通信方法、装置和系统

本申请要求于 2024 年 3 月 21 日提交中国国家知识产权局、申请号为 202410332551.1、申请名称为“通信方法、装置和系统”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

本申请涉及通信技术领域，并且更具体地，涉及一种通信方法、装置和系统。

背景技术

大模型的蓬勃发展催生了各行各业的应用。云侧大模型可以指部署在云侧设备或边侧设备的大模型。端侧大模型可以指部署在端侧设备的大模型。端侧大模型可以满足用户保护端侧隐私数据的需求，从而备受关注。然而，端侧设备的存储空间有限，仅能存储参数量较小的大模型，而小参数大模型的性能远低于大参数量的大模型的性能，从而不利于大模型在端侧的应用。

因此，在端侧大模型推理过程中，如何应对大模型参数量过大与存储空间过小的矛盾，进而提高端侧大模型的推理性能，是亟待解决的问题。

发明内容

本申请提供一种通信方法、装置和系统，能够提高端侧大模型的推理性能。

第一方面，提供了一种通信方法，该方法可以由第一装置（例如，终端设备或端侧设备）执行，也可以由第一装置中的组件（例如，处理器、芯片、或芯片系统等）执行，或者是由能实现全部或部分第一装置的功能的逻辑模块或软件执行。

该方法包括：接收第一信息，该第一信息用于指示第一活性参数，该第一活性参数是第一模型中的 N 个第一模块的活性参数，N 为正整数；根据该第一活性参数，基于该 N 个第一模块进行推理。

通过上述实施例，用于推理的第一装置可以不存储第一模块的全量参数，而是通过第一信息获取第一模块的全量参数中的活性参数。与将全量参数存储在第一装置本地，并从全量参数中读取出活性参数的方案相比，上述实施例能够节约第一装置的存储空间，使得存储空间有限的第一装置能够运行大参数量的大模型，从而提高了大模型的推理性能。例如，第一模块可以是变换器（transformer）结构的大模型中的前馈神经网络（feedforward neural network，FNN）模块，FNN 模块的全量参数与多头注意力（multi-head attention，MHA）模块的全量参数的比例约为 2:1，如果 FNN 模块的全量参数不存储端侧设备中，将节约端侧设备三分之二的闪存空间。节约出的闪存空间可以存储其他模块的参数，例如 MHA 模块的参数，这样，大模型的参数量可以提高三倍，从而有效提高大模型的性能。

在一些实现方式中，该 N 个第一模块包括 N1 个第一模块，该 N1 个第一模块是该第一模型中的至少一个层中的第一个层的模块，N1 是小于或等于 N 的正整数。

通过上述实施例，第一装置可以基于第一信息指示的活性参数，基于第一模型中的第一个层中的模块进行推理。与根据全量参数和第一个层中的模块进行推理的方案相比，本申请实施例无需将第一个层中的模块的全量参数加载到内存，从而节约了第一装置的内存空间，从而进一步提升了第一模型的推理性能。

在一些实现方式中，该 N1 个第一模块是该第一模型的第一个层中的多个模块中的第二个，N1=1，该方法还包括：根据第二输入值，预测该第一活性参数的位置的信息，其中，该第二输入值为该第一模型的第一个层中的多个模块中的第二个的输入值或该第一模型的第一个层中的多个模块中的第一个的输出值。

通过上述实施例，第一装置可以预测第一模块中的第一个层中的多个模块中的第二个的活性参数，从而根据多个模块中的第二个的活性参数，基于多个模块中的第二个进行推理。与根据全量参数和第一个层中的多个模块中的第二个进行推理的方案相比，本申请实施例无需将第一个层中的多个模块中的第二个的全量参数加载到内存，从而节约了第一装置的内存空间，从而进一步提升了第一模型的推理性能。

在一些实现方式中，该方法还包括：根据第一输入值，预测第二活性参数的位置的信息，该第二活性参数是该第一模型的第一个层中的多个模块中的第一个的活性参数，该第一输入值是该第一模型的第一个层中的多个模块中的第一个的输入值。

通过上述实施例，第一装置可以预测第一模块中的第一个层中的第一个的活性参数，从而根据多个模块中的第一个的活性参数，基于多个模块中的第一个进行推理。与根据全量参数和第一个层中的多个模块中的第一个进行推理的方案相比，本申请实施例无需将第一个层中的多个模块中的第一个的全量参数加载到内存，从而节约了第一装置的内存空间，从而进一步提升了第一模型的推理性能。

在一些实现方式中，该 N 个第一模块还包括 N2 个第一模块，该第一模型包括 N2 个层，该 N2 个层中的每个层包括该 N2 个第一模块中的一个第一模块，N2≥2，N2 是小于或等于 N 的正整数。

通过上述实施例，第二装置可以一次下发多个层的活性参数，减少了下发活性参数的次数，从而减少了频繁下发活性参数所导致的信令开销。

在一些实现方式中，该第一模型包括 M 个层，M 是大于 N2 的正整数，该 N2 个层在该 M 个层中的编号不连续。

通过上述实施例，第一信息所指示的第一活性参数可以包括编号不连续的层的活性参数，即可以这些活性参数可以跳过一个或多个层，从而减少了传输第一信息的信息量，降低了时延。

在一些实现方式中，N2 与用于传输该第一信息的资源和/或推理时延要求相关。

通过上述实施例，N2 的取值可以与用于传输该第一信息的资源和/或推理时延要求相关，从而使得第一信息的传输可以满足资源限制和/或推理时延要求。上述方案能够保障第一信息能够顺利传输，或者，能够保障第一信息的传输可以满足第一模型的推理时延要求，从而进一步保障第一模型的推理性能。

在一些实现方式中，该方法还包括：根据该 N2 个层中第一个层的上一个层的输入值，预测该第一活性参数的位置的信息。

通过上述实施例，一方面，第一装置可以预测多个层分别包括的多个第一模块的活性参数的位置的信息。与一次仅预测一个模块的活性参数的方案相比，上述实施例能够节约计算资源。另一方面，第一装置预测 N2 个层是根据 N2 个层中第一个层的上一个层的输入值，而不是根据 N2 个层中的第一个层的输入值。这样，第一装置可以在推理 N2 个层中第一个层的上一个层时，预测 N2 个层中第一模块的活性参数的位置的信息。也就是说，第一装置预测活性参数的位置所针对的层不是正在推理的层，因此上述实施例不会影响正在推理的层，即不会降低大模型的推理性能。

在一些实现方式中，该 N2 个层中的每个层还包括一个第二模块，该一个第二模块是该第一模型中的 N2 个第二模块之一，该方法还包括：根据该 N2 个层中第一个层的上一个层的输入值，预测第三活性参数的位置的信息，该第三活性参数是该 N2 个第二模块的活性参数。

通过上述实施例，一方面，第一装置可以预测多个层分别包括的多个第二模块的活性参数的位置的信息。与一次仅预测一个模块的活性参数的方案相比，上述实施例能够节约计算资源。另一方面，第一装置预测活性参数的位置所针对的层不是正在推理的层，因此上述实施例不会影响正在推理的层，即不会降低大模型的推理性能。

在一些实现方式中，该第二模块为 MHA 模块。

在一些实现方式中，该 N2 个层中的每个层还包括一个第二模块，该方法还包括：根据该 N2 个层中的第 k 层中的第一模块的输出值与该第 k 层中的第二模块的输出值，确定第三输入值，该第三输入值用于该第 k 层的下一层中的第一模块的输入，以及该第 k 层的该下一层中的第二模块的输入，k 是小于或等于 N2 的正整数。

通过上述实施例，N2 个层中的一个或多个层可以采取平行推理的形式，也就是一个层的第一模块在进行推理时，这个层的第二模块也可以进行推理。顺序推理是将一个层的第二模块推理结束后得到的输出，作为这个层的第二模块的输入。采用平行推理的层，相比于采用顺序推理的层，输入的数量可以从 2 个减少为 1 个。输入的数量减少，相邻的输入的变化率一定，可以增加能够选取的 N2 的数值，使得第二装置下发的第一信息所指示的活性参数可以是更多个层的活性参数，换言之，第二装置可以一次性下发更多的层的活性参数，从而进一步减少了第二装置向第一装置传输的时延。

在一些实现方式中，该方法还包括：在该第一模型的输出值的置信度大于或等于第一阈值的情况下，发送第二信息，该第二信息用于指示停止传输该第一信息。

通过上述实施例，在第一模型的输出值的置信度满足条件的情况下，第一装置可以发送第二信息，

第二信息可以指示停止传输第一信息，从而减少传输第一信息所导致的资源浪费。此外，上述实施例可以应用于推理提前退出（early exiting）的场景，在第一模型提前退出推理的场景下，第一模型的退出推理不是由于第一模型的所有层都完成了推理，而是由于第一模型的输出值的置信度满足条件。推理提前退出可以使得第一模型中的一部分层不参与推理，从而减少了推理时延。上述实施例能够在第一模型提前退出推理的情况下，避免第二装置传输指示用于推理的活性参数的信息，从而节约了传输资源。

在一些实现方式中，该方法还包括：发送第三信息，该第三信息包括该第一活性参数的位置的指示信息，该第三信息用于请求该第一活性参数。

通过上述实施例，第一装置可以请求下发第一活性参数，第一装置发送的请求中可以携带第一活性参数的位置的指示信息，从而使得第一信息所指示的活性参数的取值准确。

在一些实现方式中，该第一模块为 FNN 模块。

第二方面，提供了一种通信方法，该方法可以由第二装置（例如，网络设备、核心网设备、云侧设备或边侧设备）执行，也可以由第二装置中的组件（例如，处理器、芯片、或芯片系统等）执行，或者是由能实现全部或部分第二装置的功能的逻辑模块或软件执行。

该方法包括：生成第一信息，该第一信息用于指示第一活性参数，该第一活性参数是第一模型中的 N 个第一模块的活性参数，该第一活性参数用于基于该 N 个第一模块进行推理，N 为正整数；发送该第一信息。

在一些实现方式中，该 N 个第一模块包括 N1 个第一模块，该 N1 个第一模块是该第一模型中的至少一个层中的第一个层的模块，N1 是小于或等于 N 的正整数。

在一些实现方式中，该 N 个第一模块还包括 N2 个第一模块，该第一模型包括 N2 个层，该 N2 个层中的每个层包括该 N2 个第一模块中的一个第一模块，N2≥2，N2 是小于或等于 N 的正整数。

在一些实现方式中，该第一模型包括 M 个层，M 是大于 N2 的正整数，该 N2 个层在该 M 个层中的编号不连续。

在一些实现方式中，N2 与用于传输该第一信息的资源和/或推理时延要求相关。

在一些实现方式中，该方法还包括：接收第二信息，该第二信息用于指示停止传输该第一信息。

在一些实现方式中，该方法还包括：接收第三信息，该第三信息包括该第一活性参数的位置的指示信息。

在一些实现方式中，发送第一信息，包括：响应于该第三信息，发送该第一信息。

在一些实现方式中，第一模块为 FNN 模块。

第三方面，提供了一种通信装置，包括处理电路（或者处理器）和输入输出接口（也可以称为接口电路），该输入输出接口用于输入和/或输出信号，该处理电路用于执行第一方面以及第一方面的任一种可能的方法，或者，该处理电路用于执行第二方面以及第二方面的任一种可能的方法。

在一些实现方式中，处理电路用于通过接口电路与其它装置通信，并执行上述第一方面以及第一方面的任一种可能的方法，或者执行第二方面以及第二方面的任一种可能的方法。

第四方面，提供了一种通信装置。该通信装置可以包括用于执行通信装置功能的单元或者模块等。

在一些实现方式中，该通信装置可以包括用于执行第一方面以及第一方面的任一种可能的实现方式所描述的方法/操作/步骤/动作所一一对应的模块或单元或手段（means），该模块或单元或手段可以是硬件电路，也可是软件，也可以是硬件电路结合软件实现。

在一些实现方式中，该通信装置可以包括用于执行第二方面以及第二方面的任一种可能的实现方式所描述的方法/操作/步骤/动作所一一对应的模块或单元或手段，该模块或单元或手段可以是硬件电路，也可是软件，也可以是硬件电路结合软件实现。

第五方面，提供了一种计算机可读存储介质，该计算机可读存储介质上存储有计算机程序或指令，当该计算机程序或该指令在计算机上运行时，使得第一方面以及第一方面的任一种可能的方法被执行，或者，使得第二方面以及第二方面的任一种可能的方法被执行。

第六方面，提供了一种计算机程序产品，包含计算机程序或指令，当该计算机程序或指令在计算机上运行时，使得第一方面以及第一方面的任一种可能的方法被执行（或被实现），或者，使得第二方面以及第二方面的任一种可能的方法被执行（或被实现）。

第七方面，提供一种通信装置，包括处理器，用于通过执行存储器中存储的计算机程序（或计算机可执行指令），和/或，通过逻辑电路，使得该装置执行上述第一方面的任一种可能的方法，或者，使得

该装置执行上述第二方面的任一种可能的方法。

在一种可能的实现中，该装置还包括存储器。在一种可能的实现中，处理器和存储器集成在一起。在另一种可能的实现中，该存储器位于该通信装置之外。该处理器可以包括一个或多个。

在一种可能的实现中，该通信装置还包括通信接口，该通信接口用于该通信装置与其他设备进行通信，例如数据和/或信号的发送或接收。示例性的，通信接口可以是收发器、电路、总线、模块或其它类型的通信接口。

