

一种人工智能系统中数据处理方法及装置

本申请要求于 2020 年 12 月 30 日提交中国知识产权局、申请号为 202011610237.3、申请名称为“一种人工智能系统中数据处理方法及装置”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。

技术领域

本申请涉及人工智能 (artificial intelligence, AI) 领域，特别涉及一种人工智能系统中数据处理的方法及装置。

背景技术

人工智能，是指由人制造出来的机器所表现出来的智能。人工智能有着多种实现方式，深度学习 (deep learning) 是其中的一种实现方式。深度学习是利用人工神经网络为架构，对数据、资料进行表征学习的算法。人工智能框架，也称深度学习框架，是为数据科学家、开发人员和研究人员提供的一种高级编程语言，专门用于对神经网络进行训练、验证以及利用神经网络进行推理等操作。

当前常用的人工智能框架包括 TensorFlow, Caffe 和 PyTorch 等。以 TensorFlow 为例，该模型框架主要支持两种数据格式，NCHW 和 NHWC。这两种数据格式的名称均包含 N、H、W 和 C 这四个字母。其中，这四个字母的意思分别是：N 是指一个批处理 (Batch) 中图像的数量，H 是指图像中垂直高度方向的像素个数，W 是指图像中水平宽度方向的像素个数，C 是指图像包含的通道数，例如灰度图像的通道数为 1，而采用红绿蓝 (Red Green Blue, RGB) 格式的彩色图像的通道数为 3。

NCHW 和 NHWC 格式的主要区别在于垂直高度方向的像素、水平宽度的像素以及通道的排列顺序不同，这也导致了两种数据格式各有优势和劣势。其中，NHWC 格式的数据局部性较好，缓存利用率高，更适合采用中央处理器 (central processor unit, CPU) 来处理；而 NCHW 格式的数据则格式采用图像处理器 (graphic processor unit, GPU) 进行处理。由于 TensorFlow 的早期开发更多地使用 CPU 来处理数据，在这种情况下使用 NHWC 格式的数据会比使用 NCHW 格式的数据效率更高，因此 TensorFlow 选择将 NHWC 格式作为默认格式。随着 GPU 在人工智能的训练任务和处理任务中越来越多的被使用，NCHW 格式也成为需要人工智能框架所支持的数据格式。

数据格式会影响到 AI 系统处理进行数据处理的一些处理逻辑和程序代码。当前的 AI 系统，包括主机和加速卡，往往只支持某一种特定格式的数据。当该 AI 系统需要适配新的数据格式时，需要对 AI 系统中的程序代码等进行较大的改动，从而导致工程量和验证量的增加以及开发时间增加等问题的产生。

发明内容

针对现有技术中 AI 系统适配新的数据格式需要对 AI 系统中的程序代码进行较大改动的问题，本申请提供一种 AI 系统以及相应的数据处理方法和装置，用于对来自用户的输入数据进行格式转换，既在成本不大的前提下扩展了 AI 系统用于处理 AI 任务的数据格式，也可以使 AI 系统选择合适的数据格式处理 AI 任务，从而增加了 AI 系统的效率以及该 AI 系统的应用范围。

第一方面，本申请提供一种 AI 系统，该 AI 系统包括主机和加速卡，其中，主机用于：获取来自用户的 AI 任务以及与 AI 任务对应的第一数据，该第一数据采用第一格式；根据 AI 任务，向加速卡发送第一指令，该第一指令用于指示加速卡将第一数据转化为采用第二格式的第二数据；加速卡用于：根据接收到的第一指令，将第一数据转换为第二数据，并利用第二数据执行 AI 任务。

在该方案中，AI 系统中的主机可以对来自用户的第一数据进行格式转换，并利用格式转换后的数据来执行 AI 任务，从而可以在不对 AI 系统中的程序代码进行较大改动的情况下，使 AI 系统支持采用新的格式的数据来执行 AI 任务，从而增加了 AI 系统的效率和应用范围。

根据第一方面，在本申请一种可能的实现方式中，主机具体用于，根据 AI 任务包括的算子类型，向加速卡发送第一指令。

在该方案中，主机根据 AI 任务中的算子类型，确定格式转换的目标格式为第二格式，并通过第一指令指示加速卡将第一数据转换为采用第二格式的第二数据。由于采用特定的数据格式会提高亲和性算子的处理效率，当根据 AI 任务的算子类型确定格式转换的目标格式时，有利于提高 AI 系统处理 AI 任务时的效率。

根据第一方面，在本申请另一种可能的实现方式中，AI 任务中包含第二指令，该第二指令用于指示采用第二格式的数据执行 AI 任务；主机具体用于，根据该第二指令，向加速卡发送第一指令。

在该方案中，用户在 AI 任务中可以指定 AI 系统处理 AI 任务时所用的数据格式，有利于提高 AI 系统处理 AI 任务的效率。

根据第一方面，在本申请另一种可能的实现方式中，主机还用于，建立 AI 任务包括的算子类型和第二格式之间的对应关系。

在该方案中，主机预先获取算子类型和数据格式之间的亲和性对应关系，从而在接收到 AI 任务后，可以根据该 AI 任务所包括的算子类型以及主机中的亲和性对应关系，确定所应该采用的目标数据格式，从而增加了 AI 系统处理 AI 任务的效率。

根据第一方面，在本申请另一种可能的实现方式中，主机存储将第一格式的数据转换为第二格式的数据的第一映射关系，并将该第一映射关系包含在第一指令中发送给加速卡，而加速卡则根据接收到的映射关系将第一数据转换为第二格式。

在该方案中，主机中预存了格式转换所需要的映射关系，并在确定需要格式转换时将该相应的映射关系发送给加速卡，使得加速卡可以根据主机中存储的映射关系进行格式转换，从而减小了加速卡所需存储的数据，提高了 AI 系统的效率。

根据第一方面，在本申请另一种可能的实现方式中，该 AI 系统位于公有云中。

在该方案中，AI 系统可以用于为用户提供云服务，因此拓展了本方案的使用场景，从而增加了 AI 系统的应用范围。

根据第一方面，在本申请另一种可能的实现方式中，该 AI 系统处理至少一种 AI 框架，且该至少一个 AI 框架均支持第一格式的数据。

在该方案中，AI 系统所支持的用户输入数据的格式与该 AI 系统所处理的 AI 框架相关，即 AI 系统所支持的输入数据的格式需要被相关的 AI 框架所支持。这使得无论 AI 任务对应的是哪个 AI 框架，AI 系统都可以处理该 AI 任务对应的数据，从而增加了 AI 系统的稳定性。

根据第一方面，在本申请另一种可能的实现方式中，主机还存储第二映射关系，该第二映射关系用于将第一格式的数据转换为第三格式的数据。

在该方案中，主机中可以存储多种映射关系，这些映射关系可以用于将采用第一格式的数据转换成多种目标格式中的任一种格式的数据。在这种情况下，AI 系统可以支持多种数据转换的目标格式的类型，从而扩展了 AI 系统的使用场景，并增加了该 AI 系统的使用范围。

根据第一方面，在本申请另一种可能的实现方式中，主机还存储第三映射关系，该第三映射关系用于将第四格式的数据转换为第二格式的数据。

在该方案中，主机可以存储多种映射关系，这些映射关系使得 AI 系统支持多种来自用户的数据的格式，并可以用于将任一格式的数据转换为第二格式的数据。该方案增加了 AI 系统所支持的输入数据的格式的数量，从而扩展了 AI 系统的使用场景，并增加了该 AI 系统的使用范围。

根据第一方面，在本申请另一种可能的实现方式中，主机还用于，获取第一数据的格式信息，并根据该格式信息，确定第一数据采用的是第一格式。

在该方案中，该 AI 系统可以支持一种或多种输入数据的格式，而由于主机无法自适应的识别来自用户的第一数据的格式，因此，可以通过用户输入格式信息等方式确定第一数据的格式，从而对第一数据进行后续处理。该方案确保 AI 系统可以识别来自用户的第一数据的格式，从而增加了系统的稳定性。

根据第一方面，在本申请另一种可能的实现方式中，主机还用于，输出提示信息，该提示信息用于提示用户输入第一格式的数据。

在该方案中，主机会输出该主机所支持的输入数据的格式的信息，从而提示用户提供主机所能支持的格式的输入数据，从而增加了系统的稳定性。

第二方面，本申请提供一种 AI 系统，该 AI 系统包括主机和加速卡，该主机用于：获取来自用户的 AI 任务以及与该 AI 任务对应的第一数据，该第一数据采用第一格式；根据该 AI 任务，将第一数据转换为采用第二格式的第二数据，并向加速卡发送该第二数据；加速卡用于：接收第二数据，并利用该第二数据执行 AI 任务。

