TiES Revista de
Tecnología e Innovación
en Educación Superior

GPU Y NPU PARA EL DESARROLLO DE LA INTELIGENCIA ARTIFICIAL. EL CASO NPU ASCEND HUAWEI EN LINUX

DOI: 10.22201/dgtic.26832968e.2023.8.3

José Fabián Romo Zamudio (fabian.romo@unam.mx)
Universidad Nacional Autónoma de México, Dirección General de Cómputo y de Tecnologías de Información y Comunicación. Ciudad de México, México.
ORCID: 0009-0003-9269-8185

www.ties.unam.mx

Fecha de recepción: julio de 2023 • Fecha de publicación: noviembre, 2023

Noviembre 2023 | número de revista 8 • ISSN 2683-2968

Noviembre 2023 | número de revista 8 • ISSN 2683-2968
www.ties.unam.mx

GPU Y NPU PARA EL DESARROLLO DE LA INTELIGENCIA ARTIFICIAL. EL CASO NPU ASCEND HUAWEI EN LINUX

Resumen

Los rápidos avances en Inteligencia Artificial (IA) se han potenciados por el desarrollo de hardware de computadora más capaz y algoritmos innovadores. Las Unidades de Procesamiento Gráfico (GPU) emergieron como un componente esencial para la investigación y el desarrollo de la IA, otorgando un inmenso poder computacional para acelerar el entrenamiento y la inferencia de las redes neuronales. Este artículo se adentra en el uso de las GPU para la investigación y desarrollo de nuevos modelos de redes neuronales, partiendo de los esquemas más conocidos, en lo que corresponde a los proyectos del espacio de innovación para el desarrollo de habilidades digitales Alianza UNAM – Huawei, con un enfoque específico en los servidores Linux que contienen las Unidades de Procesamiento Neuronal (NPU) Ascend de la marca Huawei y sus capacidades de aceleración de los cálculos para la IA.

Palabras clave:

Arquitectura DaVinci, Contenedor, GPU, Huawei, Inteligencia Artificial, Linux, Mindspore, NPU, Pytorch, Redes Neuronales, Software libre, TensorFlow, Unidad de Procesamiento Gráfico, Unidad de Procesamiento Neuronal.

GPU AND NPU FOR ARTIFICIAL INTELLIGENCE DEVELOPMENT. THE CASE HUAWEI'S ASCEND NPU ON LINUX

Abstract

The fast advancements in artificial intelligence (AI) have been largely fueled by the development of powerful computing hardware and innovative algorithms. Graphic Processing Units (GPUs) have emerged as a vital component in AI research and development, providing immense computational power to accelerate neural network training and inference. This article delves into the usage of GPUs for AI research and the development of new neural network models, starting from the most well-known schemas regarding the projects under the space for innovation for digital capabilities development inside the UNAM-Huawei Alliance, with a specific focus on Linux-based servers and the Huawei NPU Ascend in supporting AI workloads and accelerating its computations.

Keywords:

DaVinci Architecture, Docker, GPU, Huawei, Artificial Intelligence, Linux, Mindspore, NPU, Pytorch, Neuronal Networks, Open source, TensorFlow, Graphic Processing Unit, Neuronal Processing Unit.

Noviembre 2023 | número de revista 8 • ISSN 2683-2968
www.ties.unam.mx

GPU Y NPU PARA EL DESARROLLO DE LA INTELIGENCIA ARTIFICIAL. EL CASO NPU ASCEND HUAWEI EN LINUX

1. Introducción

La Inteligencia Artificial (IA) ha revolucionado múltiples campos del saber, la producción y la educación al facilitar que los algoritmos para aprendizaje de máquina puedan abordar problemas complejos de muy diversos tipos. El crecimiento exponencial de la IA está íntimamente ligado a la disponibilidad de cómputo de alto rendimiento, capaz de manejar grandes volúmenes de datos y resolver operaciones con mayor demanda de cómputo. En este contexto, los microprocesadores conocidos como GPU[1] se han consolidado como elementos clave en el ámbito de la IA, proporcionando capacidades superiores, tanto computacionales como de procesamiento en paralelo. Adicionalmente, dos factores han incidido en ese desarrollo acelerado de la IA: la evolución de tecnologías para el aprendizaje profundo y la alta disponibilidad de grandes volúmenes de datos. Sin embargo, procesar esas cantidades de datos implica un entrenamiento de redes neuronales profundas con sustanciales recursos computacionales sustanciales, lo que de por sí es una labor matemáticamente intensa. Los microprocesadores tradicionales o CPU[2] son insuficientes para cumplir con las demandas del entrenamiento de esas redes neuronales complejas debido a su naturaleza secuencial y escalar. En consecuencia, las GPU se han convertido en el mejor soporte de hardware para el entrenamiento de modelos de aprendizaje profundo, que son la base inobjetable de la inteligencia artificial actual. Las GPU, diseñadas inicialmente para la renderización, han sido por muchos años la mejor alternativa para acelerar los cálculos computacionales que requiere la IA.

