La Reestructuración de Arquitecturas Pre-Establecidas y su Efecto en Ingenieros, Desarrolladores, e Investigadores: Unidades de Procesamiento Tensorial
Restructuration of Pre-Established Architectures, and it's Effect on Engineers, Developers, and Researchers: Tensor Processing Units

1ro Jorge Andrés Mogotocoro Fajardo
Escuela de Ingeniería de Sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
jorgemogotocoro05@gmail.com

2do Jheyson Arley Jaimes Esteban
Escuela de Ingeniería de Sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
jheyson_2998@Outlook.com

3ro Luis Carlos Jimenez Arciniegas
Escuela de Ingeniería de Sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
lucaja999@gmail.com

4to Juan Pablo Moreno Ríos
Escuela de Ingeniería de Sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
raikeon0@gmail.com

Abstract

Los avances en los campos de investigacíon y desarrollo traen consigo la necesidad de adaptar nuevos métodos operacionales tanto en software como hardware, lo cual a su vez crea el reto de reestructurar arquitecturas convencionales ya existentes para acomodar susodichos avances. Uno de estos casos se da con el deep learning, y por consiguiente, las redes neuronales; dando paso a una nueva arquitectura de procesador conocida como TPU, dado que las CPUs y GPUs ya establecidas por un tiempo no están desarrolladas en especifico para esta trata de datos, por lo cual asumen un mayor costo (energia y tiempo) al tratar con estos nuevos desarrollos.

Abstract

The technological advancements in the research and development fields bring with themselves the necesity of adapting new operation methods in both software and hardware areas, which in itself generates a challenge, a challenge of restructuring previously established conventional architectures to accomodate such advancements. One of these cases is deep learning, and therefore neural networks; giving way for a new type of architecture to develop known as TPU (Tensor Processing Unit), as already existing CPUs and GPUs are not specifically tailored to deal with the operations involving neural networks, which sets expectations for a bigger cost (power consumption and time) when creating datacenters for storage and processing of neural-network-generated-data.

Index Terms

TPUs, GPUs, CPUs, deep learning, redes neuronales, tensorflow

I. Introducción

A medida que nuevos métodos y algoritmos de resolución de problemas surgen en este mundo siempre cambiante, se deben analizar las arquitecturas previamente establecidas, y preguntarse si se podría desarrollar una más eficaz, eficiente, y apropiada para estos algoritmos.

Este es el caso con las unidades de procesamiento tensor (TPUs), las cuales optimizan y aceleran el rendimiento de los procesos llevados a cabo por redes neuronales en el ámbito del machine learning, dentro del campo de la inteligencia artificial.

Al haberse planteado esta nueva arquitectura se tenía la necesidad de reducir la latencia de las tareas llevadas acabo por una red neuronal, ya que con procesadores tanto gráficos como centrales el tiempo de ejecución de estas era costosamente largo.

Cabe añadir que la invención de las TPUs se dió también debido a la preocupación creciente del uso mayor que se pronosticaba para las redes neuronales, lo que implicaba un crecimiento de la demanda por más y más centros de datos para procesar estas redes, esto significaria un aumento en la necesidad por fondos economicos para crear estos centros de datos, lo cual fomentó la busqueda de una manera en la que se solucionaran tantos los problemas de rendimiento de los procesadores en ese entonces, y a su vez reducir la creciente demanda economica.
Estos nuevos procesadores, en su primera generación, fueron supuestamente entre 15 y 30 veces más rápidos que GPUs y CPUs del estado del arte en ese entonces, con insinuaciones de ser aún más veloces si tuvieran la misma cantidad de memoria que estos.

Lo anterior conlleva a muchas preguntas, tales como: ¿Qué hace que estos procesadores tengan tanta ventaja sobre los ya establecidos? ¿Cuál será el significado de una nueva arquitectura de procesamiento para programadores, desarrolladores, e ingenieros? Esto será tratado con detalle en las siguientes secciones.

