



UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE CIENCIA DE LA COMPUTACIÓN

*Entrenamiento de una Deep Neural Network
mediante el uso de GPGPU*

SEMINARIO DE TESIS 1

Autor: Víctor Jesús Sotelo Chico

Asesor: Cesar Lara Avila

Diciembre, 2017

Resumen

Índice general

Resumen	III
1. Introducción	2
1.1. Motivación	2
1.2. Objetivos	3
1.3. Estructura del Seminario	4
2. Estado del Arte	5
2.1. GPU computing	5
2.1.1. The GPU computing Era	5
2.2. Machine learning aplicado a GPU's	6
2.2.1. Uso de redes neuronales para encontrar el rendimiento de una GPU	6
2.3. Deep Learning	6
2.3.1. Deep Machine Learning - A New Frontier in Artificial Intelligence	6
2.4. Conclusiones	7
3. Marco Teórico	8
3.1. GPGPU computing	8
3.1.1. CPU	8
3.1.2. GPU	8
3.1.3. Procesamiento en Paralelo	9
3.1.4. GPGPU	9
3.2. Redes Neuronales	9
3.2.1. Neuronas	9

3.2.2.	Redes Neuronales Artificiales	10
3.2.3.	Redes Neuronales Profundas	10
3.3.	Herramientas	11
3.3.1.	TensorFlow	11
4.	Conclusiones y Trabajo Futuro	14
4.1.	Conclusiones	14
4.2.	Trabajo Futuro	14
A.	Título del apéndice	18

Índice de figuras

3.1. Esquema GPGPU	9
3.2. ReLu	11
3.3. TensorFlow	12

Índice de Acrónimos

API	Application Programming Interface
AJAX	Asynchronous JavaScript And XML
ETC	Etcétera

Agradecimientos

Agradezco a mis padres por todo el apoyo incondicional en estos años de estudio Asimismo agradecer a mis compañeros de estudio en especial a Briggette y Rogger que lo largo de estos años fueron un apoyo en todo el proceso formativo.

Y un especial agradecimeinto al profesor Lara cuyo apoyo ayudo a culminar este seminario.

Capítulo 1

Introducción

En el campo de la inteligencia artificial las redes neuronales profundas tienen un papel muy importante debido a que estas son el camino para que las computadoras realicen tareas que nuestros cerebros realizan de manera natural, tareas como el reconocimiento de voz, imágenes y patrones. En la actualidad empresas importantes utilizan las redes neuronales profundas uno ejemplo de esto es google con el reconocimiento de voz e imágenes. Una característica de las redes neuronales profundas es que están compuestas por una gran cantidad de capas lo cual dificulta el entrenamiento en computadoras que solo usan el CPU. Una manera de resolver este problema es mediante el uso de las GPU's debido que las tareas de entrenamiento son paralelizables podemos usar las GPU'S para acelerar el proceso de entrenamientos de nuestra red neuronal profunda.

1.1. Motivación

La inteligencia artificial constituye una base muy importante en el campo de la computación, mezcla un conjunto de disciplinas como la estadística y ciencia de la computación con el objetivo de construir modelos que puedan permitir a las computadoras realizar tareas que hace años hubiese sido considerado imposible. El hecho de lograr que las computadoras sean capaces de reconocer objetos, clasificarlos lo cual ha permitido que la industria de la robótica desarrolle de manera acelerada en las últimas décadas. Hoy en día existen muchas herramientas que nos permiten desarrollar este tema y profundizarlo pero a medida que aumenta la complejidad del problema, el

costo computacional incrementa lo cual se convierte un problema importante. Una de las soluciones que apareció fue el uso de las GPU's para acelerar procesos como el entrenamiento de una red neural con muchas capas ocultas, las GPU's representan una solución muy eficaz debido a que en el campo de la inteligencia artificial existen muchas tareas que son paralelizables.

Actualmente el mercado de GPU's evoluciona muy rápido debido a su gran demanda en la industria de los videojuegos este mercado está dominado por NVIDIA y AMD esta competencia y la alta demanda permite que las GPU's tengan mejor rendimiento, además actualmente la nube ofrece otra posibilidad para el desarrollo de GPGPU computing debido a que nos permite tener acceso a mejores recursos. La evolución de GPU's nos permite usarlos para mejorar el procesamiento de grandes cantidades de datos.

