

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2018 年 10 月 18 日 (18.10.2018)



(10) 国际公布号
WO 2018/188463 A1

(51) 国际专利分类号:
G06F 8/41 (2018.01) **G06N 3/02** (2006.01)

(21) 国际申请号: PCT/CN2018/079954

(22) 国际申请日: 2018 年 3 月 22 日 (22.03.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710231894.9 2017年4月11日 (11.04.2017) CN

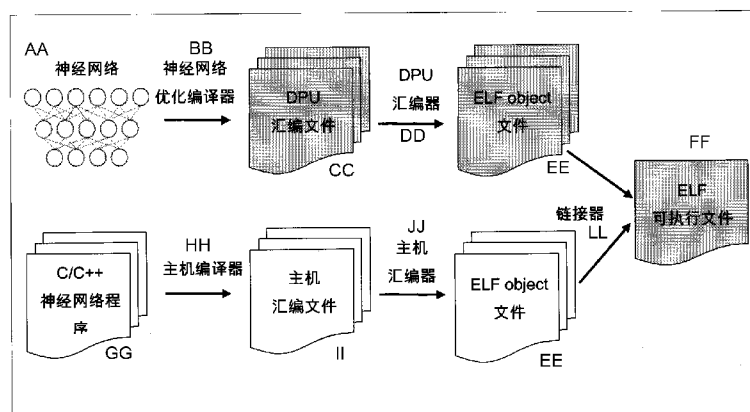
(71) 申请人: 北京深鉴智能科技有限公司(BEIJING DEEPHI INTELLIGENT TECHNOLOGY CO., LTD.) [CN/CN]; 中国北京市海淀区王庄路1号院4号楼17层1706号, Beijing 100083 (CN)。

(72) 发明人: 孙晓明(SUN, Xiaoming); 中国北京市海淀区王庄路1号院4号楼8层807号, Beijing 100083 (CN)。 隋凌志(SUL, Lingzhi); 中国北京市海淀区王庄路1号院4号楼8层807号, Beijing 100083 (CN)。 罗洪(LUO, Hong); 中国北京市海淀区王庄路1号院4号楼8层807号, Beijing 100083 (CN)。 单羿(SHAN, Yi); 中国北京市海淀区王庄路1号院4号楼8层807号, Beijing 100083 (CN)。 姚颂(YAO, Song); 中国北京市海淀区王庄路1号院4号楼8层807号, Beijing 100083 (CN)。

(74) 代理人: 北京卓孚知识产权代理事务所(普通合伙)(IPFAITH PARTNERS); 中国北京市海淀区彩和坊路8号天创科技大厦907C, Beijing 100080 (CN)。

(54) Title: PROGRAMMING MODEL ORIENTED TO NEURAL NETWORK HETEROGENEOUS COMPUTING PLATFORM

(54) 发明名称: 面向神经网络异构计算平台的编程模型



AA Neural network
BB Neural network optimizing compiler
CC DPU assembling file
DD DPU assembler
EE ELF object file
FF ELF executable file
GG C/C++ neural network program
HH Host compiler
II Host assembling file
JJ Host assembler
LL Linker

图 1

(57) Abstract: Provided is a programming model oriented to a neural network heterogeneous computing platform. Specifically, provided are a compiling method and system of a heterogeneous computing platform and a program running support method and system thereof. A trained neural network model is input to a neural network (NN) optimizing compiler to generate an NN assembling file corresponding to the NN. The NN assembling file is input to an NN assembler to generate an NN binary file corresponding to the NN. A host compiler tool chain is used for compiling and assembling an NN application program developed by a user with the high-level

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

language, and orderly generating a corresponding host assembling file and a host binary file. The host linker is used for linking the NN binary file and the host binary file to generate a single mixed link executable file. The technical solution of the present invention has good computing performance, strong expandability, strong compatibility, high flexibility, etc.

(57) 摘要: 本公开提供一种面向神经网络异构计算平台的编程模型。具体地说, 本公开提出一种异构计算平台的编译方法和系统及其程序运行支持方法和系统。训练后的神经网络模型输入到神经网络(NN)优化编译器, 生成与神经网络对应的NN汇编文件。NN汇编文件输入到NN汇编器, 生成与神经网络对应的NN二进制文件。使用主机编译器工具链编译和汇编用户用高级语言开发的神经网络应用程序, 依次生成相应的主机汇编文件和主机二进制文件。使用主机链接器链接NN二进制文件和主机二进制文件, 以生成单一的混合链接可执行文件。本发明的技术方案具有计算性能好、可扩展性强、兼容性强、灵活度高等特点。

面向神经网络异构计算平台的编程模型

5 技术领域

本发明涉及异构计算平台与神经网络,更具体涉及面向神经网络异构计算平台的编程模型。

背景技术

- 10 人工智能近年得到迅速发展,极大地影响了人们的生活,世界各国都给予了高度的重视,并进行大规模研发投入。人工神经网络是人工智能应用的核心。深度学习神经网络算法是最常见的一种人工神经网络模型。它的工作负载特征是计算密集(乘加操作为 G 量级)和数据密集(参数为 M 至数百 M 字节)。基于传统通用处理器 CPU 的计算平台并不能很好地满足性能要求。
- 15 近几年,以 NVIDIA GPU 为代表的加速神经网络计算的异构平台变得流行起来。CUDA SDK 提供的编译工具链和开发包简化了用户在 CPU+GPU 异构环境下的应用开发。随着 FPGA 以及各种深度学习专用芯片 ASIC(如谷歌 TPU)等更具性价比的加速方案不断涌现,解决好 CPU+FPGA/ASIC 异构计算平台面临的以下问题变得迫切:

- 20
1. 基于流行 C/C++ 高级语言的可编程性
 2. 降低神经网络应用的开发门槛、提供编程效率
 3. 优化神经网络结构、编译生成高效的计算指令
 4. 提高数据复用、减少 CPU 与 FPGA/ASIC 之间的数据移动