Se procurará explorar aquí el papel fundamental que las GPU han tenido en la investigación en IA y el desarrollo de nuevos modelos de redes neuronales, enfocándose en el uso de las GPU en servidores con sistema operativo Linux y equipados con la tecnología de NPU Ascend desarrollada por la empresa Huawei, con la cual la Universidad Nacional Autónoma de México estableció un convenio de colaboración desde 2020 para crear el *Espacio de Innovación en el Desarrollo de Capacidades Digitales en México*[3], a través del cual diversas instituciones mexicanas han podido ejecutar proyectos que incorporan tecnologías de IA para la atención de problemas en el país y el aporte de soluciones con un verdadero impacto social. Con base en ello, se procurará dar alguna orientación acerca del significado del uso de este tipo de tecnología de hardware en la mejora de proyectos y resultados de la inteligencia artificial, espe-

[1] GPU: Graphics Processing Unit. Unidad para Procesamiento de Gráficos.

[2] CPU: Central Processing Unit. Unidad de Procesamiento Central.

[3] Para más información sobre los alcances y objetivos de la Alianza, consultar https://alianza.unam.mx/

Noviembre 2023 | número de revista 8 • ISSN 2683-2968
www.ties.unam.mx

cialmente los aprobados dentro de los proyectos de la Alianza establecida entre la UNAM y Huawei.

2. Las GPU en la investigación de IA

2.1. Arquitectura de las GPU y el cómputo paralelo

Las unidades de procesamiento gráfico son dispositivos altamente sofisticados para la ejecución de cómputo paralelo, y han experimentado un crecimiento significativo en su uso gracias a su capacidad para acelerar tareas computacionales complejas [1]. De manera general, una GPU se compone de los siguientes elementos que están diseñados para llevar a cabo cálculos paralelos de manera altamente eficiente:

A) **Multiprocesadores de secuencias** (*Streaming Multiprocessors* o SM). representan los componentes fundamentales en la estructura de una GPU. Cada SM se compone de varias unidades de procesamiento conocidas como núcleos CUDA[4] o núcleos de sombreado. La función primordial de los SM radica en la ejecución de instrucciones y el procesamiento de datos en un entorno paralelo.

B) **Memoria global**. Representa el principal depósito de memoria en una GPU, donde se almacenan los datos accesibles por todas las SM. Aunque la Memoria Global tiene una capacidad considerable, su característica distintiva radica en una latencia relativamente elevada en comparación con otros tipos de memoria.

C) **Memoria compartida**. Se presenta como un tipo de memoria más ágil y compacta localizada en cada SM. Su finalidad reside en facilitar una comunicación eficaz y en permitir el intercambio de datos entre los diversos subprocesos o hilos que operan en un mismo bloque.

D) **Memoria de texturas y cachés**. Está optimizada para ubicar puntos en dos dimensiones y soportar el manejo de caché con el fin de tener ciclos de texturizado más eficientes. Adicionalmente, las GPU emplean diversos niveles de caché, tales como el L1 y el L2 para reducir la latencia en el acceso a la memoria.

E) **Interfaz de memoria**. Facilita la comunicación entre la GPU y la memoria del sistema. Define el ancho de banda y la latencia en las transferencias de datos con la memoria.

Por otra parte, el manejo eficiente de la memoria por las GPU es un elemento crítico para obtener resultados más precisos en tiempo considerablemente menor, a la vez que mejora los patrones de acceso a los datos y las transferencias de información. Además de las categorías de memoria global y compartida, las GPU operan con otras categorías que los investigadores en IA deben considerar al usar este tipo de procesadores [2]:

A) **Memoria constante**. Es un espacio de memoria solo de lectura que almacena datos preservados de manera idéntica a lo largo de toda la ejecución de proceso. Con ello se puede tener acceso rápido y eficiente a valores constantes frecuentemente utilizados.

B) **Memoria local**. Es el espacio de memoria reservado por subproceso o hilo, que se ubica dentro de la Memoria Global. Se emplea para almacenar variables locales y grupos de llamadas a funciones.

