英特尔推荐技术平台清单

1. OpenVINO™

这是一个用于优化和部署人工智能推理的开源工具套件。可以提高计算机视觉、大语言模型、生成式 AI、自动语音识别、自然语言处理和其他常见任务的深度学习性能；应用于使用 TensorFlow、PyTorch、Paddlepaddle 等广泛使用的人工智能框架训练的模型；减少算力资源需求，并在从边缘到云的一系列英特尔硬件平台上高效部署。

OpenVINO™开源版本包括几个组件：即 Model Converter、OpenVINO™ Runtime、Neural Network Compression Framework，以及 CPU、GPU、NPU、GNA、多设备和异构架构的插件，以加速英特尔 CPU、英特尔图像处理器以及 NPU 上的深度学习推理。它支持 Open Model Zoo 的预训练模型，以及众多流行格式的开源和公共模型，如 TensorFlow、ONNX、飞桨、MXNet、Caffe、Kaldi。

开源地址：https://github.com/openvinotoolkit

英特尔分发版 OpenVINO™ 工具套件地址：
https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.htm

2. oneAPI

oneAPI 是英特尔遵循开放标准推出的一个统一的软件开发套件，旨在使软件开发人员能够使用单一的代码库在不同的异构计算平台上开发应用程序。

它提供了一组标准化的应用编程接口（API），让软件开发人员在不同的计算平台上使用相同的代码开发应用程序。通过对于直接编程语言的支持及 API 函数方式支持，借助 C++/SYCL、Python 等语言，可以更好地发挥 CPU、GPU、FPGA、AI 加速器等相关硬件的性能。

oneAPI 包含了一系列领域专属的工具，适用于不同的技术领域及应用场景，助力软件开发人员在数字化转型及创新的过程中更便捷地调试、优化和部署解决方案。

开源地址： https://github.com/oneapi-src/

英特尔 oneAPI 产品及各领域专属工具套件产品信息及下载地址：
https://www.intel.com/content/www/us/en/developer/tools/oneapi/overview.html

3. Intel® LLM library for Pytorch (IPEX-LLM)

IPEX-LLM 是一个基于 Pytorch 的大语言模型推理及微调的加速库。它帮助用户在英特尔硬件平台上，利用更少的计算与存储资源，加速开发使用基于大语言模型的人工智能应用。IPEX-LLM 支持多种低精度（例如 INT4 /NF4 /FP4 /INT5 /INT8 /FP8）实现，并和大语言社区生态友好结合，从而方便用户将大语言模型纳入到各类应用中。IPEX-LLM 已经支持国内外多种模型家族（例如 llama, gptneox, bloomz，ChatGLM, Baichuan，Qwen 等），并提供了超过 50 个模型的示例代码。

开源地址：https://github.com/intel-analytics/ipex-llm

4. Python

Python 是一种广泛使用的高级和通用且开放的编程语言，常用于各种用的快速开发。英特尔 Python 分发版，有效利用台式机、笔记本及服务器处理器中的所有核心，为高性能数值和科学计算提供了更接近原生代码的性能，

开源地址：https://www.python.org/

英特尔 Python 分发版的技术信息及下载地址：
https://www.intel.com/content/www/us/en/developer/tools/oneapi/distribution-for-python.html

5. PyTorch

PyTorch 是一个开源的应用于人工智能领域等领域的框架。广泛运用于机器学习、计算机视觉和自然语言处理等场景。英特尔 PyTorch 扩展为特定场景中运用 CPU 和 GPU 等硬件进行训练和推理提供了更好的性能。

开源地址： https://github.com/intel/intel-extension-for-pytorch

英特尔 PyTorch 扩展的技术信息及下载地址：
https://www.intel.com/content/www/us/en/developer/tools/oneapi/optimization-for-pytorch.html

6. TensorFlow

TensorFlow 是一个端到端开源机器学习平台。助力研究人员推动先进机器学习技术的发展，并使开发者能够轻松地构建和部署由机器学习提供支持的应用。英特尔 TensorFlow 扩展为特定场景中运用 CPU 和 GPU 等硬件进行训练和推理提供了更好的性能。