- 25 因此,需要一种面向神经网络异构计算平台的编程模型,有效解决 CPU+神经网络专用处理器的异构环境下神经网络应用的开发、编译、部署和运行阶段面临的各种困难。

发明内容

- 30 根据以上的讨论,本发明主要关注面向神经网络异构计算平台上的一系列高效编译和运行时支持技术。更具体地说,本发明的目的在于提供一种面

向神经网络异构计算平台的编程模型，从而创新性地有效解决 CPU+神经网络专用处理器异构环境下神经网络应用的开发、编译、部署和运行阶段面临的各种困难。

- 5 为实现上述目的，根据本发明的第一方面，提供一种异构计算平台的编译方法，包括：将训练后的神经网络模型输入到神经网络（NN）优化编译器，生成与神经网络对应的 NN 汇编文件；将 NN 汇编文件输入到 NN 汇编器，生成与神经网络对应的 NN 二进制文件；使用主机编译器工具链编译和汇编用户用高级语言开发的神经网络应用程序，依次生成相应的主机汇编文件
10 和主机二进制文件；使用主机链接器链接 NN 二进制文件和主机二进制文件，以生成单一的混合链接可执行文件。

 根据本发明第一方面的异构计算平台的编译方法，优选地，所述的二进制文件是 ELF 二进制文件，所述的单一的混合链接可执行文件是单一的混合
15 链接 ELF 可执行文件。

 根据本发明第一方面的异构计算平台的编译方法，优选地，所述高级语言是 C/C++。

- 20 根据本发明第一方面的异构计算平台的编译方法，优选地，所述的单一的混合链接可执行文件包括：ELF 标准可执行文件部分和 NN 扩展部分。ELF 标准可执行文件部分进一步包括：ELF 文件头，段表，代码段，数据段。NN 扩展部分进一步包括：NN 头，代码段，数据段。

- 25 根据本发明的第二方面，提供一种异构计算平台的程序运行支持方法，包括：用户通过调用神经网络专用处理器开发库的高级语言应用编程接口（API），实现神经网络专用处理器内核的初始化操作和运行时的控制；启动
根据本发明第一方面的异构计算平台的编译方法所得到的混合链接可执行文件；所述可执行文件中的主机的代码段和数据段被操作系统的标准加载器
30 加载到主机的内存空间；所述可执行文件中的神经网络专用处理器的代码段和数据段被神经网络专用处理器加载器加载到神经网络专用处理器的内存

空间；主机代码段中的指令和神经网络专用处理器代码段中的指令按照神经网络应用设定的顺序串行执行。

根据本发明的第三方面，提供一种异构计算平台的编译系统，包括：神经网络（NN）优化编译器，用于接收训练后的神经网络模型，生成与神经网络对应的 NN 汇编文件；NN 汇编器，用于接收 NN 汇编文件，生成与神经网络对应的 NN 二进制文件；主机编译器，用于编译用户用高级语言开发的神经网络应用程序，生成相应的主机汇编文件；主机汇编器，用于汇编主机汇编文件，生成主机二进制文件；主机链接器，用于链接 NN 二进制文件和主机二进制文件，以生成单一的混合链接可执行文件。

根据本发明第三方面的异构计算平台的编译系统，优选地，所述的二进制文件是 ELF 二进制文件，所述的单一的混合链接可执行文件是单一的混合链接 ELF 可执行文件。

根据本发明第三方面的异构计算平台的编译系统，优选地，所述高级语言是 C/C++。

根据本发明第三方面的异构计算平台的编译系统，优选地，所述的单一的混合链接可执行文件包括：ELF 标准可执行文件部分和 NN 扩展部分。ELF 标准可执行文件部分进一步包括：ELF 文件头，段表，代码段，数据段。NN 扩展部分进一步包括：NN 头，代码段，数据段。

根据本发明的第四方面，提供一种异构计算平台的程序运行支持系统，包括：主机；神经网络专用处理器；神经网络应用程序；神经网络专用处理器开发库，用于封装神经网络专用处理器的各种高级语言应用编程接口，以便提供给用户开发神经网络应用时调用，实现神经网络专用处理器内核的初始化操作和运行时的控制；根据本发明第三方面的异构计算平台的编译系统，用于生成混合链接可执行文件；操作系统的标准加载器，用于将所述可执行文件中的主机的代码段和数据段加载到主机的内存空间；神经网络专用处理器加载器，用于将所述可执行文件中的神经网络专用处理器的代码段和数据

段加载到神经网络专用处理器的内存空间；主机操作系统；神经网络专用处理器驱动程序，用于与主机操作系统相配合，使主机代码段中的指令和神经网络专用处理器代码段中的指令按照神经网络应用设定的顺序串行执行。

- 5 本发明的技术方案具有计算性能好、可扩展性强、兼容性强、灵活度高等特点。

附图说明

下面参考附图结合实施例说明本发明。在附图中：

- 10 图 1 是 CPU+神经网络专用处理器异构计算平台混合编译模型；
图 2 是根据本发明的第一实施例的异构计算平台的编译方法的流程图；
图 3 是 CPU+神经网络专用处理器混合链接 ELF 可执行文件的示意图；
图 4 是根据本发明的第二实施例的异构计算平台的程序运行支持模型。

15 具体实施方式

附图仅用于示例说明，不能理解为对本专利的限制；下面结合附图和实施例对本发明的技术方案做进一步的说明。

- 20 与单一计算平台（即只有主机或 CPU 的计算平台）相比，本发明还包括了以下的功能部件：1）神经网络优化编译器；2）NN 汇编器；3）CPU+神经网络专用处理器异构混合（hybrid）链接 ELF（Executable and Linking Format）可执行文件；4）神经网络专用处理器驱动程序；5）运行时神经网络专用处理器加载器（Loader）；6）神经网络专用处理器开发库（Library）。

