2 所示。其中，表 1 表示二阶模型的比较，表 2 表示三阶模型的比较。二阶模式表示，模型的特征组合项的最高阶数是二阶。三阶阶模式表示，模型的特征组合项的最高阶数是三阶。

表 1

模型（model）	公开数据集 Avazu					
	AUC	Log loss	Top	time（s）	Search+re-train cost(min)	Rel.Impr
FM	0.7793	0.3805	100%	0.51	0+3	0
FwFM	0.7822	0.3784	100%	0.52	0+4	0.37%
AFM	0.7806	0.3794	100%	1.92	0+14	0.17%
FFM	0.7831	0.3781	100%	0.24	0+6	0.49%
DeepFM	0.7836	0.3776	100%	0.76	0+6	0.55%
GBDT+LR	0.7721	0.3841	100%	0.45	8+3	-0.92%
GBDT+FFM	0.7835	0.3777	100%	2.66	6+21	0.54%
AutoFM（2nd）	0.7831	0.3778	29%	0.23	4+2	0.49%
AutoDeepFM（2nd）	0.7852	0.3765	24%	0.48	7+4	0.76%

表 2

模型（model）	公开数据集 Avazu					
	AUC	Log loss	Top	time（s）	Search+re-train cost(min)	Rel.Impr
FM(3rd)	0.7843	0.3772	100%	5.70	0+21	0.64%
DeepFM(3rd)	0.7854	0.3765	100%	5.97	0+23	0.78%
AutoFM(3rd)	0.7860	0.3762	25%/2%	0.33	22+5	0.86%
AutoDeepFM(3rd)	0.7870	0.3756	21%/10%	0.94	24+10	0.99%

在表 1 与表 2 中，横向表头的含义如下。

AUC (area under curve) 表示曲线下方的面积。Log loss 表示损失值的 log。Top 表示进行特征组合项筛选后所保留的特征组合项的比例。Time 表示模型推理两百万个样本的时间。Search+re-train cost 表示，搜索耗时+重新训练耗时，其中，搜索耗时表示上文实施例步骤 S320 与步骤 S330 的耗时，重新训练耗时表示上文实施例中步骤 S340 的耗时。Rel.Impr 表示相对提升值。

在表 1 中，纵向表头的含义如下。

FM、FwFM、AFM、FFM、DeepFM 表示现有技术中的基于 FM 的模型。GBDT+LR 与 GBDT+FFM 表示现有技术中采用手工特征组合筛选的模型。

AutoFM(2nd) 表示使用本申请实施例提供的方案获取的二阶的 FM 模型。AutoDeepFM(2nd) 表示使用本申请实施例提供的方案获取的三阶的 DeepFM 模型。

在表 2 中，纵向表头的含义如下。

FM(3rd) 表示现有技术中的三阶的 FM 模型。DeepFM(3rd) 表示现有技术中的三阶的 DeepFM 模型。

AutoFM(3rd) 表示使用本申请实施例提供的方案获取的三阶的 FM 模型。AutoDeepFM(3rd) 表示使用本申请实施例提供的方案获取的三阶的 DeepFM 模型。

从表 1 与表 2 可知，相对于现有技术，采用本申请实施例提供的方案获取的 FM 模型或 DeepFM 模型，进行 CTR 预测，可以显著提升 CTR 预测精度，并且可以有效降低推理时间与能耗。

基于上文描述可知，在本申请实施例中，通过在基于 FM 的模型中引入架构参数，通过对架构参数的优化学习，可以实现特征组合项的筛选。也就是说，本申请只需执行架构参数的优化学习，这样的一次模型训练过程，就可以实现特征组合项筛选，而无需现有技术中针对多个候选子集的多遍训练，因此，可以有效减小特征组合筛选的计算量，从而可以节省算力，同时可以提高特征组合筛选的效率。

此外，采用本申请实施例提供的方案，可以支持基于 FM 的模型中的特征组合项扩展到更高阶。

本文中描述的各个实施例可以为独立的方案，也可以根据内在逻辑进行组合，这些方案都落入本申请的保护范围中。

上文描述了本申请提供的方法实施例，下文将描述本申请提供的装置实施例。应理解，装置实施例的描述与方法实施例的描述相互对应，因此，未详细描述的内容可以参见上文方法实施例，为了简洁，这里不再赘述。

如图 7 所示，本申请实施例还提供一种数据处理的装置 700。该装置 700 包括如下单元。

第一处理单元 710，用于为第一模型中的每个特征组合项添加架构参数，得到第二模型，其中，第一模型为基于因子分解机 FM 的模型，架构参数用于表征相应特征组合项的重要性。