在一种实现方式中，上述第三方面、第四方面或第七方面的通信装置，可以是芯片或芯片系统。

第八方面，提供一种芯片，包括处理器，用于调用存储器中的计算机程序或计算机指令，以使得该处理器执行上述第一方面中的任一种实现方式，或者，以使得该处理器执行上述第二方面中的任一种实现方式。

在一些实现方式中，该处理器通过接口与该存储器耦合。

第九方面，提供一种通信系统，包括第一装置和第二装置，第一装置用于执行上述第一方面以及第一方面的任一种可能的实现方式，第二装置用于执行上述第二方面以及第二方面的任一种可能的实现方式。

关于第二方面至第九方面等中任一方面的有益效果的描述可以参照第一方面的有益效果的描述。

附图说明

图 1 是一种通信系统的示意图。

图 2 是另一种通信系统的示意性框图。

图 3 是一种推理方法的示意图。

图 4 是本申请实施例提供的一种通信方法的示意性流程图。

图 5 是本申请实施例提供的另一种通信方法的示意性流程图。

图 6 是本申请实施例提供的又一种通信方法的示意性流程图。

图 7 是本申请实施例提供的再一种通信方法的示意图。

图 8 是本申请实施例提供的再一种通信方法的示意性流程图。

图 9 是本申请实施例提供的用于平行推理的大模型的示意图。

图 10 是本申请实施例的一种通信装置的示意框图。

图 11 是本申请实施例的另一种通信装置的示意框图。

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。

本申请将围绕可包括多个设备、组件、模块等的系统来呈现各个方面、实施例或特征。应当理解和明白的是，各个系统可以包括另外的设备、组件、模块等，并且/或者可以并不包括结合附图讨论的所有设备、组件、模块等。此外，还可以使用这些方案的组合。

另外，在本申请实施例中，“示例的”“例如”等词用于表示作例子、例证或说明。本申请中被描述为“示例”的任何实施例或设计方案不应被解释为比其它实施例或设计方案更优选或更具优势。确切而言，使用示例的一词旨在以具体方式呈现概念。

本申请实施例描述的业务场景是为了更加清楚的说明本申请实施例的技术方案，并不构成对于本申请实施例提供的技术方案的限定，本领域普通技术人员可知，随着新业务场景的出现，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。

在本说明书中描述的参考“一个实施例”或“一些实施例”等意味着在本申请的一个或多个实施例中包括结合该实施例描述的特定特征、结构或特点。由此，在本说明书中的不同之处出现的语句“在一个实施例中”“在一些实施例中”“在其他一些实施例中”“在另外一些实施例中”等不是必然都参考相同的实施例，而是意味着“一个或多个但不是所有的实施例”，除非是以其他方式另外特别强调。术语“包括”“包含”“具有”及它们的变形都意味着“包括但不限于”，除非是以其他方式另外特别强调。

本申请实施例中出现的第一、第二等描述，在没有特殊说明的情况下，仅作示意与区分描述对象之用，没有次序之分，也不表示本申请实施例中对个数的特别限定，不能构成对本申请实施例的任何限制。

应理解，在本申请的各种实施例中，各过程的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本申请实施例的实施过程构成任何限定。

可以理解的是，本文中术语“和/或”，仅仅是一种描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。另外，本文中字符“/”，一般表示前后关联对象是一种“或”的关系。

为方便理解本申请，下面对本申请中可能出现的术语进行示例性的解释。

MHA 模块（block）：MHA 模块可以是大模型的变换器（transformer）结构中的多头注意力机制的模块。其中，大模型可以是人工智能（artificial intelligent，AI）模型或机器学习（machine learning，ML）模型。

FNN 模块（block）：FNN 模块可以是大模型的 transformer 结构中置于 MHA 模块后面的前馈神经网络模块。其中，FNN 模块的输入可以是 MHA 模块的输出。

变换器（transformer）层：一个 transformer 层可以包括 MHA 模块和至少一个 FNN 模块按顺序。其中，一个层中的 MHA 模块的输出可以作为这个层中 FNN 模块的输入。大模型可以包括 transformer 结构；一个 transformer 结构包括一个或多个 transformer 层。下面以一个 transformer 层包括一个 MHA 模块和一个 FNN 模块举例，但是需要说明的是，本申请对于一个 transformer 层包括多个 MHA 模块和多个 FNN 模块的情况，同样适用。

输入（input）：input 可以是大模型的输入，或者说，对大模型的第一个层的多个模块中的第一个的输入。例如，input 可以是用户对大模型发出的请求经过嵌入（embedding）后的向量或张量又例如，input 可以是大模型由事件触发或定期地从用户侧（例如端侧设备）获取的数据。input 也可以是输入到大模型的中间的一个 block（例如，FNN 模块或者 MHA 模块）的输入。例如，input 可以是输入到大模型的第一个层的多个模块中的第二个的向量或张量。又例如，input 可以是输入到大模型的第二个层或第二个层后面的任意一个层中的模块的向量或张量。

预测模块（predictor）：predictor 可以是 AI 模型或 ML 模型。例如，predictor 可以是经过训练的小型的深度神经网络（deep-learning neural network，DNN）。predictor 可以根据 input，预测 MHA 模块或者 FNN 模块的活性参数的位置。

活性参数（activated parameter）：活性参数可以包括活性神经元（activated neuron）的权重参数和/或活性头（activated head）所对应的权重参数。

活性神经元：FNN 模块的输出可以表示为 $\sigma(xW+b)$。其中，σ 表示激活函数，例如，激活函数可以是整流线性函数（rectified linear unit，ReLU）；x 可以表示 FNN 模块的输入（input）；W 可以表示权重参数（例如，权重矩阵）；b 可以表示偏置参数（例如，偏置矩阵）。对 FNN 模块来说，活性参数可以是活性神经元的权重参数（或者称为在权重矩阵中的元素）。其中，活性神经元可以是与该 FNN 模块的 input 的内积超过预设阈值的权重参数（或者称为神经元权重参数）所对应的神经元。例如，如果一个输入 x_1 根据 x_1W+b 进行计算后，得到的结果是负数，或者是很小的正数，即低于预设阈值，在该结果经过激活函数进一步处理后，得到的输出值 $\sigma(x_1W+b)$对该 FNN 模块最终的输出结果几乎没有影响。因此，用于计算该 x_1 的相关神经元可以是非活性的神经元。上述非活性的神经元在矩阵 W 中对应的元素可以置 0，得到矩阵 W’。矩阵 W’可以是绝大多数元素为 0 元素的稀疏矩阵，基于矩阵 W’进行推理能够节约计算资源、减小推理时延。又例如，如果一个输入 x_2 根据 x_2W+b 进行计算后，得到的结果超过预设阈值，那么说明该结果在经过激活函数进一步处理后，得到的输出值 $\sigma(x_2W+b)$对该 FNN 模块最终的输出结果有一定的影响。因此，用于计算该 x_2 的相关神经元是活性神经元。上述活性神经元在矩阵 W 中对应的元素可以保持不变，从而使得处理后的矩阵 W’在再次进行推理时，上述活性神经元可以参与推理。

活性头（head）：将 MHA 模块的 input 与权重参数（例如权重矩阵）W_q、W_k和 W_v分别相乘，得到 Q、K 和 V 三个矩阵。假设 MHA 模块的 head 的个数为 n，可以将矩阵 Q 划分为 n 个子矩阵，分别为 Q_1 至 Q_n，其中每个子矩阵对应一个 head，n 为正整数。例如，矩阵 Q 为 512*1（512 行 1 列）的矩阵，假设 n=8，则可以将矩阵 Q 划分为 8 个 64*1 的子矩阵，分别为子矩阵 Q_1 至 Q_8。子矩阵 Q_1 至 Q_8分别对应 8 个 head。类似地，矩阵 K 可以被划分为子矩阵 K_1 至 K_n，其中每个子矩阵对应一个 head；矩阵 V 可以被划分为子矩阵 V_1 至 V_n，其中每个子矩阵对应一个 head。这样，每个 head 对应三个子矩阵，分别为 Q_i、K_i和 V_i，i 为从 1 取到 n 的整数。每个 head 根据 input 和该 head 对应的三个子矩阵 Q_i、K_i和 V_i进行推理，得到该 head 的输出结果。将 input 与该 head 的输出结果进行内积，内积得到的结果如果超过预设阈值，

说明该 head 是活性 head，否则是非活性 head。对 MHA 模块来说，活性参数可以是至少一个活性头所对应的权重矩阵。例如，将 input 与该 head 的输出结果进行内积，如果得到的结果较大，例如超过预设阈值，则可以将该 head 所对应的三个子矩阵 Q_i、K_i和 V_i保留；如果得到的 head 参数（或者说，该 head 的输出结果）较小，例如低于预设阈值，则可以将该 head 所对应的三个子矩阵 Q_i、K_i和 V_i置 0。这样，对于包括多个 head 的 MHA 模块，多个 head 中的部分 head 的子矩阵可能被置 0，从而影响了初始的权重矩阵 W_q、W_k和 W_v的取值，得到变化后的权重矩阵 W_q'、W_k'和 W_v'。变化后的权重矩阵 W_q'、W_k'和 W_v'可以称为至少一个活性头所对应的权重矩阵。

需要说明的是，上述“超过预设阈值”，可以指大于或等于预设阈值，也可以指大于预设阈值。上述“低于预设阈值”，可以指小于预设阈值，也可以指小于或等于预设阈值。

本申请实施例的技术方案可以应用于各种通信系统，包括但不限于：长期演进（long term evolution，LTE）系统、新无线（new radio，NR）系统等第五代（5th generation，5G）移动通信系统、窄带物联网（narrow band internet of things，NB-IoT）系统、增强型机器类型通信（enhanced machine-type communication，eMTC）系统、增强移动宽带（enhanced mobile broadband，eMBB）系统、超高可靠性和低时延通信（ultra reliable low latency communications，URLLC）系统、卫星通信系统、LTE-机器到机器（LTE-machine-to-machine，LTE-M）系统、或者第六代（6th generation，6G）移动通信系统等 5G 之后演进的系统等。

需要说明的是，在本申请实施例中，术语“通信”还可以描述为“数据传输”、“信号传输”、“信息传输”或“传输”等。在本申请实施例中，传输可以包括发送或接收。示例性地，传输可以是上行传输，例如可以是终端设备向网络设备发送信号；传输也可以是下行传输，例如可以是网络设备向终端设备发送信号；传输也可以是侧行传输，例如可以是终端设备向另一终端设备发送信号。

图 1 是一种通信系统 100 的示意图。如图 1 所示，该通信系统 100 包括无线接入网 110 和核心网 120，可选的，通信系统 100 还可以包括互联网 130。其中，无线接入网 110 可以包括至少一个网络设备（如图 1 中的 111a 和 111b），还可以包括至少一个终端设备（如图 1 中的 112a-112j）。终端设备通过无线的方式与网络设备相连。网络设备通过无线或有线方式与核心网 120 连接。核心网 120 可以包括一个或多个核心网设备。其中，核心网设备与网络设备可以是独立的不同的物理设备，也可以是将核心网设备的功能与网络设备的逻辑功能集成在同一个物理设备上，还可以是一个物理设备上集成了部分核心网设备的功能和部分的网络设备的功能。终端设备和终端设备之间以及网络设备和网络设备之间可以通过有线或无线的方式相互连接。终端设备与终端设备、网络设备与网络设备、以及终端设备与网络设备之间可以通过空口资源进行无线通信。示例性地，空口资源可以包括时域资源、频域资源、码资源和空间资源中的至少一个。需要说明的是，图 1 只是示意图，该通信系统 100 中还可以包括其它网络设备，如还可以包括无线中继设备和无线回传设备，在图 1 中未画出。

网络设备可以是任意一种具有无线收发功能的设备，例如，网络设备可以为用于将终端设备接入无线接入网络（radio access network，RAN）的基站。网络设备有时也可称为接入网设备或接入网节点。可以理解的是，采用不同无线接入技术的系统中，具备网络设备功能的设备的名称可能会有所不同。为方便描述，本申请实施例将为终端设备提供无线通信接入功能的装置统称为基站。本申请实施例中，网络设备包括但不限于：各种形式的宏基站（如图 1 中的 111a）、微基站或室内站（如图 1 中的 111b）、微微基站、小站、气球站、中继站、接入点等。网络设备可以包括 LTE 中的演进型节点 B（evolved node B，eNB 或 eNodeB）、无线保真（wireless fidelity，WiFi）系统中的接入点（Access Point，AP）、无线中继节点、无线回传节点、传输点（transmission point，TP）或者发送接收点（transmission reception point，TRP）等，还可以包括 5G 系统中的下一代基站节点（next generation NodeB，gNB）或传输点（TRP 或 TP）、5G 系统中的基站的一个或一组（包括多个天线面板）天线面板、构成 gNB 或传输点的网络节点，如基带单元（baseband unit，BBU）或分布式单元（distributed unit，DU），还可以包括 6G 等 5G 之后演进的网络中的网络设备、服务器或车载设备等。网络设备还可以是完成基站部分功能的模块或单元，例如，可以是集中式单元（central unit，CU），也可以是 DU。

本申请实施例中，用于实现网络设备的功能的装置可以是网络设备，也可以是能够支持网络设备实现该功能的装置，例如芯片系统，该装置可以被安装在网络设备中。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。