II. Redes neuronales artificiales

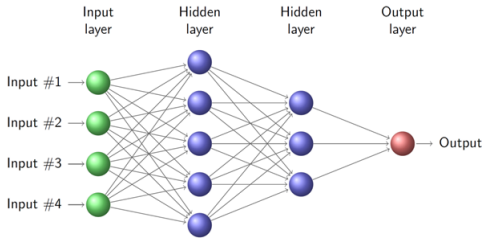


Fig. 1: Ejemplo de redes neuronales artificiales

Existe gran cantidad de procesos donde los computadores juegan un mejor papel que los humanos, por ejemplo, procesos matemáticos de gran complejidad, almacenar gran cantidad de información entre otros. Pero sin importar lo mucho que haya avanzado la tecnología en las computadoras nuestros cerebros siguen estando un paso delante de las computadoras cuando se trata del sentido común, la inspiración y la imaginación. Por esta razón, las redes neuronales artificiales (ANN, por sus siglas en inglés) están inspiradas en la estructura del cerebro humano, para así intentar lograr un sueño el cual muchos científicos tienen el, que es hacer que las computadoras para cada día sean más humanas y que estas mismas puedan razonar como lo hacemos los humanos. Ya que, los cerebros humanos siempre interpretan el contexto en el que se desarrollan las cosas de una forma en la que las computadoras aún no pueden.

Las redes neuronales se desarrollaron con el fin de simular el cerebro humano, para que una máquina pueda entender e interpretar ciertas situaciones como lo hace un humano, y también para que la máquina pueda tomar decisiones de una manera humana. En conclusión, las redes neuronales artificiales son un intento de copiar el funcionamiento del cerebro humano, para que cada día las máquinas tengan un comportamiento similar a los humanos, y así las máquinas tengan un comportamiento autonomía a la hora de tomar decisiones.

Las redes neuronales artificiales funcionan usando diferentes capas de procesamiento matemático para así dar sentido a la información que se maneja. En la mayoría de las veces una red neuronal artificial tiene desde unas cuantas neuronas artificiales hasta millones estas, está neuronas se conocen como unidades que están distribuidas en una serie de capas. La capa entrada se encarga de recibir la información que entra a la red neuronal representada en diferentes maneras, esa información que llega a la capa entrada es la que se pretende procesar. Después de entrar la información a la red neuronal y pasar por la unidad de entrada, está pasa por una o más unidades ocultas. La función que tienen las unidades ocultas en es de transformar la información de entrada en algo que la unidad de salida le pueda dar utilidad.

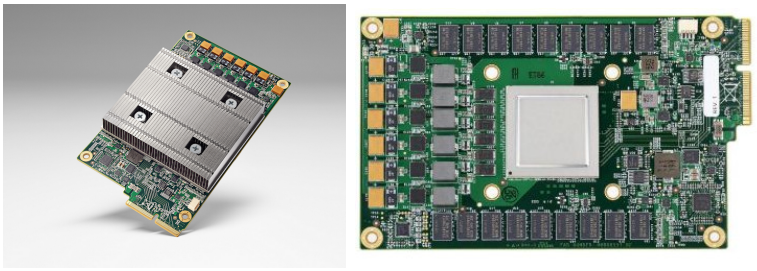
En las redes neuronales se entiende encontrar que estas están completamente conectadas que una capa a otra, se debe saber que entre mayor sea el número de las conexiones que tenga cada unidad se presentará un mejor rendimiento. Por eso a medida que iban pasando los datos por cada unidad la red neuronal aprende más sobre los datos que le suministraron. Ya por último se encuentran las unidades de salía, y estas se encargan de mostrar o responder a los datos que entraron a la red neuronal para ser procesados.

A. Aprendizaje automático (Machine learning)

El aprendizaje automático es cuando las máquinas logran descifrar patrones, aprender, tomar decisiones y hacer predicciones en base de los datos que le fueron suministrados, sin que se hubiera tenido la necesidad de programarlos para desempeñar estas actividades, o que se hubiera programado de antemano que decisiones debería tomar para determinadas situaciones. El aprendizaje automático, en forma de inteligencia artificial, automatiza el proceso de creación de modelos analíticos y permite que las máquinas se adapten a nuevas situaciones de manera independiente.

