

Atlas 200

MindSpore Studio Product Description

Issue 02
Date 2019-06-17



Copyright © Huawei Technologies Co., Ltd. 2019. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without prior written consent of Huawei Technologies Co., Ltd.

Trademarks and Permissions



HUAWEI and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: <http://e.huawei.com>

About This Document

Purpose

This document describes the overall architecture and basic functions of the Mind Studio.






Intended Audience

This document is intended for:

- Huawei Technical support engineers
- Partner Technical support engineers
- ISV Software Engineers

Symbol Conventions

The symbols that may be found in this document are defined as follows.

Symbol	Description
	Indicates an imminently hazardous situation which, if not avoided, will result in death or serious injury.
	Indicates a potentially hazardous situation which, if not avoided, could result in death or serious injury.
	Indicates a potentially hazardous situation which, if not avoided, may result in minor or moderate injury.
	Indicates a potentially hazardous situation which, if not avoided, could result in equipment damage, data loss, performance deterioration, or unanticipated results. NOTICE is used to address practices not related to personal injury.
	Calls attention to important information, best practices and tips. NOTE is used to address information not related to personal injury, equipment damage, and environment deterioration.

Change History

Issue	Date	Description
01	2019-04-30	This issue is the first official release.

Contents

About This Document.....	ii
1 Overview.....	1
2 Function Description.....	2
3 Application Development.....	4
3.1 Project Management.....	4
3.2 Graphical Service Orchestration.....	6
3.3 Model Management.....	7
3.4 Offline Model Conversion.....	9
3.5 Development of Custom Operators	12
3.6 Dataset Management.....	12
4 Performance Tuning.....	15
4.1 Performance Profiler.....	15
4.2 Log Analysis.....	17
4.3 Black Box.....	18
A Getting Help.....	19
A.1 Preparing for Contacting Huawei.....	19
A.2 Contacting Huawei Technical Support.....	19
A.3 Preparing for Debugging.....	19
A.4 Using Product Documentation.....	20
A.5 Getting Help from Website.....	20
A.6 Ways to Contact Huawei.....	20
B Glossary.....	22
C Acronyms and Abbreviations.....	29

1 Overview

MindSpore Studio is an AI full-stack development platform developed based on Huawei's neural-network processing unit (NPU). On MindSpore Studio, you can develop chip-based operators, and develop custom operators. Network migration, optimization, and analysis at the network layer is also supported. In addition, the service engine layer provides a set of visualized AI engine drag-and-drop programming services, which greatly reduces the AI engine development threshold. The entire platform provides the following four services for developers in a Web manner.

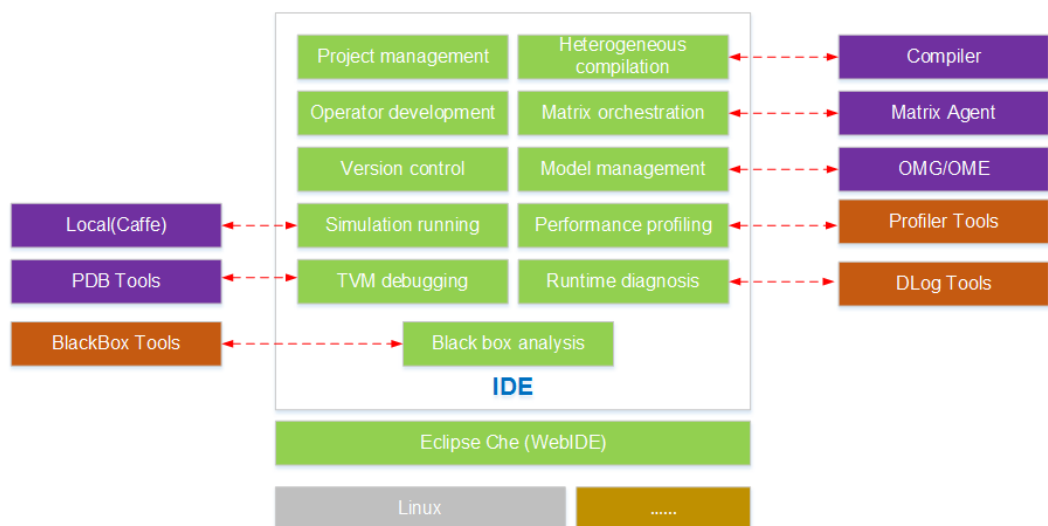
- For operator development:
MindSpore Studio provides development of a full set of operators. It supports running in a real environment, visualized debugging of heterogeneous programs that are dynamically scheduled and third-party operator development, greatly reducing the operator development threshold based on Huawei-developed NPUs, improving the efficiency of operator development and enhancing product competitiveness.
- For development at the network layer:
MindSpore Studio integrates the offline model conversion tool (OMG), model quantization tool, model running profiling tool, and log analysis tool, greatly improving the efficiency of migration, analysis and optimization of network models.
- For AI engine development:
MindSpore Studio provides the AI engine that supports visualized drag-and-drop programming and a large number of automatic algorithm code generation technologies, largely reducing the threshold for developers. MindSpore Studio is equipped with various algorithm engines, such as ResNet-18, boosting the development and migration efficiency of the user AI algorithm engine.
- For application development:
MindSpore Studio integrates various tools, such as Profile and Compiler, providing a graphically integrated development environment. Developers can perform full-process development on MindSpore Studio, covering project management, compilation, commissioning, simulation, and performance analysis, greatly enhancing the development efficiency.

2 Function Description

Overall Architecture

Figure 2-1 shows the overall architecture of MindSpore Studio.

Figure 2-1 Overall architecture of MindSpore Studio



Function Description

MindSpore Studio provides the following features:

- **User-friendly NPU-based programming GUI**
Operator developers can customize CCE development on MindSpore Studio based on the CCE programming depth to implement in-depth integration. The keywords of the extended CCE language are highlighted. You can compile heterogeneous hybrid codes in one-click mode.
- **NPU-based graphical debugging**
For the development of the operator acceleration library on the NPU, MindSpore Studio provides a graphical GUI for users to implement real-time tracking of the running status of the acceleration operators on the AI core and AI CPU.
- **Automatic offline model management**

Trained third-party models, such as Caffe models, can be imported to MindSpore Studio and converted into models supported by the system. Model interfaces are generated automatically in one-click mode, enabling interface-based model programming. For details, see 4 "Offline Model Conversion" in the *Ascend 310 MindSpore Studio Quick Start*.

- **"Zero" programming for service process orchestration**

For service process developers, MindSpore Studio provides the drag-and-drop programming mode based on service nodes. You can implement service orchestration by simply dragging and connecting service nodes. The one-stop service after orchestration, ranging from compilation and running to result display, makes process development smarter. "Zero" programming is involved. In this way, you can get started quickly without extra learning costs. For details, see 3 "Building the First Machine Learning App" in the *Ascend 310 MindSpore Studio Quick Start*.

- **Graphical TE programming**

MindSpore Studio provides the industry's first integrated development environment based on the TVM-based Tensor Engine (TE) for programming development. Operators can be transplanted quickly across platforms, enabling instant NPU adaption.

- **Log Analysis**

MindSpore Studio provides a system-wide log collection and analysis solution for the NPU platform, improving the efficiency of locating runtime algorithm problems. A unified log format is adopted. Visualized analysis of cross-platform logs and runtime diagnosis runs in Web mode, improving the usability of the log analysis system.

- **Performance analysis**

MindSpore Studio provides graphical user interfaces (GUIs) and command-line interfaces (CLIs) to implement efficient, easy-to-use, and flexible performance profiling on the multi-node and multi-module heterogeneous system on the host and device. Synchronous analysis of performance and power consumption of the NPU device is implemented, which meets the requirements of algorithm optimization for system performance analysis.

