Tomorrow's Computational Innovation

innovación computacional del mañana

1st Daniel Alejandro Perez Altamar
Escuela de ingeniería de sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
dapeal42@gmail.com

2nd Duvan Fernando Pinto Diaz
Escuela de ingeniería de sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
duvanpidi@hotmail.com

3rd Edinson Camilo Porras Melgarejo
Escuela de ingeniería de sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
camiloporras98@gmail.com

4th Jean Carlos Portilla Mora
Escuela de ingeniería de sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
jeancarlosportilla@hotmail.com

5th Sebastian Cardenas Acevedo
Escuela de ingeniería de sistemas
Universidad Industrial de Santander
Bucaramanga, Colombia
Cardenassebastian10@gmail.com

Abstract

In the architecture of computers they infer several elements that affect the profitability of the machine, these in turn play a very important role in the design since they allow to analyze the performance, the cost, the operation that it handles, its use, towards whom it goes oriented or simply improve the capabilities of the machines that are handled today; For this, machine learning allows the design of new programs that improve the performance and processing capacity of the information a computer has; This article will discuss the benefits of implementing this scientific discipline in the architecture of the new computers that are to be established in the global market.

En la arquitectura de computadores infieren varios elementos que afectan en la rentabilidad de la máquina, estos a su vez juegan un papel muy importante en el diseño ya que permiten analizar el rendimiento, el coste, la operativa que maneja, su uso, hacia quien va orientada o simplemente mejorar las capacidades de las maquinas que se manejan en la actualidad; para esto último el machine learning permite diseñar nuevos programas que mejoren el rendimiento y la capacidad del procesamiento de la información que tiene una computadora; en este artículo se hablara sobre los beneficios que trae implementar esta disciplina científica en la arquitectura de las nuevas computadoras que están por establecerse en el mercado global.

Index Terms

ML (Machine Learning),IA (Inteligencia Artificial),Deep learning, PCA (Principal Component Analysis),K-means clustering, Unsupervised learning, Regresión y clasificación, procesador IPC(instructions per cycle), Semi-supervised learning, troughput, SVM(Support vector machine), Arboles de decisión, Redes bayesianas.

I. Introducción

En los tiempos modernos los seres humanos siempre se hacen la típica pregunta de ¿Puede las cosas mejorar más de lo que ya están?, en el paso de los años el hombre ha hecho avances importantes para poder mejorar de la calidad de vida en cada época, como por ejemplo desde la prehistoria el hombre buscaba fuego para poder disfrutar sus alimentos, iluminarse de noche y alejar a depredadores. En esta época estamos rodeados de la tecnología que es suministrado por diferentes máquinas para mejorar nuestra vidas desde los computadores que hacen nuestra vida supremamente sencilla facilitándonos tareas de la vida diaria como realizar cuentas, documentos como cartas, informes, presentaciones y mucho más, además de hacer que la gente se sienta a gusto e incluso entretenerse con actividades de ver series y películas, jugar uno que otro videojuego, o navegar en redes sociales haciendo diferentes tareas para poder comunicarse con amigos, dar opiniones, mostrar fotos, etc.

También se han creado supercomputadores para ya tareas complejas que van de la mano con empresas multinacionales para el ámbito profesional, usadas en la NASA para trabajos espaciales por lo que es una arquitectura muy especial, y hasta esto se ha podido llegar las ideas del hombre, a tal nivel de crear maquinas con su propia inteligencia, lo que simplemente conocemos en el mercado tecnológico como inteligencia artificial (IA) que ha entrado en el mercado mundial por medio de programas, aplicaciones, maquinas, robots e incluso gracias empresas como Samsung hoy en día se está trabajando en el mundo de los androides, algo que caracteriza a este campo es que muchos de estos resultados, salvo los humanoides o androides de las compañías han sido especialmente realizadas para una tarea específica ya sea caminar, mover cosas, escribir, hablar, repetir, etc. pero si este realiza otra actividad diferente a la que ha sido asignado va a resultar en rotundo fracaso, estas máquinas las denominan a pertenecientes a la inteligencia artificial débil, los trabajos

que realizan estas máquinas son aquellas denominadas pertenecientes a la inteligencia artificial fuerte que lastimosamente todavía no han salido a la luz aunque se han planteado diferentes ideas y proyectos que tiene como objetivo crear una máquina que sea multitarea y las realice por sí mismo, y se esperan en los próximos años para abrirse a la tecnología del futuro.