El aprendizaje automático se genera cuando el software es capaz de tomar de forma autónoma decisiones, es decir, tiene la habilidad de predecir y reaccionar de manera correcta ante determinadas situaciones basándose en resultados anteriores que no fueron influenciados por el aporte humano. En resumen, el aprendizaje automático es una rama de la inteligencia artificial, que consiste en que las maquinas pueden aprender de datos, identificar patrones y tomar decisiones con una mínima ayuda humana.

B. Google supera las tareas de aprendizaje



(a) (b)

Fig. 2: Placa de la unidad de procesamiento del tensor

El aprendizaje automático es la idea o proceso que está detrás de muchas de las aplicaciones de Google. Pero hay algo que se debe tener claro, todo gran software es mejor cuando se ejecuta en un gran hardware. Por esa razón, Google desde hace varios años empezó a trabajar en crear sus propios aceleradores personalizados para aplicaciones de aprendizaje automático.

El resultado de que las investigaciones se denomina unidad de procesamiento de tensor (TPU), un ASIC personalizado que fue construido para el aprendizaje automático y adaptado al TensorFlow. Google ha ido trabajando con la TPU por un buen tiempo, y usando este hardware se dieron cuenta que ofrecía un mejor rendimiento al momento de trabajar con el aprendizaje automático.

El TPU logra adaptarse a las aplicaciones de aprendizaje automático, por eso mismo se permite que Street y se más toleran con la precisión computacional, es decir, que no necesita muchos transistores por operación. Por eso mismo se pueden incluir más operaciones por segundo, usar modelos de aprendizaje automático más sofisticados, potentes y rápidos. Para que así los usuarios que hagan uso de estas herramienta obtengan resultados de una manera más rápida y con mayor detalle.

El objetivo de Google es liderar toda la industria del aprendizaje automático y poner en el mercado está nueva innovación. Incorporar TPU en la infraestructura de Google permitirá llevar a los desarrolladores el poder de Google a través de software como TensorFlow y Cloud Machine Learning con capacidades avanzadas de aceleración. El aprendizaje automático a ido transformando la forma con que los desarrolladores tienden a crear sus aplicaciones inteligentes, ya que con las herramientas que ofrece Google le permiten al desarrollador crear sus aplicaciones inteligentes de una forma más simple y efectiva.

C. Unidades de procesamiento de tensor de nube

Cloud TPU es una herramienta que permite ejecutar las cargas de trabajo de aprendizaje automático en el hardware acelerador TPU de Google mediante TensorFlow. Cloud TPU está diseñado para ofrecer un gran rendimiento, ayudando a investigadores y desarrolladores. Las API de Tensorflow de alto nivel le ayudan a obtener modelos que se ejecutan en el hardware de TPU en la nube.

Al contar con los TPU de nube, se tiene la oportunidad de usar los aceleradores de aprendizaje automático de última generación directamente en su infraestructura de producción y beneficiarse de la capacidad informática acelerada, y de forma

gratuita. Al contar con las TPU de la nube se logrará que salgan beneficiadas varias aplicaciones de aprendizaje automático. Google ofrece varios tipos de hardware en Google Cloud para que se pueda elegir los aceleradores que mejor se adapten a la necesidad de los desarrolladores o investigadores que usan esta herramienta.

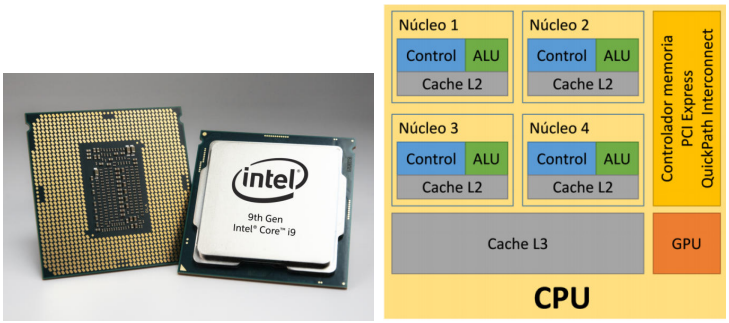
Los TPU de nube, son rápidos a la hora de realizar cálculos de vectores densos y matrices. El problema surge en la transferencia de datos entre Cloud TPU y la memoria del host, ya que esta es lenta si se compara con la velocidad de cálculo, es decir, que puede llegar a usarse el dispositivo de una manera muy ineficiente, pues el dispositivo estaría la mayor parte del tiempo esperando que le llegue los datos a procesar. Un único chip de TPU en la nube contiene 2 núcleos, cada uno contiene múltiples unidades de matriz (MXU) que fueron diseñadas con el propósito de acelerar los programas dominados por multiplicaciones de matriz densa y convoluciones.