- 25 本领域技术人员应当理解，本申请中所使用的术语“神经网络专用处理器”，也可简称为“神经网络处理器”或“NN 处理器”。由于深度学习是神经网络技术中目前最为流行的一个技术分类，因此神经网络专用处理器可以被实现为深度学习专用处理器或深度学习处理器。但是，本领域技术人员应该明白，神经网络还有其他的技术分支，例如深度神经网络（DNN，Deep
30 Neutral Network），因此神经网络专用处理器也可以被实现为深度神经网络专用处理器或深度神经网络处理器（DNN 处理器）。也就是说，有关“深度学

习处理器”或“深度神经网络处理器”在异构计算平台中的编译和运行支持方面的技术也在本发明的范围之内。

如下面参考图 1 所述的内容中所提到的，上述的功能部件 1)、2)、3) 提供神经网络在 CPU+神经网络专用处理器异构平台上的编译支持，即涉及本发明的第一实施例。如下面参考图 4 所述的内容中所提到的，功能部件 4)、5)、6) 为 CPU+神经网络专用处理器异构神经网络应用提供相关程序运行支持，即涉及本发明的第二实施例。

第一实施例：

下面，将首先结合图 1 来描述第一实施例中的组成部件的功能。

图 1 是 CPU+神经网络专用处理器异构计算平台混合编译模型。

在图 1 中，神经网络 (Neural Network, 可简称为 NN) 优化编译器 (或称为“NN 优化编译器”) 以神经网络模型为输入，通过分析网络的拓扑结构获取模型中控制流和数据流信息，以此为基础在模型上施加各种优化变换技术。具体地说，该编译器合并神经网络模型中不同网络层之间的计算操作、降低计算强度。对于结构化和非结构化稀疏网络，该编译器将消除稀疏值带来的不必要计算和数据移动。此外，该编译器将充分复用神经网络专用处理器片上内存中存放的网络参数和特征图 (feature map) 数值，降低访存带宽需求，从而减少功耗。该编译器输出 NN 汇编文件到 NN 汇编器。

NN 汇编器接收来自 NN 优化编译器的 NN 汇编文件。接收到的 NN 汇编文件 (汇编指令) 被 NN 汇编器转换而生成对应的二进制机器编码。此后，神经网络专用处理器的机器编码和数据被集成在一起生成标准格式的 ELF 二进制文件。

主机链接器用于生成 CPU+神经网络专用处理器异构混合链接 ELF 可执行文件。在本发明的优选实施例中，可以扩展 ELF 可执行文件的格式，增

加神经网络专用处理器的支持。具体地说，把神经网络专用处理器和主机两种 ELF 二进制文件链接在一起，生成单一可执行文件（ELF 可执行文件），方便神经网络应用程序的部署和运行。图 1 中，神经网络专用处理器和主机 ELF 二进制文件都被示出为 ELF object 文件。图 3 给出了 CPU+神经网络专用处理器混合链接 ELF 可执行文件的一个示例，下文将进一步详细描述。

主机链接器所使用的主机 ELF 二进制文件是通过主机编译器对 C/C++ 神经网络程序进行编译，通过主机汇编器对生成的主机汇编文件进行汇编而得到的。其中，主机编译器、主机汇编器、主机链接器都是单一计算平台的编译模型中已有的功能部件。

下面从方法步骤的角度来进一步描述根据本发明第一实施例的编译模型的工作状态。

图 2 是根据本发明的第一实施例的异构计算平台的编译方法的流程图。

如图 2 中所示，根据本发明的第一实施例的异构计算平台的编译方法 200 开始于步骤 S201。在该步骤 S201，将训练后的神经网络模型输入到 NN 优化编译器，生成与神经网络对应的 NN 汇编文件。

在步骤 S203，将 NN 汇编文件输入到 NN 汇编器，生成与神经网络对应的 NN 二进制文件。

在步骤 S205，使用主机编译器工具链编译和汇编用户用高级语言开发的神经网络应用程序，依次生成相应的主机汇编文件和主机二进制文件。

在本发明的优选实施例中，以及参考图 1 的描述，将以上提到的高级语言都具体化为 C/C++。本领域技术人员应该理解，C/C++是本领域中典型且常用的高级语言，从而使本发明的可编程性得到提升。但是，本领域技术人员也应该理解，这里所述的高级语言也可以是除了 C/C++以外的其他高级语言，例如，包括但不限于：Java、Pascal、Python、Perl 等等面向对象的高级

语言或可能在本发明之前的编程平台上典型或常用，也可能在本发明的发明人所能预见到的未来的编程平台上典型或常用的其他高级语言。

在步骤 S207, 使用主机链接器链接 NN 二进制文件和主机二进制文件，
5 以生成单一的混合链接可执行文件。

在本发明的优选实施例中，以及参考图 1 的描述，将以上提到的二进制文件都具体化为 ELF 二进制文件（或称为 ELF 文件）。本领域技术人员应该理解，ELF（Executable and Linking Format）文件使用典型且常用的标准文件
10 文件格式，从而使本发明的可编程性得到提升。但是，本领域技术人员也应该理解，这里所述的二进制文件也可以是除了 ELF 文件以外的其他文件格式，其可能在本发明之前的编程平台上典型或常用，也可能在本发明的发明人所能预见到的未来的编程平台上典型或常用。

15 此外，在本发明的优选实施方式中，在步骤 S207 生成的单一的混合链接可执行文件可以是单一的混合链接 ELF 可执行文件。图 3 给出了这样的一种示例。

图 3 是 CPU+神经网络专用处理器混合链接 ELF 可执行文件的示意图。
20 如图 3 所示，在步骤 S207 所生成的单一的混合链接可执行文件可以包括两个部分：ELF 标准可执行文件部分和 NN 扩展部分。其中，ELF 标准可执行文件部分与普通可执行文件一样，其进一步包括：ELF 文件头（图 3 示为“ELF Header”），段表（图 3 示为“Section Header Table”），代码段（图 3 示为“.text”），数据段（图 3 示为“.data”）。NN 扩展部分是对 ELF 标准可执行文件部分的
25 扩展，其与 ELF 标准可执行文件部分链接在一起，进一步包括：NN 头（图 3 示为“NN Header”），代码段（图 3 示为“.text”），数据段（图 3 示为“.data”）。