- **Simulation**

Function-level simulation execution libraries for the AI core under the Caffe framework are provided. You can call AI core simulation by using the program.

3 Application Development

This topic describes the main functions of MindSpore Studio in application development.

[3.1 Project Management](#)

[3.2 Graphical Service Orchestration](#)

[3.3 Model Management](#)

[3.4 Offline Model Conversion](#)

[3.5 Development of Custom Operators](#)

[3.6 Dataset Management](#)

3.1 Project Management

MindSpore Studio supports the following projects:

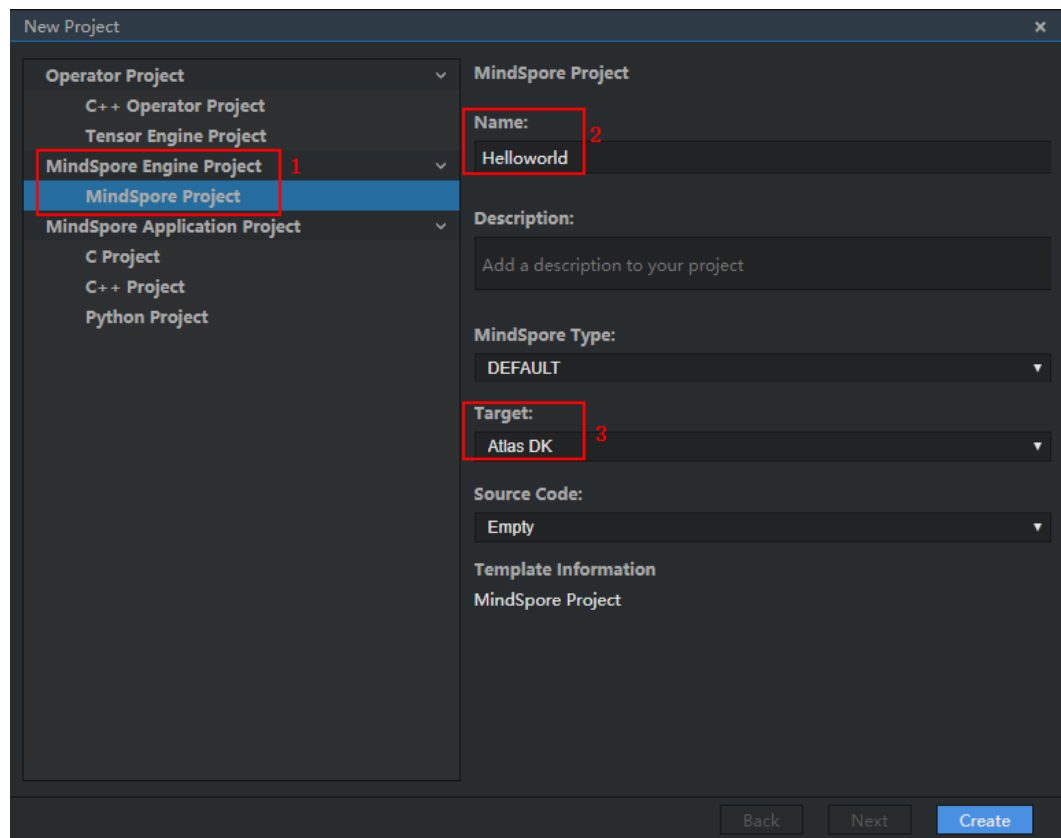
- Python projects
- C/C++ projects
- Matrix orchestration projects (Mind projects)
- C/C++ projects developed based on offline models
- Tensor Engine projects

You can perform the following project management operations:

- Creating/Deleting a project
- Uploading/Downloading a project
- Opening/Closing a project
- Creating/Deleting a file
- Uploading a file/folder

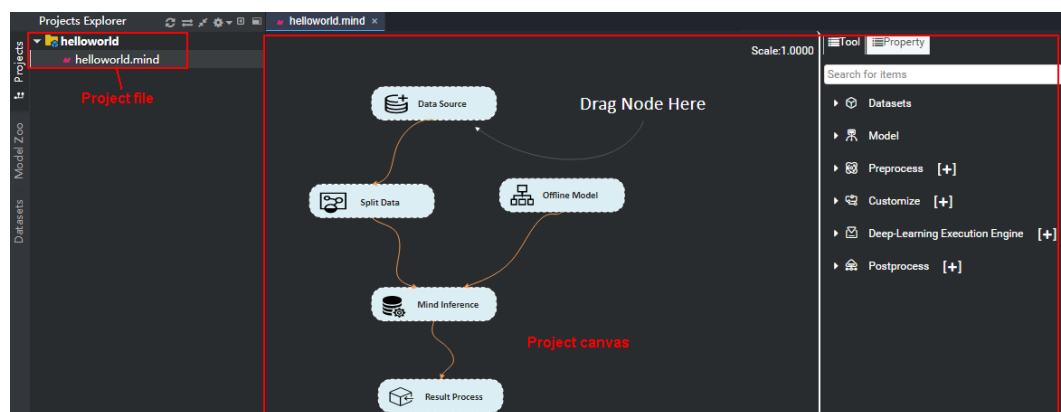
Figure 3-1 shows the dialog box for creating a project.

Figure 3-1 Creating a project



After a project is created, a .mind file with the same name as the project name is generated. Figure 3-2 shows the workspace after a project is created.

Figure 3-2 Workspace



Currently, the following 12 keyboard shortcuts are not supported on the canvas.

Table 3-1 Keyboard shortcuts not supported on the canvas

Keyboard Shortcut	Description	Keyboard Shortcut	Description
Ctrl+E	Open the last opened file.	Alt+F12	Open the terminal operation window.
Alt+W	Close the last opened file.	Shift+F10	Check the command line options.
Ctrl+Shift+F	Open the search dialog box.	Alt+Shift+F9	Open the debug configuration.
Ctrl+Shift+A	Query an action.	Alt+←	Go to the previous file.
Ctrl+Alt+N	Query a file.	Alt+→	Go to the next file.
Alt+G	Open the compilation configuration.	Alt+O	Modify a file.

3.2 Graphical Service Orchestration

MindSpore Studio provides Matrix, a graphical service orchestration tool. With this tool, you are allowed to orchestrate projects by dragging nodes. The process code is automatically generated by the DSL, requiring "zero" human programming and greatly reducing the bug introduction risks. MindSpore Studio provides various visualized views, covering data flows, models, result information, and system analysis.

Figure 3-3 and **Table 3-2** describe the nodes supported by MMindSpore Studio.