III. Unidades de procesamiento CPU y GPU

la producción y construcción de hardware suele ser costosa lo cual ha llevado a que la mayoría sean producidos en grandes volúmenes y que que se realicen para propósitos generales de control y procesamiento de datos, estos pueden usarse casi en cualquier proceso que se requiera sin embargo para algunos de estos su desempeño es demasiado bajo llevando un tiempo de demoras excesivas o altos consumos de energía, un problema bastante molesto que genera un mayor costo en terminos economicos y de tiempo, para solucionarlo se han desarrollado diferentes procesadores especializados para la realizacion de dichas operacion buslando claramente que sean sotenibles y eficientes en terminos de consumo de energia vs opereaciondes de punto flotante por segundo .

A. CPU (Unidad central de procesamiento)

la unidad central de procesamiento es un circuito integrado de uso general diseñados para realizar el procesamiento de datos,de los cuales no se tiene en cuenta el volumen de información a tratar ni la complejidad de los algoritmos a aplicar. Estos circuitos tiene la siguiente arquitectura :



(a) CPU intel i9.

(b) Representacion del arquitectura de una CPU.

Fig. 3: Unidad de procesamiento central

Se puede apreciar que por cada núcleo hay una ALU o unidad aritmético lógico que encarga de las funciones de procesamiento de datos, también para cada núcleo hay una unidad de control y una caché l2, la primera extrae instrucciones de memoria, las decodifica y ejecuta llamados a la ALU si lo requiere y la segunda extrae datos de uso frecuente siendo más rápida que la memoria ram además de una caché l3 fuera de los núcleos la cual genera una copia a la L2, siendo más lenta que esta pero mas rapida que la memoria ram.

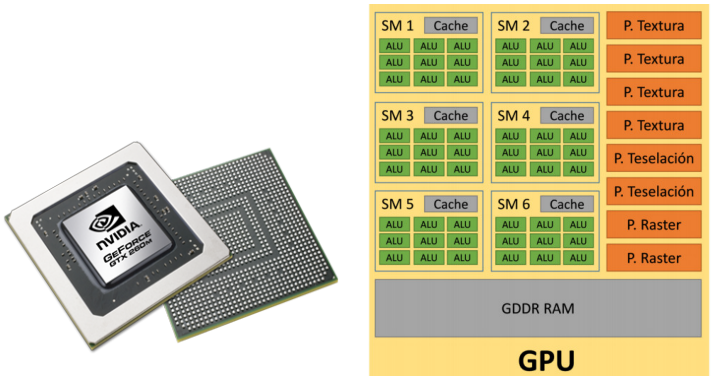
Estos procesadores fueron especialmente diseñados para realizar operaciones secuenciales, sin embargo con el pasar de los años y la creciente necesidad por más potencia de procesamiento su nivel de paralelismo se a venido incrementado mediante diversos métodos, inicialmente se usó la segmentación del cauce el cual descompone el proceso en fases o etapas que permitan su ejecución simultánea, y actualmente esto se hace mediante la introducción de múltiples núcleos,de propósito general, los cuales han hecho al procesador central más rápido y eficiente sin embargo su principal función sigue siendo el mantener una eficiencia homogénea en la ejecución todas las operaciones , lo cual hace que su tiempo de ejecución de procesos que requieran un alto grado de paralelización sea bastante elevado.

Un ejemplo de lo anterior sería el aprendizaje de máquina el cual requiere miles o incluso millones de muestras de datos para realizar un entrenamiento adecuado que permita obtener un proceso de predicción o inferencia con un alto grado de precisión o uno mas concreto la multiplicacion de matrices donde una imagen matriz de 224×224 píxeles, incluyendo la propagación

hacia adelante y hacia atrás, con ResNET-200, una red profunda con 200 capas, es de más de 22 segundos en una CPU Dual XeonE5-2630.

