United States Patent Application Publication	20250258715
Kind Code	A1
Publication Date	August 14, 2025
Inventor(s)	PARK; Yongjun et al.

APPARATUS AND METHOD FOR PROVIDING EXECUTION PLANS OF MULTIPLE MIXED-PRECISION DEEP LEARNING MODELS BASED ON MULTI-PRECISION NPU

Abstract

Disclosed herein are an apparatus and method for providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU. The apparatus includes a memory, and a processor electrically connected to the memory. The processor is configured to form a multi-precision Neural Processing Unit (NPU) including a processing element (PE) composed of multiple Micro-PEs, to generate multiple mixed-precision deep learning models that perform multiple precision operations requiring different degrees of precision in a model execution process, and to generate the execution plans for executing the multiple mixed-precision deep learning models on the multi-precision NPU.

Inventors: **PARK; Yongjun (Seoul, KR), JUNG; Kiung (Cheongju-si, KR)**

Applicant: **UIF (UNIVERSITY INDUSTRY FOUNDATION), YONSEI UNIVERSITY** (Seoul, KR)

Family ID: **1000007784874**

Assignee: **UIF (UNIVERSITY INDUSTRY FOUNDATION), YONSEI UNIVERSITY (Seoul, KR)**

Appl. No.: **18/617846**

Filed: **March 27, 2024**

Foreign Application Priority Data

KR	10-2024-0020617	Feb. 13, 2024

Publication Classification

Int. Cl.: **G06F9/50** (20060101); **G06F15/80** (20060101)

U.S. Cl.:

CPC **G06F9/5038** (20130101); **G06F15/80** (20130101); G06F2209/5021 (20130101); G06F2209/508 (20130101)

Background/Summary

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims the benefit of priority to Korean Patent Application No. 10-2024-0020617 filed on Feb. 13, 2024. The disclosures of the above-listed applications are hereby incorporated by reference herein in their entirety.

TECHNICAL FIELD

[0002] The present disclosure relates to execution planning technology for mixed-precision deep learning models and, more particularly, to technology for providing an execution plan of a mixed-precision deep learning model, which is intended to efficiently execute multiple mixed-precision deep learning models in a multi-precision NPU.

BACKGROUND

[0003] A Neural Processing Unit (NPU) may refer to a processor optimized for deep learning algorithm operations, which may be said to be the core of artificial intelligence, and may be used to accelerate the learning and inference tasks of a neural network model. Unlike an existing central processing unit (CPU) and graphic processing unit (GPU), the NPU may be specialized for deep learning operations to efficiently perform related tasks.

[0004] Mixed-precision may refer to the representation precision of data, which is used when learning or inferring a deep learning model. Precision may be expressed in a bit number representing data, and may generally use single precision (float32) and half precision (float16). That is, the mixed-precision may refer to technology that reduces the computational workload of the deep learning model and optimizes memory usage by performing operations using both the single precision and the half precision during model learning and inference processes.

[0005] The mixed-precision deep learning model may correspond to a deep learning model that is built through mixed-precision-based training. The training process of the deep learning model may be executed on the NPU, and the multi-precision NPU may correspond to a hardware accelerator designed to efficiently support multiple precision operations that require different degrees of precision for model training.

[0006] The execution plan of the deep learning model on the NPU may refer to a plan for operation and resource allocation to execute the deep learning model, and may refer to a plan needed to ensure that the deep learning model can be efficiently executed on the NPU and achieve optimum performance. The execution plan may be established considering the structure of the given deep learning model, input data, available hardware resource, etc., and may include the execution order for each layer or operation, a data transfer method, the parallelization degree of operation, memory and storage allocation, etc.

SUMMARY

[0007] In view of the above, the present disclosure provides an apparatus and method for providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU, which can provide the execution plans of the mixed-precision deep learning models to efficiently execute the multiple mixed-precision deep learning models on the multi-precision NPU.

[0008] The present disclosure provides an apparatus for providing execution plans of multiple

mixed-precision deep learning models based on a multi-precision NPU, the apparatus including a memory, and a processor electrically connected to the memory. The processor is configured to form a multi-precision Neural Processing Unit (NPU) including a processing element (PE) composed of multiple Micro-PEs, to generate multiple mixed-precision deep learning models that perform multiple precision operations requiring different degrees of precision in a model execution process, and to generate the execution plans for executing the multiple mixed-precision deep learning models on the multi-precision NPU.

[0009] The processor may generate the multiple mixed-precision deep learning models through Hardware-Aware Mixed-precision Quantization (HAWQ).

[0010] The processor may generate the execution plan to ensure efficient distribution of precision operations for each model, taking into account a structure and characteristics of the multiple mixed-precision deep learning models.

[0011] The processor may generate the execution plan to ensure efficient resource utilization while minimizing execution time of each of the multiple mixed-precision deep learning models using a dynamic programming method.