在另一种可能的场景中，由多个网络设备协作协助终端实现无线接入，不同网络设备分别实现基站

的部分功能。例如，网络设备可以是 CU，DU，CU-控制面（control plane，CP），CU-用户面（user plane，UP），或者无线单元（radio unit，RU）等。CU 和 DU 可以是单独设置，或者也可以包括在同一个网元中，例如 BBU 中。RU 可以包括在射频设备或者射频单元中，例如包括在射频拉远单元（remote radio unit，RRU）、有源天线处理单元（active antenna unit，AAU）或远程射频头（remote radio head，RRH）中。

在不同系统中，CU（或 CU-CP 和 CU-UP）、DU 或 RU 也可以有不同的名称，但是本领域的技术人员可以理解其含义。例如，在开放式无线接入网（open radio access network，O-RAN）系统中，CU 也可以称为 O-CU（开放式 CU），DU 也可以称为 O-DU，CU-CP 也可以称为 O-CU-CP，CU-UP 也可以称为 O-CU-UP，RU 也可以称为 O-RU。为描述方便，本申请中以 CU，CU-CP，CU-UP、DU 和 RU 为例进行描述。本申请中的 CU（或 CU-CP、CU-UP）、DU 和 RU 中的任一单元，可以是通过软件模块、硬件模块、或者软件模块与硬件模块结合来实现。本申请的实施例对网络设备所采用的具体技术和具体设备形态不做限定。

终端设备可以是向用户提供语音和/或数据连通性的设备；终端设备也可以是具有无线连接功能的设备。终端设备可以部署在陆地上，包括室内或室外、手持或车载；也可以部署在水面上（如轮船等）；还可以部署在空中（例如飞机、气球和卫星上等）。终端设备也可以称为用户设备（user equipment，UE）、接入终端、终端、用户单元（subscriber unit）、用户站、移动站、移动台（mobile station，MS）、移动终端（mobile terminal，MT）、远方站、远程终端、移动设备、用户终端、无线网络设备、用户代理或用户装置。本申请实施例中，终端设备包括但不限于：蜂窝电话（cellular phone）、手机（mobile phone）、无线数据卡、无线调制解调器（modem）、平板型电脑（pad）、膝上型电脑（laptop computer）、笔记本电脑、掌上电脑、移动互联网设备（mobile internet device，MID）、带无线收发功能的电脑、无绳电话、会话启动协议（session initiation protocol，SIP）电话、智能电话（smart phone）、无线本地环路（wireless local loop，WLL）站、个人数字处理（personal digital assistant，PDA）、具有无线通信功能的手持设备（handset）、计算设备或连接到无线调制解调器的其它设备、车载设备（例如，汽车、自行车、电动车、飞机、船舶、火车、高铁等）、可穿戴设备（例如智能手表、智能手环、计步器、智能眼镜等）、卫星终端、物联网或车联网中的终端设备，以及未来网络中的任意形态的终端、中继用户设备或者未来演进的公用陆地移动通信网络（public land mobile network，PLMN）中的终端等。终端设备还可以是虚拟现实（virtual reality，VR）设备、增强现实（augmented reality，AR）设备、智能销售点（point of sale，POS）机、客户终端设备（customer-premises equipment，CPE）、轻型终端设备（light UE）、能力降低的用户设备（reduced capability UE，REDCAP UE）、机器类型通信（machine type communication，MTC）终端、工业控制（industrial control）中的终端设备、无人驾驶（self driving）中的终端设备、远程医疗（remote medical）中的终端设备、智能电网（smart grid）中的终端设备、运输安全（transportation safety）中的无线终端、智慧城市（smart city）中的终端设备、智慧家庭（smart home）中的终端设备、触觉终端设备、智能家居设备（例如，冰箱、电视、空调、电表等）、智能机器人、机械臂、车间设备、无人驾驶中的无线终端、或飞行设备（例如，智能机器人、热气球、无人机、飞机）等。终端设备还可以是车辆装置，例如整车装置、车载模组、车载芯片、车载单元（on board unit，OBU）或车联网终端盒子（telematics box，T-BOX）等，终端设备还可以是其他具有终端功能的设备，例如，终端设备还可以是设备到设备（device to device，D2D）通信中担任终端功能的设备。本申请实施例对此并不限定。

本申请实施例中，用于实现终端设备的功能的装置可以是终端设备，也可以是能够支持终端设备实现该功能的装置，例如芯片或芯片系统，该装置可以被安装在终端设备中。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。本申请实施例的技术方案中，以用于实现终端设备的功能的装置是终端设备，也可以称为终端（terminal），下面可能以终端设备是 UE 为例，描述本申请实施例提供的技术方案。

基站和终端的角色可以是相对的，例如，图 1 中的直升机或无人机 112i 可以被配置成移动基站，对于那些通过 112i 接入到无线接入网 110 的终端 112j 来说，终端 112i 是基站；但对于基站 111a 来说，112i 是终端，即 111a 与 112i 之间是通过无线空口协议进行通信的。当然，111a 与 112i 之间也可以是通过基站与基站之间的接口协议进行通信的，此时，相对于 111a 来说，112i 也是基站。因此，基站和终端都可以统一称为通信装置，图 1 中的 111a 和 111b 可以称为具有基站功能的通信装置，图 1 中的 112a-112j 可

以称为具有终端功能的通信装置。

网络设备与终端设备之间可以通过无线链路通信。网络设备到终端设备的传输链路可以称为下行链路（downlink，DL）或下行信道，用于传输下行信号。终端设备到网络设备的传输链路可以称为上行链路（uplink，UL）或上行信道，用于传输上行信号。

图 2 是另一种通信系统 200 的示意性框图。该通信系统 200 也可以称为云系统，或具有其他名称。

通信系统 200 包括中心云 210、边缘云 220 和端侧 230。其中，边缘云 220 可以与中心云 210（或者称为传统的云平台）不同。例如，中心云 210（或者称为传统的云平台）可以包括资源规模化且位置集中的数据中心，而边缘云 220 可以包括分布在不同区域的多个边缘节，其中，单个边缘节点的资源规模较小。边缘云 220 覆盖的网络范围更广，距离端侧 230 中的端侧设备更近。端侧 230 中的端侧设备和边缘云 220 中的边缘节点可以通过路由系统或其他方式进行通信。

边缘节点可以对外提供各种资源，例如，中央处理器（central processing unit，CPU）、图像处理器（graphics processing unit，GPU）、张量处理器（tensor processing unit，TPU）等具有计算能力的计算资源；又例如，内存、硬盘等具有存储功能的存储资源以及带宽等网络资源。多个边缘节点中的两个边缘节点包括的资源的种类和数量可以相同，也可以不同。多个边缘节点可以是同一个互联网服务提供商部署，也可以分别由不同的互联网服务提供商提供。边缘云 220 中的边缘节点中能够部署提供云计算的服务，即在边缘节点部署可提供服务的应用。例如能够在边缘节点中部署大模型功能。

中心云 210 也可以称为云侧设备。边缘云 220 或者其中的边缘节点也可以称为边侧设备。示例性地，端侧设备可以是终端设备，例如，图 1 所示出的终端设备。

在一些可能的实现方式中，中心云 210 可以部署在核心网 120 中。或者说，核心网 120 可以具备中心云 210 的功能。在一些可能的实现方式中，边缘云 220 可以部署在网络设备中。或者说，网络设备可以包括边缘节点。

大模型的蓬勃发展催生了各行各业的应用。云侧大模型可以指部署在云侧设备或边侧设备的大模型。端侧大模型可以指部署在端侧设备的大模型。端侧大模型可以满足用户保护端侧隐私数据的需求，从而备受关注。然而，端侧设备的存储空间有限，仅能存储参数量较小的大模型，而小参数大模型的性能远低于大参数量的大模型的性能，从而不利于大模型在端侧的应用。

图 3 是一种推理方法的示意图。如图 3 所示的大模型可以部署在端侧，该大模型可以采用 transformer 结构。

图 3 右侧部分为左侧部分的图例。图 3 的左侧部分示出了该大模型的前三层的推理过程。其中，每一层用短线圆角矩形表示，每一层包括一个 MHA 模块和一个 FNN 模块。例如，MHA 模块可以包括 4 个 head，4 个 head 在图 3 中可以用 4 个长方形表示；MHA 模块的 input 在图 3 中用虚线（dashed line）箭头表示。FNN 模块可以包括 5 个神经元，5 个神经元用 5 个正方形表示；FNN 模块的 input 用点线（dotted line）箭头表示。预测模块可以根据 input，预测 MHA 模块或者 FNN 模块的活性参数的位置，预测的过程用弧形实线箭头表示。图 3 中的 P1 至 P4 分别表示预测模块的第一次至第四次预测。例如，第一次预测是根据第一层的 FNN 模块的输出，对第二层的 MHA 模块的活性参数进行的预测。

具体而言，首次输入可以是第一层的 MHA 模块的输入。端侧设备可以根据全量参数和第一层的 MHA 模块对用户的 input 进行推理，得到输出结果。也就是说，虽然在第一层的 MHA 模块中，不是所有 head 都是活性 head，但第一层的 MHA 模块进行推理时，需要加载所有 head 的子矩阵到端侧设备的内存。第一层的 MHA 模块的输出结果可以作为第一层的 FNN 模块的 input。端侧设备可以根据全量参数和第一层的 FNN 模块对 input 进行推理，得到输出结果。也就是说，虽然在第一层的 FNN 模块中，不是所有神经元都是活性神经元，但第一层的 FNN 模块进行推理时，端侧设备需要加载所有的权重参数到端侧设备的内存。在第一层的 FNN 模块进行推理的同时，预测模块可以根据第一层的 FNN 模块的 input 对第二层的 MHA 模块的活性参数的位置进行预测，在图 3 中以 P1 和 P1 相关的弧形实线箭头表示。

第一层的 FNN 模块的输出结果可以作为第二层的 MHA 模块的 input。端侧设备根据预测模块的预测到的第二层的 MHA 模块的活性参数的位置，在端侧设备的闪存中获取活性参数，并根据活性参数和第二层的 MHA 模块对 input 进行推理，得到输出结果。也就是说，第二层的 MHA 模块中，只有活性 head 的活性参数被加载到内存，例如，图 3 中的左数第一个位置和第三个位置的 head 为第二层的 MHA 模块的活性 head，对应地，这两个活性 head 所对应的权重矩阵为活性参数；并且，由于端侧设备无法预知全量参数中的哪些参数是活性参数，因此端侧设备的闪存中需要存储 MHA 模块的全量参数，以便通过活性参

数的位置确定全量参数中的活性参数。在第二层的 MHA 模块进行推理的同时，预测模块可以根据第二层的 MHA 模块的 input 对第二层的 FNN 模块的活性参数进行预测，在图 3 中以 P2 和 P2 相关的弧形实线箭头表示。第二层的 MHA 模块的输出结果可以作为第二层的 FNN 模块的 input。端侧设备可以根据预测模块预测到的第二层的 FNN 模块的活性参数的位置，在端侧设备的闪存中获取活性参数，并根据活性参数和第二层的 FNN 模块对 input 进行推理，得到输出结果。也就是说，第二层的 FNN 模块中，只有活性神经元的权重参数被加载到内存，例如，图 3 中的左数第三个位置和第五个位置的神经元为活性神经元，对应地，这两个神经元在权重矩阵中对应的元素为活性参数；并且，由于端侧设备无法预知全量参数中的哪些参数是活性参数，因此端侧设备的闪存中需要存储 FNN 模块的全量参数，以便通过活性参数的位置确定全量参数中的活性参数。在第二层的 FNN 模块进行推理的同时，预测模块可以根据第二层的 FNN 模块的 input 对第三层的 MHA 模块的活性参数进行预测，在图 3 中以 P3 和 P3 相关的弧形实线箭头表示。关于第三层的 MHA 模块和 FNN 模块的描述参照上面对第二层的描述，不再赘述。

端侧设备往往有较小的闪存空间。根据图 3 示出的方案，端侧设备在进行推理的过程中，需要在闪存中读取所需的活性参数，因此，端侧设备需要将 MHA 模块的全部权重矩阵和 FNN 模块的全部权重参数存储在闪存中。受限于闪存空间，端侧设备只能存储较小参数量的大模型，例如，7 十亿（billion，B）参数量的大模型用单精度浮点存储，需要 7 B × 32 比特（bit）/ 8 bit 每字节（byte）= 28 吉字节（GB）的闪存空间。端侧设备无法存储大参数量的大模型，例如，端侧设备无法存储超过百亿的参数量的大模型。本领域技术人员可以理解，小参数量的大模型的性能低于大参数量的大模型，使得端侧大模型无法发挥与云侧大模型相近的性能，从而影响用户体验。

端侧设备往往有较小的内存空间。根据图 3 示出的方案，第一个 transformer 层采取全量参数进行推理。例如，生成式预训练变换器（generative pre-trained transformer，GPT）3 包括 96 个 transformer 层，每一个 transformer 层的全量参数所占空间为 175 B / 96 x 4 = 7.3 GB。也就是说，GPT3 在进行第一个 transformer 层的推理时，需要将 7.3 GB 的参数全部加载到内存中。然而，大多数内存空间有限的端侧设备难以应对如此高的内存占用，例如，手机的内存可能只有 8 GB。内存空间的有限导致端侧大模型的性能降低，使得端侧大模型无法发挥与云侧大模型相近的性能，从而影响用户体验。