B. GPU (Unidad de procesamiento gráfico)

a unidad gráfica de procesamiento es un coprocesador el cual a sido diseñado para realizar tareas específicas tales como la representación de videos, manejo de gráficos 3d y operaciones de punto flotante y ayudar a la cpu con operaciones que requieran un alto grado de paralelización, su arquitectura es la siguiente:



(a) GPU NVIDIA GeForce GTX.

(b) Representacion del arquitectura de una GPU.

Fig. 4: Unidad de precesamiento gradfico

como es evidenciable su arquitectura muy diferente a la de la cpu,siendo notorio el hecho que esta está compuesta por una gran cantidad de núcleos que pueden manejar miles de subprocesos simultáneamente, la cantidad de núcleos contenidos en promedio por una GPU son del orden de los miles, los cuales son relativamente simples y estan especializados en realizar unas pocas operaciones con tipos de datos fijos,además su velocidad de reloj suele ser relativamente baja en comparación a la de una cpu, operan en promedio a 1 GHz.

Las GPUs ofrecen un soporte a la unidad central de procesamiento alivianando su carga y acelerando la ejecución de muchos programas, las gpu tienen un tamaño mayor al de las cpus e igualmente un mayor número de componentes, todo esto se refleja en un gran rendimiento al trabajar con operaciones paralelas pero no muy optima en operaciones secuenciales. esto ya que fueron ideadas para el tratamiento de texturas , calculo de tranformaciones sobre los vértices geometricos y prosesamiento me millones de pixeles. Retomando el ejemplo de la multiplicacion matricial echo en para una CPU Dual XeonE5-2630 y ya que la multiplicacion de matricas es altamente paralelizable, el uso de la GPU nos permitira reducir los 22 segundos antes indicados para una CPU a entre 0.2 y 0.5 segundos, dependiendo de la GPU usada.

IV. Funcionamiento de una TPU

A. Concepción y Nacimiento

La unidad de procesamiento tensorial nace en un periodo de desarrollo de 15 meses, debido a la discusión por demandas de un circuito integrado de aplicación específica que reduzca los costos de construcción de abastecimiento de centros de datos para la trata de redes neuronales; aunque inicialmente dada en 2006, no se tomó seriamente sino hasta el 2013, cuando se estimó que los usos de las redes neuronales crecerían de manera rápida y a gran volumen en los próximos años. La TPU es diseñada entonces como un Co-procesador, para facilitar su integración en centros de datos ya construidos, y su implementación en nuevos centros de datos; recibe instrucciones a ejecutar de un servidor, mas no las solicita.

B. Arquitectura

La arquitectura de una TPU está basada en una matriz multiplicadora, que en su primera generación trata sólo con números enteros. Esta matriz contiene 65536 MACs que operan con multiplicación y adición de 8 bits, los productos en 16 bits son almacenados en una memoria relativamente grande de 28 MiB. El funcionamiento de la máquina se puede observar más a fondo en la figura 1.

Esta matriz física de 256x256 es llamada MXU, y es ideal para la trata de datos de redes neuronales, puesto que se pueden procesar pesos y nodos de una manera efectiva y eficaz, por lo tanto, esta es la parte central de una TPU, y la parte que se mantiene más activa durante su uso. Un problema de cuello de botella dado en la TPU es a causa del angosto ancho de banda

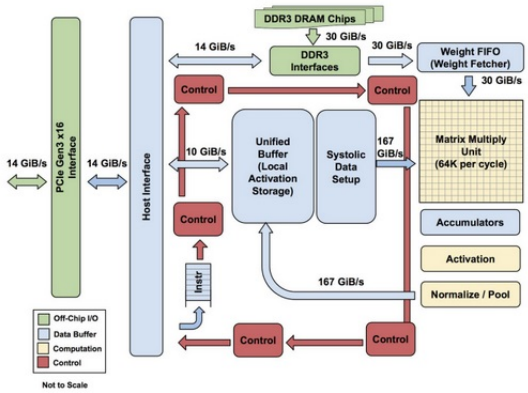


Fig. 5: Arquitectura de la primera generación de una TPU.

de los cables y puertos PCIe, puesto que la capacidad de procesamiento de datos en masa de la TPU es masiva, pero la tasa a la que se transmiten los datos de un servidor a la TPU aumenta los periodos de tiempo entre procesos.

