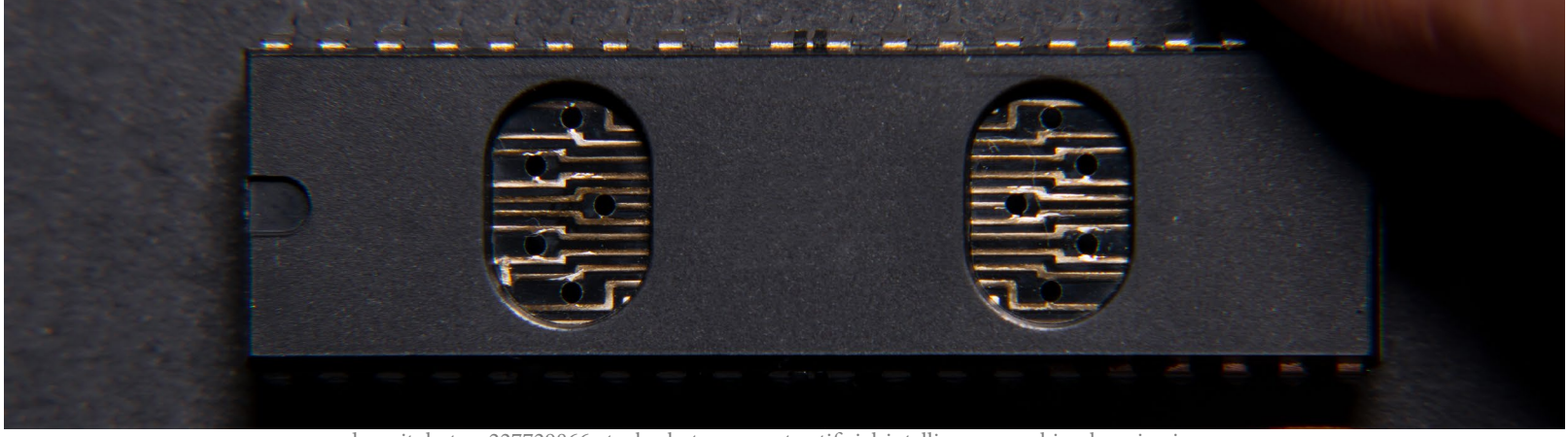


depositphotos_227729866-stock-photo-concept-artificial-intelligence-machine-learning.jpg

Entrenamiento a bajo costo de Redes Neuronales Artificiales mediante GPU y TPU, análisis exploratorio de hardware

Eduardo Berra Villaseñor
Mónica Pérez Castañeda
José Rodrigo Cuautle Parodi

Introducción

El desarrollo de la Inteligencia Artificial (IA) es un hito que marca un cambio de paradigma significativo en los procesos de creación de algoritmos para *deep learning* (aprendizaje profundo) (Basogain, 2005). La parte esencial de este proceso radica en entrenar redes neuronales artificiales, donde el reconocimiento de imágenes y objetos en tiempo real es necesario para que la toma de decisiones quede a cargo de la máquina. De igual manera, el hardware utilizado para el procesamiento de estos algoritmos tiene una alta exigencia de rapidez de procesamiento, el cual ha tenido una evolución en las capacidades y costos, tal es el caso de las tarjetas GPU (Graphics Processing Unit) de las marcas Ati y Nvidia.

En el presente trabajo se realiza un análisis exploratorio de las tendencias sobre el hardware utilizado para entrenamiento de redes neuronales artificiales, enfocadas a deep learning, obteniendo una comparación entre estas opciones que permitirán tomar una decisión entre las herramientas disponibles en el mercado, con la finalidad de experimentar sobre redes neuronales artificiales a bajo costo y con las tecnologías más apropiadas para su uso.

Metodología

Se analizarán tres tecnologías de las principales marcas de hardware que pueden ser utilizadas para ser enfocadas a entrenamiento de redes neuronales artificiales a bajo costo y dos especializadas

en aceleración de entrenamiento para redes neuronales artificiales, tres GPU y dos TPU (Tensor Processing Unit), tomando en cuenta los siguientes factores: El desempeño del hardware en procesos paralelos, arquitectura de procesamiento, capacidad de cálculos de procesamiento y costo.

