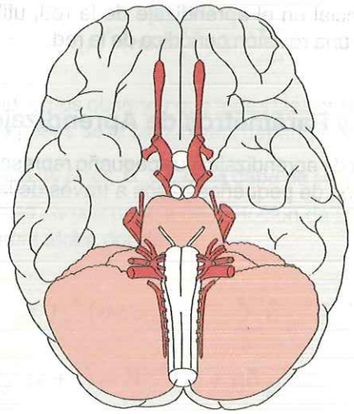
¿REDES NEURONALES HIBRIDAS?

ING. CESAR DARIO GUERRERO
Maestría en Ciencias Computacionales
E-mail: cguerrer@bumanga.unab.edu.co

Desde 1.956 cuando en la "Conferencia de Darmouth" se mostró al mundo informático la Inteligencia Artificial como una alternativa eficaz para la solución de problemas lo suficientemente complejos como para que los realicen las personas e inclusive los algoritmos tradicionales; los ojos de muchas universidades y centros de investigación como M.I.T., Stanford, Carnegie-Mellon, Laboratorios AT&T, Bell, entre otros, se volcaron hacia la profundización e innovación de nuevas alternativas de procesamiento de información.



Tal es el caso de Frank Rosenblatt de la Universidad de Cornell quien durante la década de los sesenta enfocó su estudio hacia una nueva forma de procesamiento fundamentada en los principios de organización y funcionamiento cerebrales, denominada Red Neuronal Artificial.

Basado posiblemente en investigaciones previas emprendidas en 1.943 por Warren S. MacCulloch y Walter H. Pitts quienes formularon el modelo de una red lógica compuesta por unidades elementales con características digitales, Rosenblatt planteó una red para aprendizaje de patrones llamada **perceptrón**. Dicha red consta de una sola neurona que recibe valores binarios (simulando señales inhibitorias o excitatorias propias de neuronas cerebrales) que son ponderados mediante un vector de conexiones (sinapsis cerebral) para finalmente generar mediante una función no lineal de tipo escalón una señal de salida que indique la pertenencia o no del patrón de entrada a una clase preestablecida de patrones.

Este modelo sin embargo, presentó serias limitaciones al no poder solucionar un problema tan elemental como la simulación de una compuerta OR exclusiva y fue así como empezaron a surgir severas críticas por parte de científicos de renombre como Minsky y Papert. A pesar de ello las investigaciones no cesaron. Hoy en día, las redes neuronales artificiales presentan una alternativa muy atractiva para la solución de diferentes problemas donde por un lado no existen reglas muy claras para el procesamiento de extensos volúmenes de información y por otro, porque entran a jugar muchas variables que lograrían resquebrajar cualquier esquema de programación tradicional.

Gracias a que van modificando su estructura a partir de ejemplos claves iniciales para obtener una respuesta deseada, podemos decir que poseen capacidad de aprendizaje. Además, la arquitectura y funcionamiento de la red presentes en sus diferentes canales de información, moldean una filosofía de procesamiento en paralelo.

Estas características y otras más han llevado con éxito a aplicarlas en problemas de reconocimiento y clasificación de señales, tratamiento de imágenes, modelamiento y en general, para resolver problemas de optimización.

Sin embargo una red neuronal como tal, presenta algunas limitaciones en su funcionamiento. Una de ellas radica en que para cada problema se debe dar un nuevo proceso de entrenamiento, perdiéndose de ésta manera el conocimiento adquirido anteriormente. Es por eso que la etapa de aprendizaje en la red, debe entenderse más como una síntesis algorítmica que continúa durante el entrenamiento y muere al llegar a la respuesta deseada de un problema específico. Todo lo contrario a lo que sucede en los sistemas expertos donde el aprendizaje no muere con el problema, sino que entra a formar parte de los heurísticos que ayudan a construir la base de conocimiento.

Otra importante desventaja se basa en el hecho de que la red es un sistema de caja negra, que se limita a dar una solución ante determinado problema, sin poseer reglas o procesamiento lógico que le permita razonar y justificar su respuesta ante el usuario. Los sistemas expertos a pesar de que no siempre llegan a una solución, cuando lo hacen tienen los mecanismos necesarios para indicar la forma en que se hizo, dando mayor seguridad y comprensión a dicho usuario.

Es por esto, que la Inteligencia Artificial se debe enmarcar dentro del enfoque sistemático donde no se trata de tener alternativas aisladas sino de buscar las posibles interacciones que pueden existir entre sus diferentes ramas en pro de la búsqueda de soluciones a problemas de diferente índole que puedan surgir del mundo real.