Las redes neuronales nos permiten realizar tareas

1.2. Objetivos

El objetivo de este seminario es el de mostrar las ventajas que presenta el uso de GPGPU computing frente a otros tipos de recursos en el entrenamiento de una deep neural network.

Específicamente, los objetivos de este trabajo con respecto al sistema son:

- Entender el funcionamiento de las redes neuronales profundas
- Mostrar los resultados de distintas GPU's en el entrenamiento de una red neuronal profunda.
- Realizar un análisis comparando el rendimiento de cada GPU al momento de entrenar la deep neural network.

Y los objetivos con respecto a las competencias académicas desplegadas en el trabajo son:

- Desarrollar un manejo adecuado del uso GPU computing.

- Lograr un entendimiento de la tareas que son paralelizables en el entrenamiento de la red neuronal profunda.
- Desarrollar un manejo adecuado de tensorflow como herramienta para el reconocimiento de imágenes.

1.3. Estructura del Seminario

■ Introducción:

En este capítulo introductorio se comentará sobre las motivaciones, objetivos con los cual se penso el presente seminario.

■ Estado del Arte:

En este capítulo observaremos los trabajos e investigaciones previa que tiene el presente seminario , además se expondrá el aporte que realizaremos.

Capítulo 2

Estado del Arte

En este capítulo describiremos anteriores investigaciones de la relación entre GPGPU computing y machine learning, además de sus aplicaciones. También mostraremos investigaciones referentes a deep learning exclusivamente nos enfocaremos a Convolutional Neural Network(CNN) que son parte del tema de estudio en este seminario.

2.1. GPU computing

El uso de las GPU's han permitido lograr aplicaciones que antes podriamos imaginar que eran imposibles debido a su largo tiempo de ejecución. Hoy en día las GPU's son altamente usadas debido que cuentan con cientos de núcleos de procesadores en paralelo que permiten resolver rápidamente los problemas que son altamente paralelizables.

2.1.1. The GPU computing Era

En artículo se enfoca principalmente en describir la evolución que sufrieron las arquitecturas de GPU's, además de mostrar la importancia del uso de las GPU's para un mayor rendimiento y eficiencia que antes hubiesen sido consideradas imposibles debido al alto tiempo de ejecución que requerian. Además nos muestra que la escalabilidad es la principal característica que ha permitido que las GPU's aumenten su paralelismo y rendimiento.

2.2. Machine learning aplicado a GPU's

El uso de machine learning representa una gran ventaja para empresas que manejan gran cantidad de datos debido a que permiten descubrir patrones y analizar los datos.

2.2.1. Uso de redes neuronales para encontrar el rendimiento de una GPU

En la actualidad existen empresas dedicadas a la creación de GPU's, en el proceso una parte fundamental es la verificación del rendimiento de las GPU's actualmente existen simuladores conocidos como GPGPU-SIM que permiten estimaciones precisas pero estos poseen algunas dificultades como el tiempo empleado en configurarlos en base al hardware real además que este proceso esta propenso a errores. Un equipo de equipo conformado por investigadores de AMD y The University of Texas at Austin, quienes propusieron el uso de redes neuronales para predecir el rendimiento.

2.3. Deep Learning

Dentro del área de machine learning encontramos deep learning o aprendizaje profundo el cual consiste en un conjunto de algoritmos que modela abstracciones de alto nivel.

2.3.1. Deep Machine Learning - A New Frontier in Artificial Intelligence

Este trabajo de investigación fue realizado por investigadores oak Ridge National Laboratory y University of tennessee, el objetivo principal de este trabajo fue presentarnos el aprendizaje profundo como un camino para la imitación del cerebro humano y sus principales cualidad como el reconocimientos de objetos, rostros, etc.

Además de mostrarnos las aplicaciones del aprendizaje profundo, como : análisis de documentos, detección de voz, rostro, procesamiento natural del lenguaje, etc.

Actualmente existen algunas empresas privadas que apoyan el campo de deep learning con el objetivo de buscar sus aplicaciones comerciales entre estas empresas tenemos: Numenta y Binatix.

2.4. Conclusiones

Hemos visto la necesidad....

Poner unas conclusiones del capítulo y lo más importante, donde se enfoca tu trabajo y lo que se diferencia del resto

Capítulo 3

Marco Teórico

En este capítulo veremos los principales conocimientos que son necesario para el entendimiento del presente seminario.