C) **Archivo de registros**. Cada subproceso en un GPU tiene acceso a su propio archivo de registros, que se usan para almacenar variables y valores intermedios durante el cómputo. Es crucial optimizar el uso del registro para evitar fallas de anotación y mejorar el rendimiento.

2.2. Técnicas de aceleración por GPU para redes neuronales

Cuando se trata de acelerar el procesamiento y mejora de comportamiento de redes neuronales que usan las GPU, se pueden emplear varias técnicas:

A) **Procesamiento por lotes**. Las GPU están especialmente diseñadas para procesar extensos conjuntos de datos de forma simultánea. Aumentar el tamaño de estos lotes conlleva el beneficio de mejorar el paralelismo y, por consiguiente, acelerar el tiempo de entrenamiento de manera general. No obstante, es esencial encontrar un equilibrio adecuado, ya que una elección de lote excesivamente grande podría ocasionar errores relacionados con los límites de memoria..

B) **Gestión de la memoria de la GPU**. Las redes neuronales con modelos y conjuntos de datos de gran envergadura a menudo superan la capacidad de memoria de las GPU. Para mitigar este desafío, se requieren estrategias como el modelado recortado o la cuantificación por volúmenes diferenciados, que pueden significativamente reducir la carga en la me-

[4] CUDA: Compute Unified Device Architecture. Es un tipo de plataforma de cómputo, propietaria de la empresa Nvidia, con la cual se pueden programar las GPU usando una variación del lenguaje C, llamado CUDA C.

Noviembre 2023 | número de revista 8 • ISSN 2683-2968
www.ties.unam.mx

moria. Además, la adopción de técnicas de entrenamiento de precisión mixta, como el uso de FP16, puede disminuir aún más los requisitos de memoria adicionales, manteniendo un alto rendimiento.

C) **Cómputo en paralelo**. Al utilizar técnicas de paralelización, tales como el paralelismo de datos y el paralelismo de modelos, se pueden distribuir las cargas de trabajo entre múltiples GPU o núcleos de GPU, reduciendo el tiempo de entrenamiento de la red neuronal que da solución al problema de Inteligencia artificial del que se trate. El *paralelismo de datos* implica replicar el modelo entre múltiples GPU y dividir los datos de entrada. El *paralelismo de modelos* involucra dividir en modelo entre varias GPU y con ello ejecutar los cálculos en paralelo.

D) **Librerías y esquemas de programación para las GPU**. Al utilizar librerías y esquemas de programación optimizados se pueden obtener incrementos de velocidad en el entrenamiento de la red neuronal. Librerías como *cuDNN* (*CUDA Deep Neuronal Network*) facilitan implementar operaciones de aceleración que son comunes entre diversos tipos de redes neuronales, tales como las convoluciones. Los marcos de referencia para el aprendizaje profundo como *Tensorflow* y *Pytorch* se desarrollaron en ese sentido para mejorar la operación de las GPU.

E) **Acumulación de gradientes**. En ocasiones, incluso después de aplicar técnicas de optimización, la memoria de la GPU podría seguir siendo insuficiente. Para abordar este desafío, se puede recurrir a la técnica de acumulación de gradientes. Aquí, los gradientes se calculan y acumulan tras la ejecución de pequeños lotes antes de llevar a cabo actualizaciones más significativas en los pesos del modelo. Esta estrategia posibilita el empleo posterior de lotes mucho más amplios y eficaces sin aumentar la demanda de memoria.

F) **Núcleos de tensores**. Las modernas GPU incluyen la capacidad para trabajar con hardware especializado denominado "*núcleo de tensor*" que pueden acelerar las operaciones de matrices con precisión mixta. Estos núcleos de tensores tienen una mayor capacidad de salida de datos, combinada con una menor precisión matemática, lo cual deriva en entrenamientos e inferencias de los modelos más rápidos, sin sacrificar demasiado a la precisión.

G) **Entrenamiento distribuido**. Cuando se trabaja con modelos o conjuntos de datos demasiado grandes, se puede utilizar el entrenamiento distribuido para usar múltiples GPU ubicados en varios nodos de cómputo. Tecnologías como *Horovod*, la Estrategia Distribuida de *TensorFlow* o la Capacidad de Paralelismo de Datos de *Pytorch* permiten el entrenamiento distribuido de forma eficiente, a la vez que se reduce el tiempo total de entrenamiento del modelo de red neuronal que sustente el proyecto de inteligencia artificial.