[0012] The processor may generate the execution plan as a result of applying the dynamic programming method based on a result of measuring execution times of all precision operations for each mixed-precision deep learning model through a pre-simulation method or an actual execution method.

[0013] The processor may dynamically allocate at least one Micro-PE that executes each of the multiple precision operations in a process of executing the multiple mixed-precision deep learning models according to the execution plan.

[0014] The processor may control precision operations of a model with high execution priority among the multiple mixed-precision deep learning models to be preferentially executed according to the execution plan.

[0015] The processor may dynamically adjust the execution plan by tracking and monitoring the execution time and resource usage for precision operations of each mixed-precision deep learning model.

[0016] The present disclosure provides a method for providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU, the method being performed in a computing device including a memory, and a processor electrically connected to the memory, the method including forming a multi-precision Neural Processing Unit (NPU) including a processing element (PE) composed of multiple Micro-PEs, through the processor; generating multiple mixed-precision deep learning models that perform multiple precision operations requiring different degrees of precision in a model execution process, through the processor; and generating the execution plans for executing the multiple mixed-precision deep learning models on the multi-precision NPU, through the processor.

[0017] The generating the multiple mixed-precision deep learning models may include generating the multiple mixed-precision deep learning models through Hardware-Aware Mixed-precision Quantization (HAWQ).

[0018] The generating the execution plans may include generating the execution plan to ensure efficient distribution of precision operations for each model, taking into account a structure and characteristics of the multiple mixed-precision deep learning models.

[0019] The generating the execution plans may include generating the execution plan to ensure efficient resource utilization while minimizing execution time of each of the multiple mixed-precision deep learning models using a dynamic programming method.

[0020] The generating the execution plans may include generating the execution plan as a result of applying the dynamic programming method based on a result of measuring execution times of all precision operations for each mixed-precision deep learning model through a pre-simulation method or an actual execution method.

[0021] The generating the execution plans may include analyzing data dependency that occurs during the execution of each mixed-precision deep learning model and then optimizing the execution plan, taking into account the data dependency.
[0022] The generating the execution plans may include detecting errors and exceptions that occur during the execution of the mixed-precision deep learning model and adding processing plans for the detected errors and exceptions to the execution plan.
[0023] The method may further include dynamically allocating at least one Micro-PE that executes each of the multiple precision operations in a process of executing the multiple mixed-precision deep learning models according to the execution plan.
[0024] The present disclosure provides a computer-readable recording medium for storing a computer program, wherein the computer program, when executed by a processor, includes instructions for causing the processor to perform an operation including forming a multi-precision Neural Processing Unit (NPU) including a processing element (PE) composed of multiple Micro-PEs; generating multiple mixed-precision deep learning models that perform multiple precision operations requiring different degrees of precision in a model execution process; and generating the execution plans for executing the multiple mixed-precision deep learning models on the multi-precision NPU.
Advantageous Effects
[0025] The disclosed technology may have the following effects. However, since it does not mean that a specific embodiment should include all of the following effects or only the following effects, the scope of rights of the disclosed technology should not be understood as being limited thereto.
[0026] An apparatus and method for providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU according to an embodiment of the present disclosure can provide the execution plans of the mixed-precision deep learning models to efficiently execute the multiple mixed-precision deep learning models on the multi-precision NPU.
[0027] Particularly, the present disclosure can execute mixed-precision deep learning models produced through other technologies as well as HAWQ without depending on mixed-precision deep learning models produced through HAWQ. Further, the present disclosure does not simply execute the mixed-precision deep learning models several times, but can simultaneously execute multiple mixed-precision deep learning models to achieve energy efficiency and fast execution time. The present disclosure is advantageous in that additional programs or hardware are not required to provide execution plans of the multiple mixed-precision deep learning models.

Description

BRIEF DESCRIPTION OF THE DRAWINGS
[0028] FIG. **1** is a diagram illustrating a system for providing an execution plan according to the present disclosure.
[0029] FIG. **2** is a diagram illustrating the system configuration of an execution plan providing apparatus of FIG. **1**.
[0030] FIG. **3** is a diagram illustrating the functional component of a processor of FIG. **2**.
[0031] FIG. **4** is a flowchart illustrating a method for providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU according to the present disclosure.
[0032] FIG. **5** is a diagram illustrating the structure of a framework providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU according to the present disclosure.
[0033] FIG. **6** is a diagram illustrating an embodiment of an execution plan according to the present disclosure.
[0034] FIG. **7** is a diagram illustrating another embodiment of an execution plan according to the

