

模型的运行

技术领域

[01] 本说明书一个或多个实施例涉及通信技术，尤其涉及模型的运行方法和装置。

背景技术

[02] 在人工智能（Artificial Intelligence, AI）及机器学习等技术中，会建立大量的模型，并利用模型来完成一系列的计算处理。

[03] 随着隐私保护与效率的要求，越来越多的模型不再被部署在服务器中，而被下发到终端设备中运行。但是，终端设备的运行环境相对开放，模型在终端设备上运行，则会面临更多的被逆向与窃取的风险，从而大大降低了模型的安全性，甚至导致整个系统的安全受到威胁。

发明内容

[04] 本说明书一个或多个实施例描述了模型的运行方法和装置，能够提高模型运行的安全性。

[05] 根据第一方面，提供了一种模型的运行方法，其中，包括：对原始模型进行拆分，得到基础模型以及 TEE（Trusted Execution Environment，可信执行环境）模型；将所述基础模型的数据以及所述 TEE 模型的数据下发到终端设备，以使所述终端设备中的 REE（Rich Execution Environment，富执行环境）运行所述基础模型的数据，且所述终端设备中的 TEE 运行所述 TEE 模型的数据。

[06] 其中，所述对原始模型进行拆分，包括如下中的至少一项：根据预设的拆分比例对原始模型进行拆分；将所述原始模型使用的神经网络中的前 M 层拆分为所述基础模型，后 N 层拆分为所述 TEE 模型；其中，所述神经网络的总层数为 M 加 N；根据终端设备中 TEE 的运行空间的大小，对原始模型进行拆分。

[07] 所述将所述基础模型的数据以及所述 TEE 模型的数据下发到终端设备，包括：将所述基础模型的数据下发到终端设备中的 REE；将所述 TEE 模型的数据通过 TEE 专用通道下发到终端设备中的 TEE。

[08] 在得到 TEE 模型之后，并在将 TEE 模型的数据下发到终端设备之前，进一步包括：根据预先获取的加密密钥，对 TEE 模型进行加密，得到加密模型；所述 TEE 模型的数据，包括：加密模型。

[09] 所述将所述基础模型的数据以及所述 TEE 模型的数据下发到终端设备，包括：将所

述加密模型封装成 TEE 算子；根据所述基础模型以及所述 TEE 算子，生成重塑模型；将所述重塑模型下发到所述终端设备中的 REE。

[10]根据第二方面，提供了一种模型的运行方法，其中，包括：接收基础模型的数据以及 TEE 模型的数据；其中，所述基础模型的数据以及所述 TEE 模型的数据是对原始模型进行拆分之后得到的；通过 REE 运行所述基础模型；通过 TEE 运行所述 TEE 模型。

[11]所述接收基础模型的数据以及 TEE 模型的数据，包括：由所述 REE 接收基础模型的数据；有所述 TEE 通过 TEE 专用通道接收所述 TEE 模型的数据。

[12]所述 TEE 模型的数据，包括：对所述 TEE 模型加密后生成的加密模型。

[13]所述接收基础模型的数据以及 TEE 模型的数据，包括：REE 接收重塑模型；该重塑模型根据所述基础模型以及所述加密模型封装出的 TEE 算子生成；所述通过 REE 运行所述基础模型，包括：REE 运行该重塑模型中的基础模型；所述通过 TEE 运行所述 TEE 模型，包括：REE 运行到重塑模型中的 TEE 算子时，无法识别该 TEE 算子，则将该 TEE 算子发送给 TEE；TEE 对 TEE 算子解封装，得到所述加密模型；TEE 根据预先得到的解密密钥对所述加密模型进行解密，得到 TEE 模型；TEE 运行该 TEE 模型。

[14]其中，在所述通过 TEE 运行所述 TEE 模型之后，进一步包括：TEE 将运行 TEE 模型后得到的特征向量输出到 TEE 的外部；或者，TEE 根据运行 TEE 模型后得到的特征向量获取识别结果，将该识别结果输出到 TEE 的外部。

[15]根据第三方面，提供了一种模型的运行装置，其中，包括：模型重塑模块，配置为对原始模型进行拆分，得到基础模型以及 TEE 模型；模型下发模块，配置为将基础模型的数据以及 TEE 模型的数据下发到终端设备，以使所述终端设备中的 REE 运行基础模型的数据，且终端设备中的 TEE 运行 TEE 模型的数据。