La eficacia de las técnicas mencionadas, previamente detalladas a los participantes de los proyectos aprobados en el marco de la Alianza UNAM – Huawei, está influida por diversos factores. Estos incluyen la arquitectura del modelo de red neuronal, el tamaño del conjunto de datos utilizado para nutrir el modelo y la configuración específica de los servidores dentro del Laboratorio del Espacio de Innovación. En las dos convocatorias publicadas hasta 2023, la mayoría de los líderes de proyecto optaron por configurar sus entornos de operación como contenedores (Docker). Esta elección les proporciona un mayor control sobre múltiples variables de la plataforma durante el entrenamiento de sus modelos de inteligencia artificial, sin depender exclusivamente de las configuraciones genéricas de los nodos de cómputo en el Laboratorio del Espacio de Innovación.

2.3. Ventajas de usar GPU en la investigación de IA

El uso de las unidades de procesamiento gráfico en la investigación en inteligencia artificial tiene diversas ventajas [3]. Entre las más importantes están:

A) **Capacidad de procesamiento en paralelo**. Las GPU pueden acelerar significativamente el entrenamiento y la inferencia, comparados con las CPU tradicionales.

B) **Tiempos de entrenamiento más rápidos**. Los modelos de aprendizaje profundo requieren, en ocasiones, un entrenamiento extenso con grandes volúmenes de datos. Las GPU permiten un entrenamiento más rápido. Este incremento en la velocidad permite a los investigadores iterar de forma más rápida sus modelos y experimentar con diferentes estructuras.

C) **Complejidad extendida de los modelos**. Los investigadores pueden entrenar modelos con millones o miles de millones de parámetros, permitiéndoles resolver los retos que imponen las tareas de inteligencia artificial, con un rendimiento superior al promedio que se alcanza con las tradicionales CPU.

D) **Técnicas para la optimización de los modelos.** Técnicas como las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN) han demostrado un notable éxito en la manipulación de datos de imágenes y secuencias de datos, respectivamente. Gracias a la eficacia de las GPU, estas estrategias de optimización han propiciado una auténtica revolución en la investigación de la inteligencia artificial, influyendo de manera significativa en campos como la visión por computadora, el procesamiento del lenguaje natural y el reconocimiento de voz.

E) **Procesamiento de datos a gran escala**. La investigación en inteligencia artificial a menudo implica el procesamiento de ingentes cantidades de datos, que pueden variar desde terabytes (TB) hasta petabytes (PB) en magnitud. Las GPU son capaces de afrontar con éxito las demandas computacionales asociadas con la tarea de procesar y analizar tales volúmenes de información. Esta capacidad se torna especialmente valiosa en aplicaciones que involucran redes neuronales de aprendizaje profundo, como el análisis de grandes conjuntos de imágenes o volúmenes extensos de texto.

F) **Disponibilidad y eficiencia en costos**. Las GPU se encuentran ampliamente disponibles y son de fácil acceso para investigadores, presentando costos significativamente más bajos en comparación con hardware especializado, como las TPU (Unidades de Procesamiento de Tensores), en muchas de las tareas de investigación en inteligencia artificial. Además, debido al aumento en la demanda de GPU en diversas áreas, como los videojuegos y la minería de criptomonedas, la disponibilidad y los precios de las GPU se han vuelto más asequibles, convirtiéndolas en una elección cada vez más popular entre los investigadores en inteligencia artificial.

3. Servidores basados en Linux para procesamiento de IA

3.1. El software libre en la IA

El software de fuente libre (*open-source*) ha tenido un impacto significativo en el desarrollo de la inteligencia artificial, al establecer un fundamento sólido que propicia la colaboración y la innovación. Algunos de los elementos más relevantes que sustentan el impacto de del software libre en la IA son los siguientes:

A) **Flexibilidad y personalización**. El código abierto garantiza la flexibilidad y la capacidad de personalización, lo que habilita a los investigadores y desarrolladores en el campo de la inteligencia artificial a adaptar los entornos de cómputo según sus necesidades particulares.

B) **Accesibilidad y disponibilidad**. Los populares esquemas de programación en inteligencia artificial, como *Tensorflow* y *PyTorch*, son de código abierto, lo que significa que cualquiera puede utilizar, modificar y contribuir a su evolución. Esta accesibilidad ha democratizado la investigación y el desarrollo en inteligencia artificial, al mismo tiempo que ha fomentado la colaboración y el intercambio de conocimientos entre la comunidad académica y la industria.