Figure 3-3 Node types supported by MindSpore Studio

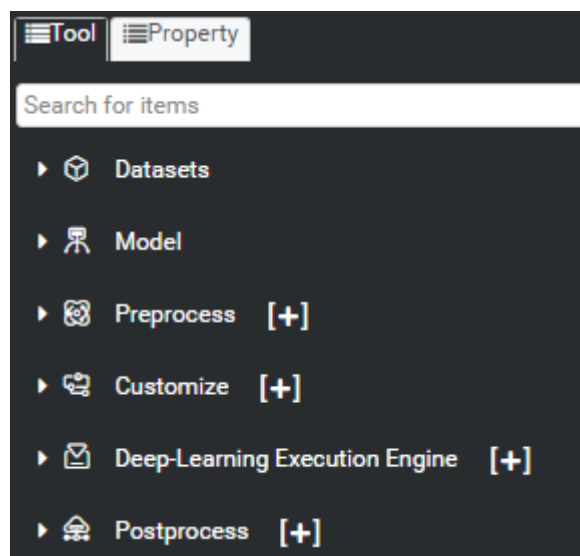


Table 3-2 Nodes supported by MindSpore Studio

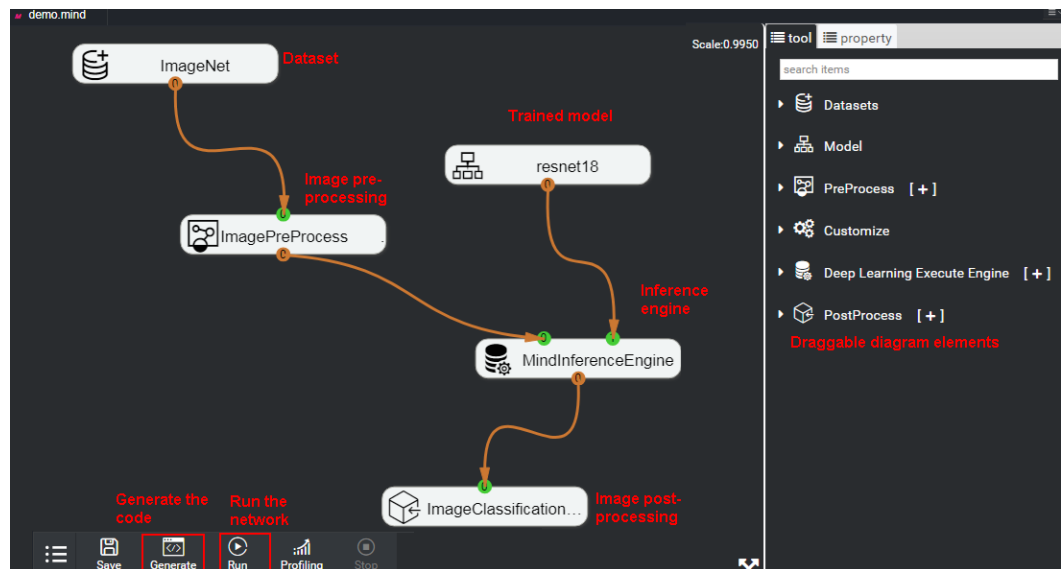
Node Type	Description
Datasets	Dataset nodes
Model	Model nodes
PreProcess	Pre-processing nodes
Customize	Custom nodes
Deep Learning Execute Engine	Deep learning network execution engine (DLEE)
PostProcess	Post-processing nodes

The basic operations in process orchestration are as follows:

- Node dragging: Drag a node on the **tool** tab page to the canvas. Select a node in the canvas to set the properties of the node on the **property** tab page.
- Node connecting: Select a node and drag the cursor to another node to connect the process. The output from a node serves as the input to another node.

Figure 3-4 shows a process orchestration example.

Figure 3-4 Process orchestration example



3.3 Model Management

Models are classified into built-in models, custom models, and Caffe models.

- **BuiltIn Models**

Built-in models are preset in MindSpore Studio and exist before a workspace is created. You can use them but cannot add, delete, or modify them.

The currently available built-in model is ResNet-18.

- **My Models**

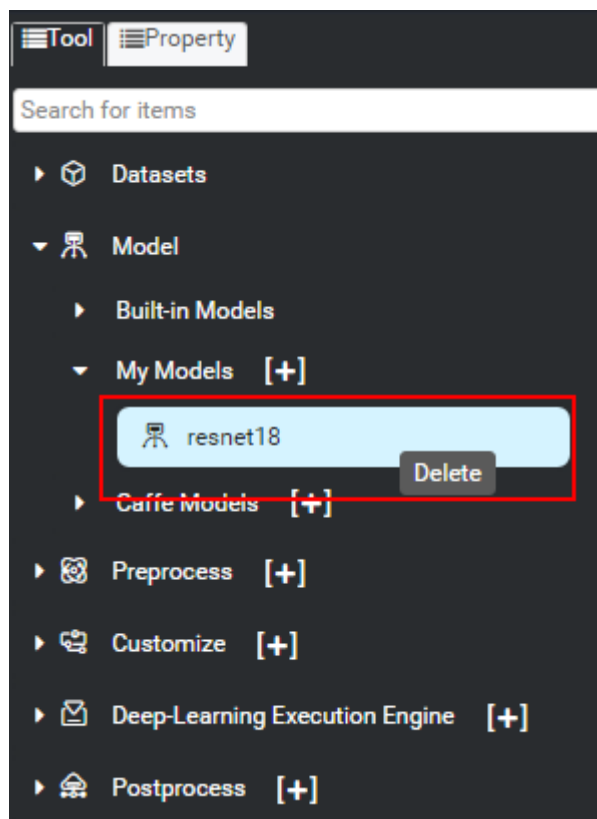
A custom model (such as a Caffe model) can be added to MindSpore Studio by using the offline conversion function for future use. A newly created workspace has no custom models. You can add a custom model by model conversion or simply adding one.

- **Caffe Models**

After a Caffe model is added on the orchestration window, the Caffe model is added to **Model Zoo**. A newly created workspace has no Caffe models.

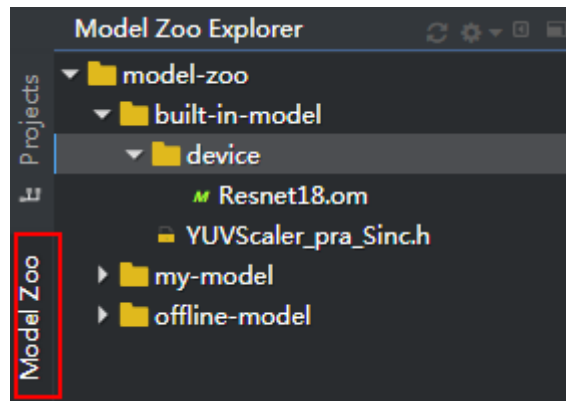
As shown in **Figure 3-5**, available models are automatically displayed as diagram elements on the **tool** tab page on the right. You can drag the nodes to the canvas for process orchestration.

Figure 3-5 Diagram elements ready to be dragged



As shown in **Figure 3-6**, you can view the file structure of a model in the **Model Zoo** area on the left.

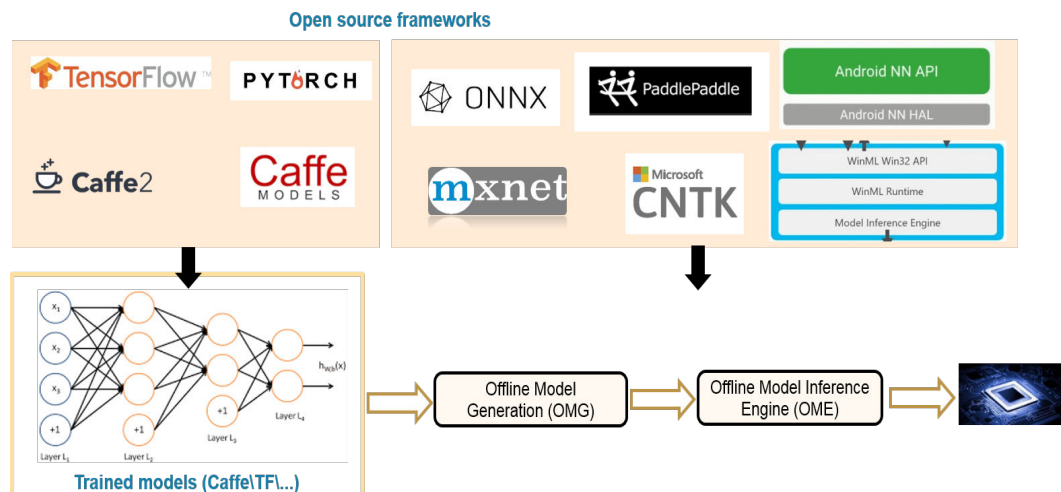
Figure 3-6 Model Zoo area



3.4 Offline Model Conversion

You can convert an open-source neural network model such as a Caffe model into a model supported by the Huawei NPU. Figure 3-7 shows the overall solution.