尽管在图 3 中，NN 扩展部分链接在 ELF 标准可执行文件之后，但本领域
30 技术人员应该理解，CPU+神经网络专用处理器混合链接可执行文件的链接方式和组成结构也可以是除了图 3 以外的形式。也就是说，根据本发明的

第一实施例所述的单一的混合链接可执行文件不限于图3中所给出的示例的形式，还可以是任意链接形式和内容构成的CPU+神经网络专用处理器混合链接可执行文件。

5 回到图2，步骤S207执行完之后，编译方法200可视为结束。生成的可执行文件可以被启动，从而在整个面向神经网络异构计算平台的编程模型的支持下，得到运行。

10 在本发明的第一实施例中，除了上述的异构计算平台的编译方法，还提出了一种异构计算平台的编译系统。

参看图1和图2，根据本发明的第一实施例的异构计算平台的编译系统可以包括NN优化编译器，其用于接收训练后的神经网络模型，生成与神经网络对应的NN汇编文件。

15

编译系统中的NN汇编器用于从NN优化编译器接收NN汇编文件，从而生成与神经网络对应的NN二进制文件。

20 另一方面，针对主机方面，编译系统的主机编译器用于编译用户用高级语言开发的神经网络应用程序，生成相应的主机汇编文件。而主机汇编器则用于汇编主机汇编文件，生成主机二进制文件。所述高级语言可以是C/C++，也可以是用于提升系统可编程性的其他高级语言。

25 编译系统仍然使用主机链接器来链接NN二进制文件和主机二进制文件，但是生成的是单一的混合链接可执行文件。

如上在对图2的描述中所解释的，所述的二进制文件可以是ELF二进制文件，所述的单一的混合链接可执行文件可以是单一的混合链接ELF可执行文件。

30

如前所述，所述的单一的混合链接可执行文件可以包括：ELF标准可执

行文件部分和 NN 扩展部分。其中，ELF 标准可执行文件部分可以进一步包括：ELF 文件头，段表，代码段，数据段。NN 扩展部分则可以进一步包括：NN 头，代码段，数据段。可以再次参看图 3 所给出的示例。但是，根据本
5 发明的第一实施例所述的单一的混合链接可执行文件不限于图 3 中所给出的示例的形式，还可以是任意链接形式和内容构成的 CPU+神经网络专用处理器混合链接可执行文件。

第二实施例：

10 下面将结合图 4 来描述本发明的编程环境，即程序运行支持模型。

图 4 是根据本发明的第二实施例的异构计算平台的程序运行支持模型。

15 在图 4 中，将神经网络专用处理器简称为 DPU，以与主机 CPU 相区别。本领域技术人员应该理解，这样的命名并不影响神经网络专用处理器的一般性。也就是说，本说明书与附图中，“神经网络专用处理器”与“DPU”是可以互换使用的术语，用于表示异构计算平台上与 CPU 相异的另一处理器。

20 如图 4 中所示，神经网络专用处理器驱动程序（图 4 中示出为“DPU 驱动程序”）工作在操作系统内核态（kernel space），负责神经网络专用处理器资源运行时的动态管理和调度，并且负责运行在神经网络专用处理器上神经网络代码（以下称为神经网络专用处理器内核）的管理和控制，负责神经网络专用处理器内存空间的管理和分配。

25

神经网络专用处理器加载器（图 4 中示为“DPU loader”），在运行时，工作在操作系统用户态（user space），负责神经网络专用处理器内核的动态加载和释放，并且负责神经网络专用处理器内核代码的重定位。

30 神经网络专用处理器开发库（图 4 中示为“DPU Library”）工作在操作系统用户态（user space），用于封装神经网络专用处理器的各种 C/C++ 应用

编程接口（API）提供给用户开发各种神经网络应用时调用，从而提高神经网络专用处理器上的编程效率。

此外，再加上本领域技术人员应该知道的主机（或 CPU）、神经网络专用处理器（图 4 中示为“DPU”）、主机操作系统和操作系统的标准加载器（OS Loader），用于配合以上提到的服务于神经网络专用处理器的部件，一起提供执行神经网络应用程序的运行支持环境。换句话说，实现神经网络应用程序的神经网络专用处理器加速。

下面分别从方法步骤的角度和系统构成的角度来进一步描述根据本发明第二实施例的异构计算平台的程序运行支持模型的工作状态。

CPU+神经网络专用处理器异构计算平台开发的神经网络应用程序，主要有两部分组成：C/C++代码、训练后的神经网络模型（网络拓扑结构和神经网络参数）。从开发、编译到部署和运行，典型的工作流程如下所示：

步骤 1. 用户通过调用神经网络专用处理器开发库的 C/C++ API 编程接口，实现神经网络专用处理器内核的初始化操作和运行时的各种控制，例如 dpuInit(), dpuFini(), dpuCreateKernel(), dpuDestroyKernel(), dpuRunKernel() 等。其中的“dpu”表示神经网络专用处理器。

步骤 2. 输入训练后的神经网络模型给 NN 优化编译器，生成与网络对应的 NN 汇编文件。

步骤 3. 输入 NN 汇编文件给 NN 汇编器，生成与网络对应的 ELF 二进制文件。

步骤 4. 使用主机编译器工具链编译用户开发的 C/C++神经网络应用程序，生成相应的主机汇编文件和主机 ELF 二进制文件。

步骤 5. 使用主机链接器链接神经网络专用处理器和主机的所有 ELF 二

进制文件，生成单一的混合链接 ELF 可执行文件。

5 步骤 6. 编译完生成的单一混合链接 ELF 可执行文件包含了主机和神经网络专用处理器运行所需的全部信息，执行方式与普通的主机 ELF 可执行文件完全一样，可实现一键式部署。