En este artículo hablaremos de como imaginamos los computadores del futuro ya habiendo mencionado un poco de la inteligencia artificial que igualmente se explicará mejor en el marco teórico, vamos a relacionar o imaginar cómo sería un computador y que arquitectura tendría que tener con diferencia a las PC de hoy en día, y por eso miraremos comportamientos que se han visto en aplicaciones u otras máquinas que en estos días están funcionando y pueden ser fundamentales para ver equipos más innovadores y que puedan venderse en gran cantidad al público en general y genere satisfacción a los usuarios y sean implementados exitosamente en la maquinas del futuro.

Seguramente habrán escuchado acerca de cómo una maquina pueda aprender procedimiento por medio de algoritmos, pues eso es el Machine learning, este hace parte o es un fragmento de la inteligencia artificial que se centra en especializar a los equipos a aprender, en sí, son usados últimamente en algunos procesos de seguridad, como con reconocimiento de rostros, pero el ML no solo se enfoca en que la maquina aprenda a reconocer un rostro, este va más allá, busca que la maquina en si sepa y aprenda acerca de cómo es una cara y pueda incluso identificar rasgos de la cara y decirnos quien es por medio de una base de datos almacenados acerca de rasgos similares, y que pueda identificar esta caras en cualquier posición que se encuentre y decirnos de quien es la cara que se escaneó y que información se sabe de aquel persona, bueno pero ya llegando a un aspecto más del común y más de entretenimiento, muchos de los usuarios han usado aplicaciones de edición de imágenes en la cámara tal como la muy conocida Instagram se han visto en sus diferentes historias con los usuarios el uso de plantillas para cara o en el mundo de las redes conocido popularmente como filtros, estos editaran lo que se esté grabando en ese instante, se usa procedimiento para identificar las caras que este dentro de la cámara por medio de un algoritmo que se haya usado para el aprendizaje del rostro humano y al ser reconocido este se hace la realización del filtro de la imagen mostrando el filtro usado por encima de la cara del usuario simulando ser otra persona, especie o cosa. También es usado el ML en Facebook para aquellas imágenes que se postean en el perfil de cada usuario, se identifica un rostro que este en esa imagen y puedas decir quién es, entonces te manda la opción de etiquetar a esa persona con la que estas para que también vea esa foto y reaccione a esta, pues en si eso es lo que busca el ML que aprenda a identificar cosas del mundo en el que estamos.

Fig. 1. Lectura de rostro por medio de algoritmos en redes sociales.

Ha sido el avance con la implementación de ML que gracias a este hay maquinas que pueden sorprendentemente reconocer nuestras propias voces pero no solo eso, las maquinas están preparadas para que entienda lo que les decimos y puedan obedecer a una orden, respondernos una pregunta que nosotros le planteemos o lo más sorprendente mantener una conversación con nosotros por medio de una codificación de lo que se extrajo en lo que le dijimos, procesa el lenguaje binario y lo almacena en una memoria en donde por medio al aprendizaje que este posee o las diferentes alternativas que se aprendió devuelve un binario de respuesta que se vuelve a codificar en un comando de voz en donde nos muestra su respuesta y finalmente nos la comunica, de esto hay muchos ejemplos en el mercado como Alexa, siri, google asistant, etc. Pero para todas estas tareas complejas de aprendizaje están dentro de una sub-rama del ML que es el aprendizaje profundo o Deep learning que nos ayuda a estas tareas más complejas, y ´poder innovar con ayuda de centros de datos masivos como la big data, eso buscaremos como se pueden hacer dispositivos o computadores con esta práctica para así crear los computadores del mañana.

Fig. 2. Asistentes de voz mas conocidos en el mercado.

II. Marco teórico

La evolución de los computadores viene dada por los continuos cambios en la perspectiva de la fabricación de los mismos y de las nuevas tecnologías que han ido surgiendo con el paso del tiempo, algunos de estos cambios han sido el incremento de la velocidad del procesador, la disminución de componentes y tamaño, así como el aumento en la memoria. La historia de esta evolución viene desde los tubos de vacío, los transistores y finalmente encontramos los circuitos integrados.