Las pruebas de rendimiento serán tomadas de diversas fuentes involucradas en el análisis de rendimiento de procesadores GPU, de igual forma, la TPU NVIDIA Jetson Nano, tecnologías utilizadas por su alto desempeño en el procesamiento de información en paralelo, haciendo notar que la tecnología de Movidius será testeada con un algoritmo similar a los ejecutados en las GPU ya que se cuenta con una unidad para pruebas. Por último, se mostrará una tabla comparativa de las cinco tecnologías y su desempeño.

Objetivos

Desarrollar un estudio comparativo sobre tarjetas GPU y TPU, así como recabar las características tomadas del datasheet del fabricante del hardware que se encuentran en el mercado para entrenamiento de redes neuronales artificiales.

Mediante las gráficas correspondientes a dos variables de desempeño (Velocidad de entrenamiento muestras por segundo y Gflops reales por segundo), se realizará una tabla comparativa entre las mismas tecnologías.

Definir los requerimientos mínimos para realizar deep learning entrenando redes neuronales artificiales con hardware a bajo costo.

Planteamientos

La tecnología y los sistemas que hoy en día se desarrollan distan, por mucho, con respecto a la programación que se realizaba antaño. El cambio de paradigma se está dando desde el año 2000 en forma acelerada, con la democratización de la IA y el deep learning; el software que se requiere desarrollar con lleva la toma de decisiones y el entendimiento del mundo, esto no solo como un objeto, más bien, como *un todo* sobre el cual hay que accionar.

El denominado deep learning es un campo de Inteligencia Artificial (IA) que permite que las computadoras aprendan con experiencia y entiendan el mundo en términos de jerarquía de conceptos, con cada concepto definido por su relación con conceptos más simples (Florencio, et al., 2019). Sin embargo, las redes neuronales artificiales, algoritmos bases para la IA, tienen la dificultad de requerir un proceso de entrenamiento tardado y que generalmente exige muchos recursos de procesamiento que, a su vez,

consumen tiempo en demasía (Guo, et al., 2019).

El entrenamiento de las redes neuronales es un proceso que puede variar dependiendo del problema tratado, dichos problemas pueden ir desde la identificación y clasificación de objetos hasta la detección de enfermedades analizando síntomas y padecimientos. Básicamente el desarrollo de una red neuronal artificial estándar consta de los siguientes puntos:

1. Diseño de un modelo de red neuronal artificial que puede caer en los siguientes tipos:

a) Perceptron Simple, b) Red de Hopfield, c) Perceptrón Multicapa, d) Competitiva Simple, e) Online ART1, f) competitivas ART2 y g) Autoorganizadas-*Mapas de Kohonen.*

2. Entrenamiento de la red neuronal artificial el cual consta de las siguientes etapas:

- Recabar la base que la red neuronal artificial analizará.
- Definir los parámetros sobre los que la red neuronal artificial accionará: imágenes, videos o información (por ejemplo, para identificar peces mediante imágenes las cuales son representativas de la especie).
- Ejecutar un conjunto de iteraciones con los datos, alimentando al software que clasificará o actuará basado en el modelo establecido previamente.
- Por último, maximizar las interacciones para obtener un modelo más refinado.

Este consumo exagerado que provoca el procesamiento de las interacciones de entrenamiento de la red neuronal artificial, ha llevado a la utilización de hardware que permita acelerar las iteraciones matemáticas, e incluso, a usar hardware previamente desarrollado para otros propósitos de aceleración. Un ejemplo de uso para aceleración de cálculos matemáticos son las GPU que se utilizan para acelerar las ordenes de tratamiento de video que requieren muchos cálculos matemáticos (Kayid, 2018). Para obtener modelos más rápidos y precisos se exige una agilidad en el entrenamiento de los mismos, por lo que el hardware que no fue pensado para entrenar y generar redes neuronales artificiales, puede acelerar los procesos computacionales debido a su arquitectura, estos a su vez se tornan como una opción más de procesamiento ágil.