10 步骤 7. 启动混合链接 ELF 可执行文件，ELF 中主机的代码段和数据段会被操作系统的标准加载器加载到主机的内存空间；ELF 中神经网络专用处理器的代码段和数据段由神经网络专用处理器加载器负责加载到神经网络专用处理器的内存空间（其中神经网络专用处理器的内存空间由神经网络专用处理器驱动程序管理）。

15 步骤 8. 主机代码段中的指令和神经网络专用处理器代码段中的指令按照神经网络应用设定的顺序串行执行，计算和数据密集的神经网络算法以神经网络专用处理器内核的形态被主机卸载（offload）至神经网络专用处理器上加速运行，更加高效地运行。

可以将以上的过程总结一下。

20 在根据本发明的第二实施例的异构计算平台的程序运行支持方法中，用户通过调用神经网络专用处理器开发库的高级语言应用编程接口（API），实现神经网络专用处理器内核的初始化操作和运行时的控制。

25 用户根据本发明第一实施例的异构计算平台的编译方法得到了混合链接可执行文件之后，启动该可执行文件。所述可执行文件中的主机的代码段和数据段被操作系统的标准加载器加载到主机的内存空间；而所述可执行文件中的神经网络专用处理器的代码段和数据段被神经网络专用处理器加载器加载到神经网络专用处理器的内存空间。

30 然后，主机操作系统与神经网络专用处理器驱动程序相配合，将主机代码段中的指令和神经网络专用处理器代码段中的指令按照神经网络应用设

定的顺序串行执行。

在本发明的第二实施例中，除了上述的异构计算平台的程序运行支持方法，还提出了一种异构计算平台的程序运行支持系统。

5

如图 4 所示，根据本发明第二实施例的异构计算平台的程序运行支持系统不仅包括单一计算平台中常见的主机、主机操作系统、操作系统的标准加载器，还包括神经网络专用处理器、神经网络应用程序、神经网络专用处理器开发库、编译系统、神经网络专用处理器加载器、神经网络专用处理器驱动程序。

10

在本发明的程序运行支持系统中，神经网络专用处理器开发库用于封装神经网络专用处理器的各种高级语言（例如 C/C++）应用编程接口（API），以便提供给用户开发神经网络应用时调用，实现神经网络专用处理器内核的初始化操作和运行时的控制。

15

程序运行支持系统中的编译系统，就是根据本发明第一实施例的异构计算平台的编译系统，用于生成混合链接可执行文件。

20

操作系统的标准加载器用于将所述可执行文件中的主机的代码段和数据段加载到主机的内存空间；而神经网络专用处理器加载器用于将所述可执行文件中的神经网络专用处理器的代码段和数据段加载到神经网络专用处理器的内存空间。

25

神经网络专用处理器驱动程序用于与主机操作系统相配合，使主机代码段中的指令和神经网络专用处理器代码段中的指令按照神经网络应用设定的顺序串行执行。

以下对本发明所提到的创新点进行进一步讨论。

30

目前面向神经网络领域专用处理器的设计和研发正处于百花齐放的初

期阶段，如何构建面向神经网络异构计算平台的专用编译工具链在学术界和工业界还没有形成标准和规范，也没有相关的完整技术方案公布，因此本发明具有一定的先进性和创新性，总结起来具有以下优点：

5 1. 完备性好：本发明提出的面向神经网络异构计算平台的编译和程序运行支持技术，覆盖了神经网络应用的开发、编译、部署和运行各阶段，是一套完整的、实用性强的技术解决方案。

10 2. 性能好：本发明中的神经网络优化编译器，能够针对神经网络专用处理器中计算单元和片上内存的资源配置，平衡数据读写负载和计算操作负载，充分发挥神经网络专用处理器硬件资源并行度，精确挖掘神经网络专用处理器的计算能力、生成最高效的指令代码。

15 3. 灵活性高：本发明中编译工具链能把各种神经网络模型编译生成神经网络专用处理器上可运行指令码，不受限于某种特定的神经网络算法，适用于神经网络专用处理器平台各种神经网络应用程序的开发，具有较好的灵活性。

20 4. 易用性强：通过神经网络专用处理器开发库向用户提供标准 C/C++ API 编程接口，与 CUDA、OpenCL 通过扩展 C/C++ 语法提供异构编程的方式相比，免除了用户熟悉新语法的额外负担，因此可降低神经网络专用处理器平台的开发门槛、提高编程效率。

25 上面已经描述了本发明的各种实施例和实施情形。但是，本发明的精神和范围不限于此。本领域技术人员将能够根据本发明的教导而做出更多的应用，而这些应用都在本发明的范围之内。

30 也就是说，本发明的上述实施例仅仅是为清楚说明本发明所做的举例，而非对本发明实施方式的限定。对于所属领域的普通技术人员来说，在上述说明的基础上还可以做出其他不同形式的变化或变动。这里无需也无法对所有的实施方式予以穷举。凡在本发明的精神和原则之内所作的任何修改、替

换或改进等，均应包含在本发明权利要求的保护范围之内。

权 利 要 求 书

1、一种异构计算平台的编译方法，包括：

5 将训练后的神经网络模型输入到神经网络（NN）优化编译器，生成与神经网络对应的 NN 汇编文件；

将 NN 汇编文件输入到 NN 汇编器，生成与神经网络对应的 NN 二进制文件；

使用主机编译器工具链编译和汇编用户用高级语言开发的神经网络应用程序，依次生成相应的主机汇编文件和主机二进制文件；

10 使用主机链接器链接 NN 二进制文件和主机二进制文件，以生成单一的混合链接可执行文件。

2、根据权利要求 1 所述的异构计算平台的编译方法，其中，所述的二进制文件是 ELF 二进制文件，所述的单一的混合链接可执行文件是单一的混合链接 ELF 可执行文件。