[16]根据第四方面，提供了一种模型的运行装置，其中，包括：接收模块，配置为接收基础模型的数据以及 TEE 模型的数据；其中，所述基础模型的数据以及所述 TEE 模型的数据是对原始模型进行拆分之后得到的；REE，配置为运行所述基础模型；TEE，配置为运行所述 TEE 模型。

[17]根据第五方面，提供了一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现本说明书任一实施例所述的方法。

[18]本说明书实施例提供的模型的运行方法及装置，在服务器侧，将模型进行拆分，拆分出基础模型部分以及 TEE 模型部分。相应地，在终端设备侧，由 REE 运行基础模型，由 TEE 运行 TEE 模型。因为 TEE 在终端设备中构建了一个可信执行环境，因此能够保证 TEE 模型运行时的安全性。同时，由于不是将完整的原始模型放入 TEE 运行，因此，

即使终端设备中安装了众多的原始模型，也能够减轻 TEE 的运行压力。

附图说明

[19]为了更清楚地说明本说明书实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本说明书的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[20]图 1 是本说明书一个实施例所应用的系统架构的示意图。

[21]图 2 是本说明书一个实施例中在服务器侧实现模型运行的方法的流程图。

[22]图 3 是本说明书一个实施例中实现模型运行的方法的示意图。

[23]图 4 是本说明书一个实施例中在终端设备侧实现模型运行的方法流程图。

[24]图 5 是本说明书一个实施例中模型运行装置的结构示意图。

[25]图 6 是本说明书另一个实施例中模型运行装置的结构示意图。

具体实施方式

[26]如前所述，模型在终端设备上部署及运行时，会面临更多的被逆向与窃取的风险，从而大大降低了模型的安全性。

[27]下面结合附图，对本说明书提供的方案进行描述。

[28]为了方便对本说明书的理解，首先对本说明书所应用的系统架构进行描述。如图 1 中所示，该系统架构主要包括服务器及终端设备。其中，终端设备可以是任意一种位于用户侧的设备，比如手机、电脑以及物联网终端设备比如基于物联网的室外广告投放机或基于物联网的自动售卖设备等。服务器及终端设备通过网络交互。其中，网络可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等。

[29]参见图 1，本说明书的核心思想是：在服务器侧，将模型进行拆分，拆分出基础模型部分以及 TEE 模型部分。相应地，在终端设备侧，由 REE 运行基础模型，由 TEE 运行 TEE 模型。因为 TEE 在终端设备中构建了一个可信执行环境，因此能够保证 TEE 模型运行时的安全性。同时，由于不是将完整的原始模型放入 TEE 运行，因此，即使终端设备中安装了众多的原始模型，也能够减轻 TEE 的运行压力。

[30]下面用不同的实施例对服务器侧的处理以及终端设备侧的处理分别进行说明。

[31]首先针对服务器侧的处理提供实施例。

[32]图 2 是本说明书一个实施例中在服务器侧实现模型运行的方法的流程图。该方法的执行主体为模型的运行装置。该装置可以位于服务器中。可以理解，该方法也可以通过

任何具有计算、处理能力的装置、设备、平台、设备集群来执行。参见图 2，该方法包括：步骤 201：对原始模型进行拆分，得到基础模型以及 TEE 模型。

[33]步骤 203：将所述基础模型的数据以及所述 TEE 模型的数据下发到终端设备，以使所述终端设备中的 REE 运行所述基础模型的数据，且所述终端设备中的 TEE 运行所述 TEE 模型的数据。

[34]下面结合具体实施例说明图 2 中每一个步骤的处理。

[35]首先对于步骤 201，服务器对原始模型进行拆分，得到基础模型以及 TEE 模型。

[36]TEE 可以在终端设备中提供一个安全区域，确保敏感数据在隔离、可信的环境中存储、处理和保护。因此，为了提高在终端设备中运行模型时的安全性，可以考虑将模型放入终端设备中的 TEE 进行运行。但是，目前，随着终端设备处理事务的不断增加，终端设备中下载的模型的数量也越来越多，比如进行人脸识别的模型、风险评估模型、用户购物行为统计模型等等。如果将终端设备需要运行的数量众多的模型均放入 TEE 中运行，那么则会导致 TEE 的运行压力过大，降低终端设备的处理性能。因此，可以考虑将原始模型拆分为两部分，基础模型部分由终端设备中的通用环境即 REE 运行，而 TEE 模型部分由终端设备中的 TEE 运行，从而既能提高在终端设备中运行模型的安全性，又能减轻终端设备中 TEE 的运行压力。因此，在本步骤 201 中，可以以离线方式在服务器中对一个原始模型进行拆分。