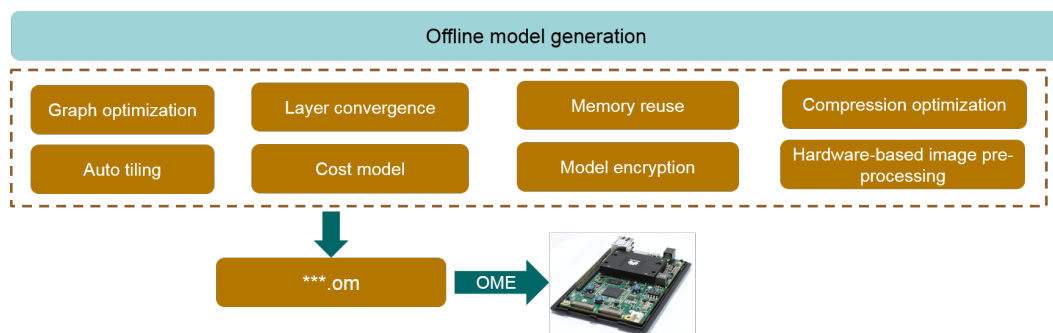
Figure 3-7 Offline model conversion solution



- Offline Model Generation (OMG): offline model generation tool, which automatically selects a proper optimization policy to generate an offline model.
- Offline Model Inference Engine (OME): model inference execution engine, which efficiently executes operators for the neural network model.

Figure 3-8 describes the key technologies involved in offline model conversion.

Figure 3-8 Key technologies involved in offline model conversion



An optimal policy is automatically selected to generate an offline OMG model.

Perform the following steps to convert a model:

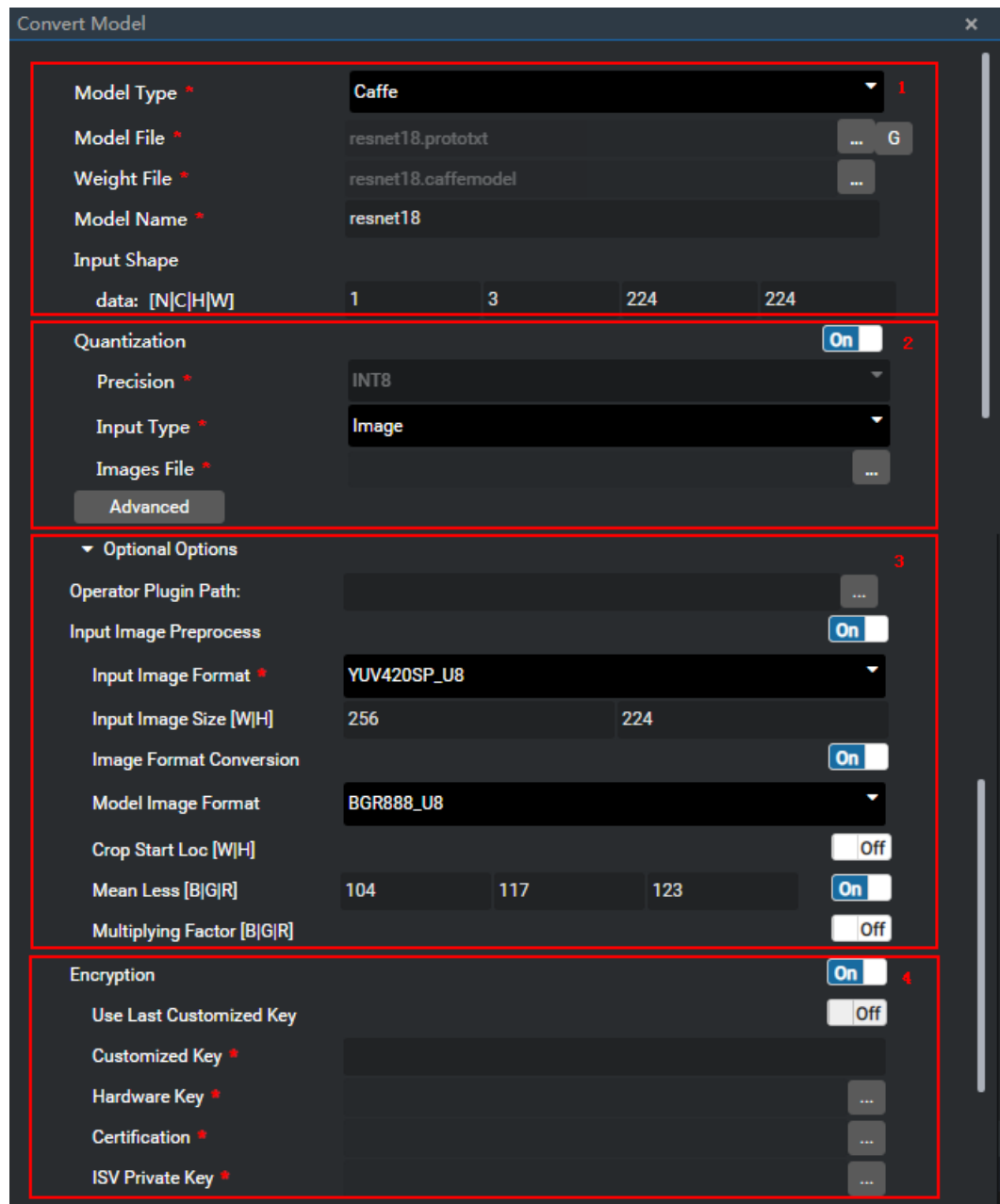
Step 1 Right-click a project and choose **Convert Model** from the shortcut menu, or choose **Tools > Convert Model** from the menu bar. The model conversion window is displayed.

MindSpore Studio offline model conversion supports only Caffe models.

Step 2 You can enable 8-bit quantization. With the verification set input, faster inference is obtained at a low memory cost.

See area 2 in [Figure 3-9](#).

Figure 3-9 Model conversion configuration



Step 3 Implement hardware-based image pre-processing during the input to the first layer in the NN, accelerating operation efficiency.

See area 3 in [Figure 3-9](#).

Step 4 You can encrypt a model by using hardware keys to support the intellectual property right of the model.

See area 4 in [Figure 3-9](#).

You can monitor the whole model conversion process in a visualized way.

After successful conversion, a message is displayed indicating the storage space occupied by the model and its runtime memory usage, helping you to identify resource risks in advance.

If the model conversion fails, you can view the operator analysis report automatically generated.