因此，在端侧大模型推理过程中，如何应对大模型参数量过大与存储空间过小的矛盾，进而提高端侧大模型的推理性能，是亟待解决的问题。

图 4 是本申请实施例提供的一种通信方法 400 的示意性流程图。方法 400 能够提高端侧大模型的推理性能。下面结合图 4 介绍方法 400。

S430，第一装置接收来自第二装置的第一信息。对应地，第二装置向第一装置发送第一信息。

示例性地，第一装置可以是终端设备或端侧设备。在 S430 中，第一信息可以由第一装置接收，也可以由第一装置中的组件（例如，处理器、芯片、或芯片系统等）接收，或者是由能实现全部或部分第一装置的功能的逻辑模块或软件接收。

示例性地，第二装置可以是网络设备、核心网设备、云侧设备或边侧设备。在 S430 中，第一信息可以由第二装置发送，也可以由第二装置中的组件（例如，处理器、芯片、或芯片系统等）发送，或者是由能实现全部或部分第二装置的功能的逻辑模块或软件发送。

与上面的描述类似，本申请中其他涉及第一装置和第二装置的操作也可以由装置本身、装置中的组件、或能实现全部或部分的装置的功能的逻辑模块或者软件执行，不再赘述。

可选地，该第一信息用于指示第一活性参数。该第一活性参数可以是第一模型中的 N 个第一模块的活性参数，N 为正整数。

其中，第一模型可以是任意的大模型、AI 模型或 ML 模型。例如，第一模型可以是 GPT1、GPT2 或 GPT3 等。第一模块可以是第一模型中的任意模块，例如，第一模块可以是 FNN 模块。从而，第一活性参数可以是 FNN 模块的活性参数。但是本申请对此不限定，例如，第一模块还可以是 MHA 模块或者其他模块。这样，第一活性参数可以是 MHA 模块或其他模块的活性参数。第一模型可以包括 N 个第一模块，换言之，第一模型可以包括一个或多个第一模块。从而，第一活性参数可以是一个或多个模块的活性参数。

本申请不限定第一模型的名称，第一模型也可以称为 AI 模型、ML 模型或具有其他名称。本申请不限定第一活性参数的名称，第一活性参数也可以称为参数、活性神经元、活性权重、活性元素或具有其他名称。

需要说明的是，本申请不限定第一模块的位置，或者说，不限定第一模块所处的层数。例如，第一模块可以处于第一层。又例如，第一模块可以处于第三层。再例如，3 个第一模块可以分别处于第二层、第三层和第四层。从而，本申请也不限定第一活性参数是哪个或哪些层的活性参数。例如，第一活性参数可以是第一层的模块的活性参数（简称第一层的活性参数，其他描述类似）。又例如，第一活性参数可以是第三层的活性参数。再例如，第一活性参数可以包括第二层的活性参数、第三层的活性参数和第四层的活性参数。

本申请对第一信息指示第一活性参数的方式不做限定。作为一个示例，第一信息可以包括第一活性参数，这样，第一装置根据第一信息的内容，可以获取第一活性参数。作为另一个示例，第一信息可以包括第一活性参数的标识。第一装置可以预先获取多个活性参数和多个标识之间的对应关系，例如，活性参数与标识的映射表。第一装置可以根据第一活性参数的标识和预先获取的对应关系，确定该标识所对应的活性参数为第一活性参数，从而获取第一活性参数。

本申请对第一信息的名称不做限定，第一信息还可以称为活性参数信息或具有其他名称。

在一些可能的实现方式中，该方法 400 还包括：（S420）第二装置生成第一信息。在一些可能的实现方式中，S420 在 S430 之前执行。在一些可能的实现方式中，S420 可以替换为：第二装置获取第一信息。例如，第二装置为基站中的射频芯片，该射频芯片可以从该基站中的基带芯片中获取第一信息。

本申请不限定第二装置生成第一信息的方式。作为一个示例，第二装置可以预先存储全部或部分的活性参数。第二装置可以根据预先存储的活性参数，确定第一活性参数，并生成用于指示该第一活性参数的第一信息。作为另一个示例，第二装置可以从其他装置获取第一活性参数，并生成用于指示该第一活性参数的第一信息。上述其他装置可以指不同于第一装置和第二装置的装置。

本申请不限定第二装置生成第一信息的触发条件。作为一个示例，第二装置可以基于第一装置的请求生成第一信息。例如，第二装置可以接收来自第一装置的请求信息，该请求信息指示第二装置下发第一活性参数；第二装置可以根据该请求信息，生成第一信息。作为另一个示例，第二装置可以基于其他装置的请求生成第一信息。例如，第二装置可以接收来自其他装置的请求信息，该请求信息指示第二装置向第一装置发送第一活性参数；第二装置可以根据该请求信息，生成第一信息。作为再一个示例，第二装置可以主动生成第一信息。也就是说，第二装置可以不基于某个装置的请求，而是自行决定生成第一信息。

在一些可能的实现方式中，该方法 400 还包括：（S440）第一装置根据该第一活性参数，基于 N 个第一模块进行推理。

可选地，第一活性参数包括一个 FNN 模块的活性参数，第一装置可以将该 FNN 模块的权重矩阵中处于活性参数的位置的元素设置为第一信息所指示的活性参数，并将权重矩阵中处于其他位置的元素置 0，从而得到处理后的权重矩阵。第一装置可以根据该 FNN 模块的 input，基于配置了处理后的权重矩阵的 FNN 模块，进行推理。

需要再次说明的是，上面是以一个 FNN 模块为例，但是本申请不限定第一活性参数只能包括一个 FNN 模块的活性参数。第一活性参数还可以包括更多 FNN 模块的活性参数。

通过上述实施例，用于推理的第一装置可以不存储第一模块的全量参数，而是通过第一信息获取第一模块的全量参数中的活性参数。与将全量参数存储在第一装置本地，并从全量参数中读取出活性参数的方案相比，上述实施例能够节约第一装置的存储空间，使得存储空间有限的第一装置能够运行大参数量的大模型，从而提高了大模型的推理性能。例如，第一模块可以是 transformer 结构的大模型中的 FNN 模块，FNN 模块的全量参数与 MHA 模块的全量参数的比例约为 2:1，如果 FNN 模块的全量参数不存储在端侧设备中，将节约端侧设备三分之二的闪存空间。节约出的闪存空间可以存储其他模块的参数，例如 MHA 模块的参数，这样，大模型的参数量可以提高三倍，从而有效提高大模型的性能。

可选地，第一活性参数包括一个或多个 MHA 模块的活性参数。可选地，第一活性参数包括一个或多个 MHA 模块的活性参数，以及一个或多个 FNN 模块的活性参数。

在一些可能的实现方式中，该方法 400 还包括：（S410）第一装置向第二装置发送第三信息。对应地，第二装置接收来自第一装置的第三信息。

可选地，该第三信息包括该第一活性参数的位置的指示信息。第一活性参数的位置的指示信息可以指示第一活性参数的位置。作为一个示例，第一活性参数的位置的指示信息可以包括第一活性参数的具体位置。例如，第一活性参数的位置可以是 FNN 模块的权重矩阵中元素的位置。例如，对于 2*4 的权重

矩阵，第一行第二列和第二行第二列可以是第一活性参数的位置。作为另一个示例，第一活性参数的位置的指示信息可以包括第一活性参数的位置的标识。第一装置可以预先获取多个活性参数的位置与多个标识之间的对应关系。例如，第一装置可以预先获取映射表格，该映射表格包括多个活性参数的位置与多个标识之间的映射。第一装置根据第一活性参数的位置的标识，可以确定该标识所对应的第一活性参数的位置。例如，标识#1 对应第一行第二列和第二行第二列，从而，第一装置根据第一活性参数的位置的指示信息所指示的标识#1，可以确定第一活性参数的位置是第一行第二列和第二行第二列。

又例如，第一活性参数的位置可以是 MHA 模块的活性 head 的位置。例如，对于具有 8 个 head 的 MHA 模块，8 个 head 中的第三个 head 和第四个 head 可以是活性 head。这样，第一活性参数的位置的指示信息可以指示第三个 head 和第四个 head。指示第三个 head 和第四个 head 的方式可以是直接指示，也可以是间接指示，举例不再赘述。

可选地，该第三信息用于请求该第一活性参数。也就是说，第三信息可以用于请求第二装置下发第一活性参数。或者说，第三信息可以用于请求第二装置向第一装置发送第一活性参数。

通过上述实施例，第一装置可以请求下发第一活性参数，第一装置发送的请求中可以携带第一活性参数的位置的指示信息，从而使得第一信息所指示的活性参数的取值准确。

在一些可能的实现方式中，S430 包括：响应于该第三信息，第二装置向第一装置发送该第一信息。

换言之，第二装置可以基于第一装置的请求，向第一装置发送第一活性参数。

需要说明的是，本申请不限定第一装置一定要向第二装置上报第一活性参数的位置的指示信息。在另一些可能的实现方式中，第二装置可以从其他装置获取第一活性参数的位置的指示信息。其中，第一活性参数的位置的指示信息可能是直接指示信息，即指示信息包含位置信息；或者，第一活性参数的位置的指示信息可以是间接指示信息，第二装置可以根据该指示信息确定第一活性参数的位置信息。

在一些实现方式中，该方法 400 还包括：（S450）在该第一模型的输出值的置信度大于或等于第一阈值的情况下，第一装置向第二装置发送第二信息。对应地，第二装置接收来自第一装置的第二信息。

示例性地，第一模型的输出值的置信度可以根据第一模型的相邻两层的输出值的余弦相似度计算。但是本申请不限定第一模型的输出值的置信度的计算方式，第一模型的输出值的置信度还可以使用其他方式。例如，基于小型 AI 模型或 ML 模型可以训练专用线性分类器，该专用线性分类器可以利用第一模型的输出值计算其置信度。第一阈值可以是预设的。示例性地，第一阈值可以取较大的数值，例如，第一阈值可以取 95%。本申请不限定第一阈值的名称，第一阈值还可以称为置信度阈值、预设阈值、停止阈值或具有其他名称。

可选地，该第二信息用于指示停止传输该第一信息。作为一个示例，第一装置已经请求了第一活性参数，而第一装置还未接收到第一信息的情况下，可以执行 S450，即第一装置可以在满足置信度的条件的情况下，向第二装置发送第二信息，指示第一装置不再传输指示第一活性参数的信息。作为另一个示例，第一装置接收到了部分的第一信息，例如，第一装置接收到了指示 N2 个第一模块中的部分第一模块的活性参数的信息，第一装置可以执行 S450，即第一装置可以在满足置信度的条件的情况下，向第二装置发送第二信息，指示第一装置停止传输第一信息，即不在传输第一信息的剩余部分。

本申请不限定第二信息的名称，第二信息还可以称为停止信息或具有其他名称。

在一些可能的实现方式中，该 N 个第一模块包括 N1 个第一模块，该 N1 个第一模块是该第一模型中的至少一个层中的第一个层的模块，N1 是小于或等于 N 的正整数。

第一模型可以包括至少一个层。例如，第一模型可以包括至少一个 transformer 层。N1 个第一模块可以是至少一个层中的第一个层的模块。例如，N1=1，该第一模块可以是第一模型中第一个 transformer 层中的 FNN 模块。又例如，N1=1，该第一模块可以是第一模型中第一个 transformer 层中的 MHA 模块。

该 N 个第一模块包括 N1 个第一模块，则第一活性参数可以包括第一模型中 N1 个第一模块的活性参数。该 N1 个第一模块可以是第一个层中的模块，则第一活性参数可以包括第一模型中第一个层的活性参数。结合 S430，第二装置可以向第一装置发送第一模型中第一个层的活性参数。结合 S440，第一装置可以在根据第一个层中的模型进行推理时，可以基于活性参数，而不是基于全量参数。

通过上述实施例，第一装置可以基于第一信息指示的活性参数，基于第一模型中的第一个层中的模块进行推理。与根据全量参数和第一个层中的模块进行推理的方案相比，本申请实施例无需将第一个层中的模块的全量参数加载到内存，从而节约了第一装置的内存空间，从而进一步提升了第一模型的推理

性能。

图 5 是本申请实施例提供的另一种通信方法 500 的示意性流程图。方法 500 可以与方法 400 相结合。在方法 500 中，第一模型的第一个层中的多个模块中的第一个的全量参数可以存储在第一装置本地；第一模型的第一个层中的多个模块中的第二个的全量参数可以存储在第二装置。也就是说，第一模型需要向第二装置请求第一个层中的多个模块中的第二个的活性参数；第二装置可以根据第一个层中的多个模块中的第二个的全量参数，确定活性参数（即第一活性参数）。下面结合图 5 介绍方法 500。

图 5 假设第一模块包括第一模型的一个层的多个模块中的第二个。假设第二模块包括第一模块的一个层的多个模块中的第一个。但需要说明的是，图 5 的描述仅作为举例，不构成对本申请的限定。例如，第一模块可以包括第一模型的一个层的多个模块中的第一个和第二个。换言之，在本申请其他的一些实施例中，第一模块可以包括第一模型的一个层的全部模块。

S510，第一装置根据第一输入值，预测第二活性参数的位置的信息。需要说明的是，在图 5 中，S510 的操作以虚线示出，表示 S510 在与方法 400 进行结合时，S510 是可选的操作。

可选地，该 N1 个第一模块是该第一模型的第一个层中的多个模块中的第二个，N1=1。例如，第一模型是 transformer 结构的大模型。第一模型的第一个 transformer 层中的多个模块中的第一个为 MHA 模块，多个模块中的第二个为 FNN 模块。也就是说，第一模块可以是大模型的第一个 transformer 层中的 FNN 模块。