根据第二方面，在本申请一种可能的实现方式中，主机具体用于，根据 AI 任务包括的算子类型，将该第一数据转换成第二数据。

根据第二方面，在本申请另一种可能的实现方式中，AI 任务中包含第二指令，该第二

指令用于指示采用第二格式的数据执行该 AI 任务，主机具体用于，根据该第二指令，将第一数据转换为第二数据。

根据第二方面，在本申请另一种可能的实现方式中，建立 AI 任务包括的算子类型和第二格式之间的对应关系。

根据第二方面，在本申请另一种可能的实现方式中，主机存储将第一格式的数据转换为第二格式的数据的第一映射关系，主机具体用于，根据第一映射关系，将第一数据转换为第二数据。

根据第二方面，在本申请另一种可能的实现方式中，AI 系统位于公有云中。

根据第二方面，在本申请另一种可能的实现方式中，AI 系统处理至少一种 AI 框架，其中，至少一种 AI 框架均支持第一格式的数据。

根据第二方面，在本申请另一种可能的实现方式中，主机还存储第二映射关系，该第二映射关系用于将第一格式的数据转换为第三格式的数据。

根据第二方面，在本申请另一种可能的实现方式中，主机还存储第三映射关系，该第三映射关系用于将第四格式的数据转换为第二格式的数据。

根据第二方面，在本申请另一种可能的实现方式中，主机还用于获取第一数据的格式信息，并根据该格式信息，确定第一数据采用的是第一格式。

根据第二方面，在本申请另一种可能的实现方式中，主机还用于，输出提示信息，该提示信息用于指示用户输入第一格式的数据。

第三方面，本申请提供一种数据处理的方法，该方法用于 AI 系统中的主机，该 AI 系统还包括加速卡，该方法包括：获取来自用户的 AI 任务以及与 AI 任务对应的第一数据，其中，该第一数据采用第一格式；根据该 AI 任务，向加速卡发送第一指令，该第一指令用于指示加速卡将第一数据转换为采用第二格式的第二数据。

根据第三方面，在本申请一种可能的实现方式中，根据 AI 任务向加速卡发送第一指令的步骤，具体包括：根据 AI 任务包括的算子类型，向加速卡发送第一指令。

根据第三方面，在本申请另一种可能的实现方式中，AI 任务中还包括第二指令，该第二指令用于指示采用第二格式的数据执行 AI 任务；该方法还包括：根据该第二指令，向加速卡发送第一指令。

根据第三方面，在本申请另一种可能的实现方式中，该方法还包括：建立 AI 任务包括的算子类型和第二格式之间的对应关系。

根据第三方面，在本申请另一种可能的实现方式中，主机存储将第一格式的数据转换为第二格式的数据的第一映射关系，该第一指令包括该第一映射关系。

根据第三方面，在本申请另一种可能的实现方式中，AI 系统位于公有云中。

根据第三方面，在本申请另一种可能的实现方式中，AI 系统处理至少一种 AI 框架，该至少一种 AI 框架均支持第一格式的数据。

根据第三方面，在本申请另一种可能的实现方式中，该主机还存储第二映射关系，该第二映射关系用于将第一格式的数据转换为第三格式的数据。

根据第三方面，在本申请另一种可能的实现方式中，该主机还存储第三映射关系，该

第三映射关系将第四格式的数据转换为第二格式的数据。

根据第三方面，在本申请另一种可能的实现方式中，该方法还包括，获取第一数据的格式信息，并根据该格式信息，确定第一数据采用的是第一格式。

根据第三方面，在本申请另一种可能的实现方式中，该方法还包括：输出提示信息，该提示信息用于指示用于输入第一格式的数据。

第四方面，本申请提供一种数据处理的方法，该方法用于 AI 系统中的主机，该 AI 系统还包括加速卡，该方法包括：获取来自用户的 AI 任务以及与该 AI 任务对应的第一数据，其中，该第一数据采用第一格式；根据该 AI 任务，将第一数据转换为采用第二格式的第二数据，并向加速卡发送该第二数据。

根据第四方面，在本申请一种可能的实现方式中，根据 AI 任务将第一数据转换为采用第二格式的第二数据的步骤，具体包括：根据 AI 任务所包括的算子类型，将该第一数据转换为第二数据。

根据第四方面，在本申请另一种可能的实现方式中，AI 任务中包含第二指令，该第二指令用于指示采用第二格式的数据执行 AI 任务，根据 AI 任务将第一数据转换为采用第二格式的第二数据的步骤，具体包括：根据该第二指令，向加速卡发送第一指令。

根据第四方面，在本申请另一种可能的实现方式中，该方法还包括：建立 AI 任务包括的算子类型和第二格式之间的对应关系。

根据第四方面，在本申请另一种可能的实现方式中，主机存储将第一格式的数据转换为第二格式的数据的第一映射关系，该第一指令包括第一映射关系。

根据第四方面，在本申请另一种可能的实现方式中，AI 系统位于公有云中。

根据第四方面，在本申请另一种可能的实现方式中，AI 系统处理至少一种 AI 框架，该至少一种 AI 框架均支持第一格式的数据。

根据第四方面，在本申请另一种可能的实现方式中，主机还存储第二映射关系，该第二映射关系用于将第一格式的数据转换为第三格式的数据。

根据第四方面，在本申请另一种可能的实现方式中，主机还存储第三映射关系，该第三映射关系用于将第四格式的数据转换为第二格式的数据。

根据第四方面，在本申请另一种可能的实现方式中，该方法还包括：获取第一数据的格式信息，并根据该格式信息，确定第一数据采用的是第一格式。

根据第四方面，在本申请另一种可能的实现方式中，该方法还包括：输出提示信息，该提示信息用于指示用户输入第一格式的数据。

第五方面，本申请提供一种数据处理的方法，该数据处理的方法应用于 AI 系统中的加速卡，该 AI 系统还包括主机，该方法包括：接收主机发送的第一数据和第一指令，该第一数据采用第一格式；根据该第一指令，将第一数据转换为第二数据，该第二数据采用第二格式；利用第二数据执行 AI 任务。

根据第五方面，在本申请一种可能的实现方式中，第一指令包括第一映射关系，根据第一指令将第一数据转换为第二数据的步骤，具体包括：根据第一映射关系，将第一数据

转换为第二数据。

根据第五方面，在本申请另一种可能的实现方式中，第二格式与AI任务包括的算子类型相对应。

根据第五方面，在本申请另一种可能的实现方式中，AI系统位于公有云中。

根据第五方面，在本申请另一种可能的实现方式中，该方法还包括：接收主机发送的第二映射关系，该第二映射关系用于将第一格式的数据转换为第三格式的数据。

根据第五方面，在本申请另一种可能的实现方式中，该方法还包括：接收主机发送的第三映射关系，该第三映射关系用于将第四格式的数据转换为第二格式的数据。

根据第五方面，在本申请另一种可能的实现方式中，AI系统处理至少一种AI框架，该至少一种AI框架均支持第一格式的数据。

第六方面，本申请提供一种数据处理装置，该数据处理装置位于AI系统，该AI系统中还包括加速卡，该数据处理装置包括：获取模块，用于获取来自用户的AI任务以及与该AI任务对应的第一数据，该第一数据采用第一格式；传输模块，用于根据该AI任务，向加速卡发送第一指令，该第一指令用于指示加速卡将第一数据转换为采用第二格式的第二数据。

根据第六方面，在本申请一种可能的实现方式中，传输模块具体用于，根据AI任务所包括的算子类型，向加速卡发送第一指令。

根据第六方面，在本申请另一种可能的实现方式中，AI任务包括第二指令，该第二指令用于指示采用第二格式的数据执行AI任务，其中，传输模块具体用于，根据第二指令，向加速卡发送第一指令。