---End

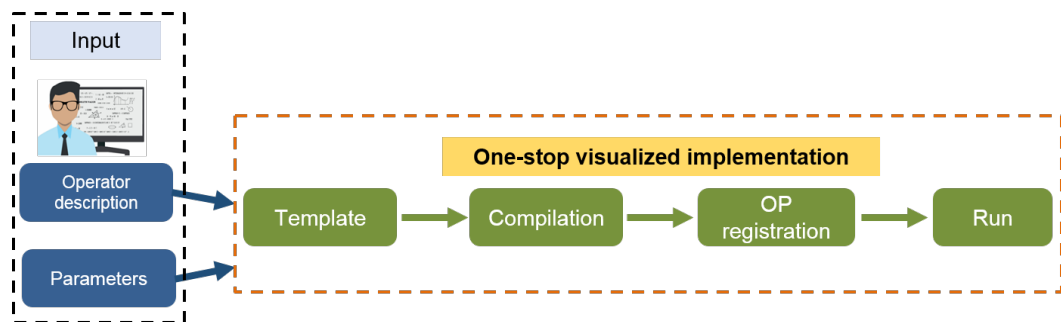
3.5 Development of Custom Operators

During model conversion, a message is displayed for an unsupported operator, that is, an operator not implemented in the acceleration operator libraries such as the CCE operator library and the AI CPU operator library and needs to be user-defined. You can add a custom operator to the operator library to facilitate the model conversion.

The MindSpore Studio provides a tool for the development of custom Tensor Engine (TE) operators. TE is a custom operator development framework based on Tensor Virtual Machine (TVM). It provides the DSL language in the Python syntax for developing custom operators.

Figure 3-10 shows the process of developing custom operators.

Figure 3-10 Process of developing custom operators



The process of model conversion using custom operators is as follows. For details, see the *Atlas 200 Tensor Engine Operator Development Guide*.

Step 1 Use the TE framework to develop an operator in MindSpore Studio.

1. Create a Mind project.
2. Use the TE framework to compile the code for operator implementation. If an operator file exists on the local host, you can choose **File > Upload Project** to upload the operator file to the custom project.
3. Build the operator and test the operator correctness.

Step 2 Develop an operator plug-in in MindSpore Studio and insert the operator developed in Step 1 into the model conversion process as a plug-in.

Step 3 Perform offline model conversion again.

---End

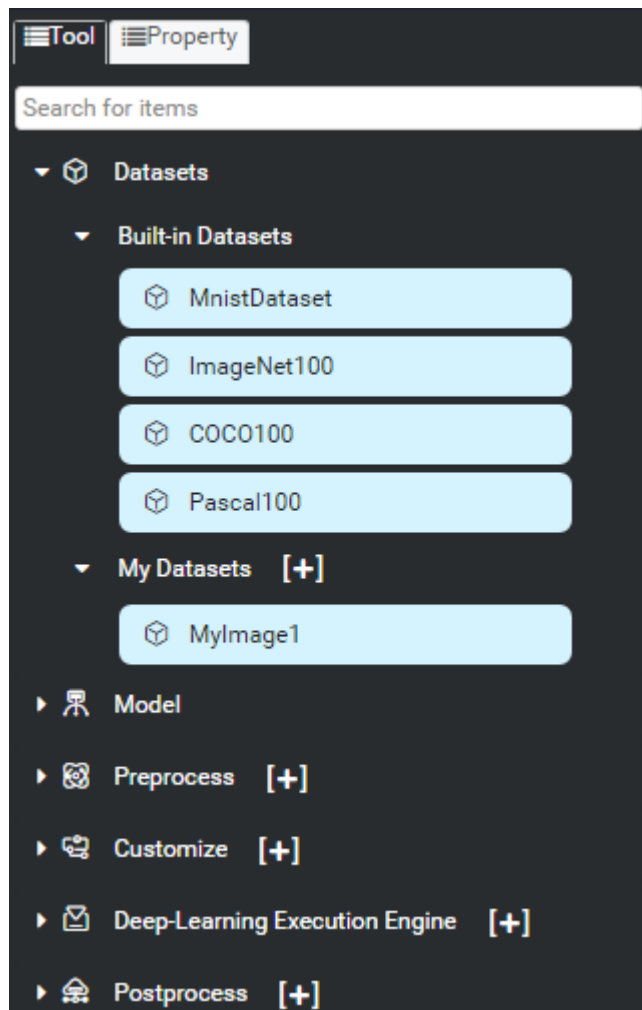
3.6 Dataset Management

Datasets are classified into built-in datasets and custom datasets (my datasets). Built-in datasets can be directly used by dragging. Custom datasets need to be manually imported.

- **BuiltIn Datasets**
They are provided by the MindSpore Studio for direct use.
- **My Datasets**
You can create your own datasets by saving sets of images as custom datasets for future use.
The images of a custom dataset can be obtained from local files, local folders, and URLs.

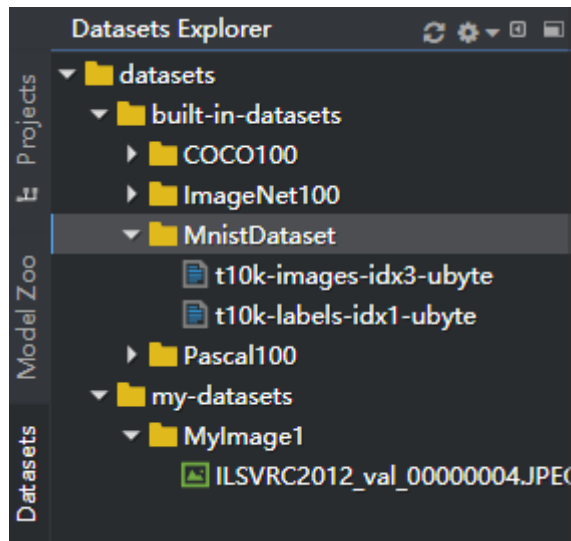
As shown in **Figure 3-11**, available datasets are automatically displayed as diagram elements on the **tool** tab page on the right. You can drag the datasets to the canvas for process orchestration.

Figure 3-11 Diagram elements ready to be dragged



As shown in **Figure 3-12**, you can browse the files contained in each dataset in **Datasets Explorer**.

Figure 3-12 Datasets Explorer



4 Performance Tuning

[4.1 Performance Profiler](#)

[4.2 Log Analysis](#)

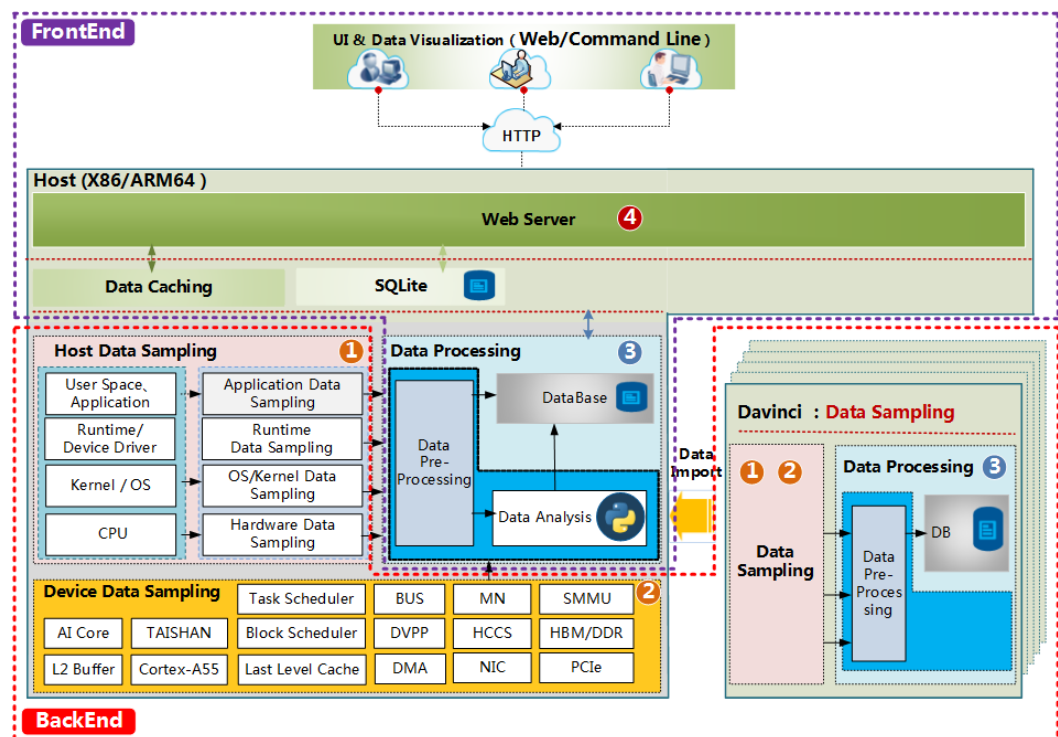
[4.3 Black Box](#)

4.1 Performance Profiler

MindSpore Studio provides graphical user interfaces (GUIs) and command-line interfaces (CLIs) to implement efficient, easy-to-use, and flexible performance profiling on the multi-node and multi-module heterogeneous system on the host and device. Synchronous analysis of performance and power consumption of the NPU device is implemented, which meets the requirements of algorithm optimization for system performance analysis.