Con la aparición de los primeros computadores analógicos y discretos electromecánicos en 1938 y 1939, y posteriormente electrónicos en 1946, se marca el inicio de la primera generación de computadores. Los relés electromecánicos son usados como dispositivos de conmutación durante los años 40 y posteriormente son reemplazados por las válvulas de vacío (bulbos) durante los años 50. Además de los elementos de conmutación usados, estos equipos se caracterizaban por estar interconectados por cables aislados. La primera computadora conocida se le llamo ENIAC, esta surgió por las necesidades militares durante la segunda guerra mundial, pesaba toneladas, además era una maquina decimal no binaria. En el año 1946 surgió un nuevo computador de programa almacenado llamado IAS, así como a su vez surgieron los primeros computadores comerciales como la UNIVAC I Y II. La unidad de control como la ALU contiene procesadores de almacenamiento llamados registros: Registro temporal de memoria (MBR) 40 bits, de dirección de memoria (MAR), registro de instrucción (IR) 8 bits, registro temporal de instrucción (IBR) 20 bits, contador de programa (PC) 12 bits, acumulador (AC) 40 bits y el multiplicador cociente (MC) 40 bits, la cual algunas de sus instrucciones son la transferencia de datos, salto incondicional, salto condicional, aritmética, modificación de instrucciones, entre otras; la estructura del procesador era bit-serie, lo que obligaba a que la aritmética se efectuará bit a bit y sin punto flotante. En estos computadores sólo se empleaba el lenguaje máquina codificado en binario. La segunda generación de computadores viene acompañada de los transistores, estos a su vez son cada vez más pequeños, disipan menos calor, es de estado sólido hecho de silicio. En esta etapa se añadieron funciones lógicas y aritméticas y unidades de control más complejas, surgió el uso de un lenguaje de programación de alto nivel y la proporción de un software del sistema en el computador; los lenguajes ensambladores siguieron utilizándose hasta la aparición de lenguajes de alto nivel como el Fortan (1957), Cobol (1959) y Algol (1960). En esta etapa también hacen su aparición los primeros circuitos impresos. En 1957 aparece DEC, con la cual surgen el desarrollo de microprocesadores. Esta computadora usa canales de datos, utiliza un multiplexor que es el punto de conexión de los canales de datos, la CPU y la memoria; los equipos electrónicos estaban compuestos de componentes discretos (Transistores).

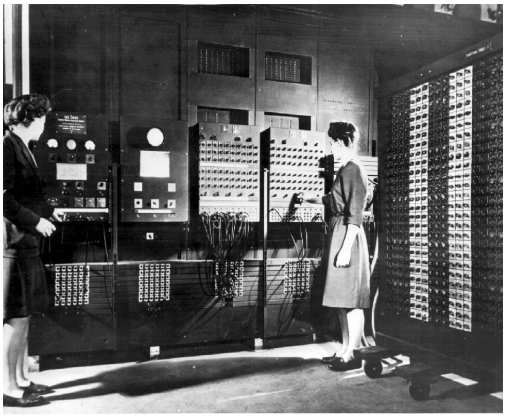


Fig. 3. Proyecto ENIAC.

En 1958 surgió la tercera generación de computadoras, Con la aparición de los circuitos integrados de pequeña escala (SSI, Small-Scale Integration) y su posterior utilización se marca el inicio de esta generación. Los elementos básicos de un computador digital tienen como prioridad el almacenamiento, procesamiento y control de funciones, tiene dos componentes principales: puertas y celdas de memoria; las puertas cumplen la función de controlar el flujo en cierta manera, la celda de memoria es un dispositivo que puede almacenar un dato de información en un bit. En estas se tienen cuatro funciones básicas; el almacenamiento de datos, el procesamiento de los mismos, la trasferencia y control de los datos de información. Podemos deducir que la computadora consta de puertas, celdas de memoria e interconexiones entre estos elementos; las puertas y celdas de memoria están constituidas por componentes electrónicos simples.

Gordon Moore, cofundador de Intel, en 1965 anticipo que la complejidad de los circuitos integrados se duplicaría

cada año con una reducción de costo conmensurable; señalo también en el año 1975 que el número de transistores en un solo chip se duplica cada dos años y hasta la fecha se sigue cumpliendo tal señalamiento. Moore además de proyectar como aumenta la complejidad de los chips (medida por transistores contenidos en un chip de computador), la ley de Moore sugiere también una disminución de los costos. A medida que los componentes y los ingredientes de las plataformas con base de silicio crecen en desempeño se vuelven exponencialmente más económicos de producir, y por lo tanto más abundantes, poderosos y transparentemente integrados en nuestras vidas diarias. Los microprocesadores de hoy se encuentran en todas partes.