根据第六方面，在本申请另一种可能的实现方式中，获取模块还用于：建立AI任务包括的算子类型和第二格式之间的对应关系。

根据第六方面，在本申请另一种可能的实现方式中，数据处理装置存储将第一格式的数据转换为第二格式的数据的第一映射关系，且第一指令包括第一映射关系。

根据第六方面，在本申请另一种可能的实现方式中，AI系统位于公有云中。

根据第六方面，在本申请另一种可能的实现方式中，AI系统处理至少一种AI框架，该至少一种AI框架均支持第一格式的数据。

根据第六方面，在本申请另一种可能的实现方式中，数据处理装置还存储第二映射关系，该第二映射关系用于将第一格式的数据转换为第三格式的数据。

根据第六方面，在本申请另一种可能的实现方式中，数据处理装置还存储第三映射关系，其中，该第三映射关系用于将第四格式的数据转换为第二格式的数据。

根据第六方面，在本申请另一种可能的实现方式中，获取模块还用于，获取第一数据的格式信息；数据处理装置还包括确定模块，该确定模块用于确定第一数据采用的是第一格式。

根据第六方面，在本申请另一种可能的实现方式中，传输模块还用于，输出提示信息，该提示信息用于指示用户输入第一格式的数据。

第七方面，本申请提供一种数据处理装置，该数据处理装置位于 AI 系统，该 AI 系统中还包括加速卡，该数据处理装置包括：获取模块，用于获取来自用户的 AI 任务以及与该 AI 任务对应的第一数据，该第一数据采用第一格式；转换模块，用于根据该 AI 任务，将第一数据转换为采用第二格式的第二数据；传输模块，用于向加速卡发送第二数据。

根据第七方面，在本申请一种可能的实现方式中，转换模块具体用于，根据 AI 任务包括的算子类型，将第一数据转换为第二数据。

根据第七方面，在本申请另一种可能的实现方式中，AI 任务中包括第二指令，该第二指令用于指示采用第二格式的数据执行该 AI 任务，其中，转换模块具体用于，根据第二指令，将第一数据转换为第二数据。

根据第七方面，在本申请另一种可能的实现方式中，获取模块还用于，建立 AI 任务所包括的算子类型和第二格式之间的对应关系。

根据第七方面，在本申请另一种可能的实现方式中，数据处理装置存储将第一格式的数据转换为第二格式的数据的第一映射关系，其中，转换模块具体用于，根据第一映射关系，将第一数据转换为第二数据。

根据第七方面，在本申请另一种可能的实现方式中，AI 系统位于公有云中。

根据第七方面，在本申请另一种可能的实现方式中，AI 系统处理至少一种 AI 框架，该至少一种 AI 框架均支持第一格式的数据。

根据第七方面，在本申请另一种可能的实现方式中，数据处理装置还存储第二映射关系，其中，该第二映射关系用于将第一格式的数据转换为第三格式的数据。

根据第七方面，在本申请另一种可能的实现方式中，数据处理装置还存储第三映射关系，其中，该第三映射关系用于将第四格式的数据转换为第二格式的数据。

根据第七方面，在本申请另一种可能的实现方式中，获取模块还用于获取第一数据的格式信息，数据处理装置还包括确定模块，该确定模块用于确定第一数据采用的是第一格式。

根据第七方面，在本申请另一种可能的实现方式中，传输模块还用于，输出提示信息，该提示信息用于指示用户输入第一格式的数据。

第八方面，本申请提供一种数据处理装置，该数据处理装置位于 AI 系统，该 AI 系统中还包括主机，该数据处理装置包括：传输模块，用于接收主机发送的第一数据和第一指令，该第一数据采用第一格式；转换模块，用于根据第一指令，将第一数据转换为第二数据，该第二数据采用第二格式；处理模块，用于利用该第二数据执行 AI 任务。

根据第八方面，在本申请一种可能的实现方式中，第一指令包括第一映射关系，转换模块具体用于，根据第一映射关系，将第一数据转换为第二数据。

根据第八方面，在本申请另一种可能的实现方式中，第二格式与 AI 任务包括的算子类型相对应。

根据第八方面，在本申请另一种可能的实现方式中，AI 系统位于公有云中。

根据第八方面，在本申请另一种可能的实现方式中，传输模块还用于，接收主机发送的第二映射关系，该第二映射关系用于将第一格式的数据转换为第三格式的数据。

根据第八方面，在本申请另一种可能的实现方式中，传输模块还用于，接收主机发送的第三映射关系，该第三映射关系用于将第四格式的数据转换为第二格式的数据。

根据第八方面，在本申请另一种可能的实现方式中，AI 系统处理至少一种 AI 框架，该至少一种 AI 框架均支持第一格式的数据。

第九方面，本申请提供一种计算机装置，该计算机装置包括处理器和存储器，该存储器用于存储程序代码，该处理器用于处理存储器中的程序代码以执行第四方面或第五方面的数据处理方法。

第十方面，本申请提供一种计算机装置，该计算机装置包括处理器和存储器，该存储器用于存储程序代码，该处理器用于处理存储器中的程序代码以执行第六方面的数据处理方法。

第十一方面，本申请提供一种计算机存储介质，该计算机存储介质包括指令，当指令在计算机装置上运行时，使得计算机装置执行第四方面或第五方面的数据处理方法。

第十二方面，本申请提供一种计算机存储介质，该计算机存储介质包括指令，当指令在计算机装置上运行时，使得计算机装置执行第六方面的数据处理方法。

第十三方面，本申请提供一种计算机程序代码，当该计算机程序代码在计算机装置上运行时，使得计算机装置执行第四方面或第五方面的数据处理方法。

第十四方面，本申请提供一种计算机程序代码，当该计算机程序代码在计算机装置上运行时，使得计算机装置执行第六方面的数据处理方法。

附图说明

图 1 是一种 NCHW 格式的数据的结构以及对 NCHW 格式的数据进行计算的示意图。

图 2 是一种 NHWC 格式的数据的结构以及对 NHWC 格式的数据进行计算的示意图。

图 3 是一种 AI 系统的架构的示意图。

图 4 是本申请所提供的一种 AI 系统的架构的示意图。

图 5 是本申请所提供的 AI 系统中数据格式处理的示意图。

图 6 是本申请所提供的一种 AI 系统处理数据的实施例的流程示意图。

图 7 是一种将 NHWC 格式的数据转换为 5HD 格式的数据的示意图。

图 8 是本申请所提供的另一种 AI 系统处理数据的实施例的流程示意图。

图 9 是本申请所提供的另一种 AI 系统处理数据的实施例的流程示意图。

图 10 是本申请所提供的一种云服务场景的界面的示意图。

图 11 是一种数据处理装置的模块构成的示意图。

图 12 是另一种数据处理装置的模块构成的示意图。

图 13 为本申请所提供的一种计算机装置的结构示意图。

具体实施方式

首先对本申请所涉及的一些术语进行解释。

算子 (Operator): 是将一个元素在向量空间中转换为另一个元素的映射。例如，对数

据等元素进行算子操作可以理解为对该数据进行某种类型的运算，例如加、减、乘、除、微分或者卷积等计算。

映射层：存储在主机中的部分程序代码，当该部分程序代码被执行时，主机执行确定特征数据在加速卡中的存储器进行存储时采用的格式的操作，并执行或者指示加速卡执行数据格式转换的操作。映射层中包含不同数据格式之间的映射关系，而映射关系是指将一种格式的数据转换成另一种格式的数据的方法。

图 1 是一种 NCHW 格式的数据的结构以及对 NCHW 格式的数据进行计算的示意图。

以 RGB 三通道的数据为例进行说明。如图 1 所示，当数据采用 NCHW 格式进行排列时，通道层排列在最外侧，每个通道内的像素则挨在一起，形成 RRRGGGBBB 样式的数据（图中以白色方框表示 R 像素，以带有十字的方框表示 G 像素，以带有交叉的方框表示 B 像素）。当对采用 NCHW 格式的数据进行彩色转灰度计算时，是将 R 通道的所有像素值分别乘以第一参数，将 G 通道的所有像素值乘以第二参数，将 B 通道的所有像素值乘以第三参数，最后将三个通道的计算结果按照对应的像素分别进行相加，得到灰度像素的数据（图中以灰色方框表示灰度像素）。

图 2 是一种 NHWC 格式的数据的结构以及对 NHWC 格式的数据进行计算的示意图。

同样以 RGB 三通道的数据为例进行说明，如图 2 所示，当数据采用 NHWC 格式进行排列时，通道层排列在最内侧，多个通道中相同的空间位置的像素则挨在一起，即形成 RGBRGBRGB 样式的数据。当对采用 NHWC 格式的数据进行彩色转灰度计算时，则是将 NHWC 格式的数据分为多个 (R, G, B) 像素组，每个像素组中的 R 像素的值乘以第一参数，G 像素的值乘以第二参数，B 像素的值乘以第三参数，直接将该像素组中的乘法结果相加后得到一个灰度像素的值，再将多个灰度像素拼接起来得到所有的灰度像素。