开源地址：https://github.com/intel/intel-extension-for-tensorflow

英特尔 TensorFlow 扩展的技术信息及下载地址：
https://www.intel.com/content/www/us/en/developer/tools/oneapi/optimization-for-tensorflow.html

7. 英特尔“芯”数字化开发套件

采用英特尔赛扬处理器 ADL-N 系列，已通过 Windows, Ubuntu* Desktop 和 OpenVINO™ 工具套件的预验证，有助于在教育方面取得更多成绩。这一组合为学生提供了在 AI、视觉处理和物联网领域培养编程技能和设计解决方案原型所需的性能。

- 多任务处理性能（例如运行编码工作负载和创作数字内容），由具有四个线程的四核处理器实现。
- 通过 24 个执行单元加速英特尔超核芯显卡的 AI 性能，这些执行单元是英特尔显卡架构中针对同时多线程处理优化的计算处理器。
- 通过教室中的 Wi-Fi[1] 提供更大的容量、速度和可靠性，以满足日益增长的访问大量在线开放式课程和数字课程的需求。
- 让学生能够使用一个外形紧凑的设备（具有 40 引脚通用输入和输出接头，可连接到传感器、灯、执行器和其他设备）学习和探索所需技能。
- (个别设备可能需要补充 WiFi 连接必要的配件设备)

除独立提供 AI 算力外，兼容匹配更多高算力 Intel 设备作为边端结合的算力组合完成项目开发。

英特尔开发者专区链接：https://www.intel.cn/content/www/cn/zh/developer/topic-technology/edge-5g/hardware/nezha-dev-kit.html

(哪吒板)

英特尔开发者专区链接：https://www.intel.cn/content/www/cn/zh/developer/topic-technology/edge-5g/hardware/lan-wa-aixboard-edge-dev-kit.html

技术文档及上手案例参考：https://www.xzsteam.com/docs/

(爱克斯板)

(灵犀板)

8. 搭载英特尔® 酷睿™ Ultra 处理器的 AI PC

搭载英特尔® 酷睿™ Ultra 处理器的 AI PC 集成了 CPU、锐炫集成显卡及 NPU 三大 AI 引擎，通过利用 OpenVINO™、IPEX-LLM 等丰富的开源软件框架及优化的 oneAPI 工具链，快速进行 AI 调优及部署，开发可落地的创新应用，为充分利用大语言模型、文生图模型、多模态大模型实现自然语言处理、代码生成、音频生成、图片视频生成等能力提供加速的基础平台及能力。

了解更多：https://www.intel.cn/content/www/cn/zh/products/docs/processors/core-ultra/ai-pc.html

9. 显卡

英特尔锐炫™ A 系列显卡，皆内置硬件光线追踪、机器学习和 AV1 硬件编码加速功能。用户可以使用英特尔的 OpenVINO™，oneAPI 工具包开发的所有软件更好的通过调用 GPU 加速 AI 代码推理。

了解更多：https://www.intel.cn/content/www/cn/zh/products/docs/discrete-gpus/arc/desktop/a-series/overview.html

10. 英特尔高性能云边协同平台

英特尔® 酷睿™ 13 代：异构平台，高主频，提供 OpenVINO™模型上获得优秀性能，从而从 “端—边—云”快速实现高性能人工智能推理。Intel Xeon W7-2475X；多核心数，高主频，高拓展性，支持第三代英特尔® 深度学习加速，加速 AI 模型的训练和推理。

支持 DDR5 RDIMM 纠错码 (ECC) 内存以及可靠性、可用性和可维护性 (RAS) 功能，可防止系统错误，从而保护关键数据完整性和系统可靠性，最大限度地延长正常运行时间。

11. FPGA 云连接套件

充分结合了 Intel Cyclone® V SoC FPGA 器件丰富的功能和云连接的优势。开发者可以轻松地通过开发基于 FPGA 的应用程序来收集、分析和响应来自 IoT 设备的数据。该开发套件已通过 Microsoft Azure 等关键云服务提供商（CSP）的认证，并附带开源设计示例，便于新用户首次体验 FPGA 作为边缘设备连接到云的过程。

了解更多：http://www.terasic.com.cn/cgi-bin/page/archive.pl?Language=China&CategoryNo=180&No=1260#contents

List of Recommended Technology Platforms

1. OpenVINO™

OpenVINO™ is an open-source toolkit for optimizing and deploying AI inference.