[Figure 4-1](#) shows the principle of the performance profiler.

Figure 4-1 Principle diagram of the performance profiler



The performance analysis of MindSpore Studio includes:

- Time slice analysis, including AI core execution, AI CPU execution, and runtime API execution time slice analysis.
- Instruction count performance.
- Memory performance specifications, including device memory usage, memory transaction loading count, and memory transaction loading throughput.
- HCCS performance specifications, including total amount of data sent and received by the HCCS, total amount of user data sent and received by the HCCS, and the HCCS overhead of sending and receiving data.
- FU performance specifications, including FU load/store execution usage, control instruction usage, and usage of feature operations such as sin and cos.
- Performance specifications of the Task Scheduler, including task execution sequence, queue status statistics, and processor load statistics.
- Bandwidth performance specifications, including statistics of the PCIe interface on the host, bus bandwidth usage, and bandwidth statistics of the DVPP input/output interface.
- System performance specifications, including system clock, memory clock, and temperature.

Figure 4-2, Figure 4-3, and Figure 4-4 show the performance analysis results.

Figure 4-2 Running status analysis

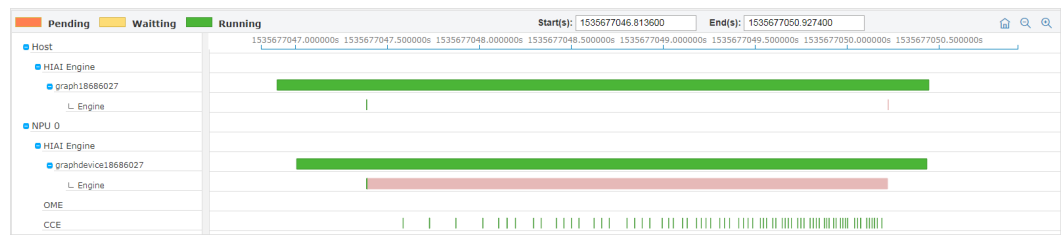
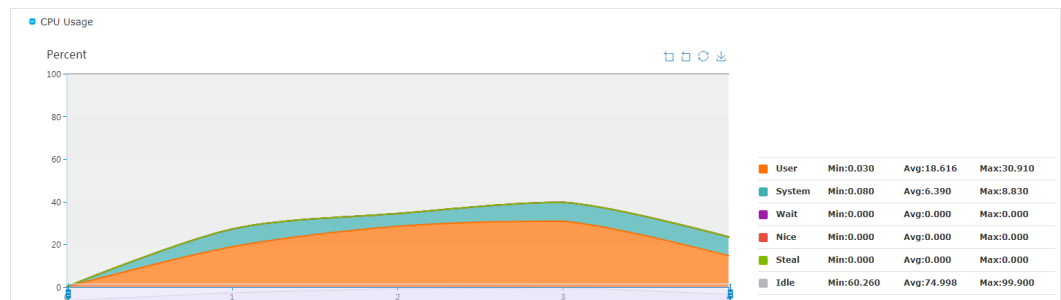


Figure 4-3 Thread analysis



Figure 4-4 CPU usage analysis

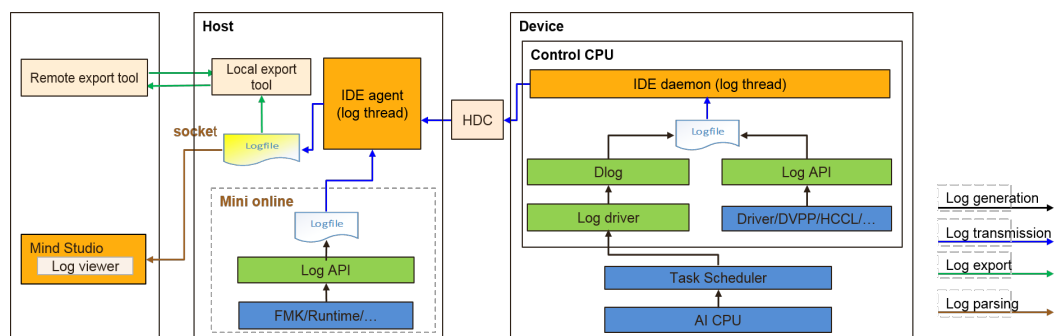


4.2 Log Analysis

MindSpore Studio provides a system-wide log collection and analysis solution for the NPU platform, improving the efficiency of locating runtime algorithm problems. A unified log format is adopted. Visualized analysis of cross-platform logs and runtime diagnosis runs in Web mode, improving the usability of the log analysis system.

Figure 4-5 shows the principle of log analysis of MindSpore Studio.

Figure 4-5 Principle diagram of log analysis



- The device generates log and transfers log through the HDC channel.
- The host dumps and compresses logs.

- MindSpore Studio parses and displays logs.
- Logs stored on the host can be exported.

Currently, logs of the following modules are supported: Dlog, Slog, IDE-daemon-host, IDE-daemon-device, Log-agent-host, HCCL, Framework, Matrix, DVPP, Runtime, CCE, HDC, Driver, MDC, DEVMM, and Kernel.

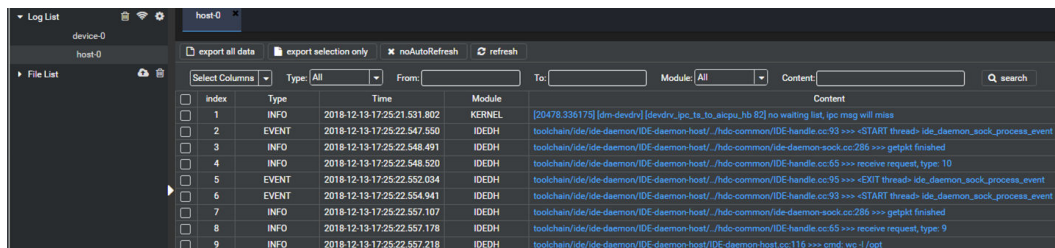
- Slog: system logs.
- Matrix: model.
- HCCL: Huawei collection communication library, which provides APIs for operations such as reduce and gather.
- MDC: self-driving, including regulation control, space perception, monitoring, and positioning.
- DEVMM: device memory management.
- Kernel: system kernel.

Logs reported by each module are displayed in the IDE in a centralized manner.

Output logs can be filtered by module, time, log type, and keyword. You can import offline logs for analysis. You can also export unfiltered logs and filtered logs.

Figure 4-6 shows the log analysis result.