从图 1 和图 2 中可以看出，在使用 NCHW 和 NHWC 这两种数据格式分别进行 RGB 像素到灰度像素的计算时，其复杂度基本上是相同的，区别点在于访存特性。在使用 NHWC 格式的数据进行计算时，每三个输入像素可以直接计算完成，得到一个输出像素，因此访存局部性较好，更适合 CPU 来执行；而 NCHW 则必须等到所有通道的输入像素都计算完成后才能计算得到最终的输出像素，需要占用较大的临时存储空间，更适合 GPU 执行。

图 3 是一种 AI 系统的架构的示意图。

如图 3 所示，AI 系统包括主机和加速卡。其中，主机包括处理器和存储器，存储器中包含逻辑层的程序代码，处理器执行存储器中的程序代码以实现逻辑层的功能。而逻辑层是用于处理接收到的 AI 任务和输入数据的程序代码，为了能够解析输入数据，通常逻辑层的代码会支持某种特定的输入数据的格式。存储器可以是随机存取存储器（Random Access Memory, RAM）等易失性存储器，也可以是闪存（flash）、只读存储器（Read Only Memory, ROM）等非易失性存储器。加速卡也包括处理器和存储器，存储器用于存储特征数据，而处理器则可以处理特征数据，得到 AI 任务的结果。主机接收客户端发送的推理或者训练任务，这些任务中带有特征数据。主机将特征数据发送给加速卡，并指示该特征数据在加速卡的存储器中的排布方式。特征数据在加速卡的存储器中的排布方式与逻辑层中指定的数据格式是相同的。而逻辑层作为连通客户端和加速卡的桥梁，从客户端中接收的数据格式也需要是逻辑层中指定的数据格式。也就是说，如果在逻辑层中确定了采用 NCHW 的数据格式存

储特征数据,那么客户端发送给主机的也只能是 NCHW 的特征数据,且加速卡也将采用 NCHW 的数据格式存储该特征数据。

图 3 所示的 AI 系统可以较好地支撑采用某一种固定格式的特征数据的处理。然而,当需要 AI 系统支持新的数据格式时,需要修改主机和加速卡中的程序代码进行适配,而这涉及到如下多个方面:

第一,从逻辑层面上,需要修改关于张量的逻辑和代码,包括规格、步长(stride)和形状等相关信息和模块;第二,当将特征数据写入加速卡中的存储器时,需要修改张量在存储器中的数据格式,以及关于数据读取和写入等相关信息和模块;第三,在算子层面,需要修改逻辑数据的输入、输出等相关逻辑;第四,在用户层面,需要新增内存排布转换逻辑,从而可以在不同场景中使用不同的数据格式。同时,加速卡制造商为了适配不同的 AI 框架,需要针对不同的模型和数据存储格式分别进行开发,且逻辑层面、存储层面和用户层面的资源无法被复用,从而造成了大量的资源浪费。

图 4 是本申请所提供的一种 AI 系统的架构的示意图。

如图 4 所示, AI 系统包括主机和加速卡。与图 3 中的 AI 系统相比,主机中的存储器包括了映射层的程序代码,这段代码被主机中的处理器执行时,使得主机可以将采用一种格式的数据转换成采用另一格式的数据,或者指示加速卡将采用一种格式的数据转换为采用另一种格式的数据。另外,映射层还可以包括确定加速卡应该采用什么样的格式来存储数据的代码,本申请不对映射层的程序代码的形式进行限定。

主机和加速卡之间通过总线或者网络进行连接。当通过总线进行连接时,主机可以通过周边部件高速互联(Peripheral Component Interconnect Express, PCIe)、计算快速互联(compute express link, CXL)、通用串行总线(Universal Serial Bus, USB)协议等与加速卡进行数据传输。而通过网络进行连接时,主机可以通过网线等有线网络与加速卡进行连接,也可以通过无线热点(WiFi)或者蓝牙(Bluetooth)等无线网络与加速卡进行连接。另一方面,加速卡可以直接插在主机的主板上的卡槽中,也可以通过线缆与主机进行连接;既可以位于主机的内部,也可以位于主机的外部,本申请不对加速卡与主机的连接方式和位置关系进行限定。

加速卡包括处理器、存储器等部件。其中,处理器包括负责对 AI 任务进行调度的任务调度器(task scheduler)以及用于处理 AI 任务的 AI 核心。AI 核心进一步包括加载/存储单元(Load/Store Unit, LSU)、标量(Scalar)计算单元、向量(Vector)计算单元等模块,其中,标量计算单元使一种单指令流单数据流(Single instruction Single data, SISD)的处理器,该类型处理器在同一时间内只处理一条数据(通常为整数或浮点数);向量计算单元又称数组处理器,是可以实现直接操作一组数组或向量进行计算的处理器,加载/存储单元用于加载待处理的数据以及存储处理后的数据。

当采用图 4 所示的 AI 系统时,主机可以从客户端接收 AI 任务以及相应的特征数据,并将该 AI 任务以及特征数据交给加速卡进行处理。其中,客户端既可以是主机以外的计算机装置,该计算机装置和主机通过网络等方式进行连接;客户端也可以是运行在主机上的软件程序,该软件程序在运行时为用户提供界面,用户可以通过该界面输入 AI 任务以及上传相应的特征数据,本申请不对客户端的形态进行限定。

图5是本申请所提供的AI系统中数据格式处理的示意图。

如图5所示，逻辑层包含所支持的从客户端接收的数据的格式的类型，例如，该AI系统的逻辑层仅包括NCHW格式相关的程序代码，从而该AI系统仅能处理用户通过客户端传输的NCHW格式的数据。而映射层则包含AI系统支持的输入数据的格式与多种其他数据格式的映射关系，例如5HD格式、NZ格式、FZ格式和NHWC格式，这意味着逻辑层的程序代码被运行时，可以将从客户端处接收的NCHW格式的数据转换为这些格式的数据或者指示加速卡根据映射关系进行该数据转换的操作。而进行格式转换后的数据可以以新的格式存储在加速卡的存储器中，并进行后续的计算。

其中，5HD格式也称为NC1HWC0格式，该种格式相对于NHWC格式而言，是将NHWC格式中的通道参数拆分为2个参数，C1和C0，而原先的通道参数C的值等于C1乘以C0。FZ格式也称为Fractal Z，是将数据块内部按照列进行排序，块之间按照行进行排序，采用小n大Z的方式；而NZ与FZ相对，采用大N小z的方式进行排序，即数据块内部按照行进行排序，块之间按照列进行排序。需要指出的是，上述数据格式的类型仅为举例说明，本申请不对AI系统所支持的输入数据的格式和转换后的目标数据格式的类型进行限定。

同时，AI系统，特别是其中的主机，通常会预设设置功能，供管理员通过该设置功能对AI系统进行设置。其中，这里的管理员指的主要是维护该AI系统的人，但在某些场景下，也可以是指发送AI任务给AI系统的用户或者其他与AI系统相关的使用者，本申请不对管理员的具体范围进行限定。

如下是对一些管理员对AI系统进行设置的示例：

第一，管理员可以确定AI系统所支持的来自客户端的数据的格式。

对于AI系统来说，当处理接收用户通过客户端输入的数据时，不仅需要在逻辑层有着处理一定格式的输入数据的程序代码，还需要在映射层具有将该输入数据转换成其他格式的数据的映射关系。因此，AI系统需要明确该AI系统所能支持的输入数据的格式，并将该格式的信息输出给用户，使得用户可以知悉AI系统所支持的输入数据的格式，并正确地输入数据。

当管理员确定所支持的输入数据的格式时，可以根据该AI系统所处理、涉及的全部或者主要的AI框架来进行。例如，当该AI系统主要用于处理PyTorch框架的AI任务时，而PyTorch框架的特征数据大多数采用NCHW格式，因此管理员可以将NCHW格式确定为该AI系统所支持的输入数据的格式。

第二，管理员可以维护AI系统保存的亲和性对应关系。

不同的算子，或者说计算类型，在某种数据格式下会计算的更快。因此，本申请可以根据在执行AI任务中输入数据所需要进行的计算的类型，确定输入数据在加速卡中存储的格式。而AI系统本身往往不能自行确认计算类型和数据格式的亲和性对应关系，因此，可以接收管理员发送的维护指令，该指令用于指示主机建立、修改或者删除计算类型和数据格式的亲和性对应关系，主机接收到指令后，根据该指令建立新的亲和性对应关系，或者对现有的某项亲和性对应关系进行修改或者删除。