- It can enhance the performance of deep learning for computer vision, large language models, generative AI, automatic speech recognition, natural language processing, and other common tasks.
- Use models trained with popular frameworks like TensorFlow, PyTorch and more.
- Reduce resource demands and efficiently deploy on a range of Intel® platforms from edge to cloud.

This open-source version includes several components: namely Model Optimizer, OpenVINO™ Runtime, Neural Network Compression Framework, as well as CPU, GPU, GNA, multi device and heterogeneous plugins to accelerate deep learning inference on Intel® CPUs, Intel® Processor Graphics, and Deep Learning Inference on NPU. It supports pre-trained models from Open Model Zoo, along with many open source and public models in popular formats such as TensorFlow, ONNX, PaddlePaddle, MXNet, Caffe, Kaldi.

Opensource Access: https://github.com/openvinotoolkit

Intel Distribution of OpenVINO™ Toolkit:

https://www.intel.com/content/www/us/en/developer/tools/openvino-toolkit/overview.html

2. oneAPI

oneAPI is a unified software development toolkit introduced by Intel in accordance with open standards, designed to enable software developers to use a single code base to develop applications on different heterogeneous computing platforms.

It offers the standard based Application Program Interface (API) enable software developers to develop and run the same code across different computing platforms. Through support for direct programming languages and API function methods,

leveraging languages such as C++/SYCL and Python, one can better harness the performance of hardware such as CPUs, GPUs, FPGAs, and AI accelerators.

oneAPI includes a suite of domain-specific tools that cater to various technical fields and application scenarios, facilitating software developers to more conveniently debug, optimize, and deploy solutions in the process of digital transformation and innovation.

Opensource website: https://github.com/oneapi-src/

Intel oneAPI product information and domain specific toolkit download:

https://www.intel.com/content/www/us/en/developer/tools/oneapi/overview.html

3. Intel® LLM Library for Pytorch (IPEX-LLM)

IPEX-LLM is an acceleration library for large language model (LLM) inference and fine-tuning based on Pytorch. It helps users to accelerate the development of AI applications based on large language models with less computational and storage resources on Intel hardware platforms. IPEX-LLM supports a variety of low-precision implementations (such as INT4/NF4/FP4/INT5/INT8/FP8) and is friendly with the large language community ecosystem, making it convenient for users to integrate large language models into various applications. IPEX-LLM has already supported multiple model families (such as llama, gptneox, bloomz, ChatGLM, Baichuan, Qwen, etc.) and provides example code for more than 50 models.

Opensource website: https://github.com/intel-analytics/ipex-llm

4. Python

Python is a widely used high-level general-purpose open programming language, often used for rapid development of various applications. The Intel Distribution for Python efficiently utilizes multiple cores from the processor on desktop, notebook and server while deliver the performance close to native code in high-performance numerical and scientific computing.

Opensource website: https://www.python.org/

Intel Distribution for Python product information and download:

https://www.intel.com/content/www/us/en/developer/tools/oneapi/distribution-forpython.html

5. PyTorch

PyTorch is an opensource framework for artificial intelligence. It is widely used in scenarios such as machine learning, computer vision, and natural language processing. Intel Extension for PyTorch deliver enhanced performance under specific scenarios for training and inference running on CPUs and GPUs.

Opensource website: https://github.com/intel/intel-extension-for-pytorch

Intel PyTorch extension product information and download: https://www.intel.com/content/www/us/en/developer/tools/oneapi/optimization-forpytorch.html

6. TensorFlow

TensorFlow is an end-to-end opensource machine learning and artificial intelligence platform. Empowering researchers to advance the state of art machine learning technologies and enabling developers to easily build and deploy applications. Intel Extension for TensorFlow deliver enhanced performance under specific scenarios for training and inference running on CPUs and GPUs.