可选地，S510 中的该第二活性参数是该第一模型的第一个层中的多个模块中的第一个的活性参数。例如，第一模块可以是第一模型的第一个层中的多个模块中的第二个，则第二活性参数可以是第一模块在第一模型的第一个层中的上一个模块的活性参数。又例如，第一模型的第一个层中的多个模块中的第一个可以是 MHA 模块。也就是说，第二活性参数可以是第一模型的第一层中的 MHA 模块的活性参数。

本申请不限定第二活性参数的名称，第二活性参数也可以称为参数、活性 head、活性权重、活性矩阵或具有其他名称。

可选地，该第一输入值是该第一模型的第一个层中的多个模块中的第一个的输入值。例如，第一输入值可以是第一模型的第一个 transformer 层中的 MHA 模块的输入值。又例如，第一输入值可以是用户对大模型发出的请求经过嵌入（embedding）后的向量或张量。

示例性地，第一装置可以使用预测模块进行预测。具体而言，预测模块可以根据第一输入值，预测第二活性参数的位置的信息。可选地，预测模块部署在第一装置中；换言之，第一装置包括预测模块。示例性地，预测模块可以是第二模型。例如，第二模型可以是 AI 模型或 ML 模型等。又例如，第二模型可以是 DNN 模型。

在一些可能的实现方式中，方法 500 还包括：（S520）第一装置根据第二活性参数的位置的信息，确定第二活性参数。例如，第一装置可以存储第一模型的第一个层中的多个模块中的第一个（例如 MHA 模块）的全量参数。这样，第一装置可以根据第二活性参数的位置的信息和存储的全量参数，确定第二活性参数的取值。在另一种可能的实现方式中，第一装置向第二装置发送请求信息；第二装置接收请求信息。请求信息可以用于请求第二活性参数。在这种情况下，第二活性参数可以是从第二装置获取的。第二装置可以预先存储第一模型的第一个层的多个模块中的第一个（例如 MHA 模块）的全量参数，从而根据第一装置的请求信息，在存储的全量参数中，确定被请求的活性参数。

在一些可能的实现方式中，方法 500 还包括：（S530）第一装置根据第二活性参数，基于第一模型中的第一个层中的多个模块中的第一个进行推理。例如，第二活性参数为第一模型的第一个层中的 MHA 模块的活性参数。第一装置可以将该 MHA 模块的权重矩阵用第二活性参数进行更新，具体而言，MHA 模块中的非活性 head 的权重子矩阵被置 0，活性 head 的权重子矩阵则不受影响。第一装置可以根据该第一输入值，基于第二活性参数（即更新后的 MHA 模块的非零权重子矩阵，或者更新后的 MHA 模块的权重矩阵），进行推理。

示例性地，方法 500 可以应用于数智人聊天的场景。例如，用户可以在手机上向虚拟的数智人发送文字消息、图片消息或语音消息，并接收该数智人回复的文字消息、图片消息或语音消息。在数智人聊天的场景下，虚拟的数智人可以由第一模型驱动。例如，用户向虚拟数智人发送的消息可以被输入到第一模型中；第一模型根据上述输入进行推理后，可以理解并输出用于回复用户的消息。在数智人聊天的场景下，输入第一模型的数据可能涉及用户的聊天记录等信息，因此大模型可以部署在手机等端侧设备中，并且端侧设备的数据不上传，从而保护用户的隐私。并且，手机等端侧设备的存储空间（例如，内

存或闪存）有限，存在难以部署大参数量的大模型的问题。此外，用户与虚拟的数智人聊天，对虚拟数智人回复消息的延迟要求不高，即用户不要求虚拟数智人立即回复消息。

通过上述实施例，第一装置可以预测第一模型中的第一个层中的多个模块中的第一个的活性参数，从而根据多个模块中的第一个的活性参数，基于多个模块中的第一个进行推理。与根据全量参数和第一个层中的多个模块中的第一个进行推理的方案相比，本申请实施例无需将第一个层中的多个模块中的第一个的全量参数加载到内存，从而节约了第一装置的内存空间，从而进一步提升了第一模型的推理性能。

S540，第一装置根据第二输入值，预测该第一活性参数的位置的信息。需要说明的是，在图 5 中，S540 的操作以虚线示出，表示 S540 在与方法 400 进行结合时，S540 是可选的操作。

可选地，该第二输入值可以为该第一模型的第一个层中的多个模块中的第二个的输入值或该第一模型的第一个层中的多个模块中的第一个的输出值。例如，第二输入值可以是第一模型的第一个 transformer 层中的 MHA 模块的输出值。又例如，第二输入值可以是第一模型的第一个 transformer 层中的 FNN 模块的输入值。示例性地，第一模型的第一个层中的多个模块中的第二个可以是 FNN 模块。示例性地，第一模型的第一个层中的多个模块中的第二个可以是前述第一模块。

示例性地，第一装置可以使用预测模块进行预测。具体而言，预测模块可以根据第二输入值，预测第一活性参数的位置的信息。可选地，预测模块部署在第一装置中；换言之，第一装置包括预测模块。示例性地，预测模块可以是第二模型。例如，第二模型可以是 AI 模型或 ML 模型等。又例如，第二模型可以是 DNN 模型。

在一些可能的实现方式中，S410 在 S540 之后执行。例如，第三信息所包括的第一活性参数的位置的指示信息可以是由第一装置根据第二输入值预测的。

通过上述实施例，第一装置可以预测第一模块中的第一个层中的多个模块中的第二个的活性参数，从而根据多个模块中的第二个的活性参数，基于多个模块中的第二个进行推理。与根据全量参数和第一个层中的多个模块中的第二个进行推理的方案相比，本申请实施例无需将第一个层中的多个模块中的第二个的全量参数加载到内存，从而节约了第一装置的内存空间，从而进一步提升了第一模型的推理性能。

在一些可能的实现方式中，该 N 个第一模块还包括 N2 个第一模块，该第一模型包括 N2 个层，该 N2 个层中的每个层包括该 N2 个第一模块中的一个第一模块，N2≥2，N2 是小于或等于 N 的正整数。

N 个第一模块可以包括 N2 个第一模块。这样，第一活性参数可以包括 N2 个第一模块的活性参数。N2≥2，也就是说，第一活性参数可以包括多个第一模块的活性参数。上述多个第一模块可以分别位于不同的层，因此，第一活性参数可以包括多个层的活性参数。为了减小传输时延，第二装置可以一次下发多个层的活性参数，减少了下发活性参数的次数，从而减少了频繁下发活性参数所导致的信令开销。可选地，N 可以取自 1 至 8 的整数。

需要说明的是，第一模型可以包括 N2 个层，但不限于 N2 个层。例如，第一模型可以包括多于 N2 个层。与第一模型中的其他层不同的是，第一模型中的 N2 个层可以与 N2 个第一模块一一对应。例如，第一模型的 N2 个层中的第 m 个层可以包括 N2 个第一模块中的第 m 个第一模块，m 为从 1 取至 N2 的整数。这样，第二装置下发的第一活性参数可以包括 N2 个层中的 N2 个第一模块的活性参数。第一模型中除 N2 个层之外的层，可以包括第一模块，也可以不包括第一模块。换言之，第二装置下发的第一活性参数可以包括第一模型中除 N2 个层之外的层中的模块的活性参数，也可以不包括第一模型中除 N2 个层之外的层中的模块的活性参数。

本申请不限定 N2 个层在第一模型中所有的层中的位置。例如，N2 个层可以是第一模型的前 N2 个层。也就是说，N2 个层可以包括第一模型中第一个层至第 N2 个层。又例如，N2 个层可以是第一模型的中间的 N2 个层。也就是说，N2 个层可以包括第一模型中第 x 个层至第 x+N2-1 个层，x 可以是大于 1 的整数。在上面两个例子中，N2 个层在第一模型的所有的层中编号是连续的；换言之，这 N2 个层中没有插入其他的层（即可能包括第一模块，也可能不包括第一模块的层）。上述第一模块可以理解为，第二装置下发的活性参数所属的模块。但是，本申请对此不做限定，N2 个层在第一模型的所有的层中编号也可以是不连续的；换言之，这 N2 个层中可以插入其他的层。

通过上述实施例，第二装置可以一次下发多个层的活性参数，减少了下发活性参数的次数，从而减少了频繁下发活性参数所导致的信令开销。

结合图 5 示出的方法 500，N 个第一模块可以包括 N1 个第一模块和 N2 个第一模块。这样，第一活性参数可以包括 N1 个第一模块和 N2 个第一模块的活性参数。需要说明的是，在一些可能的实现方式中，

第一信息可以传输两次，在其中一次传输中，第一信息用于指示第一活性参数包括的一部分活性参数，例如，N1 个第一模块的活性参数。在另一次传输中，第一信息用于指示第一活性参数包括的另一部分活性参数，例如，N2 个第一模块的活性参数。但是，以上仅仅是举例，不作为对申请的限定。例如，在另一些可能的实现方式中，第一信息可以仅传输一次。第一信息所指示的第一活性参数可以包括 N1 个第一模块和 N2 个第一模块的活性参数。又例如，在另一些可能的实现方式中，第一信息可以传输更多的次数。每次传输的第一信息所指示的第一活性参数，相比于前述两次传输的第一信息所分别指示的第一活性参数，可以包括更少量的活性参数。

图 6 是本申请实施例提供的又一种通信方法 600 的示意性流程图。方法 600 可以与方法 400、方法 500 相结合。在方法 600 中，第一模型的 N2 个层中的第二模块的全量参数可以存储在第一装置本地；第一模型的 N2 个层中的第一模块的全量参数可以存储在第二装置。也就是说，第一模型需要向第二装置请求 N2 个层中的第一模块的活性参数；第二装置可以根据 N2 个层中的第一模块的全量参数，确定活性参数（即第一活性参数）。下面结合图 6 介绍方法 600。

图 6 示出的方案假设第一模块包括第一模型的多个层中每个层的多个模块中的第二个。假设第二模块包括第一模块的多个层中每个层的多个模块中的第一个。但需要说明的是，图 6 的描述仅作为举例，不构成对本申请的限定。例如，第一模块可以包括第一模型的多个层中每个层的多个模块中的第一个和第二个。换言之，在本申请其他的一些实施例中，第一模块可以包括第一模块的多个层中的全部模块。

S610，第一装置根据 N2 个层中第一个层的上一个层的输入值，预测该第一活性参数的位置的信息。需要说明的是，在图 6 中，S610 的操作以虚线示出，表示 S610 在与方法 400、方法 500 进行结合时，S610 是可选的操作。

可以理解的是，N2 个层之前，第一模型还包括其他层。例如，第一模型包括 16 个层，N2 个层是这 16 个层中的第 2 至 9 个层。示例性地，预测模块可以根据 N2 个层中第一个层的上一个层的输入值，预测该第一活性参数的位置的信息。举例不再赘述，参见上面方法 400 和方法 500 的相关描述。

N2 个层中第一个层的上一个层的输入值，可以理解为 N2 个层中第一个层的上一个层的多个模块中的第一个的输入值。

在一些可能的实现方式中，在执行 S610 的同时，N2 个层中第一个层的上一个层可以根据其输入值，进行推理。这样，在 N2 个层中第一个层的上一个层进行推理的同时，预测模块可以预测第一活性参数的位置的信息，即预测正在推理的层后面的 N2 个层中的 N2 个第一模块的活性参数。预测活性参数所针对的层与模型推理所处的层不同步，因此上述方案也可以称为异步预测。

本领域技术人员可以理解，由于残差连接（residual connections），transformer 结构的大模型中每一层的输出的变化十分缓慢。即，相邻层的输出具有高度相似性，这使得上述异步预测多层的方案几乎不会降低预测活性参数的位置的准确性。在一种可能的实现方式中，N2 可以取 8，即第一装置根据一层的 input，预测该层下面 8 层中的第一模块的活性参数的位置消息。

通过上述实施例，一方面，第一装置可以预测多个层分别包括的多个第一模块的活性参数的位置的信息。与一次仅预测一个模块的活性参数的方案相比，上述实施例能够节约计算资源。另一方面，第一装置预测 N2 个层是根据 N2 个层中第一个层的上一个层的输入值，而不是根据 N2 个层中的第一个层的输入值。这样，第一装置可以在推理 N2 个层中第一个层的上一个层时，预测 N2 个层中第一模块的活性参数的位置的信息。也就是说，第一装置预测活性参数的位置所针对的层不是正在推理的层，因此上述实施例不会影响正在推理的层，即不会降低大模型的推理性能。

在另一些可能的实现方式中，S610 可以替换为：第一装置根据 N2 个层中第一个层的输入值，预测该第一活性参数的位置的信息。本领域技术人员可以理解，第一装置得到了 N2 个层中第一个层的输入值，表明模型推理正处于 N2 个层中第一个层，与此同时，预测模块正在预测第一活性参数的位置，即预测正在推理的层以及该层后面的 N2-1 个层中的 N2 个第一模块的活性参数。预测活性参数所针对的层与模型推理所处的层同步，因此上述方案也可以称为同步预测。在另一种理解中，由于上述方案也预测了模型推理所处的层的后面的层，因此上述方案也可以称为异步预测。

结合方法 400，在一些可能的实现方式中，S410 在 S610 之后执行。例如，第三信息所包括的指示信息所指示的第一活性参数的位置可以是由第一装置根据 N2 个层中第一个层的上一个层的输入值（或者，N2 个层中第一个层的输入值）预测的。结合方法 400，在一些可能的实现方式中，S430 可以在 S610 之后执行。