[37]在本说明书的一个实施例中，本步骤 201 至少存在如下的三种实现方式：方式一、根据预设的拆分比例对原始模型进行拆分。

[38]比如，预设的拆分比例为基础模型与 TEE 模型各占一半，那么采用方式一时，则可以将原始模型文件按照文件大小分为前后两部分，可以将原始模型的前半部分作为基础模型，将原始模型的后半部分作为 TEE 模型。当然，也可以有其他的拆分方式，比如按照原始模型文件的大小，将原始模型的中间 1/3 部分作为基础模型，将原始模型的前 1/3 部分及后 1/3 部分作为 TEE 模型等。

[39]方式二、将原始模型使用的神经网络中的前 M 层拆分为基础模型，后 N 层拆分为 TEE 模型；其中，神经网络的总层数为 M 加 N。其中，M 和 N 均为正整数。

[40]一个原始模型通常采用神经网络实现。神经网络包括多个层。比如卷积神经网络（CNN）包括：输入层、卷积层、激活函数、池化层、全连接层。在神经网络的各个层中，通常越靠后的层越可能涉及到该神经网络在训练过程中的参数调整，其对于该原始模型的贡献越大，也就是说，对安全性的要求越高。因此，在本说明书的一个实施例中，可以采用方式二将原始模型使用的神经网络的层数依次划分为 M 加 N 层，前 M 层因为

对安全性的要求相对较低，因此，该前 M 层可以作为基础模型；后 N 层通常涉及参数调整，对安全性的要求相对较高，因此，该后 N 层可以作为 TEE 模型。

[41]方式三、根据终端设备中 TEE 的运行空间的大小，对原始模型进行拆分。

[42]终端设备中之前可能会下载有数量众多的模型，那么在 TEE 中需要运行的模型的数量也会众多。因此，每当终端设备需要下载新的模型时，需要考虑终端设备中 TEE 的运行空间的大小，以便保证 TEE 的运行。因此，在本说明书的一个实施例中，可以采用方式三，根据终端设备中 TEE 的运行空间的大小，对原始模型进行拆分，以使得拆分出的 TEE 模型满足该 TEE 的运行空间大小的要求，从而能够在 TEE 中运行。

[43]在本说明书的实施例中，上述三种方式中的任意两种或者三种也可以结合使用。比如，结合上述方式二和方式三，同时考虑终端设备中运行空间的大小、原始模型的神经网络中涉及参数调整的后 N 层的数量，来对原始模型进行拆分。

[44]接下来在步骤 203 中，服务器将基础模型的数据以及所述 TEE 模型的数据下发到终端设备，以使所述终端设备中的 REE 运行所述基础模型的数据，且所述终端设备中的 TEE 运行所述 TEE 模型的数据。

[45]步骤 203 的具体实现过程包括至少如下的实现方式：方式 A、直接各自分发。

[46]服务器将拆分出的基础模型的数据直接下发到终端设备中的 REE，并将拆分出的 TEE 模型的数据通过与 TEE 之间的专用通道，直接下发到终端设备的 TEE。

[47]方式 B、加密后统一分发。

[48]服务器将拆分出的基础模型的数据以及 TEE 模型加密后得到的加密模型统一下发到 REE。

[49]在原有的业务下发逻辑中，服务器是将原始模型直接下发到终端设备中的 REE。因此，为了尽可能的减少对业务下发逻辑的改动，可以采用该方式 B。

[50]参见图 3，基于方式 B，步骤 203 的一种具体实现过程包括：步骤 301：服务器根据预先获取的加密密钥，对 TEE 模型进行加密，得到加密模型。

[51]在后续过程中，因为会将 TEE 模型相关的数据也发送到终端设备中的 REE。如果发送明文形式的 TEE 模型，那么，可能会造成 TEE 模型在 REE 中的泄露。因此，为了提高安全性，可以对 TEE 模型进行加密，并将得到的加密模型的相关数据发给 REE，从而避免 TEE 模型在 REE 中的泄露，进一步提高模型运行的安全性。