En contraparte, existen los procesadores TPU pensados específicamente para generar y maximizar el procesamiento requerido por las interacciones, tal es el caso de procesadores como Movidius, que se especializa en el entrenamiento de las redes neuronales artificiales, permitiendo una amplia gama de aplicaciones, desde la navegación interior basada en cámara de video hasta el escaneo 3D (Moloney, et al., 2014). En la Figura 1 se muestra físicamente una GPU y una TPU.

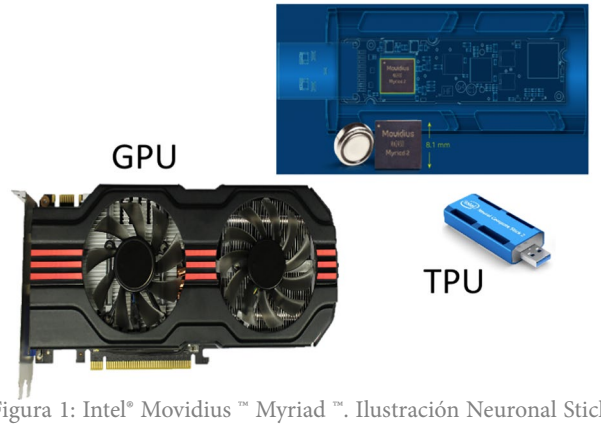


Figura 1: Intel® Movidius ™ Myriad ™. Ilustración Neuronal Stick 2, GPU ATI. Recuperado de https://www.movidius.com/myriad2

Resultados

Se seleccionaron tres tarjetas gráficas de los siguientes modelos GeForce GTX 1070, AMD Radeon RX 580, HD Graphics 4400. Todas las tarjetas GPU tienen una capacidad de 8Gb de RAM gddr5. En cuanto a los TPU se seleccionaron las siguientes Movidius neuronalstick 2 de Intel y NVIDIA Jetson Nano basados en el costo similar a las de las GPU.

Según las características del fabricante, se obtuvieron los siguientes datos, para cada modelo de GPU como se puede observar en la Tabla 1.

Tabla 1.

Modelo	GeForce GTX 1070	Radeon RX 580	Graphics 4400
Memoria Gddr5	8 Gb	8Gb	6,8 Gb
Procesador	GP104	AMD 580	Haswell GT2
Nanómetros de fabricación	16 nm	28 nm	22 nm
Consumo de energía	150 W	60 W	20 W
Punto Flotante	6,738 Gflops	1,389 Gflops	4,6 Gflops
Número de los transitores	7,200 Millones	1,870 millones	392 millones
Bus	PCIe 3.0	PCIe 3.0	PCIe 3.0
Precio	US$389.00	US$401.89	US$425.32

Datos obtenidos de datasheet de cada proveedor. Fuente: Elaboración propia.

Como se puede observar, la GPU Intel HD Graphics 4400 puede tener una variante de memoria de 6Gb, para este estudio las características y datos de rendimiento se tomaron con la variante de 8Gb.

Las GPU son contempladas bajo el nivel de tecnología y precios adecuados al costo de equipos de cómputo convencionales, ya que existen versiones de gama ultra alta que superan el 300% en costo, más no existe una diferencia significativa en cuanto al procesamiento de entrenamiento para redes neuronales artificiales. La Figura 2 muestra físicamente estos equipos.

Figura 2. GPUs seleccionadas. Fuente: NCG proveedores.