Dicho esto, resulta evidente pensar que las limitaciones impuestas a las redes neuronales en su estado natural, podrían ser ocultadas bajo el concepto de Red Neuronal Híbrida. De hecho, en los sistemas computacionales de la siguiente generación, conceptos como lógica difusa, redes neuronales, sistemas expertos y métodos de aprendizaje genéticos tendrán gran relevancia no sólo al nivel de software, sino en cuanto al mismo hardware se refiere. El abaratamiento de los diferentes recursos informáticos y la disponibilidad de altas tecnologías de integración electrónica permitirán la implantación de éstas metodologías aprovechando muy posiblemente las ventajas que presta el procesamiento en paralelo.

De esta forma, la red neuronal podría intercambiar información con el experto, que le permitirá por ejemplo dar un soporte a la justificación de sus respuestas o tomar futuras decisiones con base a los heurísticos previamente introducidos en la base de conocimiento. De igual forma, la potencia de la red como clasificador, permitirá identificar para los sistemas *"fuzzy"* una determinada función de membrecía. Por su parte, los métodos de aprendizaje genéticos suministrarían nuevas reglas a la base de conocimiento del experto. (Ver figura 1)

Figura 1. Esquema de procesamiento de una red neuronal híbrida que comparte Información con otras tecnologías de Inteligencia Artificial

Así, una red neuronal híbrida no sólo se refiere a la información que pueda obtener de otros esquemas computacionales sino también al aporte que ella pueda dar a estos en busca de la solución óptima a determinado problema.

"Divide y Vencerás", ésta será muy posiblemente la filosofía que tomarán los sistemas computacionales de futuras generaciones, fusionando para este fin arquitecturas de paralelismo real con nuevas tecnologías de procesamiento como las que proporciona la Inteligencia Artificial.

COMPUTACION PARALELA CONCEPTOS BÁSICOS

ING. FERNANDO ROJAS MORALES
MAESTRIA EN CIENCIAS COMPUTACIONALES
E- mail: frojas@bumanga.unab.edu.co

El paralelismo

El paralelismo se presenta en un sistema cuando se realizan varias tareas al mismo tiempo.

Un caso de la vida real:

Un hombre que conduce un automóvil debe hacerse cargo de lograr que la máquina automotor funcione correctamente en el aspecto mecánico, aunque no sepa de mecánica, es decir hacer el cambio de velocidad adecuadamente y presionar los pedales en la forma correcta para lograr que la máquina se desplace sin sobresaltos; a su vez debe orientar el volante, a derecha o a izquierda, avanzar o detener la máquina de acuerdo a sus intereses de desplazamiento; y a su vez debe estar pendiente de los elementos que se mueven en su entorno, los demás automotores, las señales de tránsito, los huecos que hay en el pavimento, y los transeúntes entre otros. Al tiempo algunos pueden estar manteniendo una conversación con sus tripulantes, escuchando música, ajustándose el cinturón de seguridad y saboreando un caramelo.

Podemos decir que existe procesamiento en paralelo en cuanto al hecho de que el conductor está realizando varias tareas a la vez sin embargo tenemos un sólo cerebro controlando todas las actividades.

Existe otro ejemplo un poco más patético: un grupo de trabajadores que levantan una construcción:

Primero algunos levantan las columnas de soporte, luego por grupos unos levantan el muro oriental mientras otros levantan el muro occidental, a la vez otros levantan el muro norte y otros el muro sur, finalmente todos se reúnen a trabajar en la postura del techo.

Factores que limitan el paralelismo

En los dos casos podemos ver algunos factores limitantes propios del procesamiento paralelo:

- ***Secuencialidad Intrínseca:*** Es decir algunas tareas solo podrán hacerse cuando otras que las preceden sean terminadas. En el caso uno, se debe dar arranque al motor antes de avanzar el auto. En el segundo caso la segunda línea de ladrillos solo se podrá poner después de haber puesto la primera.
- ***Concurrencia:*** Algunos recursos van a intentar ser utilizados a la vez. En el primer caso el conductor no puede mirar adelante y atrás a la vez. En el caso dos, dos grupos de obreros pueden intentar hacer uso del cemento o de las herramientas, a la vez.
- ***Sincronización:*** Ciertos grupos de tareas deben ser programados en el tiempo, tanto en el aspecto ya mencionado de la Secuencialidad como en el aspecto de maduración u obtención de respuesta. Para el caso primero: El conductor debe esperar a que el motor haga el embrague para poder hacer el cambio de velocidad, a pesar de que el auto puede estar avanzando. En el segundo caso, los obreros tendrán que esperar hasta que todos los muros estén terminados antes de poder poner el techo de la construcción.