第三，管理员可以调整可支持的输入数据的格式。

上文提到，管理员可以根据AI系统所处理的AI框架来确定所支持的输入数据的格式。

而当 AI 系统处理多种类型的 AI 框架时，则需要综合这些 AI 框架所支持的数据格式，并从中选择一个所有的 AI 框架均支持的数据格式；或者，使该 AI 系统支持多种输入数据的格式，而这多种输入数据的格式可以覆盖 AI 系统所处理的 AI 框架所支持的数据格式。

在这种情况下，当该 AI 系统所处理的 AI 框架发生变化时，AI 系统所支持的输入数据的格式可可能发生相应的变化。管理员可以给 AI 系统中的主机发送指令，通过该指令来使主机调整所支持的输入数据的格式，包括增加、修改或者删除所支持的数据格式。管理员还可以调整主机所存储的逻辑层和映射层的程序代码，一方面使得逻辑层能够支持调整后的输入数据的格式，另一方面也使得映射层包含调整后的输入数据的格式与 AI 系统所支持的转换后的目标数据的格式之间的映射关系。

第四，管理员可以调整数据转换的目标格式的类型。

随着科学技术的发展，可能会在 AI 任务中出现新的算子类型，并出现新的数据格式用于处理该类型的算子；也可能会有新的数据格式被创造，且新的数据格式被发现可以更高效率地处理现有的算子类型。在这种情况下，AI 系统可以根据管理员的指令，调整可供转换的目标格式。

图 6 是本申请所提供的一种 AI 系统处理数据的实施例的流程示意图。

S601：主机接收来自用户的输入数据。

AI 系统主要用来执行 AI 任务，包括训练 (training) 任务和推理 (inference) 任务。训练任务是指将大量的训练用的特征数据输入给 AI 系统，AI 系统利用神经网络分别对接收到的特征数据进行判断，并将得到的判断结果与真实结果进行对比，根据对比结果来调整神经网络中使用的模型，从而得到更准确的模型。而推理任务则是利用训练好的模型来解决真实的问题，例如图像识别、智能视频监控等。而无论是训练任务或者推理任务，往往都需要用户提供大量的输入数据，其中，输入数据包括表征特征图 (feature map) 的特征数据，在某些情况下也可以包括由用户提供的模型数据，例如模型的权值等。AI 系统将这些数据作为训练任务或者推理任务的输入，并后续处理这些数据。

AI 系统的主机从客户端处接收来自用户的 AI 任务，以及该 AI 任务对应的特征数据。其中，主机中的映射层包含了处理特征数据的逻辑。在一种可能的实现方式中，主机会输出主机所支持的输入数据的格式，从而知会用户这一信息。例如，在客户端给主机发送 AI 任务的界面中显示主机所支持的数据格式的信息。并且，主机可以解析用户发送的输入数据，并判断该特征数据是否满足格式要求。当该输入数据的格式不属于主机所支持的输入数据的格式时，主机向客户端返回错误提示，该错误提示用于指示客户端所发送的数据的格式有错误。

S602：主机将输入数据发送给加速卡。

当主机接收到来自用户的输入数据后，将该输入数据发送给加速卡。由于输入数据往往比较大，为了提高整个 AI 系统的效率，主机可以将接收到的输入数据直接透明传输给加速卡。

S603：确定输入数据需要转换的目标数据格式。

在本实施例中，主机中的存储器不仅包括了逻辑层对应的程序代码，还包括了映射层程序代码。而映射层的功能首先是确定 AI 系统接收到的输入数据应该采用什么样的格式存

储在加速卡中，以保证这部分输入数据能被加速卡正确、高效地处理。

针对输入数据所转换的数据格式，在本申请一种可能的实现方式中，主机根据该部分输入数据在本次 AI 任务中所涉及的算子类型，确定计算效率较高的一种数据格式。这是因为，某些类型的算子更适合采用某种特征的数据格式进行处理，也就是说，对于指定的计算类型而言，采用不同格式的数据进行计算会有着不同的效率，而该指定的计算类型也有着对应的效率最高的数据格式，这通常称为算子类型和数据格式之间的亲和性。例如，对于卷积计算，采用 5HD 格式时的计算效率更高；而进行矩阵乘法计算时，则采用 NZ 格式时的计算效率更高。

在这种情况下，主机可以根据输入数据所需要进行的计算类型，确定输入数据所需要转换的目标数据格式。例如，当输入数据在此次 AI 任务中主要进行卷积计算时，可以确定该输入数据需要转换成 5HD 格式的数据；而当输入数据既进行卷积计算，也进行矩阵乘法计算时，主机可以根据这两种计算类型在此次 AI 任务中的比例、或者不同数据格式在两种计算类型下的综合表现确定需要转换的目标数据格式，也可以在卷积计算时确定特征数据需要采用 5HD 格式的数据，而在完成卷积计算准备进行矩阵乘法计算时再确定将当前的数据转换为采用 NZ 格式的数据。

在本申请另一种可能的实现方式中，可以根据加速卡中的处理器类型确定输入数据所需要转换的目标数据格式。不同类型的处理器分别适合处理相应类型的格式的数据。例如，CPU 更适合处理 NWHC 格式的数据，而 GPU 更适合处理 NCHW 格式的数据。因此，主机可以获取处理此次 AI 任务的加速卡中的处理器类型的信息，并根据处理器的类型确定目标数据格式。

而在本申请另一种可能的实现方式中，可以由 AI 系统的管理员或者用户来指示主机确定输入数据需要转换的目标数据格式。例如，管理员可以在主机接收到 AI 任务后，确定输入数据需要转换的格式，并通过指令将这一信息发送给主机。主机接收到该指令后，将管理员确定的目标数据格式确定为输入数据需要转换的数据格式。

S604：主机指示加速卡对输入数据进行格式转换。

主机的映射层的另外的功能是指示加速卡对输入数据进行格式转换。当主机在步骤 S603 中确定了输入数据所采用的数据格式时，且如果该数据格式与主机所接收的数据的格式不相同，主机需要指示加速卡将先前接收的数据转换为新确定的数据格式。

映射层中包括了不同格式之间进行转换的映射关系，该映射关系可以是包括用于数据转换的函数的程序代码，当接收到一种格式的输入数据时，可以将该输入数据转换为其他格式的数据。映射层可以预设所支持的初始数据格式与一些常见的数据格式的映射关系，也可以后续跟进使用需要增加所支持的初始数据格式或者目的数据格式的类型。

以将 NHWC 格式转换为 5HD 格式为例进行说明。在该例子中，主机所接收到的输入数据的格式为 NHWC，而确定需要将该输入数据转换为 5HD 格式的数据。相应的，当进行数据格式转换时，通常的做法是：1) 先将 NHWC 格式的数据在通道维度进行分割，分出 C1 份的 NHWC0；2) 再将得到的 C1 份 NHWC0 格式的数据进行连续排布，形成 NC1HWC0 格式的数据。

图 7 是一种将 NHWC 格式的数据转换为 5HD 格式的数据的示意图。如图 7 所示，NHWC 格式的数据大小为 (1, 2, 2, 32)，也就是说，该数据的高度为 2，宽度为 2，对应一个 2*2

的矩阵，通道数为 32。而进行数据变换后，该数据在 NC1HWC0 格式下的数据大小为 (1, 2, 2, 2, 16)，其中 C1 的值为 2，C0 的值为 16。NHWC 格式的数据在进行排布时，是先排列 2*2 矩阵中第一个像素的全部 32 个通道的值，再排列第二个像素的全部 32 个通道的值，依次类推。而在 NC1HWC0 格式下，是先排列第一个像素的前 16 个通道的值，再排列第二个像素的前 16 个通道的值，直至全部像素的前 16 个通道均已排列完成，再开始排列这四个像素的后 16 个通道的值。

NHWC 格式的数据转换为 5HD 的数据可以参考以下公式：`Tensor.reshape([N, H, W, C1, C0]).transpose([0, 3, 1, 2, 4])`。类似的，NCHW 格式的数据转换为 5HD 的数据可以参考以下公式：`Tensor.reshape([N, C1, C0, H, W]).transpose([0, 1, 3, 4, 2])`。

由于典型的卷积神经网络随着层数的增加，其特征图在下采样后的长和宽的值逐渐减少，但通道数却往往随着卷积的过滤器的个数不断增加，例如，实际中经常会出现通道数为 128 或者 256 等很深的特征图。为了充分利用有限的矩阵计算单元，进行通道维度的拆分是很有必要的。