Todo esto se puede describir generalmente resumido en que las TPUs utilizan arquitectura CISC, o computador de set de instrucciones complejas, lo que quiere decir que una instrucción puede desencadenar operaciones complejas a gran escala procesadas por la MXU; la cual se diferencia fundamentalmente del orden de procesamiento de las CPUs (escalar) y GPUs (vectorial), en que su naturaleza, como ya dicho anteriormente, es de orden matricial, lo que implica que puede desarrollar cientos de miles de operaciones en un ciclo.

Al hacer operaciones con números enteros se sacrifica precisión por rendimiento.

Todas las operaciones aritméticas, las cuales se procesan en la ALU (unidad aritmética lógica), se encadenan, de modo que se reduce el número de accesos a los registros, lo cual aumenta la eficiencia a la hora de utilizar energía.

Por último, el control de la lógica se encuentra en 2% del dado de la TPU, lo cual hace que el resultado sea determinista, dandose la habilidad de predecir precisamente la latencia de ejecución.

Estas caracteristicas únicas a la TPU se pueden organizar en los siguientes puntos:

- Cuantización
- CISC
- MXU
- Matrices Sistólico
- Diseño Minimalista y Determinista

En la figura 2 se puede apreciar el dado de un procesador de la primera generación de TPUs.

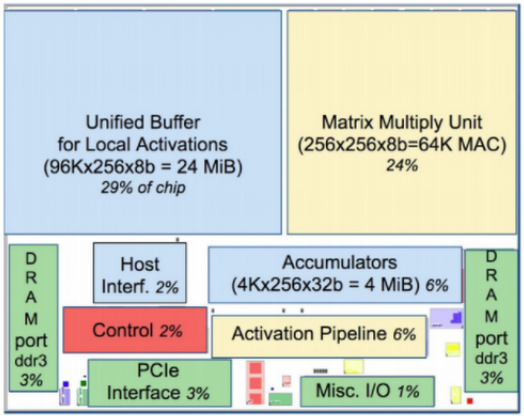


Fig. 6: Plan del dado de una TPU de primera generación.

C. *Limites*

Al tener en cuenta lo leído anteriormente, sería común preguntarse ¿Hay algo que la TPU no pueda hacer? ¿Hay algo en lo q eula TPU no tenga ventaja sobre trosoprocesad ros? y si, hay cosae que la TPU no puede hacer, las cuales son básicamente cualquier cosa que no sea srattr con deep learning y redes neuronales; el único fin de una TPU es calcular pesos y activaciones de nodos en estas redes, y por lo tanto es generalmente inútil en otros ámbitos. Entre otros problemas están la rapidez y los limitados recursos economi ocscon la que fue planeada y diseñada, lo cual dio poco tiempo para que se intagraran técnices del estado del arte, cosa que otras CPUs y GPUs tienen, débido a que la amplia dominación del mercado por Nvidia e Intel les permite utilizar tecnologías de última generación, esto puede observarse en los cuellos de botella al momento de transferirse datos, ya que la co exión PCIe no tiene el suficiente ancho de banda como para ir a la par con la tasa a la que la TPU procesa y genera resultados a partir de los datos. También se pueden observar desventajas en los tipos y modelos de memoria, ya que la TPU de primera generación posee memoria RAM DDR3, mientras que las tarjetas gráficas de Nvidia en ese momento poseían memoria DDR5. Y aunque a simple vista una arquitectura de 28 nanometros se puede descartar como grande y obsoleta, esto se recompensa con la eficiencia en el consumo de energía.