present disclosure.
[0035] FIG. **8** is a diagram illustrating a difference in performance of the multi-precision NPU depending on the application of the execution plan according to the present disclosure.
[National Research Development Project Supporting the Present Invention]
[0036] [Project Serial No.] 1711198587 [0037] [Project No.] 00277060 [0038] [Department] Ministry of Science and ICT [0039] [Project Management (Professional) Institute] Institute of Information & communication Technology Planning & Evaluation [0040] [Research Project Name] Artificial Intelligence Semiconductor SW Integrated Platform Technology Development [0041] [Research Task Name] Open-edge AI Semiconductor Design and SW Platform Technology Development [0042] [Project Performing Institute] Yonsei University [0043] [Research Period] 2023.06.01 to 2024.02.29
[National Research Development Project Supporting the Present Invention]
[0044] [Project Serial No.] 1711193986 [0045] [Project No.] 2020-0-01361-004 [0046] [Department] Ministry of Science and ICT [0047] [Project Management (Professional) Institute] Institute of Information & communication Technology Planning & Evaluation [0048] [Research Project Name] Information & Communication Broadcasting Research Development Project [0049] [Research Task Name] Artificial Intelligence Graduate School Support Project (Yonsei University) [0050] [Project Performing Institute] University Industry Foundation, Yonsei University [0051] [Research Period] 2024.01.01 to 2024.12.31
DETAILED DESCRIPTION
[0052] Specific structural or functional descriptions in the embodiments of the present disclosure introduced in this specification or application are only for description of the embodiments of the present disclosure. The descriptions should not be construed as being limited to the embodiments described in the specification or application. The present disclosure may, however, be embodied in many different forms, but should be construed as covering modifications, equivalents or alternatives falling within ideas and technical scopes of the present disclosure. Further, since effects disclosed herein do not mean that a specific embodiment should include all or only the effects, the scope of the present disclosure should not be construed as being limited thereto.
[0053] Meanwhile, the meaning of terms described herein will be understood as follows.
[0054] It will be understood that, although the terms “first”, “second”, etc. may be used herein to distinguish one element from another element, these elements should not be limited by these terms. For instance, a first element discussed below could be termed a second element without departing from the teachings of the present disclosure. Similarly, the second element could also be termed the first element.
[0055] It will be understood that when an element is referred to as being “coupled” or “connected” to another element, it can be directly coupled or connected to the other element or intervening elements may be present therebetween. In contrast, it should be understood that when an element is referred to as being “directly coupled” or “directly connected” to another element, there are no intervening elements present. Other expressions that explain the relationship between elements, such as “between”, “directly between”, “adjacent to” or “directly adjacent to” should be construed in the same way.
[0056] In the present disclosure, the singular forms are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms “comprise”, “include”, “have”, etc. when used in this specification, specify the presence of stated features, integers, steps, operations, elements, components, and/or combinations of them but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components, and/or combinations thereof.
[0057] In each step, reference characters (e.g. a, b, c, etc.) are used for the convenience of description. The reference characters do not designate the order of the steps, and the steps may be performed in a different order unless the context clearly indicates otherwise. That is, the steps may

be performed in the specified order, may be performed substantially simultaneously, or may be performed in a reverse order.

[0058] The present disclosure can be implemented as a computer-readable code on a computer-readable recording medium. The computer-readable recording medium includes all types of recording devices in which data readable by a computer system is stored. Examples of the computer-readable recording medium include ROM, RAM, CD-ROM, magnetic tape, floppy disk, an optical data storage device, etc. In addition, the computer-readable recording medium may be distributed in a computer system connected via a network, so that computer-readable codes may be stored and executed in a distributed manner.

[0059] Unless otherwise defined, all terms including technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which the present disclosure belongs. It will be further understood that terms used herein should be interpreted as having a meaning that is consistent with their meaning in the context of this specification and the relevant art and will not be interpreted in an idealized or overly formal sense unless expressly so defined herein.

[0060] FIG. **1** is a diagram illustrating a system for providing an execution plan according to the present disclosure.

[0061] Referring to FIG. **1**, the execution plan providing system **100** may include a user terminal **110**, an execution plan providing apparatus **130**, and a database **150**.

[0062] The user terminal **110** may correspond to a computing device operated by a user. In this regard, a user may be an entity that executes and manages a model on a multi-precision NPU according to execution plans for multiple mixed-precision deep learning models. The user terminal **110** may be implemented as a smartphone, a laptop, or a computer, but may also be implemented as a variety of devices such as a tablet PC without being necessarily limited thereto. Further, the user terminal **110** may be implemented as a device forming the execution plan providing system **100** according to the present disclosure, and the execution plan providing system **100** may be implemented in various forms depending on the purpose of providing the execution plan.

[0063] The user terminal **110** may be connected to the execution plan providing apparatus **130** via a network, and multiple user terminals **110** may be simultaneously connected to the execution plan providing apparatus **130**. The user terminal may install and execute a dedicated program or application for interworking with the execution plan providing apparatus **130**.

[0064] The execution plan providing apparatus **130** may be implemented as a computer or server that performs a method for providing the execution plans of the multiple mixed-precision deep learning models based on the multi-precision NPU according to the present disclosure. In this regard, the method for providing the execution plans of the multiple mixed-precision deep learning models based on the multi-precision NPU according to the present disclosure may perform a series of processes for generating the execution plan when a user request for providing the execution plan is received from the user terminal **110**, and may include a plurality of operation steps defined in the process of providing the resulting execution plan to the user terminal **110**.