第二处理单元 720，用于对第二模型中的架构参数进行优化学习，获得优化后的架构参数。

第三处理单元 730, 用于根据优化后的架构参数, 基于第一模型或第二模型, 获得经过特征组合项删减后的第三模型。

可选地, 第二处理单元 720 对架构参数进行的优化学习使得, 优化后的架构参数呈现稀疏性。

- 5 在本实施例中, 第三处理单元 730 用于, 基于第一模型或第二模型, 将优化后的架构参数中取值小于阈值的架构参数所对应的特征组合项删减, 获得第三模型。

可选地, 第二处理单元 720 对架构参数进行的优化学习使得, 至少一个特征组合项的架构参数的取值在优化结束后等于零。

- 10 例如, 第三处理单元 730 用于, 使用 GRDA 优化器对第二模型中的架构参数进行优化, GRDA 优化器使得至少一个特征组合项的架构参数的取值在优化过程中趋向于零。

可选地, 第二处理单元 720 还用于, 对第二模型中的模型参数进行优化学习, 该优化学习包括, 对第二模型中的模型参数进行标量化处理。

例如, 第二处理单元 720 用于, 对第二模型中的模型参数进行批量归一化 BN 处理。

- 15 可选地, 第二处理单元 720 用于, 使用相同的训练数据, 对第二模型中的架构参数与模型参数同时进行优化学习, 获得优化后的架构参数。

可选地, 该装置 700 还包括训练单元 740, 用于训练第三模型。

可选地, 训练单元 740 用于, 训练第三模型, 以得到 CTR 预估模型或 CVR 预估模型。

该装置 700 可以集成在终端设备、网络设备或芯片上。

该装置 700 可以部署在相关设备的计算节点上。

- 20 如图 8 所示, 本申请实施例还提供一种图像处理的装置 800。装置 800 包括如下单元。

第一处理单元 810 用于, 将目标对象的数据输入点击率 CTR 预估模型或转化率 CVR 预估模型, 获得所述目标对象的预估结果。

第一处理单元 820 用于, 根据所述目标对象的预估结果, 确定所述目标对象的推荐情况。

- 25 其中, 所述 CTR 预估模型或所述 CVR 预估模型通过上文实施例中的方法 300 或 500 获得。

其中, 训练第三模型的步骤包括: 使用该目标对象的训练样本, 训练第三模型, 以得到 CTR 预估模型或 CVR 预估模型。

- 30 可选地, 对架构参数进行优化学习的步骤包括: 使用所述目标对象的训练样本中相同的训练数据, 对所述第二模型中的架构参数与模型参数同时进行优化学习, 获得优化后的架构参数。

该装置 800 可以集成在终端设备、网络设备或芯片上。

该装置 800 可以部署在相关设备的计算节点上。

- 35 如图 9 所示, 本申请实施例还提供一种图像处理的装置 900。该装置 900 包括处理器 910, 处理器 910 与存储器 920 耦合, 存储器 920 用于存储计算机程序或指令, 处理器 910 用于执行存储器 920 存储的计算机程序或指令, 使得上文方法实施例中的方法被执行。

可选地, 如图 9 所示, 该装置 900 还可以包括存储器 920。

可选地, 如图 9 所示, 该装置 900 还可以包括数据接口 930, 数据接口 930 用于与外界进行数据的传输。

可选地，作为一种方案，该装置 900 用于实现上文实施例中的方法 300。

可选地，作为另一种方案，该装置 900 用于实现上文实施例中的方法 500。

可选地，作为又一种方案，该装置 900 用于实现上文实施例中的方法 600。

本申请实施例还提供一种计算机可读介质，该计算机可读介质存储用于设备执行的程序代码，该程序代码包括用于执行上述实施例的方法。

本申请实施例还提供一种包含指令的计算机程序产品，当该计算机程序产品在计算机上运行时，使得计算机执行上述实施例的方法。

本申请实施例还提供一种芯片，所述芯片包括处理器与数据接口，所述处理器通过所述数据接口读取存储器上存储的指令，执行上述实施例的方法。

可选地，作为一种实现方式，所述芯片还可以包括存储器，所述存储器中存储有指令，所述处理器用于执行所述存储器上存储的指令，当所述指令被执行时，所述处理器用于执行上述实施例中的方法。