需要指出的是，本申请仅以 NHWC 和 NCHW 格式的数据转换为 5HD 格式的数据说明格式转换的过程，实际上，本申请不限定转换前的数据格式，也不限定转换后的数据格式。

另外，在本申请中，主机既可以在接收到来自客户端的数据后，先将数据直接发送给加速卡，再在确定数据所应该采用的存储格式后再指示客户端进行数据格式转换操作；也可以等确定数据所应该采用的数据格式后，再将数据格式转换的操作指令与操作指令所对应的这部分输入数据一起发送给加速卡，也就是说，步骤 S602 可以在步骤 S604 之前，也可以是与 S602 和 S604 一起进行的，本申请不对这几个步骤的先后顺序进行限定。

S605：指示加速卡不需要对输入数据进行格式转换。

当步骤 S604 中主机确定目标数据格式与主机所接收的输入数据的格式一致时，不需要加速卡再对输入数据进行格式转换。在这种情况下，主机可以给加速卡发送一条指令，指示加速卡在执行 AI 任务时不需要再对输入数据的格式进行转换；也可以不给加速卡发送指令，当加速卡接收到一段时间或者开始执行 AI 任务时还没有接收到主机发送的对输入数据进行格式转换的指令，就认为不需要对该输入数据进行格式转换，而直接对该输入数据进行处理。

图 8 是本申请所提供的另一种 AI 系统处理数据的实施例的流程示意图。

S801：主机接收来自用户的输入数据。

主机接收输入数据的操作可以参考步骤 S601，在此不再赘述。

S802：主机确定输入数据需要转换的数据格式。

主机确定输入数据需要转换的数据格式的操作可以参考步骤 S602，在此不再赘述。

S803：主机对输入数据进行格式转换。

当主机确定输入数据在加速卡中存储应该采用的数据格式后，根据接收到的数据的格式以及应该采用的数据格式，进一步确定是否需要输入数据进行转换。当接收到的数据的格式与应该采用数据格式不一致时，则主机确定对输入数据进行转换。

在图 6 所示的实施例中，是由主机指示加速卡完成输入数据的格式转换。而在本实施例中，主机直接根据映射层中接收到的数据的格式和确定的目标数据格式之间的映射关系，

对输入数据进行格式转换。

S804: 主机将转换格式后的数据发送给加速卡。

当完成格式转换后，主机将采用目标格式的输入数据发送给加速卡。加速卡直接存储主机发送的进行格式转换后的数据，并在后续处理该部分数据时不需要再进行格式转换。

图 6 和图 8 的实施例分别是由主机指示加速卡进行输入数据的格式转换以及主机自身进行输入数据的格式转换。需要指出的是，在本申请一种可能的实施方式中，这两个方式也可以结合起来，即当需要对输入数据完成多次数据转换时，可以分别由主机和加速卡执行一次或者多次的格式转换的操作。例如，当输入数据后续需要进行多种类型的计算时，可能会对输入数据进行多次格式转换。在这种情况下，可以由主机完成第一次的格式转换，并将格式转换后的数据发送给加速卡。而加速卡接收到该数据后，存储并使用格式转换后的数据完成相应类型的计算，并根据主机的指示再次转换数据的格式。

图 9 是本申请所提供的另一种 AI 系统处理数据的实施例的流程示意图。

S901: 加速卡接收来自主机的输入数据。

由于 AI 任务最终是由加速卡来执行，因此，当主机从客户端处接收到 AI 任务后，会将该 AI 任务以及对应的输入数据发送给加速卡。在一种可能的实现方式中，为了增加 AI 系统进行数据处理的效率，当主机接收到输入数据时，直接将该输入数据透明传输给加速卡。在这种情况下，加速卡接收到的输入数据的格式是初始的数据格式，主机没有进行格式转换操作。而在另一种可能的实现方式中，主机确定从客户端接收的输入数据需要进行多次格式转换，而主机先执行一次格式转换的操作，并将格式转换后的输入数据发送给加速卡，由加速卡在适当的时间点完成剩下的格式转换的操作。

S902: 加速卡接收主机的指令，该指令用于指示加速卡对输入数据进行格式转换。

主机的逻辑层具有确定输入数据需要转换的目标数据格式以及指示加速卡进行输入数据的格式转换的功能。当主机确定输入数据需要转换的数据格式和当前输入数据的数据格式不一致时，将向加速卡发送指令，该指令指示加速卡对输入数据进行格式转换。同时，指令中也包括当前输入数据的数据格式和目标数据格式之间的映射关系，关于映射关系的说明可以参考步骤 S504，此处不再赘述。

S903: 加速卡根据主机的指令，对输入数据进行格式转换。

加速卡接收到主机的指令之后，根据指令中所指示的需要转换的目标数据格式以及该目标数据格式与输入数据的格式的映射关系，对输入数据进行格式转换，并将完成格式转换后的输入数据写入存储器中。

S904: 加速卡处理完成格式转换后的数据以执行 AI 任务。

对输入数据进行格式转换的目的主要在于提高处理 AI 任务的效率。因此，当加速卡完成对输入数据的格式转换后，使用转换后的输入数据执行 AI 任务所涉及的计算。

本申请所提供的 AI 系统可以具有多种使用场景。在一种可能的实现方式中，该 AI 系统被部署在有 AI 方面业务、需要处理 AI 任务的公司内，专门用于执行该公司的 AI 任务。在这种情况下，该公司的员工作为用户通过客户端将 AI 任务以及对应的特征数据发送给该 AI 系统。而在另一种可能的实现方式中，可以将该 AI 系统所提供的 AI 方面的算力作为云服务提供的用户。在这种情况下，具有少量 AI 业务的公司不需要购买并部署 AI 系统，而

是可以直接购买云服务来处理这部分 AI 任务。而在云服务场景下，既可以将主机和加速卡都部署在共有云上，用户通过软件、网页等客户端远程向主机发送 AI 任务，也可以仅将加速卡部署在公有云上，主机与加速卡通过网络进行连接，而用户可以直接操作主机，本申请不对云服务场景下 AI 系统的架构进行限定。

图 10 是本申请所提供的一种云服务场景的界面的示意图。如图 10 所示，云服务场景的界面可以供用户进行如下操作：首先，用户可以选择所需要执行的 AI 任务的类型，例如是推理任务或者训练任务。其次，用户可以将 AI 任务涉及的数据上传到云服务系统。其中，这部分数据可以包括特征数据，也可以包括与模型相关的权值数据等。其中，云服务系统也可以提供部分模型，在这种情况下，用户可以不上传模型相关的数据，而是从云服务系统所提供的模型中选择此处 AI 任务所使用的模型。云服务系统可以指定所接收的数据的格式，例如支持用户输入 NCHW 格式的数据或者 NHWC 格式的数据，并检查用户所输入的数据的格式是否满足要求。当 AI 系统支持用户输入多种数据格式时，主机可以提供选项或者表单，供用户选择或者填写所输入的数据的格式。最后，用户可以启动 AI 任务，而这一步骤相当于使 AI 系统中的主机接收到了 AI 任务以及相应的数据，从而使得 AI 系统可以执行图 6、图 8 或者图 9 中的方法流程。

图 11 是一种数据处理装置 1100 的模块构成的示意图。其中，该数据处理装置 1100 为图 4 中的主机或者主机的一部分。如图 11 所示，该数据处理装置 1100 包括如下模块：

获取模块 1110，用于获取来自用户的 AI 任务以及与该 AI 任务对应的第一数据，该第一数据采用第一格式；

传输模块 1120，用于根据该 AI 任务，向加速卡发送第一指令，该第一指令用于指示加速卡将第一数据转换为采用第二格式的第二数据；

确定模块 1130，用于根据获取模块 1110 所获取的第一数据的格式信息，确定第一数据采用的是第一格式。

数据处理装置 1100 还可以包括转换模块 1140，用于根据 AI 任务，将第一数据转换为采用第二格式的第二数据。

其中，数据处理装置 1100 中的获取模块 1110、传输模块 1120、确定模块 1130 以及转换模块 1140 可以用于执行如图 6 和图 8 所示的流程。具体来说，获取模块 1110 用于执行图 6 中的步骤 S601 以及图 8 中的步骤 S801，传输模块 1120 用于执行图 6 中的步骤 S602、S604、S605 以及图 8 中的步骤 S804，确定模块用于执行图 6 中的步骤 S603 和图 8 中的步骤 S802，转换模块用于执行图 8 中的步骤 S803，此处不再展开赘述。