En pocas palabras, las desventejas de una TPU en contraste con sus competidores son:

- Memoria limitada; ancho de banda limitado
- Tamaño del dado
- Baja accesibilidad a tecnología del estado del arte

V. Impacto generado por la TPU

Dados los relativamente recientes anuncios de Google sobre el lanzamiento de TPUs enfocadas en tareas especificas otras empresas que se mueven también en el medio de los procesadores empezaron a desear esta nueva tecnología que había creado, esto cambia el panoramoma total que se tenia anteriormente, en mayor medida a lo referente con aprendizaje de maquina, porque se tenia un mercado bastante competitivo donde se discudia acerca de si las GPU, FPGAs o VPU eran mas adecuadas. Habrá varias empresas poderosas que quieran estar interesadas en abordar el mismo trayecto para lograr un significativo rendimiento y eficiencia. Viendo ese panorama que se nos presenta el fruto de esta competencia, será la creación de nuevos dispositivos que le faciliten los procesos o aplicaciones, a los investigadores y desarrolladores. Aunque Google con el TPU en la nube, ofrece esta herramienta para que los desarrolladores de programas inteligentes puedan usar esa nueva tecnología que ellos están implementando, para hacer que sus aplicaciones sean mas efectivas.

Actualmente es notoria la fuerza que ha ejercido Google bajo el uso de esta nueva arquitectura, la cual son el poder detras de muchos servicios que son usando diariamente como lo son su buscador, Street View, Traductor entre otros. En palabras del gigante de Mountain View:

”Si consideraramos un escenario en el que las personas utilizan la busqueda por voz de Google durante tres minutos al dia y ejecutamos redes neuronales profundas para nuestro sistema de reconocimiento de voz en las unidades de procesamiento que estabamos usando, ¡habriamos tenido que duplicar el numero de cestros de datos de Google!”

Fig. 7: Gran variedad de los servicios que ofrece Google entre ellos cabe destacar aquellos que se vieron impulsados con el lanzamiento de las TPU como lo son Street View y Traductor.

Plasmado en estas declaraciones se encuentra presente como ha evolucionado las necesidades en cuanto a procesamiento de datos a gran escala (big data) y como ha sido necesario incurrir en el desarrollo de nuevas tecnologias capases de soportar tareas sumamente complejas lo cual repercute enormemente en los planteamientos que le concierne a competidores corporativos que basaban sus desarrollos en otras tecnologias.

Se tiene en paralelo como resultado una fuerte influencia sobre el consumo energetico que pueda requerir suministrarse al sistema que emplee TPU, claro esta que en cargas de trabajo de produccion de AI para el cual es requerido la inferencia de redes neuronales, el TPU se considera mas rapido que las GPU y CPU contemporaneos y frente a chips convencionales la inclucion de TPU tambien logra mucho mejor eficiencia energetica tomando una medida de calculo por vatio de energia consumuda como es apreciado en la grafica siguiente.

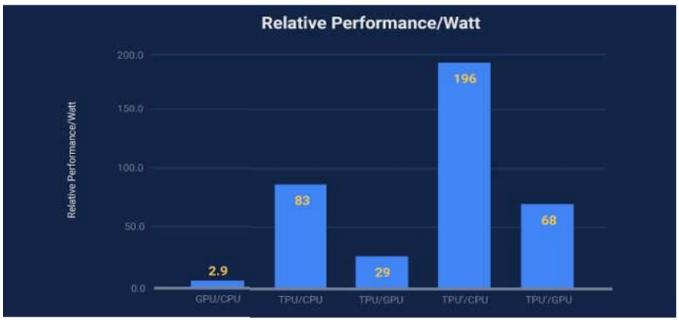


Fig. 8: Grafico de barras que muestra la comparativa del rendimiento relativo por watt entre GPU,CPU y TPU

Por tanto al tomar en consideracion la importancia que toman los TPU frente los competidores predentes en el grafico da por sentado que el desarrollo de unidades de procesamiento enfocadas a tareas en concreto frente al caso de la CPU como unidad de procesamiento general y la GPU como unidad de procesamiento grafico en ambientes de dsarrollo que impliquen redes neuronales, la TPU presenta un punto de quiebre como una alternativa mas viable lo cual implica un mayor atractivo frente a su uso y desarrollo.