本申请实施例还提供一种电子设备，该电子设备包括上述实施例中的装置 700、装置 800 或装置 900 中的任一种或多种。

图 10 为本申请实施例提供的一种芯片硬件结构，该芯片上包括神经网络处理器 1000。该芯片可以被设置在如下任一种或多种装置或系统中：

如图 7 所示的装置 700、如图 8 所示的装置 800、如图 9 中所示的装置 900。

上文方法实施例中的方法 300、500 或 600 均可在如图 10 所示的芯片中得以实现。

神经网络处理器 1000 作为协处理器挂载到主处理器（Host CPU）上，由主 CPU 分配任务。神经网络处理器 1000 的核心部分为运算电路 1003，控制器 1004 控制运算电路 1003 获取存储器（权重存储器 1002 或输入存储器 1001）中的数据并进行运算。

在一些实现中，运算电路 1003 内部包括多个处理单元（process engine，PE）。在一些实现中，运算电路 1003 是二维脉动阵列。运算电路 1003 还可以是一维脉动阵列或者能够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 1003 是通用的矩阵处理器。

举例来说，假设有输入矩阵 A，权重矩阵 B，输出矩阵 C。运算电路 1003 从权重存储器 1002 中取矩阵 B 相应的数据，并缓存在运算电路 1003 中每一个 PE 上。运算电路 1003 从输入存储器 1001 中取矩阵 A 数据与矩阵 B 进行矩阵运算，得到的矩阵的部分结果或最终结果，保存在累加器（accumulator）1008 中。

向量计算单元 1007 可以对运算电路 1003 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。例如，向量计算单元 1007 可以用于神经网络中非卷积/非 FC 层的网络计算，如池化（pooling），批归一化（batch normalization），局部响应归一化（local response normalization）等。

在一些实现种，向量计算单元 1007 将经处理的输出的向量存储到统一存储器（也可称为统一缓存器）1006。例如，向量计算单元 1007 可以将非线性函数应用到运算电路 1003 的输出，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 1007 生成归一化的值、合并值，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 1003 的激活输入，例如用于在神经网络中的后续层中的使用。

上文方法实施例中的方法 300、500 或 600 可以由 1003 或 1007 执行。

统一存储器 1006 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 1005 (direct memory access controller, DMAC) 将外部存储器中的输入数据搬运到输入存储器 1001 和/或统一存储器 1006、将外部存储器中的权重数据存入权重存储器 1002, 以及将统一存储器 1006 中的数据存入外部存储器。

总线接口单元 (bus interface unit, BIU) 1010, 用于通过总线实现主 CPU、DMAC 和取指存储器 1009 之间进行交互。

与控制器 1004 连接的取指存储器(instruction fetch buffer)1009, 用于存储控制器 1004 使用的指令;

控制器 1004, 用于调用指存储器 1009 中缓存的指令, 实现控制该运算加速器的工作过程。

在本申请实施例中, 这里的数据可以是待处理的图像数据。

一般地, 统一存储器 1006, 输入存储器 1001, 权重存储器 1002 以及取指存储器 1009 均为片上 (On-Chip) 存储器, 外部存储器为该 NPU 外部的存储器, 该外部存储器可以为双倍数据率同步动态随机存储器 (double data rate synchronous dynamic random access memory, DDR SDRAM)、高带宽存储器 (high bandwidth memory, HBM) 或其他可读可写的存储器。

除非另有定义, 本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中在本申请的说明书中所使用的术语只是为了描述具体的实施例的目的, 不是旨在限制本申请。

需要说明的是, 本文中涉及的第一、第二、第三或第四等各种数字编号仅为描述方便进行的区分, 并不用来限制本申请实施例的范围。

本领域普通技术人员可以意识到, 结合本文中所公开的实施例描述的各示例的单元及算法步骤, 能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行, 取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能, 但是这种实现不应认为超出本申请的范围。

所属领域的技术人员可以清楚地了解到, 为描述的方便和简洁, 上述描述的系统、装置和单元的具体工作过程, 可以参考前述方法实施例中的对应过程, 在此不再赘述。

在本申请所提供的几个实施例中, 应该理解到, 所揭露的系统、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, 所述单元的划分, 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, 例如多个单元或组件可以结合或者可以集成到另一个系统, 或一些特征可以忽略, 或不执行。另一点, 所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口, 装置或单元的间接耦合或通信连接, 可以是电性, 机械或其它的形式。

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单元来实现本实施例方案的目的。

另外, 在本申请各个实施例中的各功能单元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。

所述功能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，可以存储在一个计算机可读取存储介质中。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或者该技术方案的部分可以以软件产品的形式体现出来，该计算机

5 软件产品存储在一个存储介质中，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，服务器，或者网络设备）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：通用串行总线闪存盘（USB flash disk，UFD）（UFD也可以简称为U盘或者优盘）、移动硬盘、只读存储器（read-only memory，ROM）、随机存取存储器（random access memory，RAM）、磁碟或者光盘等各种可以存储程序代码的介质。

10 以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

15