图 12 是另一种数据处理装置 1200 的模块构成的示意图。其中，该数据处理装置 1200 为图 4 中的加速卡或者加速卡的一部分。如图 12 所示，该数据处理装置 1200 包括以下模块：

传输模块 1210，用于接收主机发送的第一数据和第一指令，该第一数据采用第一格式；

转换模块 1220，用于根据第一指令，将第一数据转换为第二数据，该第二数据采用第二格式；

处理模块 1230，用于利用该第二数据执行 AI 任务。

其中，数据处理装置 1200 中的传输模块 1210、转换模块 1220 以及处理模块 1230 可

用于执行如图9所示的流程。具体来说,传输模块1210用于执行图9中的步骤S901和S902,转换模块1220用于执行图9中的步骤S903,处理模块1230用于执行图9中的步骤S904,此处不再赘述。

图13为本申请所提供的一种计算机装置1300的结构示意图。

本实施例中的计算机装置1300可以是上述各实施例中的计算机装置的其中一种具体实现方式,既可以为图4中的主机,也可以为图4中的加速卡。

如图13所示,计算机装置1300包括处理器1301,处理器1301与存储器1305连接。处理器1301可以为现场可编程门阵列(英文全称:Field Programmable Gate Array,缩写:FPGA),或数字信号处理器(英文全称:Digital Signal Processor,缩写:DSP)等计算逻辑或以上任意计算逻辑的组合。处理器1101也可以为单核处理器或多核处理器。

存储器1305可以是RAM存储器、闪存、ROM存储器、EPROM存储器、EEPROM存储器、寄存器或者本领域熟知的任何其它形式的存储介质,存储器可以用于存储程序指令,该程序指令被处理器1301执行时,处理器1301执行上述实施例中的所述的方法。

连接线1309用于在通信装置的各部件之间传递信息,连接线1309可以使用有线的连接方式或采用无线的连接方式,本申请并不对此进行限定。连接线1309还连接有网络接口1304。

网络接口1304使用例如但不限于电缆或电绞线一类的连接装置,来实现与其他设备或网络1311之间的通信,网络接口1304还可以通过无线的形式与网络1311互连。

本申请实施例的一些特征可以由处理器1301执行存储器1305中的程序指令或者软件代码来完成/支持。存储器1305上在加载的软件组件可以从功能或者逻辑上进行概括,例如,图11所示的获取模块1110、传输模块1120、确定模块1130以及转换模块1140或者图12所示的传输模块1210、转换模块1220以及处理模块1230。

在本申请的一个实施例中,当存储器1305加载程序指令后,处理器1301执行存储器1305中的上述功能/逻辑模块相关的事务。

此外,图13仅仅是一个计算机装置1300的例子,计算机装置1300可能包含相比于图13展示的更多或者更少的组件,或者有不同的组件配置方式。同时,图13中展示的各种组件可以用硬件、软件或者硬件与软件的结合方式实施。例如,存储器和处理器可以在一个模块中实现,存储器中的指令可以是预先写入存储器的,也可以是后续处理器在执行的过程中加载的,本申请不对此进行限定。

权 利 要 求 书

1. 一种人工智能 AI 系统，其特征在于，所述 AI 系统包括主机和加速卡，
所述主机用于：获取来自用户的 AI 任务以及与所述 AI 任务对应的第一数据，其中，所述
第一数据采用第一格式；
根据所述 AI 任务，向所述加速卡发送第一指令，所述第一指令用于指示所述加速卡将所
述第一数据转换为采用第二格式的第二数据；
所述加速卡用于：根据所述第一指令，将所述第一数据转换为所述第二数据；
利用所述第二数据执行所述 AI 任务。
2. 根据权利要求 1 所述的 AI 系统，其特征在于，
所述主机用于：根据所述 AI 任务包括的算子类型，向所述加速卡发送所述第一指令。
3. 根据权利要求 1 所述的 AI 系统，其特征在于，所述 AI 任务中包含第二指令，所述第
二指令用于指示采用所述第二格式数据执行所述 AI 任务；
所述主机用于：根据所述第二指令，向所述加速卡发送所述第一指令。
4. 根据权利要求 2 所述的 AI 系统，其特征在于，
所述主机还用于：建立所述 AI 任务包括的算子类型与所述第二格式之间的对应关系。
5. 根据权利要求 1-4 任一项所述的 AI 系统，其特征在于，所述主机存储将所述第一格
式的数据转换为所述第二格式的数据的第一映射关系，所述第一指令包含所述第一映射关系，
所述加速卡用于：根据所述第一映射关系，将所述第一数据转换为所述第二数据。
6. 根据权利要求 1-5 任一项所述的 AI 系统，其特征在于，
所述 AI 系统位于公有云中。
7. 根据权利要求 1-6 任一项所述的 AI 系统，其特征在于，
所述 AI 系统处理至少一种 AI 框架，其中，所述至少一种 AI 框架均支持所述第一格式
的数据。
8. 根据权利要求 1-7 任一项所述的 AI 系统，其特征在于，
所述主机还存储第二映射关系，其中，所述第二映射关系用于将所述第一格式的数据转
换为第三格式的数据。
9. 根据权利要求 1-8 任一项所述的 AI 系统，其特征在于，
所述主机还存储第三映射关系，其中，所述第三映射关系用于将第四格式的数据转换为
所述第二格式的数据。
10. 根据权利要求 1-9 任一项所述的 AI 系统，其特征在于，
所述主机还用于：获取所述第一数据的格式信息；
根据所述格式信息，确定所述第一数据采用的是所述第一格式。
11. 根据权利要求 1-10 任一项所述的 AI 系统，其特征在于，
所述主机还用于：输出提示信息，所述提示信息用于指示用户输入所述第一格式的数据。
12. 一种人工智能 AI 系统，其特征在于，所述 AI 系统包括主机和加速卡，
所述主机用于：获取来自用户的 AI 任务以及与所述 AI 任务对应的第一数据，其中，所
述第一数据采用第一格式；
根据所述 AI 任务，将所述第一数据转换为采用第二格式的第二数据；
向所述加速卡发送第二数据；

所述加速卡用于：接收所述第二数据；

利用所述第二数据执行所述 AI 任务。

13. 根据权利要求 12 所述的 AI 系统，其特征在于，

所述主机用于：根据所述 AI 任务包括的算子类型，将所述第一数据转换为所述第二数据。

14. 根据权利要求 12 所述的 AI 系统，其特征在于，所述 AI 任务中包含第二指令，所述第二指令用于指示采用所述第二格式的数据执行所述 AI 任务；

所述主机用于：根据所述第二指令，将所述第一数据转换为所述第二数据。

15. 根据权利要求 13 所述的 AI 系统，其特征在于，所述主机还用于：建立所述 AI 任务包括的算子类型与所述第二格式之间的对应关系。

16. 根据权利要求 12-15 任一项所述的 AI 系统，其特征在于，所述主机存储将所述第一格式的数据转换为所述第二格式的数据的第一映射关系，

所述主机用于：根据所述第一映射关系，将所述第一数据转换为所述第二数据。

17. 根据权利要求 12-16 任一项所述的 AI 系统，其特征在于，

所述 AI 系统位于所述公有云中。

18. 根据权利要求 12-17 任一项所述的 AI 系统，其特征在于，

所述 AI 系统处理至少一种 AI 框架，其中，所述至少一种 AI 框架均支持所述第一格式的数据。

19. 根据权利要求 12-18 任一项所述的 AI 系统，其特征在于，

所述主机还存储第二映射关系，其中，所述第二映射关系用于将所述第一格式的数据转换为第三格式的数据。

20. 根据权利要求 12-19 任一项所述的 AI 系统，其特征在于，

所述主机还存储第三映射关系，其中，所述第三映射关系用于将第四格式的数据转换为所述第二格式的数据。

21. 根据权利要求 12-20 任一项所述的 AI 系统，其特征在于，

所述主机还用于：获取所述第一数据的格式信息；

根据所述格式信息，确定所述第一数据采用的是所述第一格式。

22. 根据权利要求 12-21 任一项所述的 AI 系统，其特征在于，

所述主机还用于：输出提示信息，所述提示信息用于指示用户输入所述第一格式的数据。

23. 一种数据处理的方法，其特征在于，所述方法用于人工智能 AI 系统中的主机，所述 AI 系统还包括加速卡，所述方法包括：

获取来自用户的 AI 任务以及与所述 AI 任务对应的第一数据，其中，所述第一数据采用第一格式；

根据所述 AI 任务，向所述加速卡发送第一指令，所述第一指令用于指示所述加速卡将所述第一数据转换为采用第二格式的第二数据。

24. 根据权利要求 23 所述的方法，其特征在于，所述根据所述 AI 任务，向所述加速卡发送第一指令，包括：

根据所述 AI 任务包括的算子类型，向所述加速卡发送所述第一指令。

25. 根据权利要求 23 所述的方法，其特征在于，所述 AI 任务中包含第二指令，所述第

二指令用于指示采用所述第二格式的数据执行所述 AI 任务；

所述根据所述 AI 任务，向所述加速卡发送第一指令，包括：

根据所述第二指令，向所述加速卡发送所述第一指令。

26. 根据权利要求 24 所述的方法，其特征在于，所述方法还包括：

建立所述 AI 任务包括的算子类型与所述第二格式之间的对应关系。

27. 根据权利要求 23-26 任一项所述的方法，其特征在于，所述主机存储将所述第一格式的数据转换为所述第二格式的数据的第一映射关系，所述第一指令包括所述第一映射关系。