[52]在本说明书的一个实施例中，本步骤 301 可以采用对称加密算法进行加密。

[53]可以预先使服务器获取加密密钥，使终端设备中的 TEE 获取解密密钥。获取的方法可以是静态获取，比如，服务器跟 TEE 预先约定好一个密钥。获取的方法也可以是动

态获取方法,比如由服务器生成密钥,并通过与 TEE 之间的专用通道,将密钥发给 TEE。

[54]步骤 303: 服务器将加密模型封装成 TEE 算子。

[55]本步骤 303 的目的是:将 TEE 模型加密后得到的加密模型从形式上更改为一个算子,从而能够将拆分出的两部分模型重新组合成一个重塑模型,并按照原有业务逻辑将该一个重塑模型下发给终端设备的 REE,使得 REE 感知不到对原始模型的改动,也就是说,在业务链路层面,服务器如何下发,REE 如何执行接收到的模型都没有改变,从而减少了对整体业务实现逻辑的改动,使得方案更容易实现。

[56]步骤 305: 服务器根据基础模型以及 TEE 算子,生成重塑模型。

[57]在前述步骤中,在将原始模型拆分成基础模型及 TEE 模型之后,基础模型及 TEE 模型中则分别都包括了至少一个算子,比如,基础模型中有 3 个算子,TEE 模型中有 3 个算子,在将该 TEE 模型加密并封装成一个 TEE 算子后,TEE 模型对外则表现为 1 个算子。这样,经过本步骤 305 的处理,是利用基础模型中包括的 3 个算子以及封装后的 1 个 TEE 算子,生成了一个重塑模型,重塑模型对 REE 表现为 4 个算子。

[58]在本步骤 305 中,可以按照步骤 201 中的拆分顺序,来顺序组装基础模型以及 TEE 算子从而生成重塑模型。比如,在步骤 201 中,是将原始模型文件的前半部分、后半部分依次拆分为基础模型及 TEE 模型,那么在本步骤 305 中,也是按照基础模型在前 TEE 算子在后的顺序,生成重塑模型。

[59]步骤 307: 服务器将重塑模型下发到终端设备中的 REE。

[60]采用方式 B,将重塑模型下发到终端设备中的 REE 的好处在于:首先,如前所述,重塑模型下发给终端设备的 REE,使得 REE 感知不到对原始模型的改动,也就是说,在业务链路层面,服务器如何下发,REE 如何执行接收到的模型都没有改变,从而减少了对整体业务实现逻辑的改动,使得方案更容易实现;其次,一个模型中包括多个算子,各个算子之间有执行上的先后顺序。为了保证终端设备针对拆分后的两部分模型仍然能够执行原始模型中各算子的执行顺序,即进行时序控制,则需要将重塑模型下发到 REE,REE 统一执行该重塑模型,则可以保证重塑模型中各个算子的执行顺序与原始模型中各个算子的执行顺序相同。

[61]下面针对终端设备侧的处理提供实施例。

[62]图 4 是本说明书一个实施例中在终端设备侧实现模型运行的方法的流程图。该方法的执行主体为模型的运行装置。该装置可以位于终端设备中。可以理解,该方法也可以通过任何具有计算、处理能力的装置、设备、平台、设备集群来执行。参见图 4,该方法包括:步骤 401:接收基础模型的数据以及 TEE 模型的数据;其中,所述基础模型的

数据以及所述 TEE 模型的数据是对原始模型进行拆分之后得到的；步骤 403：通过 REE 运行所述基础模型；步骤 405：通过 TEE 运行所述 TEE 模型。

[63]下面分别对图 4 中的每一个步骤进行说明。

[64]首先对于步骤 401，接收基础模型的数据以及 TEE 模型的数据。

[65]参见上述对服务器侧处理的各个实施例，如果服务器采用上述步骤 203 中的方式 A，那么，在本步骤 401 中，终端设备中的 REE 接收基础模型的数据，并且终端设备中的 TEE 通过 TEE 专用通道接收 TEE 模型的数据。