[0065] Further, the execution plan providing apparatus **130** may be connected to the user terminal **110** via a wired network or a wireless network such as Bluetooth, WiFi, or LTE, and may transmit and receive data with the user terminal **110** through the network. Further, the execution plan providing apparatus **130** may be implemented to operate in connection with an independent external system (not shown in FIG. **1**). For example, the execution plan providing apparatus **130** may constitute the multi-precision NPU or be linked to the independent external system that generates the mixed-precision deep learning model, and may operate in conjunction with one or more external systems that independently perform one or more steps constituting the method according to the present disclosure without being necessarily limited thereto.

[0066] The database **150** may correspond to a storage device that stores various pieces of information required during the operation of the execution plan providing apparatus **130**. The

database **150** may store tools for configuring the multi-precision NPU, and store information for generating the mixed-precision deep learning model. However, without being necessarily limited thereto, the database may store information collected or processed in various forms while the execution plan providing apparatus **130** performs the method for providing the execution plans of the multiple mixed-precision deep learning models based on the multi-precision NPU according to the present disclosure.

[0067] Further, it is shown in FIG. **1** that the database **150** is independent from the execution plan providing apparatus **130**. However, without being necessarily limited thereto, the database may be included as a logical storage device in the execution plan providing apparatus **130**.

[0068] FIG. **2** is a diagram illustrating the system configuration of the execution plan providing apparatus of FIG. **1**.

[0069] Referring to FIG. **2**, the execution plan providing apparatus **130** may include a processor **210**, a memory **230**, a user input/output unit **250**, and a network input/output unit **270**.

[0070] The processor **210** may execute a procedure for providing the execution plans of the multiple mixed-precision deep learning models based on the multi-precision NPU according to an embodiment of the present disclosure, may manage the memory **230** that is read or written in this process, and may schedule synchronization time between a volatile memory and a non-volatile memory in the memory **230**. The processor **210** may control the overall operation of the execution plan providing apparatus **130**, and may be electrically connected to the memory **230**, the user input/output unit **250**, and the network input/output unit **270** to control data flow between them. The processor **210** may be implemented as a Central Processing Unit (CPU) or a Graphics Processing Unit (GPU) of the execution plan providing apparatus **130**. In an embodiment, the processor **210** may be implemented to operate in conjunction with one or more multi-precision NPUs.

[0071] The memory **230** may include an auxiliary memory that is implemented as the non-volatile memory such as a Solid State Disk (SSD) or a Hard Disk Drive (HDD) and is used to store all data required for the execution plan providing apparatus **130**, and may include a main memory that is implemented as the volatile memory such as a Random Access Memory (RAM). Further, the memory **230** may be executed by the electrically connected processor **210** to store a set of instructions that execute the method for providing the execution plans of the multiple mixed-precision deep learning models based on the multi-precision NPU according to the present disclosure.

[0072] The user input/output unit **250** may include an environment for receiving user input and an environment for outputting specific information to the user, and may include, for example, an input device including an adapter such as a touch pad, a touch screen, an on-screen keyboard, or a pointing device, and an output device including an adapter such as a monitor or a touch screen. In an embodiment, the user input/output unit **250** may correspond to a computing device connected through a remote connection. In such a case, the execution plan providing apparatus **130** may be performed as an independent server.

[0073] The network input/output unit **270** may provide a communication environment for connection to the user terminal **110** through a network, and may include an adapter for communication such as, for example, a Local Area Network (LAN), a Metropolitan Area Network (MAN), a Wide Area Network (WAN), and a Value Added Network (VAN). Further, the network input/output unit **270** may be implemented to provide short-range communication functions such as WiFi and Bluetooth or wireless communication functions of 4G or higher for wireless transmission of data.

[0074] FIG. **3** is a diagram illustrating the functional component of the processor of FIG. **2**.

[0075] Referring to FIG. **3**, the execution plan providing apparatus **130** may perform the method for providing the execution plans of the multiple mixed-precision deep learning models based on the multi-precision NPU according to the present disclosure. To this end, the processor **210** of the

execution plan providing apparatus **130** may include multiple functional components. To be more specific, the processor **210** may include an NPU former **310**, a deep learning model generator **330**, an execution plan generator **350**, a model executor **370**, and a controller **390**.

[0076] At this time, the processor **210** of the present disclosure does not need to include all of the above functional components at the same time. However, according to each embodiment, some of the above components may be omitted or some or all of the above components may be selectively included. Hereinafter, the operation of each functional component will be described in detail.