3、根据权利要求 1 所述的异构计算平台的编译方法，其中，所述高级语言是 C/C++。

20 4、根据权利要求 1 所述的异构计算平台的编译方法，其中，所述的单一的混合链接可执行文件包括：

ELF 标准可执行文件部分，其进一步包括：

ELF 文件头，

段表，

25 代码段，

数据段，

NN 扩展部分，其进一步包括：

NN 头，

代码段，

30 数据段。

5、一种异构计算平台的程序运行支持方法，包括：

用户通过调用神经网络专用处理器开发库的高级语言应用编程接口（API），实现神经网络专用处理器内核的初始化操作和运行时的控制；

5 启动根据权利要求 1 所述的异构计算平台的编译方法所得到的混合链接可执行文件；

所述可执行文件中的主机代码段和数据段被操作系统的标准加载器加载到主机的内存空间；

所述可执行文件中的神经网络专用处理器的代码段和数据段被神经网络专用处理器加载器加载到神经网络专用处理器的内存空间；

10 主机代码段中的指令和神经网络专用处理器代码段中的指令按照神经网络应用设定的顺序串行执行。

6、一种异构计算平台的编译系统，包括：

15 神经网络（NN）优化编译器，用于接收训练后的神经网络模型，生成与神经网络对应的 NN 汇编文件；

NN 汇编器，用于接收 NN 汇编文件，生成与神经网络对应的 NN 二进制文件；

主机编译器，用于编译用户用高级语言开发的神经网络应用程序，生成相应的主机汇编文件；

20 主机汇编器，用于汇编主机汇编文件，生成主机二进制文件；

主机链接器，用于链接 NN 二进制文件和主机二进制文件，以生成单一的混合链接可执行文件。

25 7、根据权利要求 6 所述的异构计算平台的编译系统，其中，所述的二进制文件是 ELF 二进制文件，所述的单一的混合链接可执行文件是单一的混合链接 ELF 可执行文件。

8、根据权利要求 6 所述的异构计算平台的编译系统，其中，所述高级语言是 C/C++。

30

9、根据权利要求 6 所述的异构计算平台的编译系统，其中，所述的单

一的混合链接可执行文件包括：

ELF 标准可执行文件部分，其进一步包括：

ELF 文件头，

段表，

5 代码段，

数据段，

NN 扩展部分，其进一步包括：

NN 头，

代码段，

10 数据段。

10、一种异构计算平台的程序运行支持系统，包括：

主机；

神经网络专用处理器；

15 神经网络应用程序；

神经网络专用处理器开发库，用于封装神经网络专用处理器的各种高级语言应用编程接口，以便提供给用户开发神经网络应用时调用，实现神经网络专用处理器内核的初始化操作和运行时的控制；