En su artículo de test sobre entrenamiento de redes neuronales artificiales, Moloti Nakampe utiliza un modelo GroceryNet, que es un modelo CaffeNet, una red neuronal artificial más refinada para realizar las pruebas sobre GPUs la cual es una réplica de AlexNet. CaffeNet tiene una ligera ventaja computacional para AlexNet en el procesamiento (Nakampe, 2018), los resultados obtenidos en las GPUs en el entrenamiento de esta red neuronal artificial son los siguientes:

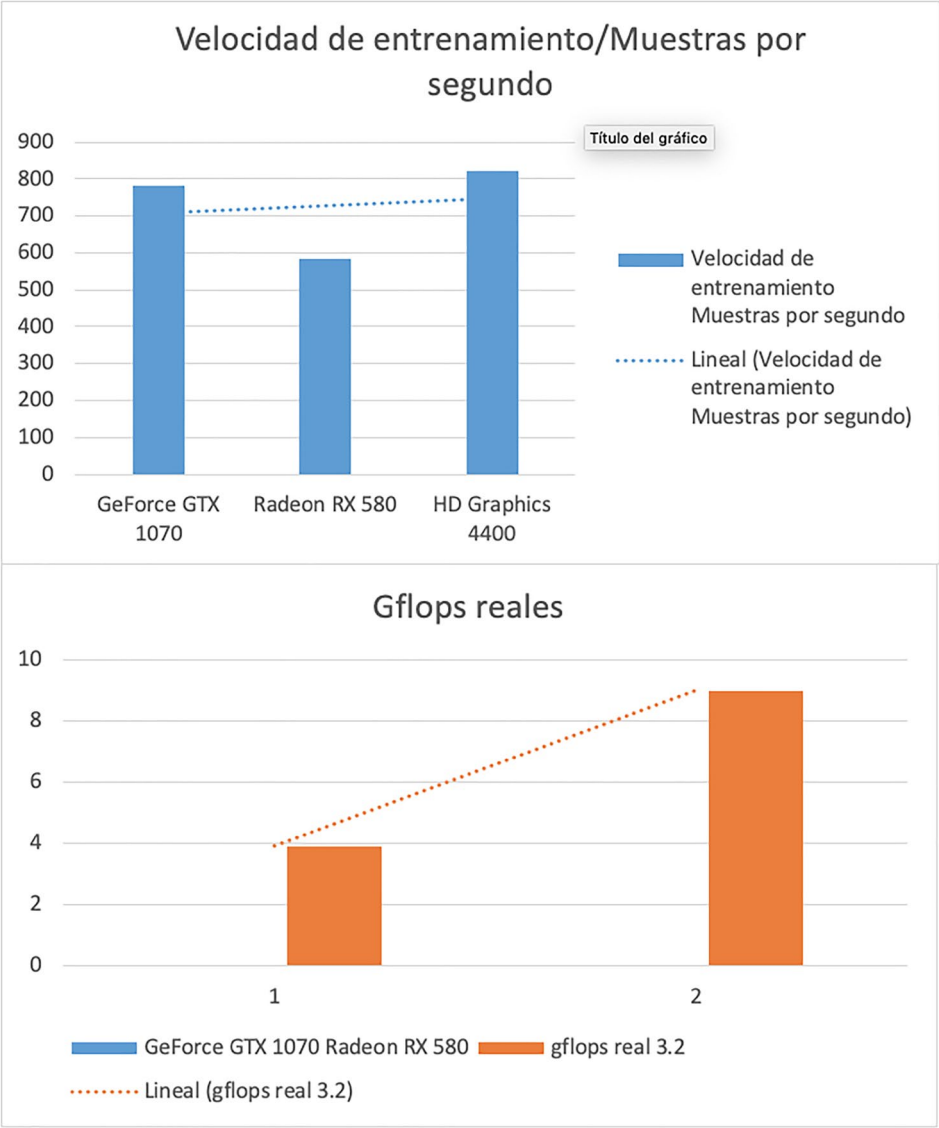


Figura 3. Resultados de prueba a GPUs datos de modelo GroceryNet. Fuente: Elaboración propia con datos de Nakampe, M. (2018). *GroceryNet at the Edge using Intel Movidius Neural Compute Stick.*

En el caso de las TPUs según las características del fabricante se obtuvieron los siguientes datos los cuales se muestran en la Tabla 2:

Tabla 2.