Opensource website: https://github.com/intel/intel-extension-for-tensorflow

Technical information and download links for Intel TensorFlow Extension: https://www.intel.com/content/www/us/en/developer/tools/oneapi/optimization-for-tensorflow.html

7. Intel Digital Dev Kit

Powered by the Intel Celeron Processor ADL-N Series, pre-validated with Windows, Ubuntu* Desktop, and the OpenVINO™ toolkit, this kit aids in achieving better educational outcomes. This combination provides students with the performance

needed to cultivate programming skills and design solution prototypes in the fields of AI, vision processing, and the Internet of Things (IoT).

- Multitasking performance (such as running coding workloads and creating digital content) is enabled by a quad-core processor with four threads.
- AI performance of Intel's core graphics is accelerated by 24 execution units, which are computing processors optimized for simultaneous multithreading in Intel's graphics architecture.
- Greater capacity, speed, and reliability in the classroom through Wi-Fi1 to meet the growing demand for access to a large number of online open courses and digital curricula.
- Enables students to learn and explore required skills using a compact device (equipped with a 40-pin general-purpose input and output header for connecting to sensors, lights, actuators, and other devices).

(Specific devices may require additional accessories for Wi-Fi connectivity)

In addition to independently provide AI computing power, Intel Digital Dev Kit is compatible to work with more Intel high performance devices as “Edge to Device” computing combination to better develop projects.

Link to Intel Developer Zone:
https://www.intel.cn/content/www/cn/zh/developer/topic-technology/edge-5g/hardware/nezha-dev-kit.html

(NeZha Board)

Link to Intel Developer Zone:
https://www.intel.cn/content/www/cn/zh/developer/topic-technology/edge-5g/hardware/lan-wa-aixboard-edge-dev-kit.html
Technical Documentation and Quick Start Examples Reference:
https://www.xzsteam.com/docs/

(Aikesi Board)

(Lingxi Board)

8. AI PC with Intel® Core™ Ultra Processor

The AI PC equipped with the Intel® Core™ Ultra processor integrates three major AI engines: CPU, Intel integrated graphics, and NPU. By leveraging a wealth of open-source software frameworks such as OpenVINO™ and IPEX-LLM, along with the optimized oneAPI toolkit, it enables rapid AI tuning and deployment. This system facilitates the development of innovative applications that can be implemented in real-world scenarios. It provides an accelerated foundational platform and capabilities for fully utilizing large language models, text-to-image models, and multimodal large models to achieve natural language processing, code generation, audio generation, and image and video generation capabilities.

Learn more: https://www.intel.cn/content/www/cn/zh/products/docs/processors/core-ultra/ai-pc.html

9. Graphic Card

Intel® Arc™ A-Series Graphics cards come with built-in hardware ray tracing, machine learning, and AV1 hardware encoding acceleration features. Users can leverage Intel's OpenVINO™ toolkit and oneAPI to develop software that better accelerates AI code inference by utilizing the GPU.

Learn more: https://www.intel.cn/content/www/cn/zh/products/docs/discrete-gpus/arc/desktop/a-series/overview.html

10. Intel High-Performance Cloud-Edge Collaborative Platform

Intel® Core™ 13th Generation: Heterogeneous platform with high clock frequency, delivering excellent performance on OpenVINO™ models, enabling fast and high-performance AI inference from edge to cloud.

Intel Xeon W7-2475X: Multi-core processor with high core count, high clock frequency, and high scalability, supporting the 3rd generation of Intel Deep Learning Boost to accelerate AI model training and inference.

Supports DDR5 RDIMM Error-Correcting Code (ECC) memory and Reliability, Availability, and Serviceability (RAS) features, preventing system errors, protecting critical data integrity and system reliability, and maximizing normal operation time.

11. The FPGA Cloud Connectivity Kit

The FPGA Cloud Connectivity Kit combines the rich functionality of the Intel Cyclone® V SoC FPGA device with the advantages of cloud connectivity. Developers can easily collect, analyze, and respond to data from IoT devices by developing FPGA-based applications. This development kit has been certified by key cloud service providers (CSPs) such as Microsoft Azure and includes open-source design examples, making it convenient for new users to experience the process of connecting FPGA as an edge device to the cloud.

Learn more: http://www.terasic.com.cn/cgi-bin/page/archive.pl?Language=China&CategoryNo=180&No=1260#contents