Justificación para la construcción de máquinas paralelas

La necesidad del Computo Paralelo surge de las limitaciones de la arquitectura de las máquinas computacionales diseñadas por Von Newmann (en dos palabras, un procesador conectado a una memoria local), de las cuales se derivan prácticamente todas las computadoras de uso diario alrededor del planeta. El desarrollo del software se quedó rezagado con respecto al desarrollo del hardware, y específicamente el de las arquitecturas paralelas (máquinas computacionales que trabajan con varios procesadores), de esta manera emerge la necesidad de rehacer los programas que corren sobre computadoras convencionales (de un solo procesador) para lograr que lo hagan sobre máquinas de varios procesadores y de esta forma obtener un mejoramiento en el rendimiento. Citando los casos de ejemplo anteriormente mencionados, la pregunta que surge es ¿en computación cual es el equivalente de los obreros que construyen paredes por grupos y al mismo tiempo?

Se presenta un caso común de ejemplo:

La mayoría de programadores para realizar el ordenamiento de una matriz por filas construimos un algoritmo que ordene las filas una por una, con un procesador a disposición no tenemos una mejor opción, pero si disponemos de varios procesadores podríamos construir un algoritmo que realice el ordenamiento de cada una de las filas sobre cada uno de los procesadores y hacer esto a la vez. Un pequeño análisis matemático nos muestra como mejora el rendimiento. Si la matriz contiene 256 filas, el ordenamiento que se hace utilizando el algoritmo convencional que corre sobre máquinas de un solo procesador gastará 256 unidades de tiempo. Pero si hacemos este ordenamiento sobre una máquina paralela que tenga 256 procesadores, el tiempo será aproximadamente 256 veces menor. Se origina entonces una nueva pregunta: ¿cómo se justifica la inversión en una máquina con 256 procesadores, cuyo costo es mucho mayor que el de una máquina con un solo procesador?

Los grandes retos que la computación paralela enfrenta y que justifican su inversión, la construcción de máquinas paralelas y el desarrollo de algoritmos que trabajen sobre estas, se encuentran en diversos campos de la ciencia que requieren de altos volúmenes de procesamiento, entre los cuales se mencionan los siguientes:

- Reducción de ruido en grabaciones magnéticas
- Diseño de medicinas
- Diseño de jets
- Mejoramiento en el diseño de motores de combustión
- Modelamiento del océano
- Tratamiento de imágenes en el área de la médicina, tomografía
- Modelamiento de la contaminación
- Estudio de la estructura de proteínas
- Imágenes sintéticas
- Apoyos educativos

Como dijimos anteriormente las máquinas que se han estado utilizando hasta hoy poseen un procesador y una memoria local, esto a su vez requiere de un canal o bus de comunicación, que al congestionarse aminora el desempeño de la máquina. Una de las primeras modificaciones es por tanto incrementar el número de buses. Pero este sencillo cambio representa rediseñar los programas, es decir un nuevo paradigma de programación: la programación en paralelo.

Clases de Arquitectura Paralela

El desarrollo de las máquinas paralelas se ha incrementado en la década de los 90. Es así como hoy por hoy, los fabricantes están compitiendo hacia el logro de lo que se denomina el desempeño 3T o la carrera hacia el teraflop, que consiste en:

- 1 Teraflop de velocidad de procesamiento
- 1 Terabyte de RAM
- 1 Terabyte/seg de I/O[1]

Existen varias clases de máquinas paralelas, que se caracterizan de acuerdo a su diseño arquitectónico de la siguiente manera:

Por la manera como administran la memoria:

- ***Máquinas que utilizan memoria compartida:*** las que utilizan un sistema global de memoria, y
- ***Máquinas que utilizan memoria distribuida:*** en las que cada procesador cuenta con su propia memoria local.

Por la manera como controlan sus procesadores:

- ***Máquinas centralizadas (SIMD[2]):*** que poseen un procesador que controla a los demás.
- ***Máquinas distribuidas (MIMD[3]):*** en la cual cada procesador puede realizar tareas diferentes y no existe un procesador que los coordine.

Máquinas con memoria compartida

Estas máquinas también se denominan **multiprocesadores**. Para evitar que el uso de la memoria se convierta en una región crítica y se genere un problema mayor que el que se está atacando la memoria se divide en bancos de memoria con una red de interconexión entre ellos.

Existen varios diseños de acuerdo a la **localización** de la memoria:

- ***UMA[4]:*** Es un sistema en el cual para que los procesadores acceden un banco de memoria se requiere la misma cantidad de tiempo. (Ver Figura 2.)

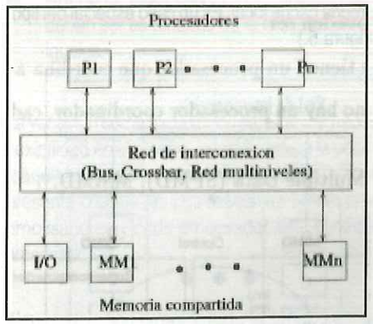


Figura 2. Modelo UMA

[1] 1 Tera = 10^{12}. 1 Terflop son 10^{12} operaciones/seg
[2] Single Instruction stream and Multiple Data streams
[3] Multiple Instruction streams and Multiple Data streams
[4] Uniform Memory Access