VI. Conclusiones

Si bien la historia misma ha sido testigo de como los avances tecnologicos han surgido para resolver necesidades y en caso de incurrir en ir contra la corriente se enfrenta no mas que la extincion. Ahora bien el TPU como arquitectura o pieza de hardware ha tenido su mayor impulso en los ultimos años sin duda gracias al gigante de Mountain View, cabe resaltar que como en muchos otros casos, las bases teoriacas y el planteamiento de la arquitectura se an desarrollado muchos años antes, sin embargo no se materializado debido a la falta de equipo tecnológicos lo suficientemente avanzados, causando en la mayoría de los casos que se retomen e implementen cuando las condiciones sean las adecuadas para ello y exista el soporte e implementacion suficientes para lograr que al pasar de la teoria a la practica esta sea viable y/o facilmente comerciable. La entrada e implementacion de estas unidades de procesamiento solo es el inicio de una carrera con muchos mas competidores esperando entrar en juego.

Aunque la TPU inicialmente entra como un competidor al mercado de los procesadores, el cambio no es muy grande debido a que no hay que desarrollar aplicaciones especificas para ella, más sólo pueden ser utilizadas por Google y colaboradores en el momento, no supone una gran amenaza para la industria ni para el dia a dia de los ingenieros y desarrolladores.

La unidad de procesamiento tensorial es una gran idea sobre la cual Google capitalizó mejoras a los rendimientos de la trata de datos de redes neuronales y machine learning, lo cual se puede percibir como uno de los grandes avances en el campo de la inteligencia artificial. Este logro conlleva a un manejo más simple y estandarizado de las complejas redes, y sus datos tanto de entrada como salida, así como también la reducción en los costos de construcción de centros de datos, dando vía para para mas densidad de trata de información e investigación en el campo.

Si la ley de Moore ciertamente ya no es válida en la contemporaneidad, el siguiente paso en cuanto a la innovación en máquinas no depende de que tan pequeño se pueda hacer el tamaño del dado, sino de que maneras se puede modificar una arquitectura para que realice ciertas tareas con una mayor eficiencia, esto abre muchas puertas al ralentizado avance en potencia de procesadores cada año.

Google hace un magnífico uso de la computación en la nube; mantiene sus TPUs al alcance de todo quien desee aprovechar sus ventajas, mientras que a su vez es capaz de monopolizar su arquitectura, manteniendo su ubicación física sólo al alcance de sus propios desarrolladores, proporcionando así un gran servicio a la comunidad.

Al ser el principal productor de este nuevo tipo de procesador, Google tiene una gran ventaja sobre competidores, al estar más al tanto de sus desventajas, y posibles soluciones a estas, por lo cual está adelante de la curva por un gran margen, esto permite que la industria y el mercado no se estanquen, ni se forme un monopolio u oligopolio en la que las empresas cobren más por sus productos a la vez que disminuyen su tasa de innovación, estancando las indutrias tanto de desarrollo, como de producción, e investigación.

REFERENCES

[1] Antonio J. Rivera, Francisco Charte, MacarenaEspinilla, and MaríaD.Pérez-Godoy,"Nuevas arquitecturas hardware de pro cesamiento de altorendimiento para aprendiza je profundo" Departamento de Informática, Universidad de Jaén, 2018.

[2] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon Google, Inc., Mountain View, CA USA ,"In-Datacenter Performance Analysis of a Tensor Processing Unit"

[3] "Tensor Processing Unit" Disponible en:https://devopedia.org/tensor-processing-unit.[Accedido: febrero-27-2019]

[4] Kaz Sato, Cliff Young y David Patterson."An in-depth look at Google's first Tensor Processing Unit (TPU)" Disponible en:https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu.[Accedido: febrero-26-2019].

[5] Bernard Marr."What Are Artificial Neural Networks - A Simple Explanation For Absolutely Anyone" Disponible en:https://www.forbes.com/sites/bernardmarr/2018/09/24/what-are-artificial-neural-networks-a-simple-explanation-for-absolutely-anyone/74b36c871245.[Accedido: febrero-26-2019].

[6] ."¿Qué es el aprendizaje automático?" Disponible en:https://www.hpe.com/mx/es/what-is/machine-learning.html.[Accedido: febrero-27-2019].

[7] The era of cpu for all tasks has long ended. Disponible en: https://www.analyticsindiamag.com/the-era-of-cpu-for-all-tasks-has-long-ended/.[Accedido: febrero-26-2019].