[0077] The NPU former **310** may form a multi-precision Neural Processing Unit (NPU) including a processing element (PE) composed of multiple Micro-PEs. In this regard, the multi-precision NPU may correspond to a hardware device specialized for performing deep learning and artificial neural network operations. Particularly, the multi-precision NPU may be implemented with an optimized structure to efficiently process matrix operations performed during model learning and inference processes and operations requiring various degrees of precisions. To be more specific, the multi-precision NPU may be composed of multiple PEs, each PE may be composed of one or more Micro-PEs, and the performance and function of the NPU may be determined based on a structure between the processing element and the Micro-PE.

[0078] Further, the processing element may correspond to one of key components of the multi-precision NPU, and may process major operations and calculations. The processing element may be designed to generate an operation result after performing a specified operation on input data. Further, the Micro-PE may correspond to a smaller unit forming the processing element, and may be designed to perform small and simple operations. The Micro-PE may be used to perform operations such as multiplication, addition, and activation functions, and the operations of the Micro-PEs may be combined to form the function of the processing element. That is, the Micro-PE is an independent sub-unit within the processing element, and may be used to improve the parallel processing and flexibility of the processing element.

[0079] Further, the NPU former **310** may provide an interface for forming the multi-precision NPU, and the interface may be designed to allow a user to define the operation of the multi-precision NPU by setting various settings and parameters. That is, the user may input configuration information about the multi-precision NPU, and the configuration information may include information such as characteristics of tasks to be executed by the NPU, operation precision, and memory allocation. The NPU former **310** may analyze the nature of task requested by the user and the requirements for necessary resources based on the configuration information input by the user and then reflect the nature and the requirements in the configuration of the NPU, and may configure the multi-precision NPU, including details of hardware resources and algorithms required for the NPU to perform the task.

[0080] The deep learning model generator **330** may generate multiple mixed-precision deep learning models that perform a plurality of precision operations that require different degrees of precision in the model execution process. The mixed-precision deep learning model may correspond to the deep learning model configured by mixing several types of precision operations, and may be designed to include different degrees of precision and detailed conditions depending on the characteristics of the operation. The precision operation may be classified according to the expression precision of the data used in the operation, and operations requiring high precision may require more operation resources at the expense of producing more accurate results. For example, 32-bit floating point operations may require less precision than 64-bit floating point operations.

[0081] Further, the deep learning model generator **330** may analyze the structure of the model and the characteristics of each operation. Thereby, the deep learning model generator **330** may identify the precision required for each operation and distinguish between an operation that requires high precision and an operation that may use low precision. Further, the deep learning model generator **330** may generate mixed-precision deep learning models by allowing each operation to use a data type of corresponding precision based on the precision of the identified operation. Meanwhile, the

deep learning model generator **330** may operate in conjunction with the user terminal **110**, and may generate multiple mixed-precision deep learning models based on configuration information set by the user.

[0082] In an embodiment, the deep learning model generator **330** may generate multiple mixed-precision deep learning models through Hardware-Aware Mixed-precision Quantization (HAWQ). That is, the deep learning model generator **330** may generate a model with minimized execution time and memory usage while maintaining model accuracy by performing additional optimization operations based on a given model or a model configured by the user. To this end, the deep learning model generator **330** may use HAWQ. Of course, without being necessarily limited thereto, the mixed-precision deep learning model may be generated not only through HAWQ but also through other technologies. Herein, a detailed description of HAWQ is omitted, and the model optimization process may include the process of adjusting and quantizing the operation of each layer with optimal precision.

[0083] The execution plan generator **350** may generate the execution plan for executing the multiple mixed-precision deep learning models on the multi-precision NPU. To this end, the execution plan generator **350** may analyze the hardware structure of the multi-precision NPU and the structure and operation characteristics of the mixed-precision deep learning models. For example, the execution plan generator **350** may extract information about the operation amount, memory requirements, and operation type of each model, and extract information about the operation speed and resource constraints of the NPU. The execution plan generator **350** may generate an optimal execution strategy based on the analysis result. At this time, the execution strategy may include the method of efficiently utilizing the resource of the multi-precision NPU to execute the multiple mixed-precision deep learning models in parallel. Further, the execution plan generator **350** may allocate resources within the NPU to process the operation of each model and schedule the operations according to the generated execution strategy. Through this process, the execution plan generator may minimize the overall execution time and maximize resource utilization by arranging the operations of each model according to the optimal execution order.

[0084] In an embodiment, the execution plan generator **350** may generate the execution plan to ensure efficient distribution of the precision operations of each model, taking into account the structure and characteristics of each of the multiple mixed-precision deep learning models. The execution plan generator **350** may determine the operation type required by each model, the parallelization possibility of operation, and memory requirements, etc., based on the model structure and characteristics, and may generate an efficient distribution strategy for the precision operations of each model by determining whether to parallelize operations between multiple models, memory access, etc. The execution plan generator **350** may generate the execution plan including the order, precision, and resource allocation of operations used when each model is executed based on the distribution strategy. That is, the execution plan generator **350** may establish the efficient distribution strategy for precision operation and generate the execution plan, taking into account the structure and characteristics of each mixed-precision deep learning model.