Modelo	Intel Neutronal Stick 2 Movidius	NVIDIA Jetson Nano
Memoria LPDDR3	4 Gb	4Gb
Procesador	Movidius 2485	NVIDIA Maxwell
Nanómetros de fabricación	28 nm	14 nm
Consumo de energía	0.9 W	5 W
Punto Flotante	100 Gflops	72 Gflops
Número de los transitores	8.50 Billones de Transistores	9 Billones de Transistores
Bus	USB 3.0	-----
Precio	US$100	US$299

Datos obtenidos de datasheet de cada proveedor. Fuente: Elaboración propia.

Como se puede observar, el consumo en Watts es inferior a los GPU, en el caso del chip Movidius claramente se aprecia el bajo consumo de menos de 1W. Además, el subsistema de memoria controlada por software permite el control detallado de diferentes cargas de trabajo si fuese necesario (Rivas-Gomez, Pena, Moloney, Laure, & Markidis, 2018). La Figura 4 muestra visualmente las TPU analizadas.

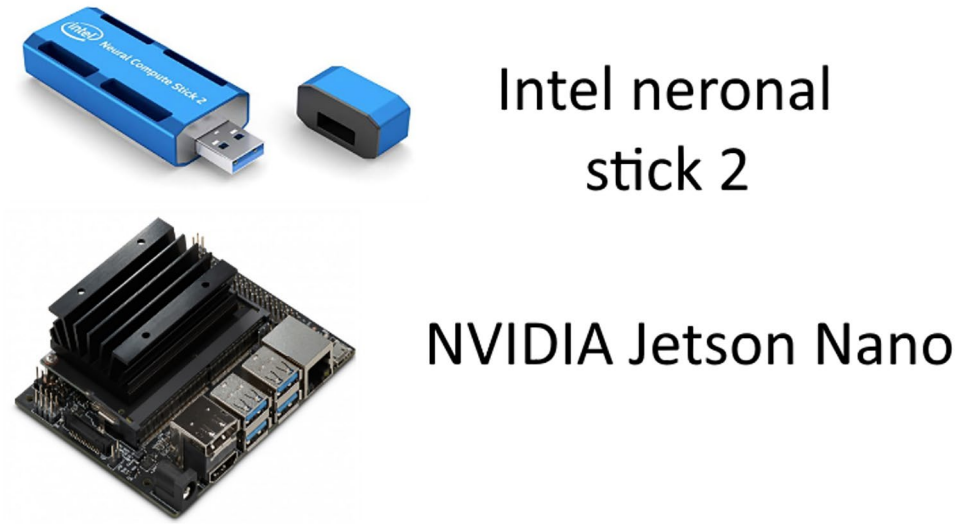
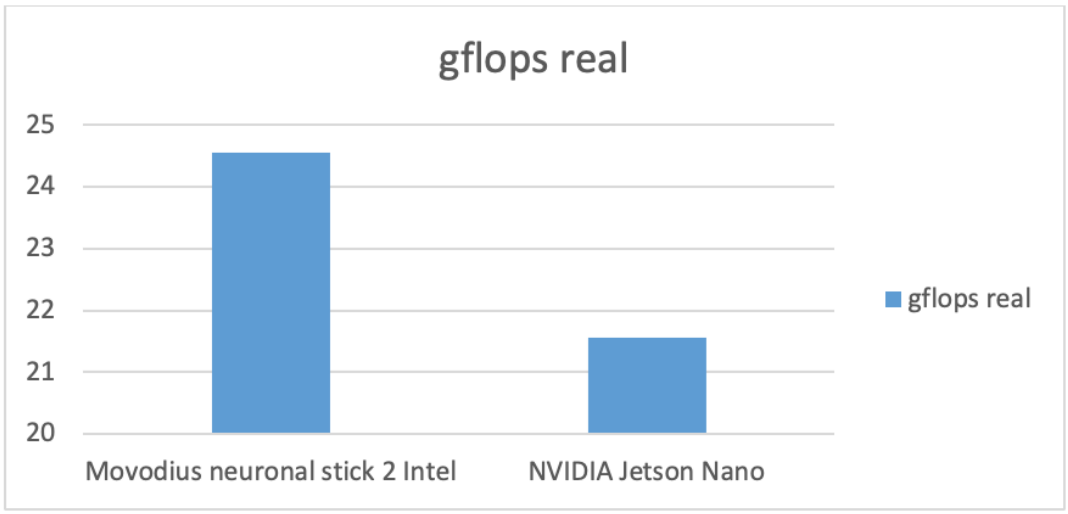