3.1. GPGPU computing

3.1.1. CPU

(Central Processing Unit) es el hardware encargado del procesamiento de tareas mediante las operaciones básicas aritméticas, logicas y de entrada y salida del sistema. Los ordenadores actuales cuenta con más de una lo cual les permite usar multiprocesamiento. Los nucleos de la cpu estan hechos para el procesamiento de las tareas de manera secuencial. Una CPU tiene 2 componentes principales: + **ALU (Unidad Aritmetica Logica):** se encarga de realizar las operaciones aritméticas y lógicas. + **CU (Unidad de Control):** extrae información de la memoria la decodifica y ejecuta llamando a ALU las veces que sea necesario.

3.1.2. GPU

Las GPU(Graphics Processing Unit) A diferencia de las CPU los GPU's estan compuestos de miles de nucleos que fueron diseñados para resolver tareas en paralelo.

3.1.3. Procesamiento en Paralelo

El procesamiento en paralelo consiste en dividir las instrucciones del programa en múltiples procesadores con el objetivo de ejecutar el programa en menos tiempo.

3.1.4. GPGPU

GPGPU (General Purpose Graphics Processing Unit) consiste en realización de los cálculos que comúnmente serían realizados por el CPU utilizando GPU. Este método permite aprovechar los beneficios de potencia en procesamiento paralelos de las GPU's.

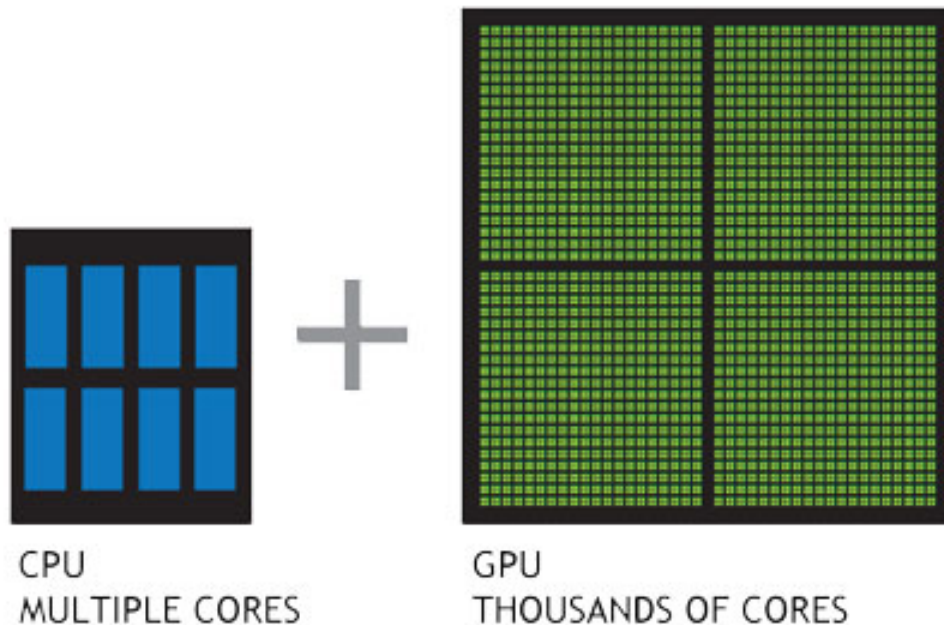


FIGURA 3.1: Esquema GPGPU

3.2. Redes Neuronales

3.2.1. Neuronas

En la biología la neurona es conocida como la unidad fundamental del cerebro humano, el cual está compuesto por millones de neuronas

interconectadas entre si. El trabajo de las neuronas consiste en recibir información, procesarla y enviarla a otras células. Este modelo fue copiado en 1943 por Warren S. McCulloch y Walter H. Pitts. Analogamente con las neuronas del cerebro humano nuestra neuro artificial toma una cantidad n de entradas $x_1, x_2, x_3, \dots, x_n$ estas entradas serán multiplicadas por pesos $w_1, w_2, w_3, \dots, w_n$ además se puede añadir una constante que llamaremos bias.

La entrada a de la neurona será la suma total de los productos $z = \sum_{i=1}^n x_i w_i$, el valor de z será la entrada a la neurona la cual la evaluará con una función f de tal forma que nuestra salida sea $y = f(z)$. Otra forma de ver esta expresión es por medio de la notación de vectores donde representaremos a las entradas como $x = [x_1 x_2 x_3 \dots x_n]$ y los pesos $w = [w_1 w_2 w_3 \dots w_n]$ de esta manera la salida de la neurona estará dada por $y = f(x \cdot w + b)$ donde b representa las bias.