[66]参见上述对服务器侧处理的各个实施例，如果服务器采用上述步骤 203 中的方式 B，那么，在本步骤 401 中，终端设备中的 REE 接收重塑模型，该重塑模型包括了基础模型以及根据 TEE 模型的加密模型所封装出的 TEE 算子。

[67]接下来，在步骤 403 中，通过 REE 运行基础模型。

[68]对应于上述方式 B，REE 接收到重塑模型之后，会将该重塑模型视为一个“原始模型”，并依次执行该重塑模型中的各个算子，从而可以顺利运行基础模型。

[69]接下来，在步骤 405 中，通过 TEE 运行 TEE 模型。

[70]对应于上述方式 B，参见图 3，本步骤 405 的具体实现过程可以包括：步骤 4051：REE 在执行重塑模型的过程中，当执行到该重新模型包括的 TEE 算子之后，无法识别该 TEE 算子。

[71]在重塑模型中，TEE 算子的得到过程包括：对 TEE 模型加密，对加密模型进行封装。虽然 TEE 算子表现为一个算子的形式，但是并非一个正常的算子，并且内部进行了加密，因此，REE 无法识别该 TEE 算子。

[72]可见，REE 既能够按照重塑模型中算子的执行顺序进行运行，又无法识别 TEE 模型，满足了业务要求。

[73]步骤 4053：REE 将该 TEE 算子发送给 TEE。

[74]这里，REE 调用对外的代理 CA（Certificate Authority，电子签证机关），由 CA 将 TEE 算子发送到 TEE 内部。

[75]这里，REE 可以将基础模型的运行结果作为 TEE 算子的输入，并将该基础模型的运行结果与 TEE 算子一起发给 TEE。

[76]步骤 4055：TEE 对 TEE 算子解封装，得到加密模型。

[77]步骤 4057：TEE 根据预先得到的解密密钥对加密模型进行解密，得到 TEE 模型。

[78]本步骤 4057 中 TEE 得到解密密钥的方法可以参见上述实施例中对步骤 301 的相关描述。

[79]步骤 4059: TEE 运行该 TEE 模型。

[80]至此, 则实现了服务器下发根据原始模型重塑的重塑模型(其包括的算子的执行顺序相同), 终端设备中由 REE 运行原始模型的一部分即基础模型, 由 TEE 运行原始模型的另一部分即 TEE 模型, 既能保证在 TEE 中运行的 TEE 模型的安全性, 又无需将完整原始模型全部放入 TEE, 减轻了 TEE 的运行压力。

[81]在本说明书的一个实施例中, 在 TEE 运行 TEE 模型之后, 会得到输出结果, 该输出结果是原始模型的输出结果。TEE 需要将该输出结果提供给外部, 比如经由 REE 返回给服务器, 从而使得服务器能够根据该输出结果进行相关业务的判断和决策。

[82]在本说明书的一个实施例中, TEE 对外提供的输出结果包括至少如下两种形式。

[83]形式 1: 原始模型的直接结果。

[84]对于模型使用的神经网络, 得到的直接结果通常为一个特征向量。TEE 可以将运行 TEE 模型后得到的该特征向量输出到 TEE 的外部。

[85]形式 2: 根据原始模型的直接结果得到的间接结果。

[86]模型使用的神经网络得到的直接结果为特征向量。如果将该特征向量直接提供给外部, 那么, 在某些场景下, 有可能会出现某些高级攻击者通过该特征向量反推出模型使用的神经网络的结构及参数, 从而导致模型泄露。因此, 为了进一步提高模型运行的安全性, TEE 也可以不输出该特征向量, 而是根据该特征向量获取一个识别结果, 比如一个分类的概率值, 分数值等, 然后 TEE 将该识别结果输出到 TEE 的外部。因为 TEE 输出到外部的是更低维度的用户可识别的识别结果, 因此进一步避免了模型泄露, 进一步提高了安全性。

[87]在本说明书的一个实施例中, 提供了一种模型的运行装置, 参见图 5, 该装置 500 可以设置在服务器中, 包括: 模型重塑模块 501, 配置为对原始模型进行拆分, 得到基础模型以及 TEE 模型; 模型下发模块 502, 配置为将基础模型的数据以及 TEE 模型的数据下发到终端设备, 以使所述终端设备中的 REE 运行基础模型的数据, 且终端设备中的 TEE 运行 TEE 模型的数据。