28. 根据权利要求 23-27 任一项所述的方法，其特征在于，所述 AI 系统位于公有云中。

29. 根据权利要求 23-28 任一项所述的方法，其特征在于，所述 AI 系统处理至少一种 AI 框架，其中，所述至少一种 AI 框架均支持所述第一格式的数据。

30. 根据权利要求 23-29 任一项所述的方法，其特征在于，所述主机还存储第二映射关系，其中，所述第二映射关系用于将所述第一格式的数据转换为第三格式的数据。

31. 根据权利要求 23-30 任一项所述的方法，其特征在于，所述主机还存储第三映射关系，其中，所述第三映射关系用于将第四格式的数据转换为所述第二格式的数据。

32. 根据权利要求 23-31 任一项所述的方法，其特征在于，所述方法还包括：

获取所述第一数据的格式信息；

根据所述格式信息，确定所述第一数据采用的是所述第一格式。

33. 根据权利要求 23-25 任一项所述的方法，其特征在于，所述方法还包括：

输出提示信息，所述提示信息用于指示用户输入所述第一格式的数据。

34. 一种数据处理的方法，其特征在于，所述方法应用于人工智能 AI 系统中的加速卡，所述 AI 系统还包括主机，所述方法包括：

接收主机发送的第一数据和第一指令，其中，所述第一数据采用第一格式；

根据所述第一指令，将所述第一数据转换为第二数据，其中，所述第二数据采用第二格式；

利用所述第二数据执行 AI 任务，其中，所述第一数据与所述 AI 任务对应。

35. 根据权利要求 34 所述的方法，其特征在于，所述第一指令包括第一映射关系，所述根据所述第一指令，将所述第一数据转换为第二数据，包括：

根据所述第一映射关系，将所述第一数据转换为所述第二数据。

36. 根据权利要求 34 或 35 所述的方法，其特征在于，所述第二格式与所述 AI 任务包括的算子类型相对应。

37. 根据权利要求 34-36 任一项所述的方法，其特征在于，所述 AI 系统位于公有云中。

38. 根据权利要求 34-37 任一项所述的方法，其特征在于，所述方法还包括：

接收所述主机发送的第二映射关系，其中，所述第二映射关系用于将所述第一格式的数据转换为第三格式的数据。

39. 根据权利要求 34-38 任一项所述的方法，其特征在于，所述方法还包括：

接收所述主机发送的第三映射关系，其中，所述第三映射关系用于将第四格式的数据转换为所述第二格式的数据。

40. 根据权利要求 34-39 任一项所述的方法，其特征在于，所述 AI 系统处理至少一种 AI 框架，其中，所述至少一种 AI 框架均支持所述第一格式的数据。

41. 一种数据处理装置，其特征在于，所述数据处理装置位于人工智能 AI 系统中，所述 AI 系统还包括加速卡，所述数据处理装置包括：

获取模块，用于获取来自用户的 AI 任务以及与所述 AI 任务对应的第一数据，所述第一数据采用第一格式；

传输模块，用于根据所述 AI 任务，向所述加速卡发送第一指令，所述第一指令用于指示所述加速卡将所述第一数据转换为采用第二格式的第二数据。

42. 根据权利要求 41 所述的数据处理装置，其特征在于，

传输模块用于，根据所述 AI 任务包括的算子类型，向所述加速卡发送所述第一指令。

43. 根据权利要求 41 所述的数据处理装置，其特征在于，所述 AI 任务包括所述第二指令，所述第二指令用于指示采用所述第二格式的数据执行所述 AI 任务，其中，

所述传输模块用于，根据所述第二指令，向所述加速卡发送所述第一指令。

44. 根据权利要求 41-43 任一项所述的数据处理装置，其特征在于，

所述获取模块还用于，建立所述 AI 任务包括的算子类型和所述第二格式之间的对应关系。

45. 根据权利要求 41-44 任一项所述的数据处理装置，其特征在于，

所述数据处理装置存储将第一格式的数据转换为所述第二格式的数据的第一映射关系，所述第一指令包括所述第一映射关系。

46. 根据权利要求 41-45 任一项所述的数据处理装置，其特征在于，

所述 AI 系统位于公有云中。

47. 根据权利要求 41-46 任一项所述的数据处理装置，其特征在于，

所述 AI 系统处理至少一种 AI 框架，所述至少一种 AI 框架均支持所述第一格式的数据。

48. 根据权利要求 41-47 任一项所述的数据处理装置，其特征在于，

所述数据处理装置还存储第二映射关系，所述第二映射关系用于将所述第一格式的数据转换为第三格式的数据。

49. 根据权利要求 41-48 任一项所述的数据处理装置，其特征在于，

所述数据处理装置还存储第三映射关系，所述第三映射关系用于将第四格式的数据转换为所述第二格式的数据。

50. 根据权利要求 41-49 任一项所述的数据处理装置，其特征在于，

所述获取模块还用于，获取第一数据的格式信息；

所述数据处理装置还包括确定模块，用于确定第一数据采用的是所述第一格式。

51. 根据权利要求 41-50 任一项所述的数据处理装置，其特征在于，

所述传输模块还用于：输出提示信息，所述提示信息用于指示用户输入所述第一格式的数据。

52. 一种数据处理装置，其特征在于，所述数据处理装置位于人工智能 AI 系统中，所述 AI 系统还包括主机，所述数据处理装置包括：

传输模块，用于：接收所述主机发送的第一数据和第一指令，所述第一数据采用所述第一格式；

转换模块，用于：根据所述第一指令，将所述第一数据转换为所述第二数据，所述第二

数据采用第二格式；

处理模块，用于：利用所述第二数据执行 AI 任务，其中，所述第一数据与所述 AI 任务对应。

53. 根据权利要求 52 所述的数据处理装置，其特征在于，所述第一指令包括第一映射关系，

所述转换模块用于：根据所述第一映射关系，将所述第一数据转换为所述第二数据。

54. 根据权利要求 52 或 53 所述的数据处理装置，其特征在于，所述第二格式与所述 AI 任务包括的算子类型相对应。

55. 根据权利要求 52-54 任一项所述的数据处理装置，其特征在于，所述 AI 系统位于公有云中。

56. 根据权利要求 52-55 任一项所述的数据处理装置，其特征在于，

所述传输模块还用于：接收所述主机发送的第二映射关系，所述第二映射关系用于将所述第一格式的数据转换为第三格式的数据。

57. 根据权利要求 52-56 任一项所述的数据处理装置，其特征在于，

所述传输模块还用于：接收所述主机发送的第三映射关系，所述第三映射关系用于将第四格式的数据转换为所述第二格式的数据。

58. 根据权利要求 52-57 任一项所述的数据处理装置，其特征在于，所述 AI 系统处理至少一种 AI 框架，所述至少一种 AI 框架均支持所述第一格式的数据。

59. 一种计算机装置，其特征在于，所述计算机装置包括存储器和处理器，所述存储器用于存储程序代码，所述处理器用于处理所述程序代码以执行如权利要求 23-33 任一项所述的数据处理方法。

60. 一种计算机装置，其特征在于，所述计算机装置包括存储器和处理器，所述存储器用于存储程序代码，所述处理器用于处理所述程序代码以执行如权利要求 34-40 任一项所述的数据处理方法。