[0085] In an embodiment, the execution plan generator **350** may generate the execution plan to ensure efficient resource utilization while minimizing the execution time of each of the multiple mixed-precision deep learning models using a dynamic programming method. In this regard, the dynamic programming method is an algorithm design technique for solving optimization problems. This may correspond to a method of dividing a problem into a plurality of sub-problems depending on the input size and generating the correct answer to the entire problem based on the correct answer to each sub-problem. To be more specific, the execution plan generator **350** may define resource allocation and scheduling problems required to perform each precision operation for each mixed-precision deep learning model as partial problems. The execution plan generator **350** may generate and combine the correct answers to sub-problems based on each calculation step of the mixed-precision deep learning model. Once all operation steps are completed, an optimal execution

plan may be created to ensure efficient resource utilization while minimizing execution time based on all operation steps.

[0086] In an embodiment, the execution plan generator **350** may generate the execution plan as the result of applying the dynamic programming method based on the results of measuring the execution times of all precision operations for each mixed-precision deep learning model through a pre-simulation method or an actual execution method. The execution plan generator **350** may selectively apply the pre-simulation method or the actual execution method to generate the optimal execution plan. At this time, the execution times of precision operations may be measured differently depending on the method used, and the execution plan generator **350** may apply the dynamic programming method by reflecting errors in execution time according to the selected method.

[0087] In an embodiment, the execution plan generator **350** may detect errors and exceptions that occur during the execution of the mixed-precision deep learning model and add processing plans for the detected errors and exceptions to the execution plan. For example, during the operation of the model, overflow may occur for very large values or underflow may occur for very small values. This may affect the numerical stability of the model. In this case, the execution plan generator **350** may add the processing plan for limiting the range of values using a method such as normalizing values to the execution plan. As another example, NaN (Not a Number) values may occur due to incorrect operation or data loss during model operation, and the execution plan generator **350** may add the processing plan, such as replacing the NaN value with another value or stopping the corresponding operation, to the execution plan.

[0088] In an embodiment, the execution plan generator **350** may analyze data dependency that occurs during the execution of each mixed-precision deep learning model and optimize the execution plan, taking into account the data dependency. In this regard, the data dependency may correspond to a case where the result of one operation affects another operation. In the case of the deep learning model, each operation may be performed on input data, and data dependency may exist between the operations. For example, the data dependency may occur between operations when the output of one operation is used as input to another operation.

[0089] To be more specific, the execution plan generator **350** may determine data dependency between operations by analyzing the structure of the mixed-precision deep learning model to be executed. That is, the execution plan generator **350** may analyze data dependency from a connection relationship between the input and output of each operation. The execution plan generator **350** may generate a dependency graph expressing the data dependency between operations. The dependency graph may represent a data flow between operations in the model by expressing each operation as a node and data dependency as an edge. Further, the execution plan generator **350** may optimize the generated execution plan by applying various techniques based on the generated dependency graph. For example, the execution plan generator **350** may apply techniques such as minimizing data loading and storage, configuring a bundle of operations that may be executed in parallel, or minimizing data movement between operations, thus optimizing the execution plan and improving the processing speed.

[0090] The model executor **370** may dynamically allocate at least one Micro-PE that executes each of multiple precision operations in the process of executing multiple mixed-precision deep learning models according to the execution plan. The model executor **370** may dynamically allocate the Micro-PE that processes operation execution, taking into account the type, precision, priority, etc. of each operation specified in the execution plan. The model executor **370** may determine the precision operation to be executed according to the execution plan and allocate one or more Micro-PEs based on the type of the operation and the required precision. At this time, the model executor **370** may determine the number of the Micro-PEs allocated, taking into account the order of operations and the number of operations that may be simultaneously executed. Further, the model executor **370** may manage the Micro-PE allocated in the operation execution process, monitor the