[88]在本说明书装置的一个实施例中, 上述模型重塑模块 501 被配置为执行如下中的至少一项: 根据预设的拆分比例对原始模型进行拆分; 将所述原始模型使用的神经网络中的前 M 层拆分为所述基础模型, 后 N 层拆分为所述 TEE 模型; 其中, 后 N 层涉及参数调整, 且所述神经网络的总层数为 M 加 N; 根据终端设备中 TEE 的运行空间的大小, 对原始模型进行拆分。

[89]在本说明书装置的一个实施例中, 模型下发模块 502 被配置为执行: 将所述基础模

型的数据下发到所述终端设备中的 REE；将所述加密模型的数据通过与 TEE 之间的专用通道下发到所述终端设备中的 TEE。

[90]在本说明书装置的一个实施例中，上述模型重塑模块 501 还可以被配置为执行：根据预先获取的加密密钥，对 TEE 模型进行加密，得到加密模型；相应地，所述 TEE 模型的数据，包括：加密模型。

[91]在本说明书装置的一个实施例中，上述模型重塑模块 501 被配置为执行：将所述加密模型封装成 TEE 算子；根据所述基础模型以及所述 TEE 算子，生成重塑模型；相应地，上述模型下发模块 502 被配置为执行：将重塑模型下发到终端设备中的 REE。

[92]在本说明书的一个实施例中，提出了一种模型的运行装置。参见图 6，该装置 600 可以被设置在终端设备中，包括：接收模块 601，配置为接收基础模型的数据以及 TEE 模型的数据；其中，所述基础模型的数据以及所述 TEE 模型的数据是对原始模型进行拆分之后得到的；REE 602，配置为运行所述基础模型；TEE 603，配置为运行所述 TEE 模型。

[93]在本说明书装置的一个实施例中，接收模块 601 包括：第一接收子模块，被设置在 REE 602 中，配置为接收服务器发来的基础模型的数据；第二接收子模块，被设置在 TEE 603 中，配置为通过 TEE 专用通道接收服务器发来的加密模型的数据。

[94]在本说明书装置的一个实施例中，TEE 模型的数据包括：对所述 TEE 模型加密后生成的加密模型的数据。

[95]在本说明书装置的一个实施例中，接收模块 601 被设置在 REE 中，被配置为接收重塑模型；该重塑模型根据所述基础模型以及所述加密模型封装出的 TEE 算子生成；相应地，REE 602 被配置为执行：运行该重塑模型中的基础模型；REE 运行到重塑模型中的 TEE 算子时，无法识别该 TEE 算子，则将该 TEE 算子发送给 TEE；相应地，TEE 603 被配置为执行：TEE 对 TEE 算子解封装，得到所述加密模型；TEE 根据预先得到的解密密钥对所述加密模型进行解密，得到 TEE 模型；TEE 运行该 TEE 模型。

[96]在本说明书装置的一个实施例中，TEE 603 进一步被配置为执行：将运行 TEE 模型后得到的特征向量输出到 TEE 的外部；或者，根据运行 TEE 模型后得到的特征向量获取识别结果，将该识别结果输出到 TEE 的外部。

[97]本说明书一个实施例提供了一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行说明书中任一实施例中的方法。

[98]本说明书一个实施例提供了一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现执行说明书中任一实施例

中的方法。

[99]可以理解的是，本说明书实施例示意的结构并不构成对本说明书实施例的装置的具体限定。在说明书的另一些实施例中，上述装置可以包括比图示更多或者更少的部件，或者组合某些部件，或者拆分某些部件，或者不同的部件布置。图示的部件可以以硬件、软件或者软件和硬件的组合来实现。

[100]上述装置、系统内的各模块之间的信息交互、执行过程等内容，由于与本说明书方法实施例基于同一构思，具体内容可参见本说明书方法实施例中的叙述，此处不再赘述。

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

[102]本领域技术人员应该可以意识到，在上述一个或多个示例中，本发明所描述的功能可以用硬件、软件、挂件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。

[103]以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本发明的保护范围之内。

权利要求书

1、一种模型的运行方法，其中，包括：

对原始模型进行拆分，得到基础模型以及可信执行环境 TEE 模型；

将所述基础模型的数据以及所述 TEE 模型的数据下发到终端设备，以使所述终端设备中的富执行环境 REE 运行所述基础模型的数据，且所述终端设备中的 TEE 运行所述 TEE 模型的数据。