20 根据权利要求 6 所述的异构计算平台的编译系统，用于生成混合链接可执行文件；

操作系统的标准加载器，用于将所述可执行文件中的主机代码段和数据段加载到主机的内存空间；

神经网络专用处理器加载器，用于将所述可执行文件中的神经网络专用处理器的代码段和数据段加载到神经网络专用处理器的内存空间；

25 主机操作系统；

神经网络专用处理器驱动程序，用于与主机操作系统相配合，使主机代码段中的指令和神经网络专用处理器代码段中的指令按照神经网络应用设定的顺序串行执行。

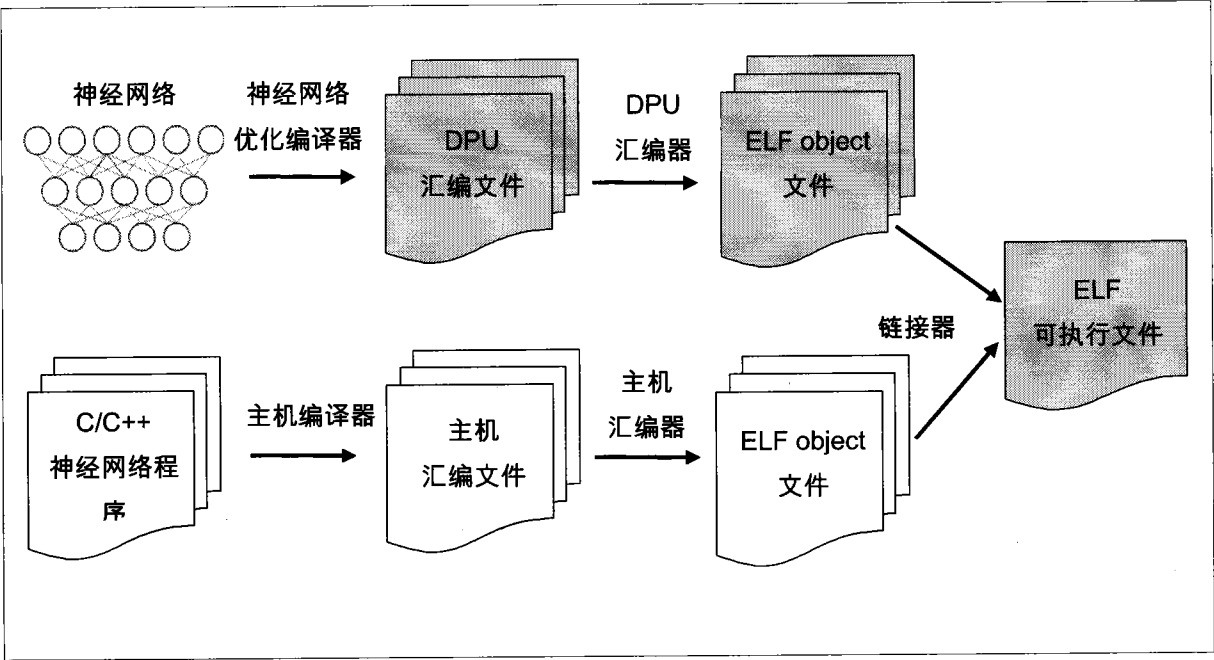


图 1

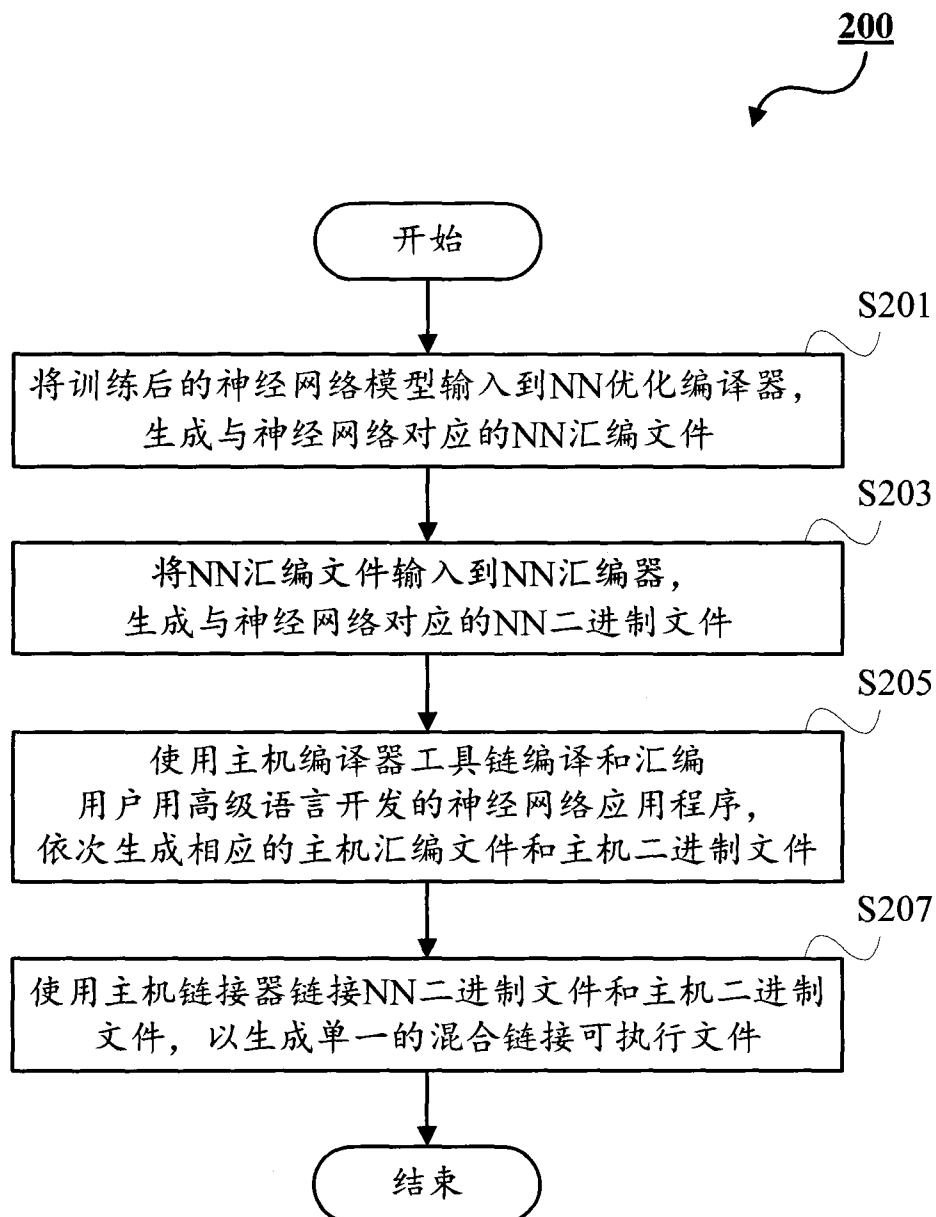


图 2

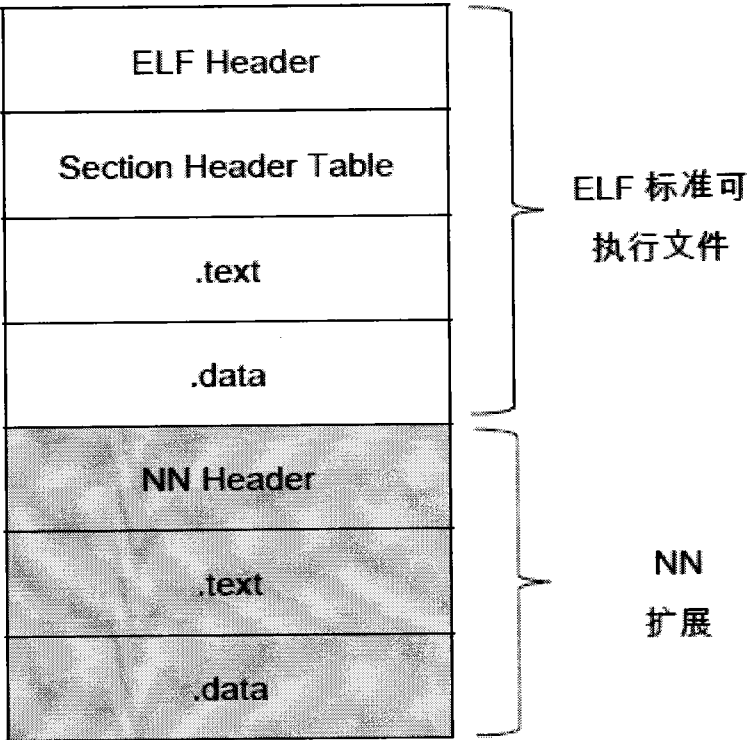


图 3

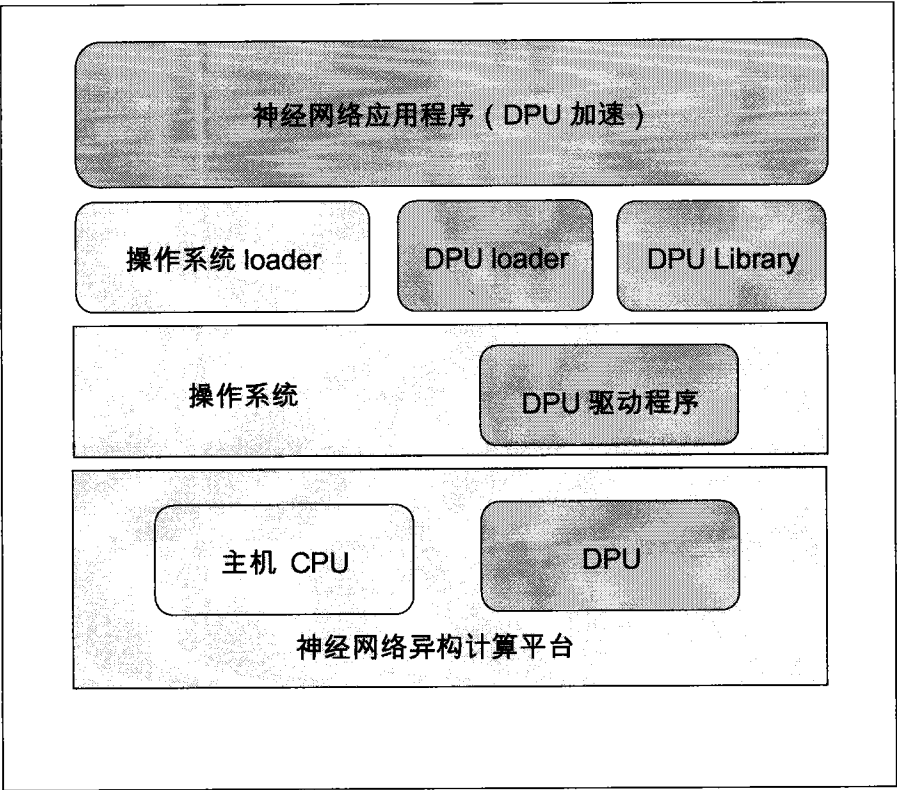


图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/079954

A. CLASSIFICATION OF SUBJECT MATTER

G06F 8/41 (2018.01) i; G06N 3/02 (2006.01) n

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI, CNABS, DWPI, SIPOABS: 异构, 编译, 神经网络, 汇编, 链接, heterogeneous, compiling, neural network, assembling, link

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 107239315 A (BEIJING DEEPHI TECH CO., LTD.) 10 October 2017 (10.10.2017), claims 1-10	1-10
X	CN 105426226 A (WUXI JIANGNAN COMPUTING TECHNOLOGY INSTITUTE) 23 March 2016 (23.03.2016), description, paragraphs [0002]-[0010] and [0041]	1-10
A	CN 101441569 A (PLA INFORMATION ENGINEERING UNIVERSITY) 27 May 2009 (27.05.2009), entire document	1-10
A	CN 106155755 A (SHANGHAI REDNEURONS INFORMATION TECHNOLOGY CO., LTD.) 23 November 2016 (23.11.2016), entire document	1-10

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
“A” document defining the general state of the art which is not considered to be of particular relevance	“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
“E” earlier application or patent but published on or after the international filing date	“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	“&” document member of the same patent family
“O” document referring to an oral disclosure, use, exhibition or other means	
“P” document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search 21 May 2018	Date of mailing of the international search report 01 June 2018
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer LIU, Haoran Telephone No. (86-10) 62411703

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/CN2018/079954

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 107239315 A	10 October 2017	None	
CN 105426226 A	23 March 2016	None	
CN 101441569 A	27 May 2009	CN 101441569 B	30 May 2012