可选地，该 N2 个层中的每个层还包括一个第二模块，该第二模块是该第一模型中的 N2 个第二模块之一。换言之，N2 个层中的每个层可以包括一个第一模块和一个第二模块。示例性地，第二模块可以为 MHA 模块。

在一些实现方式中，该方法 600 还包括：（S620）第一装置根据该 N2 个层中第一个层的上一个层的输入值，预测第三活性参数的位置的信息。

可选地，该第三活性参数是该 N2 个第二模块的活性参数。例如，第三活性参数是 N2 层中的 N2 个 MHA 模块的活性参数。本申请不限定第三活性参数的名称，第三活性参数也可以称为参数、活性神经元、活性权重、活性元素或具有其他名称。

示例性地，预测模块可以根据 N2 个层中第一个层的上一个层的输入值，预测该第三活性参数的位置的信息。举例不再赘述，参见上面方法 400 和方法 500 的相关描述。

在一些可能的实现方式中，S610 和 S620 可以同时执行，也就是说，预测模块可以根据 N2 个层中第一个层的上一个层的输入值，预测第一活性参数的位置的信息和第三活性参数的位置的信息。

在一些可能的实现方式中，在执行 S620 的同时，N2 个层中第一个层的上一个层可以根据其输入值，进行推理。这样，在 N2 个层中第一个层的上一个层进行推理的同时，预测模块可以预测第三活性参数的位置的信息，即预测正在推理的层后面的 N2 个层中的 N2 个第二模块的活性参数。预测活性参数所针对的层与模型推理所处的层不同步，因此上述方案也可以称为异步预测。

通过上述实施例，一方面，第一装置可以预测多个层分别包括的多个第二模块的活性参数的位置的信息。与一次仅预测一个模块的活性参数的方案相比，上述实施例能够节约计算资源。另一方面，第一装置预测活性参数的位置所针对的层不是正在推理的层，因此上述实施例不会影响正在推理的层，即不会降低大模型的推理性能。

在另一些可能的实现方式中，S620 可以替换为：第一装置根据 N2 个层中第一个层的输入值，预测该第三活性参数的位置的信息。本领域技术人员可以理解，第一装置得到了 N2 个层中第一个层的输入值，表明模型推理正处于 N2 个层中第一个层，与此同时，预测模块正在预测第三活性参数的位置，即预测正在推理的层以及该层后面的 N2-1 个层中的 N2 个第二模块的活性参数。预测活性参数所针对的层与模型推理所处的层同步，因此上述方案也可以称为同步预测。在另一种理解中，由于上述方案也预测了模型推理所处的层的后面的层，因此上述方案也可以称为异步预测。

在一些可能的实现方式中，方法 600 还包括：（S630）第一装置根据第三活性参数的位置的信息，确定第三活性参数。例如，第一装置可以存储第一模型的 N2 层中的 N2 个第二模块（例如 MHA 模块）的全量参数。这样，第一装置可以根据第三活性参数的位置的信息和存储的全量参数，确定第三活性参数的取值。在另一种可能的实现方式中，第一装置向第二装置发送请求信息；第二装置接收请求信息。请求信息可以用于请求第三活性参数。在这种情况下，第三活性参数可以是从第二装置获取的。第二装置可以预先存储第一模型的 N2 个层的第二模块（例如 MHA 模块）的全量参数，从而根据第一装置的请求信息，在存储的全量参数中，确定被请求的活性参数。

在一些可能的实现方式中，S630 可以在 S610 之前执行。例如，执行顺序可以为：S620、S630、S610。在另一些可能的实现方式中，S630 可以在 S610 之后执行。例如，执行顺序可以为：S610、S620、S630。又例如，执行顺序可以为：S620、S610、S630。但是本申请对此不限定，例如，S630 还可以与 S610 同时执行。

在一些可能的实现方式中，方法 600 还包括：（S640）第一装置根据第一活性参数和第三活性参数，对 N2 个层进行推理。

在一些可能的实现方式中，对 N2 个层进行推理包括：第一装置根据 N2 个层中的第 k 个层中的第一模块的输出值，作为该第 k 个层中第二模块的输入值。上述方案也可以称为顺序推理。

在一些可能的实现方式中，对 N2 个层进行推理包括：第一装置根据该 N2 个层中的第 k 个层中的第一模块的输出值与该第 k 个层中的第二模块的输出值，确定第三输入值，k 是小于或等于 N2 的正整数。上述方案也可以称为平行推理。

例如，第一装置可以将第 k 个层中的第一模块的输出值和该第 k 层中的第二模块的输出值相加，得到第三输入值。又例如，第一装置可以将第 k 个层中的第一模块的输出值和该第 k 层中的第二模块的输出值相加，并将相加后的数值乘以非零系数，得到第三输入值。

可选地，该第三输入值用于该第 k 层的下一层中的第一模块的输入，以及该第 k 层的该下一层中的

第二模块的输入。

图 7 是本申请实施例提供的再一种通信方法的示意图。图 7 示出了同步预测与异步预测相结合的实施例。简而言之，对大模型中的第 1 个层采用同步预测，对第 1 个层之后的多个层采用异步预测（或者同步预测）。需要说明的是，图 7 示出的实施例仅作为示例，不构成对本申请的限定。下面结合图 7 进行介绍。

图 7 示出了该大模型的前 2+2N2-1 层的推理过程。为方便描述，第 1 个层、第 2 个层至第 2+N2-1 个层、第 2+N2 个层至第 2+2N2-1 个层分别用短线圆角矩形表示。模块所处的层的编号用下标表示，例如，对于第一个层中的 MHA 模块，用 MHA_1 表示。关于 MHA 模块、FNN 模块和预测模块的其他描述与图 3 的相关内容类似，不再赘述。

具体而言，用户对该大模型发出的请求经过嵌入（embedding）后，生成的向量或张量可以作为图 7 中的首次“输入”。首次输入可以是第一层的 MHA 模块的输入，如图 7 中 MHA_1 以及指向 MHA_1 的虚线箭头所示。预测模块可以根据 MHA_1 的输入，预测该 MHA_1 的活性参数的位置，在图 7 中以 P1 和 P1 相关的实线箭头表示。第一装置可以在其存储空间中获取 MHA_1 的活性参数。第一装置可以根据 MHA_1 和 MHA_1 的活性参数进行推理，得到 MHA_1 的输出；也就是说，MHA_1 进行推理时，第一装置可以加载部分活性 head 的子矩阵到第一装置的内存。MHA_1 的输出可以作为第一层的 FNN 模块的输入，如图 7 中 FNN_1 以及指向 FNN_1 的点线箭头所示。预测模块可以根据 FNN_1 的输入，预测该 FNN_1 的活性参数的位置，在图 7 中以 P2 和 P2 相关的实线箭头表示。第一装置可以请求第二装置下发 FNN_1 的活性参数。第二装置可以存储 FNN 模块的全部参数。例如，第二装置可以存储 4*4 的权重矩阵中所有元素的取值及每个元素对应的位置。第二装置根据活性参数的位置，可以确定活性参数的取值。例如，活性参数的位置是第一行第三列和第二行第四列，第二装置可以在 4*4 的权重矩阵中确定第一行第三列的元素的取值，以及第二行第四列的元素的取值。第二装置可以将活性参数下发给第一装置。例如，第二装置可以向第一装置发送第一信息，第一信息用于指示 FNN_1（即 N1 个第一模块）的活性参数。又例如，第二装置可以向第一装置发送用于指示第一行第三列和第二行第四列的元素的取值的信息。第一装置请求活性参数与第二装置下发活性参数的过程如图 7 中虚线双头箭头所示。第一装置可以根据 FNN_1 和 FNN_1 的活性参数进行推理；也就是说，FNN_1 进行推理时，第一装置可以加载部分的权重参数到第一装置的内存。

本领域技术人员可以理解，transformer 结构的大模型的第一个层的两个输入的差异较大。上述方案第 1 个 transformer 层采取同步预测，用 MHA_1 的输入预测 MHA_1 的活性参数的位置，用 FNN_1 的输入预测 FNN_1 的活性参数的位置。一方面，上述方案可以节约第一装置的内存。另一方面，上述方案预测到的第一层的活性参数更加准确。

FNN_1 的输出可以作为第二层的 MHA 模块的输入，如图 7 中 MHA_2 以及指向 MHA_2 的虚线箭头所示。预测模块可以根据的 MHA_2 输入预测第 2 个层至第 2+N2-1 个层中 MHA 模块和 FNN 模块的活性参数，在图 7 中以 P3 和 P3 相关的实线箭头表示。对于 MHA 模块，例如，MHA_2 至 MHA_{2+N2-1}，第一装置可以从第一装置的存储空间中获取 MHA 模块的活性参数。对于 FNN 模块，例如，FNN_2 至 FNN_{2+N2-1}，第一装置可以请求第二装置下发 FNN_2 至 FNN_{2+N2-1} 的活性参数。第二装置可以将活性参数下发给第一装置。例如，第二装置可以向第一装置发送第一信息，第一信息用于指示 FNN_2 至 FNN_{2+N2-1}（即 N2 个第一模块）的活性参数。关于第 2+N2 个层至第 2+2N2-1 个层中的 MHA 模块和 FNN 模块的描述参照上面对第 2 个层至第 2+N2-1 个层的描述，不再赘述。

本领域技术人员可以理解，transformer 结构的大模型在第 2 个层及其之后的层的输出具有高度相似性，因此可以采用一个层的输入一次性预测后面多个层的活性参数。这样，第二装置可以一次性下发多个层的活性参数，有效减少了第一装置与第二装置之间的信息传输所带来的开销。

图 8 是本申请实施例提供的再一种通信方法 800 的示意性流程图。在图 8 中，第一装置可以包括用户操作模块、大模型推理用户界面（user interface，UI）、预测模块和端侧推理模块。可选地，用户操作模块和大模型推理 UI 也可以是和第一装置匹配使用的装置。方法 800 可以与图 7 示出的第一装置和第二装置的流程结合。需要说明的是，方法 800 仅作为举例，不构成对本申请的限定。下面介绍方法 800。

S802，用户操作模块向大模型推理用户界面（user interface，UI）发送请求信息。换言之，用户操作模块可以向大模型推理 UI 发起请求。其中，用户操作模块可以将用户的操作转换为第一装置可识别的信息。可选地，上述请求消息包括用户的输入（input）。

S804，大模型推理 UI 向预测模块发送用户的输入，预测模块接收该用户的输入。在预测模块进行预

测之前，第一装置可以将用户的输入进行嵌入（embedding）等处理，得到的向量或张量可以作为第 1 个层的 MHA_1 的输入。

S810，预测模块根据 MHA_1 的输入，预测 MHA_1 的活性参数的位置。

S812，预测模块可以向端侧推理模块发送 MHA_1 的活性参数的位置的指示信息，端侧推理模块接收该 MHA_1 的活性参数的位置的指示信息。

S814，端侧推理模块根据 MHA_1 的活性参数、MHA_1 的输入和 MHA_1 进行推理，以得到 MHA_1 的输出。其中，端侧推理模块可以先根据 MHA_1 的活性参数的位置，在第一装置的存储空间中获取 MHA_1 的活性参数，然后，根据 MHA_1 的活性参数、MHA_1 的输入和 MHA_1 进行推理。

S816，端侧推理模块向预测模块发送 MHA_1 的输出（即，FNN_1 的输入）；预测模块接收该 MHA_1 的输出（即，FNN_1 的输入）。

S818，预测模块根据 FNN_1 的输入，预测 FNN_1 的活性参数的位置。

S820，预测模块可以向第二装置发送用于请求 FNN_1 的活性参数的信息；第二装置可以接收用于请求 FNN_1 的活性参数的信息。其中，用于请求 FNN_1 的活性参数的信息可以包括 FNN_1 的活性参数的位置的指示信息。需要说明的是，方法 800 仅为方便理解，向第二装置发送请求的操作也可以由其他模块执行。例如，发送模块。

S822，第二装置向第一装置发送 FNN_1 的活性参数；第一装置接收 FNN_1 的活性参数。在一些可能的实现方式中，第一装置中的端侧推理模块可以接收 FNN_1 的活性参数。但是本申请不限定，第一装置接收的操作也可以由其他模块执行。例如，接收模块。

S824，端侧推理模块根据 FNN_1 的活性参数、FNN_1 的输入和 FNN_1 进行推理，以得到 FNN_1 的输出。

S830，端侧推理模块向预测模块发送 FNN_1 的输出（即，MHA_2 的输入）；预测模块接收 FNN_1 的输出（即，MHA_2 的输入）。

S832，预测模块根据 MHA_2 的输入，预测 MHA_2 至 MHA_{2+N2-1} 的活性参数的位置，以及 FNN_2 至 FNN_{2+N2-1} 的活性参数的位置。

S834，预测模块可以向第二装置发送用于请求 FNN_2 至 FNN_{2+N2-1} 的活性参数的信息；第二装置可以接收用于请求 FNN_2 至 FNN_{2+N2-1} 的活性参数的信息。其中，用于请求 FNN_2 至 FNN_{2+N2-1} 的活性参数的信息可以包括 FNN_2 至 FNN_{2+N2-1} 的活性参数的位置的指示信息。需要说明的是，方法 800 仅为方便理解，向第二装置发送请求的操作也可以由其他模块执行。例如，发送模块。

