

权利要求书

1. 一种基于芯片的数据处理方法，所述芯片包括嵌入式 CPU 和多个算子单元，用于执行计算任务，所述方法在芯片端执行，包括：
- 获取待处理数据；以及
- 5 通过由所述嵌入式 CPU 执行以下步骤而进行所述计算任务的计算：配置所述多个算子单元，以使得所述多个算子单元与该计算任务中包括的不同类型的计算分别对应，以及，基于所述待处理数据，通过调用所述多个算子单元，进行所述计算任务的计算，以获取处理结果。
2. 一种基于芯片的数据处理方法，所述芯片包括嵌入式 CPU 和多个算子单元，用于执行神经网络的预测计算，所述方法在芯片端执行，包括：
- 10 获取待处理数据；以及
- 基于所述待处理数据，执行所述神经网络中的每一层的计算，以获取处理结果，其中，通过由所述嵌入式 CPU 执行以下步骤而进行所述神经网络的每一层的计算：配置所述多个算子单元，以使得所述多个算子单元与该层中包括的不同类型的计算分别对应，
- 15 以及，基于神经网络上一层中各个神经元的值，通过调用所述多个算子单元，执行该层中每个神经元处的计算，以获取该层中每个神经元的值。
3. 根据权利要求 1 或 2 所述的数据处理方法，其中，所述芯片与服务器连接。
4. 根据权利要求 3 所述的数据处理方法，其中，所述待处理数据通过所述服务器传
- 送至所述芯片。
- 20 5. 根据权利要求 3 所述的数据处理方法，其中，所述嵌入式 CPU 通过运行基于程序的汇编指令而执行所述步骤，其中，所述程序由所述服务器传送至所述芯片。
6. 根据权利要求 5 所述的数据处理方法，其中，所述程序为 C 语言程序。
7. 根据权利要求 6 所述的数据处理方法，其中，调用所述多个算子单元包括，通过 C 语言程序中的函数调用，调用所述多个算子单元。
- 25 8. 根据权利要求 1 或 2 所述的数据处理方法，其中，所述芯片还包括 DMA 模块，所述方法还包括，在获取所述处理结果之后，通过所述嵌入式 CPU 控制所述 DMA 模块，以将所述处理结果写入所述服务器的内存中。
9. 根据权利要求 1 或 2 所述的数据处理方法，其中，调用所述多个算子单元包括，异步地调用所述多个算子单元。
- 30 10. 根据权利要求 1 或 2 所述的数据处理方法，其中，配置所述多个算子单元包括，对所述待处理数据的长度进行判断，并基于所述长度获取发送给各个算子单元的参数。

- 11.根据权利要求 2 所述的数据处理方法,其中,所述芯片用于执行多种神经网络的预测计算,其中,获取待处理数据包括,获取待处理数据及其对应的神经网络的类型,其中,基于所述待处理数据,执行所述神经网络中的每一层的计算,以获取处理结果包括,基于所述待处理数据及其对应的神经网络的类型,执行该类型的神经网络中的每一层的计算,以获取处理结果。
- 5 12.一种芯片,用于执行计算任务,所述芯片包括:
- 多个算子单元,以及
- 嵌入式 CPU,配置为,通过执行以下步骤而进行所述计算任务的计算:配置所述多个算子单元,以使得所述多个算子单元与该计算任务中包括的不同类型的计算分别对应,
- 10 以及,基于从所述芯片的外部输入的待处理数据,通过调用所述多个算子单元,进行所述计算任务的计算,以获取处理结果。
- 13.一种芯片,用于执行神经网络的预测计算,所述芯片包括:
- 多个算子单元,以及
- 嵌入式 CPU,配置为,基于从所述芯片的外部输入的待处理数据,执行所述神经网络中的每一层的计算,以获取处理结果,其中,通过由所述嵌入式 CPU 执行以下步骤而进行所述神经网络的每一层的计算:配置所述多个算子单元,以使得所述多个算子单元与该层中包括的不同类型的计算分别对应,以及,基于神经网络上一层中各个神经元的值,通过调用所述多个算子单元,执行该层中每个神经元处的计算,以获取该层中每个神经元的值。
- 15 20 14.根据权利要求 12 或 13 所述的芯片,其中,所述芯片与服务器连接。
- 15.根据权利要求 14 所述的芯片,其中,所述待处理数据通过所述服务器传送至所述芯片。
- 16.根据权利要求 14 所述的芯片,其中,所述嵌入式 CPU 通过运行基于程序的汇编指令而执行所述步骤,其中,所述程序由所述服务器传送至所述芯片。
- 25 17.根据权利要求 16 所述的芯片,其中,所述程序为 C 语言程序。
- 18.根据权利要求 17 所述的芯片,其中,所述嵌入式 CPU 还配置为,通过 C 语言中的函数调用,调用所述多个算子单元。
- 19.根据权利要求 12 或 13 所述的芯片,其中,所述芯片还包括 DMA 模块,所述嵌入式 CPU 还配置为,在获取所述处理结果之后,通过控制所述 DMA 模块,以将所述处理结果写入所述服务器的内存中。
- 30 20.根据权利要求 12 或 13 所述的芯片,其中,所述嵌入式 CPU 还配置为,异步地

调用所述多个算子单元。

21.根据权利要求 12 或 13 所述的芯片，其中，所述嵌入式 CPU 还配置为，对所述待处理数据的长度进行判断，并基于所述长度获取发送给各个算子单元的参数。

22.根据权利要求 12 或 13 所述的芯片，其中，所述芯片为 ASIC 芯片或 FPGA 芯片。

- 5 23.根据权利要求 13 所述的芯片，其中，所述芯片用于执行多种神经网络的预测计算，其中，所述嵌入式 CPU 还配置为，基于从所述芯片的外部输入的待处理数据及其对应的神经网络的类型，执行所述类型的神经网络中的每一层的计算，以获取处理结果。

24. 一种计算设备，包括存储器和处理器，其特征在于，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 1-11 中任一项所述的方法。