Figure 4-6 Log analysis



4.3 Black Box

The black box is used to store important running information before the system is restarted and provides debugging information for breakdown locating.

The MindSpore Studio black box function is triggered in the following scenarios:

- The system breaks down and restarts due to a software reason such as Linux panic, driver exception, secure OS exception.
- The system breaks down due to a hardware reason such as the SoC exceeds a certain temperature or the DDR bus fails to response.
- A subsystem startup failure occurs, such as a control CPU system startup failure, TS startup failure, AI CPU failure, and LPM3 startup failure.

A Getting Help

A.1 Preparing for Contacting Huawei

To better solve the problem, you need to collect troubleshooting information and make debugging preparations before contacting Huawei.

A.2 Contacting Huawei Technical Support

If a fault persists after troubleshooting, contact Huawei technical support by visiting <https://e.huawei.com>. Before you report a fault to Huawei engineers, collect the following information:

- Name and address of your organization
- Contact person and telephone number
- Time when the fault occurred
- Detailed fault symptom
- Device types, hardware models and software versions
- Any measures taken and effects
- Fault severity and expected rectification deadline

 **NOTE**

If a fault described in this document occurs and cannot be rectified by the recommended solutions, contact the local Huawei branch office or Huawei Customer Service Center for technical support.

A.3 Preparing for Debugging

When you seek Huawei technical support, Huawei technical support engineers may assist you in performing some operations to further collect fault information or rectify the fault.

Before contacting technical support engineers, prepare the spare parts for boards and port modules, screwdrivers, screws, serial cables, and network cables.

A.4 Using Product Documentation

Huawei provides the documents delivered with the equipment. This document provides guidance for you to solve common problems that occur during routine maintenance or troubleshooting.

To better rectify the fault, you are advised to use the guide before contacting Huawei technical support engineers.

A.5 Getting Help from Website

Huawei provides users with timely and efficient technical support through the regional offices, secondary technical support system, telephone technical support, remote technical support, and onsite technical support.

Technical support website

See technical documents on one of the following technical support websites:

- Huawei enterprise business website: <http://e.huawei.com>
- Huawei carrier business website: <http://carrier.huawei.com>

Huawei Technical Support

If a fault persists after taking the above measures, contact technical support at your local Huawei office. If a local Huawei office is not available, contact Huawei technical support as follows:

- Enterprise customers
Send emails to support_e@huawei.com or visit [Global Service Hotline](#).
- Carriers
Send emails to support@huawei.com or visit [Global TAC Information](#).

Knowledge Base and Self-Service Platform

If you want to further learn server knowledge and communicate with experts, perform the following operations:

- [HUAWEI Server Information Self-Service Platform](#) for specific server product documentation.
- [Huawei server intelligent Q&A system](#) for quick learning about products.
- [Huawei Enterprise Support Community \(Sever\)](#) for learning and discussion.
- You can learn server cases from [Knowledge Base](#).

A.6 Ways to Contact Huawei

Huawei Technologies Co., Ltd. provides customers with comprehensive technical support and service. For any assistance, contact our local office or company headquarters.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base Bantian, Longgang Shenzhen 518129 People's Republic of China

Website: <http://e.huawei.com>

B Glossary

A

AI Artificial intelligence (AI) is a type of computer technology which is concerned with making machines work in an intelligent way, similar to the way that the human mind works.

AIPP AI pre-processing (AIPP) supports operations such as format conversion, padding, and cropping, CSC (YUV2RGB or RGB2YUV), scaling, and channel data exchanges.

B

BIOS Firmware stored on the computer motherboard that contains basic input/output control programs, power-on self test (POST) programs, bootstraps, and system setting information. The BIOS provides hardware setting and control functions for the computer.

BIU The bus interface unit (BIU) records the memory access between an AI core and a DDR SDRAM or an L2 cache.

BTB The board-to-board connector (BTB) is used to connect printed circuit boards (PCBs).

C

CAN Controller Area Network (CAN) is a message-based protocol, designed originally for multiplex electrical wiring within automobiles, but is also used in many other contexts.

CCE The cube-based computing engine (CCE) provides acceleration for upper-layer applications (frameworks or applications supporting machine learning) through APIs.

CCE-GDB The CCE-GNU debugger (CCE-GDB) is used to debug AI applications and control the code debugging of the CPU, AI core, and AI CPU. (The debugging of the AI CPU is still under development.)

CCEC	The CCE compiler (CCEC) is a heterogeneous system compiler, or a compilation tool of the CCE heterogeneous programming language. It compiles mixed CCE codes, including CCE host codes and CCE AI CPU AI core codes, and generates executable files to be running on the CCE system.
CFM	It is a unit for measuring the gas flow rate.
CNN	The convolutional neural network (CNN) is a feedforward neural network that contains artificial neural elements capable of responding to surrounding units and supports large-scale image processing.
CPU	A central processing unit (CPU), also called a central processor or main processor, is the electronic circuitry within a computer that carries out the instructions of a computer program by performing the basic arithmetic, logic, controlling, and input/output (I/O) operations specified by the instructions.
CUBE	CUBE refers to the matrix operation.
D	
DDK	A digital development kit (DDK) is a developer suite provided by the Mind solution. After the DDK is installed, Mind Studio can obtain required components such as APIs, libraries, and tool chains for Mind development.
DDR	In computing, a computer bus operating with double data rate (DDR) transfers data on both the rising and falling edges of the clock signal.
DL	Deep learning (also known as deep structured learning or hierarchical learning) is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Learning can be supervised, semi-supervised or unsupervised.
DVPP	Digital vision pre-processing (DVPP) provides operations such as decoding and scaling of videos and images in specific formats, and encodes and outputs processed videos and images.
E	
ECC	Error checking and correction (ECC) is a technique used for detecting and correcting errors by adding check bits to the source bits.
EMMC	The embedded multi-media card (eMMC) is a memory card standard used for solid-state storage.
EP	The endpoint (EP) is a network adapter with components such as the PCIe interface and SATA controller.
EVB	The evaluation board (EVB) is used to verify the performance, reliability, and integration of chips.
F	
FLOPS	In computing, floating-point operations per second (FLOPS) is a measure of computer performance, useful in fields of scientific calculations that make heavy use of floating-point calculations. For such cases it is a more accurate measure than the generic instructions per second.

Framework	The framework contains the offline model generator (OMG) and offline model inference engine (OME), which are used to generate offline models and infer results based on models and data.
G	
GDB	The GNU debugger (GDB) is a command line debugging tool in UNIX and UNIX-like, which can execute programs, manage breakpoints, check values assigned to variables, and call functions.
GPU	A graphics processing unit (GPU) is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the creation of images in a frame buffer intended for output to a display device.
H	
HCCL	Huawei Collective Communication Library (HCCL) provides high-performance collective communication between servers for the training scenario in deep learning.
HDC	Host device communication (HDC) is a module deployed in both the host and device, which implements the communication between the host and device.
HDR	High dynamic range (HDR) is a photography term that describes media applications such as digital imaging and digital audio production.
HiAI Engine	HiAI Engine is a universal service process execution engine that consists of HiAI Engine Agent (running on the host side) and HiAI Engine Manager (running on the device side).
I	
I²C	The inter-integrated circuit (I ² C) bus is designed to allow easy communication between components which reside on the same circuit board.
IDE	The integrated development environment (IDE) is a software application that provides comprehensive facilities to computer programmers for software development.
IFU	The instruction fetch unit (IFU) records the information about each access to the I-cache.
IPC	The IP camera (IPC) is a type of digital video camera that receives control data and sends image data via the Internet.
IR	An intermediate representation (IR) is the data structure or code used internally by a compiler or virtual machine to represent source code. An IR is designed to be conducive for further processing, such as optimization and translation.
ISP	Image signal processing (ISP) is used to process the output signals of the front-end image sensor to match the image sensors of different vendors.
IVS	Intelligent video surveillance (IVS) is a Huawei-developed video surveillance system that integrates management, storage, encoding, decoding, intelligent video analysis, and applications.
J	
JPEGD	The JPEG decoder (JPEGD) decodes images in JPEG format.