C) **Comunidad de colaboración e innovación**. Numerosos proyectos de código abierto en el campo de la inteligencia artificial han progresado gracias a la colaboración intrínseca en la comunidad que los engendra. En este entorno, desarrolladores e investigadores se han unido, contribuyendo con código, documentación y mejoras a los entornos de programación. Este esfuerzo colectivo del software libre ha propulsado modelos altamente innovadores, algoritmos, redes neuronales y técnicas más allá de las limitaciones convencionales de la inteligencia artificial.

D) **Transparencia y confianza**. El software libre fomenta la transparencia y la confiabilidad en los sistemas de inteligencia artificial. Al disponer del acceso al código fuente, investigadores y desarrolladores tienen la capacidad de examinar, verificar y validar los algoritmos y modelos empleados en las aplicaciones. Esto les permite identificar y abordar cualquier desviación, vulnerabilidad o preocupación ética, además de realizar auditorías independientes y revisiones por parte de expertos, asegurando así que los sistemas de inteligencia artificial sean responsables y confiables.

E) **Adopción y estandarización**. Los ampliamente utilizados entornos de programación en el ámbito de la inteligencia artificial, como *Tensorflow* y *PyTorch*, han adquirido un estatus de facto como estándares en la industria. Estos proporcionan una plataforma unificada para el desarrollo de la inteligencia artificial, lo que a su vez fomenta la interoperabilidad y la colaboración en este campo.

F) **Integración con aceleradores de hardware**. Los esquemas de desarrollo de código abierto brindan respaldo y optimización para aceleradores de hardware, como las GPU y los NPU, permitiendo a investigadores y desarrolladores aprovechar de manera eficaz las capacidades de paralelismo de este tipo de hardware.

3.1. Linux en el desarrollo de la IA

Linux ha tenido un impacto crucial en la evolución de la inteligencia artificial, convirtiéndose en una elección destacada entre investigadores y desarrolladores en este ámbito. Entre las ventajas principales de Linux en el campo de la IA, destacan las siguientes:

A) **Fuente abierta**. El código fuente de Linux se encuentra libremente disponible para los usuarios, lo que les permite examinarlo, modificarlo y distribuirlo a su conveniencia. Esta naturaleza de código abierto de Linux promueve la colaboración y la transparencia, lo que, a su vez, concede a los investigadores en inteligencia artificial la capacidad de personalizar y ajustar los entornos de cómputo a sus necesidades específicas. Esto incluye la posibilidad de modificar el núcleo del sistema operativo Linux (*kernel*) para optimizar las configuraciones del sistema y la integración más precisa de esquemas de programación y herramientas para la inteligencia artificial.

B) **Flexibilidad y personalización**. Linux ofrece soporte para una amplia gama de arquitecturas de hardware, lo que resulta especialmente beneficioso para los desarrolladores, ya que les permite seleccionar los componentes que mejor se adapten a sus sistemas, optimizando así el manejo de las considerables cargas de trabajo relacionadas con la inteligencia artificial. El diseño modular de Linux posibilita a los usuarios la elección e instalación únicamente de los componentes necesarios, lo que se traduce en un desarrollo más eficiente y ágil.

C) **Plataforma robusta y estable**. Linux es renombrado por su estabilidad, fiabilidad y robustez. Gracias a la excelencia del núcleo de Linux en la gestión de la memoria, la programación de procesos y las operaciones de entrada y salida, se logra una ejecución notablemente eficiente y precisa de todas las cargas de trabajo relacionadas con la inteligencia artificial.

D) **Ecosistema más extens**o. Existe un amplio espectro de herramientas de código abierto, como bibliotecas, marcos de programación y entornos de desarrollo, que desempeñan un papel fundamental en la evolución e investigación en el campo de la inteligencia artificial. Los entornos de inteligencia artificial más populares, como *Tensorflow* y *PyTorch*, como se mencionó previamente, están optimizados para su funcionamiento en sistemas Linux. Esto capacita a los desarrolladores para crear algoritmos de aprendizaje automático de última generación, diseñar arquitecturas de redes neuronales innovadoras y aprovechar modelos previamente entrenados. Además, las distribuciones de Linux facilitan la instalación y gestión de todo el software relacionado con aceleradores y otros componentes clave para la innovación en modelos de inteligencia artificial, gracias a sus manejadores de paquetes y repositorios.