S836，预测模块可以向端侧推理模块发送 MHA_2 至 MHA_{2+N2-1} 的活性参数的位置的指示信息；端侧推理模块可以接收 MHA_2 至 MHA_{2+N2-1} 的活性参数的位置的指示信息。

S838，第二装置向第一装置发送 FNN_2 至 FNN_{2+N2-1} 的活性参数。在一些可能的实现方式中，第一装置中的端侧推理模块可以接收 FNN_2 至 FNN_{2+N2-1} 的活性参数。但是本申请不限定，第一装置接收的操作也可以由其他模块执行。例如，接收模块。

需要说明的是，S834 至 S838 的执行顺序不做限定。例如，按照以下顺序执行：S834、S838、S836。又例如，按照以下顺序执行：S836、S834、S838。

S840，端侧推理模块根据第 2 个层至第 2+N2-1 个层中的模块、模块的活性参数以及模块的输入进行推理，最终输出 FNN_{2+N2-1} 的输出（即 MHA_{2+N2} 的输入）。其中，对于 MHA_2 至 MHA_{2+N2-1}，端侧推理模块可以先根据 MHA_2 至 MHA_{2+N2-1} 的活性参数的位置，在第一装置的存储空间中分别获取 MHA_2 至 MHA_{2+N2-1} 的活性参数。

随后，预测模块可以根据 MHA_{2+N2} 的输入，对后面 N2 层的模块的活性参数进行预测，执行与 S832 至 S840 类似的操作，直到输出大模型的推理结果。大模型的推理结果可以返回大模型推理 UI。

在一些可能的实现方式中，第一模型包括 M 个层，M 是大于 N2 的正整数，该 N2 个层在该 M 个层中的编号不连续。换言之，第一模型包括 M 个编号连续的层，M 个层包括 N2 个层。

N2 个层在该 M 个层中的编号不连续，可以理解为，在 N2 个层中，存在两个相邻的层的编号不连续。例如，M 个层为第一模型中的前 10 个层，N2 个层分别为第 2 个层、第 3 个层、第 5 个层和第 6 个层。第 3 个层与第 5 个层相邻，但是编号 3 和编号 5 不连续。

本领域技术人员可以理解，随机跳过一些 transformer 层，例如大模型包括的所有层中的 25%的层，不会使得大模型的性能很大程度下降，并且能够降低推理的时延。本申请实施例中的大模型支持随机跳过一层的推理。例如，每 4 个 transformer 层中，可以跳过一个 transformer 层。示例性地，N2 个层分别为

第 2 个层、第 3 个层、第 5 个层和第 6 个层，第 2 个层的输出可以作为第 3 个层的输入，第 3 个层的输出可以作为第 5 个层的输入；也就是说，第 4 个层被跳过，不参与推理。又例如，N2 为 8，大模型在推理过程中，可以跳过第 4 个层和第 8 个层。

相应地，本申请实施例提供的方案在预测当前推理的层的后面的多个层时，可以跳过一些层，即不对这些层的活性参数进行预测。结合关于 S410 描述，第二装置下发的第一信息可以指示第一活性参数，即指示 N 个第一模块的活性参数。N 个第一模块可以包括 N2 个第一模块，N2 个第一模块分别属于 N2 个层。这样，第一活性参数可以包括 N2 个层中的第一模块的活性参数。N2 个层在该 M 个层中的编号不连续，表明第二装置指示的活性参数不是属于 N2 个编号连续的层，而是属于 N2 个编号不连续的层，即，第二装置指示的活性参数跳过了一个或多个层。结合关于 S610 的描述，第一装置预测的第一活性参数，不是 N2 个编号连续的层的活性参数，而是 N2 个编号不连续的层的活性参数，即，第一装置在预测活性参数时跳过了一些层，从而降低了预测的时延。

通过上述实施例，第一信息所指示的第一活性参数可以包括编号不连续的层的活性参数，即可以这些活性参数可以跳过一个或多个层，从而减少了传输第一信息的信息量，降低了时延。

在一些可能的实现方式中，N2 与用于传输该第一信息的资源和/或推理时延要求相关。

可选地，用于传输该第一信息的资源可以包括用于传输该第一信息的带宽、频谱宽度、时隙、多天线波束等。例如，用于传输该第一信息的带宽越大，N2 可以越大。反之，用于传输该第一信息的带宽越小，N2 越小。

推理时延要求可以是指第一装置进行推理的时延要求。例如，第一装置需要在 3 秒（s）内完成一个层的推理。推理时延要求越高，或者说，要求的推理时延越小，N2 越大。反之，推理时延要求越低，或者说，要求的推理时延越大，N2 可以越小。

可选地，推理时延要求可以包括第一装置获取第一信息的时延上限。例如，第一装置需要在 3 s 内完成一个层的推理，第一装置根据活性参数进行推理所花费的时间为 2 s，则第一装置获取用于指示 N2 个第一模块（包括该层和该层后面的 N2-1 个层）的活性参数的第一信息的时延上限为 1 s，则要求第一装置获取第一信息的时延不能超过 1 s，对应地，第二装置发送第一信息的时延不能超过 1 s；换言之，第一装置需要在 1 s 内获取到第一信息，从而获取 N2 第一模块的活性参数。

以 GPT3 为例，GPT3 共有 96 个 transformer 层，每一个 transformer 层的活性参数所占用的存储空间为 175 B × 0.15 / 96 × 4 = 0.09375 GB = 93.75 MB。其中，FNN 模块的活性参数占用的存储空间为 93.75 × 2 / 3 = 62.5 MB。若带宽为 200 兆字节每秒（MB/s），第一装置接收第一信息的时延上限为 1 s，则 N2 可以取 3。这样，第一信息的传输能够在给定的资源内满足推理时延要求。

通过上述实施例，N2 的取值可以与用于传输该第一信息的资源和/或推理时延要求相关，从而使得第一信息的传输可以满足资源限制和/或推理时延要求。上述方案能够保障第一信息能够顺利传输，或者，能够保障第一信息的传输可以满足第一模型的推理时延要求，从而进一步保障第一模型的推理性能。

图 9 是本申请实施例提供的用于平行推理的大模型的示意图。需要说明的是，图 9 仅作为示例，不构成对本申请的限定。

参见图 9，第 1 个层的输入可以作为 MHA_1 的输入和 FNN_1 的输入。根据 MHA_2 的输出和 FNN_2 的输出，可以确定第 1 个层的输出，即作为第 2 个层的输入。图 9 示出了的推理方式可以称为平行推理。需要说明的是，S650 中 k 是小于或等于 N2 的正整数。在一些可能的实现方式中，k 是取遍 1 至 N2 的正整数，也就是说，N2 个层中的每个层都采取上述平行推理的方式。在另一些可能的实现方式中，k 是取自 1 至 N2 中的部分正整数，也就是说，N2 个层中的一部分层（或者说，一个或多个层）采取上述平行推理的方式，另一部分层仍采用与图 3、图 7 类似的推理方式。该推理方式可以称为顺序推理，即，将一个层的输入作为这个层的 MHA 模块的输入，将 MHA 模块的输出作为这个层的 FNN 模块的输入，将 FNN 模块的输出作为这个层的输出。

相比于采用顺序推理的方式，采用平行推理的方式可以使得每一层的输入的数量减少一半。例如，对于 N2 为 10 的情况，如果 10 个层均采用顺序推理的方式，每一层存在 2 个输入，10 个层总共 20 个输入。而如果 10 个层均采用平行推理的方式，每一层仅存在一个输入，10 个层总共 10 个输入。

由于每层的输入的数量减少了一半，而相邻的输入（例如，在顺序推理中的相邻的模块的输入，或者，在平行推理中的相邻的层的输入）之间的变化率一定，因此，相比于采用顺序推理的方式，采用平行推理的方式可以使得预测模块可以预测两倍的层的活性参数的位置。例如，如果在采用顺序推理的方

式下，N2 取 8 时，即一次预测 8 层的活性参数的位置的情况下，预测到的活性参数的位置会比较准确，那么，在采用平行推理的方式下，N2 可以取到 16，即一次可以预测 16 层的活性参数的位置（假设这 16 层均采用平行推理），并且能够与采用顺序推理的情况下所预测到的 8 层的活性参数的位置同样准确。

通过上述实施例，N2 个层中的一个或多个层可以采取平行推理的形式，也就是一个层的第一模块在进行推理时，这个层的第二模块也可以进行推理。顺序推理是将一个层的第二模块推理结束后得到的输出，作为这个层的第二模块的输入。采用平行推理的层，相比于采用顺序推理的层，输入的数量可以从 2 个减少为 1 个。输入的数量减少，相邻的输入的变化率一定，可以增加能够选取的 N2 的数值，使得第二装置下发的第一信息所指示的活性参数可以是更多个层的活性参数，换言之，第二装置可以一次性下发更多的层的活性参数，从而进一步减少了第二装置向第一装置传输的时延。

结合图 8 的一些模块，在一些可能的实现中，端侧推理模块可以确定第一模型的输出值的置信度，在第一模型的输出值的置信度大于或者等于第一阈值的情况下，向预测模块发送指令，该指令用于指示预测模块停止预测。在一些可能的实现中，在第一模型的输出值的置信度大于或者等于第一阈值的情况下，第一模型可以停止推理。在一些可能的实现中，在第一模型的输出值的置信度大于或者等于第一阈值的情况下，第一模型可以向大模型推理 UI 和/或用户操作模块发送第一模型的输出值（或者称为第一模型的推理结果）。

通过上述实施例，在第一模型的输出值的置信度满足条件的情况下，第一装置可以发送第二信息，第二信息可以指示停止传输第一信息，从而减少传输第一信息所导致的资源浪费。此外，上述实施例可以应用于推理提前退出（early exiting）的场景，在第一模型提前退出推理的场景下，第一模型的退出推理不是由于第一模型的所有层都完成了推理，而是由于第一模型的输出值的置信度满足条件。推理提前退出可以使得第一模型中的一部分层不参与推理，从而减少了推理时延。上述实施例能够在第一模型提前退出推理的情况下，避免第二装置传输指示用于推理的活性参数的信息，从而节约了传输资源。

下文对本申请方法实施例对应的装置实施例进行介绍。下文仅对装置做简要介绍，方案具体实现步骤和细节可参考前文方法实施例。

为了实现本申请提供的方法中的各功能，通信装置可以包括硬件结构和/或软件模块，以硬件结构、软件模块、或硬件结构加软件模块的形式来实现上述各功能。上述各功能中的某个功能以硬件结构、软件模块、还是硬件结构加软件模块的方式来执行，取决于技术方案的特定应用和设计约束条件。

图 10 是本申请实施例的一种通信装置 1000 的示意框图。通信装置 1000 包括处理器 1010 和通信接口 1020。可选地，处理器 1010 和通信接口 1020 可以通过总线相互连接。通信装置 1000 可以是第一装置，也可以是第二装置。示例性地，第一装置可以是终端设备或端侧设备；第二装置可以是网络设备、核心网设备、云侧设备或边侧设备。

可选地，通信装置 1000 还可以包括存储器 1040。存储器 1040 包括但不限于是随机存储记忆体（random access memory，RAM）、只读存储器（read-only memory，ROM）、高速缓存（cache）、可擦除可编程只读存储器（erasable programmable read only memory，EPROM）、同步动态随机存取存储器（synchronous dynamic random access memory，SDRAM）、硬盘（hard disk drive，HDD）、固态硬盘（solid-state drive，SSD）或便携式只读存储器（compact disc read-only memory，CD-ROM），该存储器 1040 用于存储相关指令和/或数据。存储器 1040 可以与处理器 1010 集成在一起，或者分离设置。

示例性地，通信装置 1000 可以是第一装置。第一装置的内存可以包括 RAM、cache、SDRAM、或 CD-ROM。可选地，第一装置的内存可以用于加载第一活性参数。第一装置的闪存可以包括 ROM、EPROM、HDD、或 SSD。可选地，第一装置的闪存可以存储第二模块的全量参数。

示例性地，通信装置 1000 可以是第二装置。第二装置的闪存可以包括 ROM、EPROM、HDD、或 SSD。可选地，第一装置的闪存可以存储第一模块的全量参数。

处理器 1010 可以是一个或多个中央处理器（central processing unit，CPU）。在处理器 1010 是一个 CPU 的情况下，该 CPU 可以是单核 CPU，也可以是多核 CPU。但是本申请对此不限定，处理器 1010 也可以是一个或多个 GPU，还可以是一个或多个 TPU。其中，处理器 1010 可以是信号处理器、芯片，或其他可以实现本申请方法的集成电路，或者前述处理器、芯片或集成电路中的用于处理功能的部分电路。另外，通信接口 1020 也可以为输入输出接口，输入输出接口用于信号或数据的输入或输出，也可以是输入输出电路。

示例性地，通信装置 1000 为第一装置，处理器 1010 用于执行以下操作：接收第一信息，该第一信

息用于指示第一活性参数，该第一活性参数是第一模型中的 N 个第一模块的活性参数，N 为正整数；根据该第一活性参数，基于该 N 个第一模块进行推理。

示例性地，通信装置 1000 为第二装置，处理器 1010 用于执行以下操作：生成第一信息，该第一信息用于指示第一活性参数，该第一活性参数是第一模型中的 N 个第一模块的活性参数，该第一活性参数用于基于该 N 个第一模块进行推理，N 为正整数；发送该第一信息。