Los computadores de la presente generación comenzaron haciendo uso de tecnología LSI (Large-Scale Integration) y con los avances en el desarrollo de circuitos integrados de alta densidad hoy en día hacen uso de circuitos VLSI (Very Large-Scale Integration). Los lenguajes de programación se han expandido para manejar y expresar diferentes estructuras y conceptos temporales y espaciales. Los computadores comerciales hacen uso de un alto grado de multiprocesamiento a través de varios procesadores y segmentación encauzada para obtener incrementos substanciales de rendimiento y capacidades de cómputo. A inicios de la década de los 80 el concepto del procesamiento en paralelo masivo hace su aparición En las últimas generaciones de computadores se basan en avances de la tecnología de los circuitos integrados, como la memoria semiconductora. Un chip de memoria semiconductora puede llegar a contener millones de pequeños transistores o condensadores. En las computadoras modernas, la memoria principal consiste básicamente en memoria de semiconductor volátil y dinámica, a la que se le conoce como memoria dinámica o mas comúnmente como RAM. Los microprocesadores es un circuito integrado independiente; los microprocesadores modernos incorporan hasta diez millones de transistores, además de otros componentes como resistencias, diodos, condensadores y conexiones. La función básica de este microprocesador es unir todos los circuitos integrados que contenían a su vez transistores en un solo paquete, estos microprocesadores eran capaces de desarrollar todas las funciones de la unidad central de proceso. El desarrollo del microprocesador permitió la creación de los ordenadores personales (PC), este fue un concepto revolucionario que marco un cambio en la forma de trabajar para muchas personas.

Se desarrollaron los chips de Intel 8008 y 8080, de Zilog, el Z80 y de Motorola, el 6800. De esta generación cabe destacar tres grandes momentos: La aparición del kenbak I, los discos Winchister,y el 8080. Los discos duros Winchister se comercializaron a partir del año 1973, por IBM en los medlos 3340. Estos estaban provistos de un pequeño cabezal de escritura/lectura con un sistema de aire el cual cumple la función de permitirle movilizarse muy cerca de la superficie del disco. El 8080 fue el primer CPU creado por Intel en el año 1974, esta contenía 4500 transistores y podía manejar 64k de memoria RAM a través de un bus de datos de ocho bits. El 8080 fue el cerebro del primer ordenador personal de Altair, el cual provoco que se generase un interés por adquirirlo en hogares y pequeños negocios a partir del año 1975. Las siglas PC son el acrónimo de Personal Computer. El pc es un dispositivo electrónico que recibe un conjunto de instrucciones y las ejecuta realizando cálculos sobre los datos numéricos, o bien compilando y correlacionando tipos de información. En julio de 1980 IBM comenzó a desarrollar su propio ordenador personal al cual le llamaron IBM/PC, en agosto de ese mismo año se inició formalmente el desarrollo del primer prototipo con nombre código Acorn. Para el IBM/PC se eligió un procesador Intel, mas especifico el 8088 el cual tenía un bus de ocho bits y una estructura interna de 16 bits, esto con el fin de asegurarse que este nuevo equipo no compitiera con otros modelos de la empresa, ya que existía otro procesador con un bus de 16 bits. Al hallar el microprocesador ideal para el ordenador, fueron en busca de un sistema operativo adecuado, entonces se decidió agregarle un nuevo sistema operativo conocido como MS-DOS. El 12 de agosto de 1981 IBM lanzo al mercado el personal computer (IBM/PC), el cual poseía un microprocesador 8088, 16k de RAM, ampliable hasta 256k y una unidad de disquetes de 160k y una pantalla verde monocromática.

Fig. 4. Ordenador Kenbak 1.

Cada vez se hace más difícil la identificación de las generaciones de computadoras, porque los grandes avances y nuevos descubrimientos ya no nos sorprenden como sucedió a mediados del siglo XX. Hay quienes consideran que la cuarta y quinta