E) **Rendimiento y escalabilidad**. Linux se distingue por su excepcional rendimiento y escalabilidad, permitiendo aprovechar al máximo las capacidades del hardware moderno. Esto abarca desde CPU de múltiples núcleos y GPU hasta las NPU desarrolladas por Huawei, para garantizar el rendimiento óptimo en cargas de trabajo de inteligencia artificial. Asimismo, Linux ofrece una gestión eficiente de la memoria y la comunicación entre procesos, lo que capacita a los desarrolladores de inteligencia artificial para explotar plenamente los recursos computacionales disponibles y alcanzar los niveles de rendimiento más adecuados.

3.3. Distribuciones de Linux para el computo con GPUs

Están verificadas varias distribuciones de Linux para operar con GPU y NPU. En el caso particular del Laboratorio de Inteligencia Artificial UNAM – Huawei se usa Ubuntu versión 18.04, debido a que los controladores de las tarjetas con los procesadores NPU Ascend están validados en su funcionamiento con esa distribución y versión. Sin embargo, en el ámbito de Linux para el desarrollo de la IA, existen varias distribuciones con grados diversos de efectividad y compatibilidad con los circuitos aceleradores:

A) **Ubuntu**. Está ampliamente recomendada para el desarrollo de la inteligencia artificial. Se puede tener una plataforma estable y confiable para la mayoría de las cargas de trabajo, compatible con los esquemas de programación de este ámbito, tales como *Tensorflow*,

Pytorch y *Scikit Learn*. Ubuntu dispone de ediciones especializadas como *Ubuntu Studio*[5] y *Ubuntu Robotics*[6], más enfocadas en satisfacer ciertas necesidades específicas para la inteligencia artificial.

B) **CentOS**. Basada en el código fuente de Red Hat Enterprise Linux (RHEL), es una plataforma estable de nivel empresarial para el desarrollo de la inteligencia artificial. También tiene versiones con Soporte a Largo Plazo (LTS) que incluyen actualizaciones de seguridad, haciéndolo adecuado para el desarrollo de aplicaciones de inteligencia artificial en ambientes de producción. CentOS se enfoca fuertemente en la estabilidad y la confiabilidad, asegurando un rendimiento consistente, siendo compatible con la mayoría de los esquemas de desarrollo.

C) **Fedora**. Distribución administrada por la comunidad y conocida por sus características de última generación con actualizaciones rápidas. En esencia tiene un balance adecuado entre la estabilidad y los paquetes de última generación, haciéndola recomendable también para el desarrollo de proyectos de inteligencia artificial. Fedora incluye un conjunto de herramientas de desarrollo y librerías para *Tensorflow*, *Pytorch* y *Caffe*. Más aún: contiene opciones para el uso de contenedores del tipo docker y kubernetes, muy recomendables para desarrolladores que requieren construir y ampliar las aplicaciones en este ámbito de operación. Varias investigaciones en el Espacio de Innovación tomaron como referencia para el funcionamiento de contenedores las plantillas creadas en Fedora de acceso libre.

D) **Arch Linux**. Es una versión preferida por usuarios avanzados y desarrolladores. Tiene una base mínima para su instalación, lo que permite a los usuarios personalizar el ambiente de desarrollo para inteligencia artificial de acuerdo con sus necesidades específicas. También permite usar el repositorio para usuarios Arch (AUR) que contiene un rango amplio de paquetes de software para inteligencia artificial, mantenidos y desarrollados por la comunidad.

E) **Deepin**. Esta es una distribución atractiva visualmente y muy amigable con el usuario, que se enfoca en proporcionar una experiencia bastante sencilla para quien lo emplea. Ofrece un entorno de escritorio adecuado para desarrolladores de inteligencia artificial que buscan emplear interfaces simples para el trabajo con las redes neuronales. Deepin incluye una cantidad importante de esquemas de desarrollo pre-instalados para inteligencia artificial, entre los que están *TensorFlow* y *Caffe*, haciéndolo conveniente para el desarrollo de inteligencia artificial prácticamente desde que se instala.

4. Unidades de Procesamiento Neuronal (NPU) Ascend de Huawei

4.1. Características generales de las NPUs Ascend

Las NPU Ascend de Huawei representan componentes de hardware altamente especializados diseñados para potenciar el procesamiento de cálculos en inteligencia artificial y mejorar el rendimiento de dispositivos Huawei, con especial enfoque en teléfonos inteligentes y otros dispositivos móviles [4]. Algunas de las características clave de estos procesadores incluyen:

A) **Aceleración interconstruida para IA**. Los procesadores tradicionales no se concibieron para abordar de manera eficaz las exigencias computacionales de la inteligencia artificial. Por esta razón, Huawei ideó las NPU como aceleradores de hardware especializados, diseñados específicamente para ejecutar cálculos relacionados con la inteligencia artificial de manera altamente eficiente. Estas NPU se han integrado en la solución Kirin, que es un SoC[7].