Figura 4. TPUs seleccionas. Fuente: proveedores Intel-Nvidia.

Los resultados que arrojan la prueba de entrenamiento en TPU aplicados mediante el modelo de red neuronal artificial GroceryNet utilizadas de igual forma en las GPU fueron los siguientes (véase Figura 5):



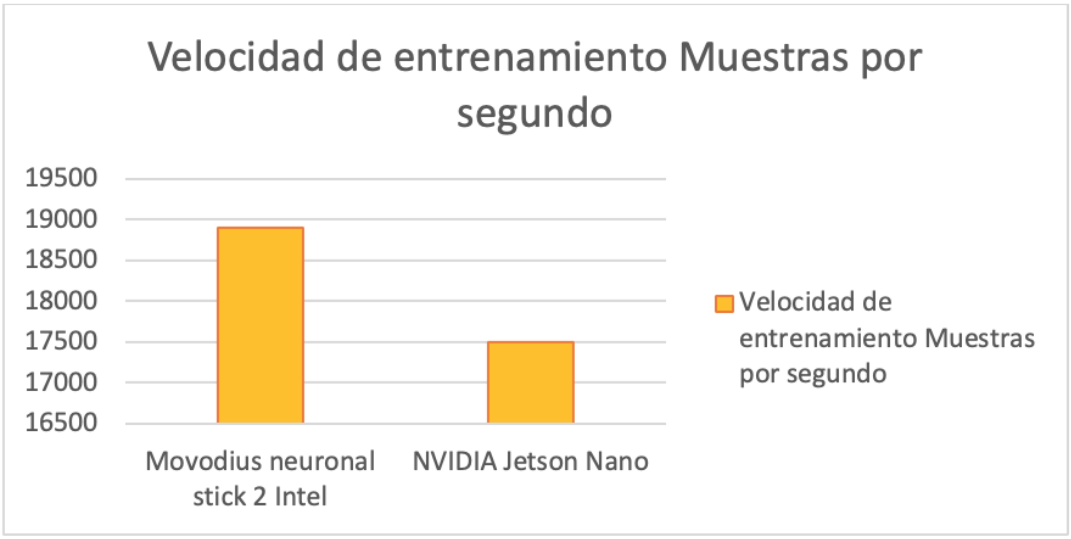


Figura 5. Resultados de pruebas a TPUs datos de modelo GroceryNet (Nakampe, 2018) y propias para Intel Movidius Neuronal Stick 2.

La velocidad de entrenamiento es notablemente mayor a menor consumo de energía, sin embargo, la capacidad de almacenamiento del modelo de la red neuronal desarrollada está limitada por la memoria interna de almacenamiento a no más de 4Gb, por lo que se tendrían que agregar módulos extras en redes neuronales artificiales de mayor tamaño, esto no es una limitante, ya que esta opción tiene como ventaja su bajo costo (cien dólares).

El deep learning, como se ha mencionado, es una técnica exigente con la aceleración de los procesos matemáticos en los equipos de computo involucrados en su desarrollo, basados

en las mediciones de una red considerada de nivel intermedio como la GroceryNet. Observando las velocidades de muestreo por segundo entre GPU y TPU podemos definir que, si se desea un entrenamiento adecuando en tiempo y a bajos costos, el mínimo requerimiento de hardware sería el siguiente:

Optando por un GPU: GPU con procesador GTX 1070 o superior 8Gb recomendado para procesamientos multicapa de red neuronal artificial debido a la capacidad de memoria disponible.