status of the model and resources, and reallocate or release the resources as needed.
[0091] In an embodiment, the model executor **370** may control precision operations of a model with high execution priority among multiple mixed-precision deep learning models to be preferentially executed according to the execution plan. That is, the model executor **370** may manage priority among multiple models according to the execution plan, determine the execution order of operations for each model according to execution priority, and adjust the execution priority of the model, taking into account changes in execution status and resource status. For example, when a new model is added or the status of the executing model is changed unlike the execution plan, the model executor **370** may readjust the execution priority and manage the execution plan according to an environment that is different from the creation time of the execution plan.
[0092] In an embodiment, the model executor **370** may dynamically adjust the execution plan by tracking and monitoring the execution time and resource usage for the precision operations of each mixed-precision deep learning model. The model executor **370** may track various precision operations required by each mixed-precision deep learning model and measure the execution time of each operation. Further, the model executor **370** may monitor the resource usage of each operation by tracking and quantifying the memory and calculation resources used in the operation process of each model. When an operation with a long execution time or an operation with high resource consumption is detected, the model executor **370** may control the execution plan to maintain optimal performance or minimize resource consumption. The model executor **370** may store tracked information and dynamic adjustment results in the database **150**, and then the collected information may be used in the process of generating the execution plans for the mixed-precision deep learning models.
[0093] The controller **390** may control the overall operation of the processor **210**, and manage the control flow or data flow between the NPU former **310**, the deep learning model generator **330**, the execution plan generator **350**, and the model executor **370**.
[0094] FIG. **4** is a flowchart illustrating a method for providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU according to the present disclosure.
[0095] Referring to FIG. **4**, the execution plan providing apparatus **130** may process a series of operational steps to perform a method for providing the execution plan through the processor **210**. To be more specific, the execution plan providing apparatus **130** may form the multi-precision Neural Processing Unit (NPU) including the processing element (PE) composed of the multiple Micro-PEs through the processor **210** (step S**410**). The execution plan providing apparatus **130** may generate multiple mixed-precision deep learning models that perform a plurality of precision operations, each requiring different degrees of precision, during the model execution process through the processor **210** (step S**430**). The execution plan providing apparatus **130** may generate the execution plan for executing the multiple mixed-precision deep learning models on the multi-precision NPU through the processor **210** (step S**450**).
[0096] FIG. **5** is a diagram illustrating the structure of a framework providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU according to the present disclosure.
[0097] Referring to FIG. **5**, the execution plan providing apparatus **130** may perform the operation of generating the mixed precision model through the HAWQ based on a given model. In this case, the operation is not dependent on HAWQ, which may be replaced by other techniques that provide a similar role. Subsequently, the execution plan providing apparatus **130** may perform the operation of generating the execution plan based on the generated mixed-precision model. The execution plan providing apparatus **130** may use open software called MLIR to perform the corresponding operation, but the present disclosure is not necessarily limited thereto. As a result, the execution plan providing apparatus **130** may efficiently execute the multiple mixed-precision models on the multi-precision NPU through the generated execution plan.
[0098] FIG. **6** is a diagram illustrating an embodiment of an execution plan according to the present

disclosure, and FIG. **7** is a diagram illustrating another embodiment of an execution plan according to the present disclosure.

[0099] Referring to FIGS. **6** and **7**, the execution plan providing apparatus **130** may provide the execution plan that efficiently executes the multiple mixed-precision deep learning models on the multi-precision NPU. Particularly, the execution plan providing apparatus **130** may utilize the multi-precision NPU energy-efficiently and improve overall execution time by eliminating the idle time of one or more Micro-PEs.

[0100] In FIG. **6**, the execution plan providing apparatus **130** may generate the execution plan so that up to four operations (A) capable of low precision may be simultaneously executed, and may generate the execution plan so that up to two operations (B) requiring a medium level of precision may be simultaneously executed. The execution plan providing apparatus **130** may generate an efficient execution plan by configuring a bundle of operations that may be simultaneously executed to eliminate the idle time of the Micro-PE.

[0101] In FIG. **7**, the vertical length of a rectangle may indicate the execution time (t**1** and t**2**) of the corresponding operation, and the top two rectangles may take longer execution time compared to operations of the bottom two rectangles. That is, when configuring the bundle of operations that may be simultaneously executed, the execution plan providing apparatus **130** may prevent the idle time from increasing after the operation of a specific Micro-PE is first completed and before a subsequent operation is started by considering the execution time of each operation.

[0102] Consequently, the execution plan providing apparatus **130** may apply the techniques shown in FIGS. **6** and **7** based on the framework shown in FIG. **5** to generate the execution plan that can quickly and energy-efficiently execute multiple mixed-precision models on the multi-precision NPU.

[0103] FIG. **8** is a diagram illustrating a difference in performance of the multi-precision NPU depending on the application of the execution plan according to the present disclosure.

[0104] Referring to FIG. **8**, the x-axis of a graph may represent the execution time. As the x-axis length becomes shorter, it may indicate that multiple mixed-precision deep learning models have been more quickly executed. In addition, the x-axis of the graph may correspond to a normalized result based on ‘no optimization applied’. That is, it can be seen in FIG. **8** that the multi-precision NPU can be more efficiently used by executing the multiple mixed-precision deep learning models according to the execution plan of the present disclosure.

[0105] To be more specific, in performance comparison experiments, eight types of mixed precision models containing various levels of precision and detailed conditions may be used, and a total of 12 types of mixed precision models may be used in combination.

[0106] As the result of the experiment, in the case of ‘no optimization applied’, the overall structure of each model in multiple mixed precision models is not understood, so it is not possible to construct a bundle of low precision operations that may be appropriately performed simultaneously. Thus, the idle time of the Micro-PE may result in inefficient resource use of the multi-precision NPU.

[0107] On the other hand, in the case of ‘optimization application’, the method according to the present disclosure is applied, and the idle time of the Micro-PE is eliminated by configuring an appropriate bundle of low-precision operations, resulting in efficient resource use of the multi-precision NPU. Thereby, the method according to the present disclosure can achieve a 1.2-fold improvement in average execution time.