2、根据权利要求 1 所述的方法，其中，所述对原始模型进行拆分，包括如下中的至少一项：

根据预设的拆分比例对原始模型进行拆分；

将所述原始模型使用的神经网络中的前 M 层拆分为所述基础模型，后 N 层拆分为所述 TEE 模型；其中，所述神经网络的总层数为 M 加 N；

根据终端设备中 TEE 的运行空间的大小，对原始模型进行拆分。

3、根据权利要求 1 所述的方法，其中，

所述将所述基础模型的数据以及所述 TEE 模型的数据下发到终端设备，包括：

将所述基础模型的数据下发到终端设备中的 REE；

将所述 TEE 模型的数据通过 TEE 专用通道下发到终端设备中的 TEE。

4、根据权利要求 1 所述的方法，其中，

在得到 TEE 模型之后，并在将 TEE 模型的数据下发到终端设备之前，进一步包括：

根据预先获取的加密密钥，对 TEE 模型进行加密，得到加密模型；

所述 TEE 模型的数据，包括：加密模型。

5、根据权利要求 4 所述的方法，其中，

所述将所述基础模型的数据以及所述 TEE 模型的数据下发到终端设备，包括：

将所述加密模型封装成 TEE 算子；

根据所述基础模型以及所述 TEE 算子，生成重塑模型；

将所述重塑模型下发到所述终端设备中的 REE。

6、一种模型的运行方法，其中，包括：

接收基础模型的数据以及可信执行环境 TEE 模型的数据；其中，所述基础模型的数据以及所述 TEE 模型的数据是对原始模型进行拆分之后得到的；

通过富执行环境 REE 运行所述基础模型；

通过 TEE 运行所述 TEE 模型。

7、根据权利要求 6 所述的方法，其中，

所述接收基础模型的数据以及 TEE 模型的数据, 包括:

由所述 REE 接收基础模型的数据;

有所述 TEE 通过 TEE 专用通道接收所述 TEE 模型的数据。

8、根据权利要求 6 所述的方法, 其中,

所述 TEE 模型的数据, 包括: 对所述 TEE 模型加密后生成的加密模型。

9、根据权利要求 8 所述的方法, 其中,

所述接收基础模型的数据以及 TEE 模型的数据, 包括: REE 接收重塑模型; 该重塑模型根据所述基础模型以及所述加密模型封装出的 TEE 算子生成;

所述通过 REE 运行所述基础模型, 包括: REE 运行该重塑模型中的基础模型;

所述通过 TEE 运行所述 TEE 模型, 包括:

REE 运行到重塑模型中的 TEE 算子时, 无法识别该 TEE 算子, 则将该 TEE 算子发送给 TEE;

TEE 对 TEE 算子解封装, 得到所述加密模型;

TEE 根据预先得到的解密密钥对所述加密模型进行解密, 得到 TEE 模型;

TEE 运行该 TEE 模型。

10、根据权利要求 6 至 9 中任一所述的方法, 其中, 在所述通过 TEE 运行所述 TEE 模型之后, 进一步包括:

TEE 将运行 TEE 模型后得到的特征向量输出到 TEE 的外部;

或者,

TEE 根据运行 TEE 模型后得到的特征向量获取识别结果, 将该识别结果输出到 TEE 的外部。

11、一种模型的运行装置, 其中, 包括:

模型重塑模块, 配置为对原始模型进行拆分, 得到基础模型以及可信执行环境 TEE 模型;

模型下发模块, 配置为将基础模型的数据以及 TEE 模型的数据下发到终端设备, 以使所述终端设备中的富执行环境 REE 运行基础模型的数据, 且终端设备中的 TEE 运行 TEE 模型的数据。

12、一种模型的运行装置, 其中, 包括:

接收模块, 配置为接收基础模型的数据以及可信执行环境 TEE 模型的数据; 其中, 所述基础模型的数据以及所述 TEE 模型的数据是对原始模型进行拆分之后得到的;

富执行环境 REE, 配置为运行所述基础模型;

TEE，配置为运行所述 TEE 模型。

13、一种计算设备，包括存储器和处理器，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现权利要求 1-10 中任一项所述的方法。