CN 106155755 A	23 November 2016	None	

国际检索报告

国际申请号

PCT/CN2018/079954

A. 主题的分类 G06F 8/41(2018.01)i; G06N 3/02(2006.01)n 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类																	
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) G06F; G06N 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用)) CNKI, CNABS, DWPI, SIPOABS: 异构、编译、神经网络、汇编、链接、heterogeneous, compiling, neural network, assembling, link																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 107239315 A (北京深鉴智能科技有限公司) 2017年 10月 10日 (2017 - 10 - 10) 权利要求1-10</td> <td>1-10</td> </tr> <tr> <td>X</td> <td>CN 105426226 A (无锡江南计算技术研究所) 2016年 3月 23日 (2016 - 03 - 23) 说明书[0002]-[0010]、[0041]段</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 101441569 A (中国人民解放军信息工程大学) 2009年 5月 27日 (2009 - 05 - 27) 全文</td> <td>1-10</td> </tr> <tr> <td>A</td> <td>CN 106155755 A (上海红神信息技术有限公司) 2016年 11月 23日 (2016 - 11 - 23) 全文</td> <td>1-10</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 107239315 A (北京深鉴智能科技有限公司) 2017年 10月 10日 (2017 - 10 - 10) 权利要求1-10	1-10	X	CN 105426226 A (无锡江南计算技术研究所) 2016年 3月 23日 (2016 - 03 - 23) 说明书[0002]-[0010]、[0041]段	1-10	A	CN 101441569 A (中国人民解放军信息工程大学) 2009年 5月 27日 (2009 - 05 - 27) 全文	1-10	A	CN 106155755 A (上海红神信息技术有限公司) 2016年 11月 23日 (2016 - 11 - 23) 全文	1-10
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
PX	CN 107239315 A (北京深鉴智能科技有限公司) 2017年 10月 10日 (2017 - 10 - 10) 权利要求1-10	1-10															
X	CN 105426226 A (无锡江南计算技术研究所) 2016年 3月 23日 (2016 - 03 - 23) 说明书[0002]-[0010]、[0041]段	1-10															
A	CN 101441569 A (中国人民解放军信息工程大学) 2009年 5月 27日 (2009 - 05 - 27) 全文	1-10															
A	CN 106155755 A (上海红神信息技术有限公司) 2016年 11月 23日 (2016 - 11 - 23) 全文	1-10															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																	
国际检索实际完成的日期 2018年 5月 21日		国际检索报告邮寄日期 2018年 6月 1日															
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 刘浩然 电话号码 86-(10)-62411703															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/079954

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	107239315	A	2017年 10月 10日	无	
CN	105426226	A	2016年 3月 23日	无	
CN	101441569	A	2009年 5月 27日	CN 101441569 B	2012年 5月 30日
CN	106155755	A	2016年 11月 23日	无	