上述内容仅作为示例性描述。通信装置 1000 是将负责执行前述方法实施例中第一装置、或第二装置相关的方法或者步骤。

一种可能的实现中，通信接口 1020 可以为收发器。收发器可以包括发送器和接收器，发送器用于执行发送操作，接收器用于执行接收操作。例如，处理器 1010 用于控制收发器进行信号的接收和/或发送。

一种可能的实现中，通信接口 1020 还可以为通信电路、管脚、输入输出接口、总线等。

需要说明的是，通信装置 1000 可以包括发送器，而不包括接收器。或者，通信装置 1000 可以包括接收器，而不包括发送器。具体可以视通信装置 1000 执行的上述方案中是否包括发送动作和接收动作。

上述描述仅是示例性描述。具体内容可以参见上述方法实施例所示的内容。图 10 中的各个操作的实现还可以对应参照图 4 至图 9 所示的方法实施例的相应描述。

例如，通信装置 1000 可以用于执行图 4 至图 9 示出的方案。

示例性地，通信装置 1000 为第一装置，通信接口 1020 可以用于接收第一信息。

示例性地，通信装置 1000 为第二装置，通信接口 1020 可以用于发送第一信息。

对于其他实现方式具体可以参阅前述图 4 至图 9 所示的实施例的详细介绍，这里不再赘述。应理解，各部件执行上述相应过程的具体过程在上述方法实施例中已经详细说明，为了简洁，在此不再赘述。

图 11 是本申请实施例的另一种通信装置 1100 的示意框图。通信装置 1100 可以为第一装置或者第二装置，也可以为第一装置或者第二装置中的芯片或模块，用于实现图 4 至图 9 所示的实施例涉及的方法，具体请参考上述方法实施例中的相关介绍。

通信装置 1100 包括收发单元 1110。下面对该收发单元 1110 进行示例性地介绍。

收发单元 1110 可以包括发送单元和接收单元。发送单元用于执行通信装置的发送动作，接收单元用于执行通信装置的接收动作。为便于描述，本申请实施例将发送单元与接收单元合为一个收发单元。在此做统一说明，后文不再赘述。收发单元 1110 可以实现相应的通信功能。收发单元 1110 还可以称为通信接口或通信模块。

需要说明的是，通信装置 1100 可以包括发送单元，而不包括接收单元。或者，通信装置 1100 可以包括接收单元，而不包括发送单元。具体可以视通信装置 1100 执行的上述方案中是否包括发送动作和接收动作。

示例性地，收发单元 1110 用于接收第一信息等。

可选地，通信装置 1100 还可以包括处理单元 1120，其用于执行通信装置 1100 涉及处理、协调等步骤的内容。示例性地，处理单元 1120 用于根据该第一活性参数，基于该 N 个第一模块进行推理。

示例性地，收发单元 1110 用于发送第一信息等。

可选地，通信装置 1100 还可以包括处理单元 1120，其用于执行通信装置 1100 涉及处理、协调等步骤的内容。示例性地，处理单元 1120 用于生成第一信息。

上述所述内容仅作为示例性描述。通信装置 1100 将负责执行前述方法实施例中相关的方法或者步骤。

可选地，通信装置 1100 还包括存储单元 1130，该存储单元 1130 用于存储用于执行前述方法的程序或者代码。或者说，存储单元 1130 可以用于存储指令和/或数据，处理单元 1120 可以读取存储单元 1130 中的指令和/或数据，以使得通信装置 1100 实现前述方法实施例。例如，通信装置 1100 可以用于执行图 4 至图 9 示出的方案。

示例性地，收发单元 1110 可以用于接收第一信息，该第一信息用于指示第一活性参数，该第一活性参数是第一模型中的 N 个第一模块的活性参数，N 为正整数；处理单元 1120 可以用于根据该第一活性参数，基于该 N 个第一模块进行推理。

示例性地，处理单元 1120 可以用于生成第一信息，该第一信息用于指示第一活性参数，该第一活性参数是第一模型中的 N 个第一模块的活性参数，该第一活性参数用于基于该 N 个第一模块进行推理，N 为正整数；收发单元 1110 可以用于发送该第一信息。

对于其他实现方式具体可以参阅前述图 4 至图 9 所示的实施例的详细介绍，这里不再赘述。应理解，

各部件执行上述相应过程的具体过程在上述方法实施例中已经详细说明，为了简洁，在此不再赘述。

当图 10 中的通信装置 1100 为芯片时，通信接口 1120 可以是该芯片的收发器、输入输出电路或通信接口。处理器 1110 可以是该芯片上集成的处理器、或者微处理器、或者集成电路。上述方法实施例中第一装置、或第二装置的发送操作可以理解为芯片的输出，上述方法实施例中第一装置、或第二装置的接收操作可以理解为芯片的输入。

当图 10 中的通信装置 1100 为芯片时，收发单元 1110 可以是该芯片的收发器、输入输出电路或通信接口。处理单元 1120 可以是该芯片上集成的处理器、或者微处理器、或者集成电路。上述方法实施例中第一装置、或第二装置的发送操作可以理解为芯片的输出，上述方法实施例中第一装置、或第二装置的接收操作可以理解为芯片的输入。

本申请还提供了一种芯片，包括处理器，用于从存储器中调用并运行所述存储器中存储的指令，使得安装有所述芯片的通信装置执行上述各示例中的方法。

本申请还提供另一种芯片，包括：输入接口、输出接口、处理器，所述输入接口、输出接口以及所述处理器之间通过内部连接通路相连，所述处理器用于执行存储器中的代码，当所述代码被执行时，所述处理器用于执行上述各示例中的方法。可选地，该芯片还包括存储器，该存储器用于存储计算机程序或者代码。

本申请还提供了一种处理器，用于与存储器耦合，用于执行上述各实施例中任一实施例中涉及通信装置或编码装置的方法和功能。

在本申请的另一实施例中提供一种包含计算机程序或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，前述实施例的方法得以实现。

本申请还提供一种计算机程序，当该计算机程序在计算机中被运行时，前述实施例的方法得以实现。

在本申请的另一实施例中提供一种计算机可读存储介质，该计算机可读存储介质存储有计算机程序，该计算机程序被计算机执行时实现前述实施例所述的方法。

本申请还提供一种通信系统，该通信系统包括第一装置和第二装置。第一装置和第二装置分别用于执行前述实施例中第一装置和第二装置所执行的方法。

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时，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，或者网络设备等）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：U 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。

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。

权 利 要 求 书

1．一种通信方法，其特征在于，所述方法包括：

接收第一信息，所述第一信息用于指示第一活性参数，所述第一活性参数是第一模型中的 N 个第一模块的活性参数，N 为正整数；

根据所述第一活性参数，基于所述 N 个第一模块进行推理。

2．根据权利要求 1 所述的方法，其特征在于，所述 N 个第一模块包括 N1 个第一模块，所述 N1 个第一模块是所述第一模型中的至少一个层中的第一个层的模块，N1 是小于或等于 N 的正整数。

3．根据权利要求 2 所述的方法，其特征在于，所述 N1 个第一模块是所述第一模型的第一个层中的多个模块中的第二个，N1=1，所述方法还包括：

根据第二输入值，预测所述第一活性参数的位置的信息，其中，所述第二输入值为所述第一模型的第一个层中的多个模块中的第二个的输入值或所述第一模型的第一个层中的多个模块中的第一个的输出值。

4．根据权利要求 2 或 3 所述的方法，其特征在于，所述方法还包括：

根据第一输入值，预测第二活性参数的位置的信息，所述第二活性参数是所述第一模型的第一个层中的多个模块中的第一个的活性参数，所述第一输入值是所述第一模型的第一个层中的多个模块中的第一个的输入值。

5．根据权利要求 1 至 4 中任一项所述的方法，其特征在于，所述 N 个第一模块还包括 N2 个第一模块，所述第一模型包括 N2 个层，所述 N2 个层中的每个层包括所述 N2 个第一模块中的一个第一模块，N2≥2，N2 是小于或等于 N 的正整数。

6．根据权利要求 5 所述的方法，其特征在于，所述第一模型包括 M 个层，M 是大于 N2 的正整数，所述 N2 个层在所述 M 个层中的编号不连续。

7．根据权利要求 5 或 6 所述的方法，其特征在于，N2 与用于传输所述第一信息的资源和/或推理时延要求相关。

8．根据权利要求 5 至 7 中任一项所述的方法，其特征在于，所述方法还包括：

根据所述 N2 个层中第一个层的上一个层的输入值，预测所述第一活性参数的位置的信息。

9．根据权利要求 5 至 8 中任一项所述的方法，其特征在于，所述 N2 个层中的每个层还包括一个第二模块，所述一个第二模块是所述第一模型中的 N2 个第二模块之一，所述方法还包括：

根据所述 N2 个层中第一个层的上一个层的输入值，预测第三活性参数的位置的信息，所述第三活性参数是所述 N2 个第二模块的活性参数。

10．根据权利要求 9 所述的方法，其特征在于，所述第二模块为多头注意力 MHA 模块。

11．根据权利要求 5 至 10 中任一项所述的方法，其特征在于，所述 N2 个层中的每个层还包括一个第二模块，所述方法还包括：

根据所述 N2 个层中的第 k 层中的第一模块的输出值与所述第 k 层中的第二模块的输出值，确定第三输入值，所述第三输入值用于所述第 k 层的下一层中的第一模块的输入，以及所述第 k 层的所述下一层中的第二模块的输入，k 是小于或等于 N2 的正整数。

12．根据权利要求 1 至 11 中任一项所述的方法，其特征在于，所述方法还包括：

在所述第一模型的输出值的置信度大于或等于第一阈值的情况下，发送第二信息，所述第二信息用于指示停止传输所述第一信息。

13．根据权利要求 1 至 12 中任一项所述的方法，其特征在于，所述方法还包括：

发送第三信息，所述第三信息包括指示所述第一活性参数的位置的信息，所述第三信息用于请求所述第一活性参数。

14．根据权利要求 1 至 13 中任一项所述的方法，其特征在于，所述第一模块为前馈神经网络 FNN 模块。

15．一种通信方法，其特征在于，所述方法包括：

生成第一信息，所述第一信息用于指示第一活性参数，所述第一活性参数是第一模型中的 N 个第一模块的活性参数，所述第一活性参数用于基于所述 N 个第一模块进行推理，N 为正整数；

发送所述第一信息。

16. 根据权利要求 15 所述的方法，其特征在于，所述 N 个第一模块包括 N1 个第一模块，所述 N1 个第一模块是所述第一模型中的至少一个层中的第一个层的模块，N1 是小于或等于 N 的正整数。

17. 根据权利要求 15 或 16 所述的方法，其特征在于，所述 N 个第一模块还包括 N2 个第一模块，所述第一模型包括 N2 个层，所述 N2 个层中的每个层包括所述 N2 个第一模块中的一个第一模块，N2≥2，N2 是小于或等于 N 的正整数。

18. 根据权利要求 17 所述的方法，其特征在于，所述第一模型包括 M 个层，M 是大于 N2 的正整数，所述 N2 个层在所述 M 个层中的编号不连续。

19. 根据权利要求 17 或 18 所述的方法，其特征在于，N2 与用于传输所述第一信息的资源和/或推理时延要求相关。

20. 根据权利要求 15 至 19 中任一项所述的方法，其特征在于，所述方法还包括：

接收第二信息，所述第二信息用于指示停止传输所述第一信息。

21. 根据权利要求 15 至 20 中任一项所述的方法，其特征在于，所述方法还包括：

接收第三信息，所述第三信息包括所述第一活性参数的位置的指示信息。

22. 根据权利要求 21 所述的方法，其特征在于，所述发送第一信息，包括：

响应于所述第三信息，发送所述第一信息。

23. 根据权利要求 15 至 22 中任一项所述的方法，其特征在于，所述第一模块为前馈神经网络 FNN 模块。

24. 一种通信装置，其特征在于，包括至少一个模块或至少一个单元，所述至少一个模块或所述至少一个单元用于执行权利要求 1 至 15 中任一项所述的方法，或者，所述至少一个模块或所述至少一个单元用于执行权利要求 16 至 23 中任一项所述的方法。

25. 一种通信装置，其特征在于，包括：处理器，所述处理器用于，通过执行计算机程序或指令，使得所述通信装置执行权利要求 1 至 14 中任一项所述的方法，或者，使得所述通信装置执行权利要求 15 至 23 中任一项所述的方法。

26. 根据权利要求 25 所述的通信装置，其特征在于，所述通信装置还包括存储器，所述存储器用于存储所述计算机程序或所述指令。

27. 一种计算机可读存储介质，其特征在于，所述计算机可读存储介质上存储有计算机程序或指令，当所述计算机程序或所述指令在计算机上运行时，使得权利要求 1 至 14 中任一项所述的方法被执行，或者，使得权利要求 15 至 23 中任一项所述的方法被执行。

28. 一种计算机程序产品，其特征在于，包括计算机程序或指令，当所述计算机程序或指令被运行时，如权利要求 1 至 14 中任一项所述的方法被实现，或者，如权利要求 15 至 23 中任一项所述的方法被实现。

29. 一种通信系统，其特征在于，包括第一装置和第二装置，其中，所述第一装置用于执行权利要求 1 至 14 中任一项所述的方法，所述第二装置用于执行权利要求 15 至 23 中任一项所述的方法。

30. 一种芯片，其特征在于，包括处理器，用于调用存储器中的计算机程序或计算机指令，以使得所述芯片执行如权利要求 1 至 14 中任一项所述的方法；或者，执行如权利要求 15 至 23 中任一项所述的方法。