3.2.2. Redes Neuronales Artificiales

Las redes neuronales artificiales(ANN)toman de ejemplo la arquitectura del cerebro como inspiración para la construcción de sistemas inteligente. Actualmente son la base para el desarrollo de la inteligencia artificial. Las redes neuronales están constituidas de las uniones de las neuronas.

3.2.3. Redes Neuronales Profundas

Las redes neuronales profundas estan constituidas principalmente de un numero de capas de convolución, No linealidad y pooling.

- **Convolución:** Un proceso importante dentro de las redes neuronales es la convolución que es usada para detectar las características de una imagen estas características pueden ser bordes, curvas, etc.
- **No linealidad:** Debido a que las convoluciones son operaciones lineales, lo cual no es adecuado para las tareas del mundo real. Debido a esto es importante introducir el ReLu que aplicará funciones no lineales a los mapas de característica producidas en las capas de convolución. Una de las funciones las común es ReLu.

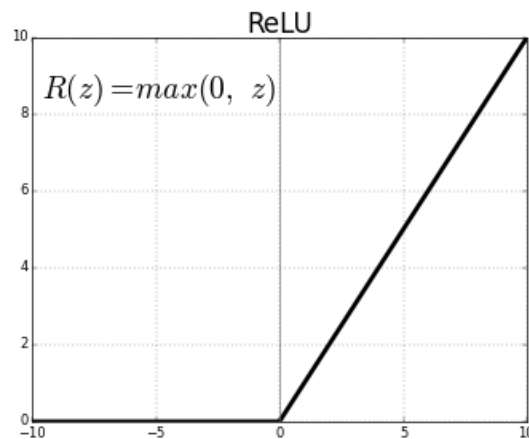


FIGURA 3.2: ReLu

- Pooling: Sirve para transformar el mapa de características en una representación de menor dimensión con el objetivo de la red sea más invariante a pequeñas transformaciones o variación de la imagen de entrada.

3.3. Herramientas

3.3.1. TensorFlow

Es un framework open source de google que nos permite realizar multiples técnicas de deep learning. TensorFlow nos provee de multiples API's en el nivel bajo de la API nos provee completo control en la programación y el nivel más alto nos permite que las tareas repetitivas sean más fáciles y consistente. Además la arquitectura flexible nos permite implementar cálculos a una o más CPU's o GPU's, actualmente este framework es usado por muchas compañías como Intel, Nvidia, UBER, etc.



FIGURA 3.3: TensorFlow

Capítulo 4

Conclusiones y Trabajo Futuro

ESTE CAPÍTULO ES UNO DE LOS MÁS IMPORTANTES, POR NO DECIR EL QUE MÁS. EN ÉL, EL JURADO VA A TENER CLARO QUÉ HA APRENDIDO EL ALUMNO Y CÓMO LO HA DESARROLLADO, LOS PROBLEMAS QUE HAN SURGIDO Y COMO LOS HA SOLUCIONADO... ADEMÁS DE QUE EL ALUMNO DEJARÁ CLARO QUE SE HA ESPECIALIZADO EN LA TEMÁTICA Y DEJARÁ EN ESCRITO TODO LO APRENDIDO Y COMO CONTINUARÁ CON LA TEMÁTICA EN POSTERIORES ESTUDIOS DEL MISMO TEMA

4.1. Conclusiones

- CONCLUSION 1: ASDFASDFASDFAS.
- CONCLUSION 2: ASDFASDFASDFAS.

Además de lo anterior

4.2. Trabajo Futuro

COMO SE VA A SEGUIR TRABAJANDO CON ESTA TEMÁTICA Y QUE FALTA POR DESARROLLAR. ADEMÁS SE ACONSEJARÁ SEGUIR UNA METODOLOGÍA PARA QUE LAS PERSONAS QUE QUIERAN SEGUIR TRABAJANDO ESTA TEMÁTICA

NASDFASDFASF

ADFASDFASDFA

NASDFASDFASF

ADFASDFASDFA

NASDFASDFASF

ADFASDFASDFA

Bibliografía

Apéndice A

Título del apéndice

Un ejemplo de los apendices