B) **Arquitectura y diseño**. Aunque los detalles precisos acerca de los componentes que integran los procesadores de Huawei están protegidos por derechos de autor y son de naturaleza propietaria, se sabe que están especialmente orientados al procesamiento avanzado de datos. Esto se logra mediante la combinación de estructuras de memoria, algoritmos especializados adaptados al hardware y unidades de procesamiento que comparten similitudes con las GPU en su capacidad para manejar cargas de cálculo matemático. Los algoritmos desempeñan un papel esencial en la interacción con el hardware, particularmente en la resolución de redes neuronales de aprendizaje profundo. Por tanto, el diseño está centrado en mejorar el rendimiento, reducir el consumo de energía y optimizar la eficiencia en general.

[5] https://ubuntustudio.org/

[6] https://wiki.ros.org/ROS/Tutorials

[7] SoC: System on a Chip. Corresponde a un circuito avanzado que incorpora en una sola tarjeta los elementos de una computadora.

C) **Desarrollo y optimización**. El desarrollo de las NPU implica una sinergia entre el diseño de hardware, desarrollo de software y la optimización de componentes, una combinación considerada de manera integral por los ingenieros de Huawei. Un componente esencial de las NPU es su capacidad para gestionar operaciones matriciales, que resulta significativamente más eficiente en comparación con las GPUs tradicionales. Esto se debe a que las NPU están especializadas en resolver problemas relacionados con redes neuronales, lo que implica una disposición tridimensional de cálculos y, como resultado, una reducción en la cantidad de ciclos necesarios para obtener los resultados en comparación con otras arquitecturas de hardware.

D) **Entrenamiento e inferencia**. Las NPU de Huawei son versátiles y admiten tanto el entrenamiento como la inferencia de modelos de inteligencia artificial. Durante la fase de entrenamiento, estas NPU aceleran las tareas que requieren un intenso poder de cómputo, lo que incluye la propagación hacia adelante y hacia atrás en el conjunto de datos, la actualización de los pesos específicos de la información y la optimización continua de los algoritmos que conforman las redes neuronales. En la etapa de inferencia, donde el modelo se especializa en predecir resultados basados en nuevos datos, las NPU de Huawei son capaces de ejecutar tareas de inteligencia artificial en tiempo real. Esto se aplica tanto en los servidores que componen el Laboratorio del Espacio de Innovación como en dispositivos periféricos.

E) **Evolución y avances**. A lo largo de los últimos años, Huawei ha realizado numerosos avances en el desarrollo de las NPU. En cada iteración del hardware, se ha trabajado en la mejora del rendimiento, la reducción del consumo de energía y la ampliación del soporte de software, al mismo tiempo que se han optimizado los algoritmos de inteligencia artificial para enriquecer la investigación respaldada por las NPU.

4.2. Arquitectura y capacidades

La arquitectura empleada en las NPU que soportan los proyectos de la Alianza UNAM – Huawei se denomina DaVinci, y es una de las más avanzadas en su tipo para el desarrollo y aceleración de los cálculos que requiere la Inteligencia Artificial, desde los teléfonos inteligentes hasta los centros de datos. Desarrollada por el segmento HiSilicon de Huawei, la arquitectura DaVinci de las NPU incluye varias características que la distinguen de las GPU y CPU ya descritas previamente en este artículo:

A) **Alto rendimiento**. Las NPU DaVinci tienen amplias capacidades de procesamiento, lo que se estima mejor para la IA, ya que facilita el aprendizaje profundo y el entrenamiento de las redes neuronales a través del aprendizaje de máquina.

B) **Versatilidad**. Esta arquitectura soporta un amplio rango de modelos de IA y marcos de desarrollo, lo que asegura la compatibilidad con los esquemas más populares (TensorFlow, Pytorch, Caffe, Keras). De esa manera los proyectos que se desarrollan en el espacio de innovación se pueden migrar de manera muy simple a otras plataformas de hardware y de software para extender las capas de entrenamiento e inferencia y con ello robustecer los modelos de IA que se están investigando.

C) **Eficiencia energética**. Las NPU consumen lo mínimo de energía necesaria para maximizar el rendimiento por watt, lo que es adecuado no solo para el entrenamiento de grandes modelos de IA que procesan volúmenes monumentales de datos, sino también para que los modelos puedan ejecutarse en dispositivos más pequeños donde la capacidad y eficiencia energéticas son claves, tales como teléfonos inteligentes, cámaras y tabletas que componen parte del universo de la Internet de las Cosas.