[0108] Although the present disclosure was provided above in relation to specific embodiments shown in the drawings, it is apparent to those skilled in the art that the present disclosure may be changed and modified in various ways without departing from the scope of the present disclosure, which is described in the following claims.

Detailed Description of Main Elements

[0109] **100**: execution plan providing system [0110] **110**: user terminal **130**: execution plan

providing apparatus [0111] **150**: database [0112] **210**: processor **230**: memory [0113] **250**: user input/output unit **270**: network input/output unit [0114] **310**: NPU former **330**: deep learning model generator [0115] **350**: execution plan generator **370**: model executor [0116] **390**: controller

Claims

1. An apparatus for providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU, the apparatus comprising: a memory; and a processor electrically connected to the memory, wherein the processor is configured to: form a multi-precision Neural Processing Unit (NPU) including a processing element (PE) composed of multiple Micro-PEs, generate multiple mixed-precision deep learning models that perform multiple precision operations requiring different degrees of precision in a model execution process, and generate the execution plans for executing the multiple mixed-precision deep learning models on the multi-precision NPU.
2. The apparatus of claim 1, wherein the processor generates the multiple mixed-precision deep learning models through Hardware-Aware Mixed-precision Quantization (HAWQ).
3. The apparatus of claim 1, wherein the processor generates the execution plan to ensure efficient distribution of precision operations for each model, taking into account a structure and characteristics of the multiple mixed-precision deep learning models.
4. The apparatus of claim 1, wherein the processor generates the execution plan to ensure efficient resource utilization while minimizing execution time of each of the multiple mixed-precision deep learning models using a dynamic programming method.
5. The apparatus of claim 4, wherein the processor generates the execution plan as a result of applying the dynamic programming method based on a result of measuring execution times of all precision operations for each mixed-precision deep learning model through a pre-simulation method or an actual execution method.
6. The apparatus of claim 1, wherein the processor dynamically allocates at least one Micro-PE that executes each of the multiple precision operations in a process of executing the multiple mixed-precision deep learning models according to the execution plan.
7. The apparatus of claim 6, wherein the processor controls precision operations of a model with high execution priority among the multiple mixed-precision deep learning models to be preferentially executed according to the execution plan.
8. The apparatus of claim 6, wherein the processor dynamically adjust the execution plan by tracking and monitoring the execution time and resource usage for precision operations of each mixed-precision deep learning model.
9. A method for providing execution plans of multiple mixed-precision deep learning models based on a multi-precision NPU, the method being performed in a computing device including a memory; and a processor electrically connected to the memory, the method comprising: forming a multi-precision Neural Processing Unit (NPU) including a processing element (PE) composed of multiple Micro-PEs, through the processor; generating multiple mixed-precision deep learning models that perform multiple precision operations requiring different degrees of precision in a model execution process, through the processor; and generating the execution plans for executing the multiple mixed-precision deep learning models on the multi-precision NPU, through the processor.
10. The method of claim 9, wherein the generating the multiple mixed-precision deep learning models comprises generating the multiple mixed-precision deep learning models through Hardware-Aware Mixed-precision Quantization (HAWQ).
11. The method of claim 9, wherein the generating the execution plans comprises generating the execution plan to ensure efficient distribution of precision operations for each model, taking into account a structure and characteristics of the multiple mixed-precision deep learning models.
12. The method of claim 9, wherein the generating the execution plans comprises generating the execution plan to ensure efficient resource utilization while minimizing execution time of each of

the multiple mixed-precision deep learning models using a dynamic programming method.

13. The method of claim 12, wherein the generating the execution plans comprises generating the execution plan as a result of applying the dynamic programming method based on a result of measuring execution times of all precision operations for each mixed-precision deep learning model through a pre-simulation method or an actual execution method.

14. The method of claim 9, wherein the generating the execution plans comprises analyzing data dependency that occurs during the execution of each mixed-precision deep learning model and then optimizing the execution plan, taking into account the data dependency.

15. The method of claim 9, wherein the generating the execution plans comprises detecting errors and exceptions that occur during the execution of the mixed-precision deep learning model and adding processing plans for the detected errors and exceptions to the execution plan.

16. The method of claim 9, further comprising: dynamically allocating at least one Micro-PE that executes each of the multiple precision operations in a process of executing the multiple mixed-precision deep learning models according to the execution plan.

17. A computer-readable recording medium for storing a computer program, wherein the computer program, when executed by a processor, comprises instructions for causing the processor to perform an operation comprising: forming a multi-precision Neural Processing Unit (NPU) including a processing element (PE) composed of multiple Micro-PEs; generating multiple mixed-precision deep learning models that perform multiple precision operations requiring different degrees of precision in a model execution process; and generating the execution plans for executing the multiple mixed-precision deep learning models on the multi-precision NPU.