JPEGE	The JPEG encoder (JPEGE) encodes and outputs images in JPEG format.
L	
LAN	The local area network (LAN) is a network formed by the computers and workstations within the coverage of a few square kilometers or within a single building, featuring high speed and low error rate.
LLC	The last level cache (LLC) refers to the shared highest-level cache, which is called before the memory access.
LPDDR4x	Low-Power DDR4x (LPDDR4x) is a communications standard designed for memories, which feature low power consumption and compact design and apply to mobile electronic products. LPDDR4X is identical to LPDDR4 except additional power is saved by reducing the I/O voltage (V _{ddq}) to 0.6 V from 1.1 V.
M	
MDC	The mobile data center (MDC) is used for self-driving.
Mic	The microphone (mic) is a transducer that converts sound into an electrical signal.
ML	Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead.
MLL	The machine learning library (MLL) is a mechanical learning library that greatly improves the performance of the OpenCV operator through algorithm optimization and NEON instructions.
MTE1	Memory transfer engine 1 (MTE1) copies the memory from an L1 buffer.
MTE2	Memory transfer engine 2 (MTE2) copies the memory from a DDR SDRAM or an L2 buffer.
MTE3	Memory transfer engine 3 (MTE3) copies the memory from a UB.
N	
NIC	The network interface controller (NIC, also known as a network interface card, network adapter, LAN adapter, and by similar terms) is a computer hardware component that connects a computer to a computer network.
NN	In the field of machine learning and cognitive science, the neural network is a mathematical model or computing model that simulates the structure and functions of a biological neural network.
NPU	The neural-network processing unit (NPU) uses the data-driven parallel computing architecture and is capable of efficiently processing massive video and image multimedia data. It is dedicated to processing a large number of computing tasks in artificial intelligence applications.
NV	Once data is written to a nonvolatile (NV) storage device, the data will not be lost even if the system is powered off. The original settings are still retained upon the next startup.
O	

OME	The offline model inference engine (OME) is used by a converted offline model for model load and inference.
OMG	The offline model generator (OMG) is used to convert a model trained by the framework such as Caffe/TensorFlow into an offline model supported by Huawei chips. This implements device-independent pre-processing functions such as operator scheduling optimization, weight data rearrangement, compression, and memory usage optimization.
OP	An operator is a symbol that tells the compiler to perform specific mathematical or logical manipulations. A common operator indicates the operations that include but are not limited to ReLU, CONV, FC, pooling, scale, and softmax of AI.
OTG	On-The-Go (OTG) is mainly used for the connections between different devices or mobile devices for data exchange.
P	
PCB	The printed circuit board (PCB) is a board used to mechanically support and electrically connect electronic components using conductive pathways, tracks, or traces, etched from copper sheets laminated onto a non-conductive substrate.
PCIe	Peripheral Component Interconnect Express (PCIe) is a high-speed serial computer expansion bus standard, designed to replace the older PCI, PCI-X and AGP bus standards. PCIe has numerous improvements over the older standards, including higher maximum system bus throughput, lower I/O pin count and smaller physical footprint, better performance scaling for bus devices, a more detailed error detection and reporting mechanism, and native hot-swap functionality.
PMU	The performance monitor unit (PMU) is a hardware unit provided by the CPU, which can read some performance data of the CPU by accessing related registers.
PNGD	The PNG decoder (PNGD) decodes images in PNG format.
PWM	Pulse-width modulation (PWM) is a method of reducing the average power delivered by an electrical signal, by effectively chopping it up into discrete parts.
R	
RAM	Semiconductor-based memory that can be read and written by the CPU or other hardware devices. The storage locations can be accessed in any order.
RC	In a PCI Express (PCIe) system, a root complex device connects the processor and memory subsystem to the PCI Express switch fabric composed of one or more switch devices. Similar to a host bridge in a PCI system, the root complex generates transaction requests on behalf of the processor, which is interconnected through a local bus. Root complex functionality may be implemented as a discrete device, or may be integrated with the processor.
Runtime	Runtime runs in the app process space and provides apps with functions such as memory management, device management, stream management, event management, and kernel execution for Ascend 310.
S	
Scalar	A scalar is an element of a field which is used to define a vector space. It is usually used to indicate a constant.

SDK	The software development kit (SDK or devkit) is typically a set of software development tools that allows the creation of applications for a certain software package, software framework, hardware platform, computer system, video game console, operating system, or similar development platform.
SoC	System on chip (SoC) is a key technology that lowers the cost of ENPs. Huawei integrates all packet forwarding functions of switch Line Processing Units (LPUs) to a chip, implementing flexibility at low cost.
SPI	The serial peripheral interface (SPI) is a synchronous serial communication interface specification used for short-distance communication, primarily in embedded systems.
T	
TE	TensorEngine (TE) is used to develop custom operators.
TEE	The trusted execution environment (TEE) is a secure area of a main processor. It guarantees code and data loaded inside to be protected with respect to confidentiality and integrity. A TEE as an isolated execution environment provides security features such as isolated execution, integrity of applications executing with the TEE, along with confidentiality of their assets.
Tensor	In mathematics, a tensor is a geometric object that maps in a multi-linear manner geometric vectors, scalars, and other tensors to a resulting tensor. Vectors and scalars which are often used in elementary physics and engineering applications, are considered as the simplest tensors. Vectors from the dual space of the vector space, which supplies the geometric vectors, are also included as tensors.
Tops	It is used to measure the computing capability of the CPU, GPU, and NPU.
TS	The task scheduler (TS) is used to distribute different kernels to the AI CPU or AI core for execution.
TVM	The tensor virtual machine (TVM) provides built-in operators and custom operators and supports open source frameworks such as Caffe and TensorFlow.
U	
USB	Universal Serial Bus (USB) is an industry standard that establishes specifications for cables, connectors and protocols for connection, communication and power supply between personal computers and their peripheral devices.
UART	The universal asynchronous receiver/transmitter (UART) is a chip for controlling computers and serial devices. It provides RS-232C DTE interfaces, so that computers can communicate with modems or other serial devices using RS-232C interfaces.
V	
VCM	Huawei's video content management (VCM) system leverages cloud-based intelligent big data analysis technology to offer a wide array of intelligent video analysis algorithms and pre-integrate multiple intelligent analysis functions, dramatically improving video usage in industry applications.
VDEC	The video decoder (VDEC) decodes videos in specific formats.
VENC	The video encoder (VENC) encodes videos in specific formats.

VECTOR Vector operation

VPC The vision preprocessing core (VPC) provides capabilities such as image scaling, color gamut conversion, bit count reduction, storage format conversion, and block conversion.

Y

YUV YUV is a color encoding system typically used as part of a color image pipeline. It encodes a color image or video taking human perception into account, allowing reduced bandwidth for chrominance components, thereby typically enabling transmission errors or compression artifacts to be more efficiently masked by the human perception.

C Acronyms and Abbreviations

A

AI Artificial Intelligence

N

NPU Neural-network Processing Unit

O

OME Offline model Inference Engine

OMG Offline model Generation

T

TE Tensor Engine

TVM Tensor Virtual Machine