D) **Escalabilidad**. Esta arquitectura está diseñada para ampliarse en función del dispositivo y los escenarios de utilización. Ya sea en teléfonos, cámaras o centros de datos, la arquitectura se reduce o amplía conforme sea necesario, sin que se requiera una reprogramación de los algoritmos.

E) **Aceleración de redes neuronales**. Se deja en la capacidad de las unidades especializadas de hardware y de los algoritmos avanzados la aceleración de las operaciones en la red neuronal, en ámbitos como las multiplicaciones matriciales o las convoluciones. Estas capacidades de aceleración incrementan la velocidad en las fases del entrenamiento y la inferencia de los modelos de IA que se aprobaron en el marco del Espacio de Innovación.

F) **Flexibilidad en el procesamiento neuronal**. Las NPU de Huawei están diseñadas para proporcionar cálculos de punto fijo y de punto flotante, lo que permite procesar múltiples tipos de datos. Esta adapta-

bilidad permite una utilización más eficiente de los recursos físicos mientras se atienden los requerimientos de las tareas de IA.

G) **Inteligencia Artificial "en el dispositivo"**. Mucho del objetivo de las NPU reside en que se puedan ejecutar realmente algoritmos de IA en cualquier dispositivo, sin depender todo el tiempo de la conectividad a grandes centros de datos o los recursos en nube pública, a la vez que se protege la privacidad e integridad de los datos recolectados localmente.

5. Conclusiones

La creación del espacio de innovación dentro de la Alianza UNAM - Huawei ha permitido el uso de componentes especializados como lo son las unidades de procesamiento neuronal (NPU), las cuales son paso evolutivo desde las GPU. Los algoritmos de inteligencia artificial actuales son altamente demandantes de ciclos de procesamiento, especialmente del tipo matricial, para lo cual fueron diseñadas las NPU que se incorporan en los nodos de cómputo del Laboratorio que sustenta el Espacio de Innovación. Los proyectos que fueron aprobados por las convocatorias de la Alianza han explotado estas capacidades de cálculo considerando elementos comunes a otros otras plataformas de investigación como lo son los esquemas de desarrollo (todos ellos apoyados en el lenguaje Python, el cual se ha convertido prácticamente en el estándar de facto para toda la innovación en inteligencia artificial) así también en redes neuronales que están disponibles a través de los repositorios con los que cuenta Huawei, desde donde se pueden descargar modelos de estas redes para adaptarlos a las necesidades específicas de los investigadores. Esta combinación de recursos de fuente abierta con tecnologías de hardware innovador ha impulsado el desarrollo de los proyectos y no solamente para la solución de problemas con enfoque social, como ha sido el objetivo de las convocatorias, sino también para la formación de recursos humanos especializados que han aprendido a utilizar de la mejor forma posible estas herramientas físicas y lógicas.

BIBLIOGRAFÍA

[1] Nickolls, J., Buck, I., Garland, M., & Skadron, K. Scalable parallel programming with CUDA. *ACM Queue*, Vol. 6 issue 2. pp. 40-53. 2008. [En línea]. Disponible en https://queue.acm.org/detail.cfm?id=1365500

[2]. Harris, M. J. Parallel prefix sum (scan) with CUDA. *NVIDIA Developer Technology*, Vol . 3 Issue 3, 1-8. 2007. [En línea]. Disponible en https://developer.nvidia.com/gpugems/gpugems3/part-vi-gpu-computing/chapter-39-parallel-prefix-sum-scan-cuda

[3]. Owens, J. D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A. E., & Purcell, T. J. A survey of general-purpose computation on graphics hardware. *Computer Graphics Forum*, Vol. 26 Issue 1, 80-113. 2007. [En línea]. Disponible en https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2189271fd54b6980b2b20ef515ac114a02854219

[4]. Huawei Technologies Co., Ltd. *Huawei Atlas AI Computing Solution*. Online ISBN 978-981-19-2879-6. 2022. https://doi.org/10.1007/978-981-19-2879-6_6

Fecha de recepción: julio, 2023
Fecha de publicación: noviembre de 2023

Cómo se cita:
J. F. Romo, "GPU y NPU para el desarrollo de la inteligencia artificial. El caso NPU Ascend Huawei en Linux," *TIES, Revista de Tecnología e Innovación en Educación Superior*, no. 8, noviembre, 2023. [En línea]. Disponible en: https://ties.unam.mx/ [Consultado en mes día, año].