Optando por un TPU: TPU Movidas Neuronal Stick 2 el mejor rendimiento por su bajo consumo de energía en procesamiento monocapa por su capacidad de 4Gb.

Conclusiones

La aparición de la Inteligencia Artificial plantea un gran reto para los procesadores convencionales, el consumo de energía y la velocidad de procesamiento son las principales carencias que tienen los desarrolladores que pretenden incursionar en el mundo del deep learning y el entrenamiento de las redes neuronales.

El presente artículo pretende ser una guía comparativa entre las principales opciones para incursionar al desarrollo de IA a bajo costo y con las mejores opciones del mercado.

Se realizó la comparación entre los tres modelos GPU que pueden proporcionar un rendimiento aceptable a costos accesibles. De igual manera, se compararon TPU dedicados solo a la aceleración de entrenamiento de redes neuronales artificiales. También se muestran las principales características que los proveedores muestran en sus datasheet.

Destacamos que las pruebas y resultados obtenidos en el entrenamiento de la red neuronal artificial del modelo propuesto por Nakampre sobre el Intel Neuronal Stik 2 fueron realizadas por los investigadores del presente trabajo, mediante las mismas pruebas realizadas a las GPU y a la TPU Nvidia Jetson Nano como fase experimental, al adquirir este dispositivo para aplicar redes neuronales artificiales en el análisis de imágenes en video requeridos en nuestras investigaciones.

Hemos observado que el trabajo de las GPU de bajo costo puede ser útiles para redes multicapa por la capacidad de almacenamiento y que, de igual manera, una vez terminado el desarrollo del hardware, puede ser reutilizado para acelerar los gráficos necesarios en aplicaciones como diseño gráfico o render 3d, mientras que las TPU proporcionan un hardware exclusivo de buen rendimiento en redes monocapa y, gracias a su tamaño y bajo consumo de energía, puede ser aplicado a cómputo móvil y robótica.

Como parte del trabajo futuro, se espera realizar el análisis de Intel Neuronal Stik 2 para la detección de plagas en cultivos, así como los procesos de pruebas bajo redes *multithrens* que permitan realizar más de una operación a la red neuronal artificial. Esto daría el panorama completo de la eficiencia de las TPU en el entrenamiento de las redes neuronales artificiales a bajo costo.

Referencias bibliográficas:

Basogain, X. O. (2005). Redes Neuronales Artificiales Y Sus Aplicaciones. *Medicina Intensiva*, 29(1), 13–20. https://doi.org/10.1016/S0210-5691(05)74198-X

Florencio, F., Sergipe, U. F. De, Sergipe, U. F. De, David, E., Ordonez, M., & Sergipe, U. F. De. (2019). Performance Analysis of Deep Learning Libraries: TensorFlow and PyTorch. *Journal Computer Science.* (May).

Guo, J., Liu, W., Wang, W., Lu, Q., Hu, S., Han, J., & Li, R. (2019). AccUDNN: A GPU *Memory Efficient Accelerator for Training Ultra-deep Deep Neural Networks,* (January). Retrieved from: http://arxiv.org/abs/1901.06773

Kayid, A. M. (2018). *Performance of CPUs/GPUs for Deep Learning workloads,* (May 2018), 25.

Moloney, D., Barry, C. B., Richmond, R., Connor, F., Brick, C., & Donohoe, D. (2014). *Movidius Myriad 2: Vision Processor.*

Nakampe, M. (2018). *GroceryNet at the Edge using Intel Movidius Neural Compute Stick.* DOI: 10.13140/RG.2.2.13745.43360

Rivas-Gomez, S., Pena, A. J., Moloney, D., Laure, E., & Markidis, S. (2018). Exploring the vision processing unit as co-processor for inference. *Proceedings - 2018 IEEE 32nd International Parallel and Distributed Processing Symposium Workshops,* IPDPSW 2018, 589–598. DOI: 10.1109/IPDPSW.2018.00098