generación han terminado, y las ubican entre los años 1971-1984 la cuarta, y entre 1984-1990 la quinta. Ellos consideran que la sexta generación está en desarrollo desde 1990 hasta la fecha. Siguiendo la pista a los acontecimientos tecnológicos en materia de computación e informática, se puede puntualizar algunas fechas y características de lo que podría ser la quinta generación de computadoras. En base en los grandes acontecimientos tecnológicos en materia de microelectrónica y computación (software) como CADI CAM, CAE, CASE, inteligencia artificial, sistemas expertos, redes neuronales, teoría del caos, algoritmos genéticos, fibras ópticas, telecomunicaciones, entre otros, a de la década de los años ochenta se establecieron las bases de lo que se puede conocer como quinta generación de computadoras. Como se dice desde hace vario tiempo la sexta generación de computadoras está en marcha desde principios de los años noventa, se debe por lo menos, conocer las características que deben tener las computadoras de esta generación. También se debe recordar de los avances tecnológicos de la última década del siglo XX y lo que se espera lograr en el siglo XXI. Las computadoras de esta generación cuentan con arquitecturas combinadas Paralelo / Vectorial, con cientos de microprocesadores vectoriales trabajando al mismo tiempo; se han creado computadoras capaces de realizar más de un millón de millones de operaciones aritméticas de punto flotante por segundo (teraflops); las redes de área mundial (Wide Area Network, WAN) seguirán creciendo desorbitadamente utilizando medios de comunicación a través de fibras ópticas y satélites, con anchos de banda impresionantes. Las tecnologías de esta generación ya han sido desarrolla das o están en ese proceso. Algunas de ellas son: inteligencia / artificial distribuida; teoría del caos, sistemas difusos, holografía, transistores ópticos, entre otras.

Debido al aumento de la capacidad y al abaratamiento de las tecnologías de la información y de los sensores, se puede producir, almacenar y enviar más datos que nunca antes en la historia. De hecho, se calcula que el 90 por ciento de los datos disponibles actualmente en el planeta se ha creado en los últimos dos años, produciéndose actualmente en torno a 2,5 quintillones (2.500.000.000.000.000.000) de bytes por día, siguiendo una tendencia fuertemente creciente. Estos datos alimentan los modelos de Machine Learning y son el impulso principal del auge que esta ciencia ha experimentado en los últimos años. Machine Learning es uno de los sub-campos de la Inteligencia Artificial y puede ser definido como: La ciencia que permite que las computadoras aprendan y actúen como lo hacen los humanos, mejorando su aprendizaje a lo largo del tiempo de una forma autónoma, alimentándolas con datos e información en forma de observaciones e interacciones con el mundo real; Los tipos de Machine Learning que se tratarán en esta serie son: Aprendizaje supervisado, Aprendizaje no supervisado y Aprendizaje profundo. El aprendizaje supervisado se refiere a un tipo de modelos de Machine Learning que se entrenan con un conjunto de ejemplos en los que los resultados de salida son conocidos. Los modelos aprenden de esos resultados conocidos y realizan ajustes en sus parámetros interiores para adaptarse a los datos de entrada. Una vez el modelo es entrenado adecuadamente, y los parámetros internos son coherentes con los datos de entrada y los resultados de la batería de datos de entrenamiento, el modelo podrá realizar predicciones adecuadas ante nuevos datos no procesados previamente. En el aprendizaje no supervisado, se maneja con datos sin etiquetar cuya estructura es desconocida. El objetivo es la extracción de información significativa, sin la referencia de variables de salida conocidas, y mediante la exploración de la estructura de dichos datos sin etiquetar. El aprendizaje profundo o Deep Learning, es un sub-campo de Machine Learning, esta usa una estructura jerárquica de redes neuronales artificiales, que se construyen de una forma similar a la estructura neuronal del cerebro humano, con los nodos de neuronas conectadas como una tela de araña. Esta arquitectura permite abordar el análisis de datos de forma no lineal.

Debido a nuevas tecnologías de cómputo, hoy en día el machine learning no es como el del pasado. Nació del reconocimiento de patrones y de la teoría que dice que las computadoras pueden aprender sin ser programadas para realizar tareas específicas; investigadores interesados en la inteligencia artificial deseaban saber si las computadoras podían aprender de datos. El aspecto iterativo del machine learning es importante porque a medida que los modelos son expuestos a nuevos datos, éstos pueden adaptarse de forma independiente. Aprenden de cálculos previos para producir decisiones y resultados confiables y repetibles. Es una ciencia que no es nueva pero que ha cobrado un nuevo impulso. Aunque muchos algoritmos de aprendizaje basado en máquina han estado entre nosotros por largo tiempo, la posibilidad de aplicar automáticamente cálculos matemáticos complejos al big data una y otra vez, cada vez más rápido es un logro reciente. El Machine Learning llevará a repensar, reestructurar y considerar nuevas posibilidades para las nuevas máquinas que se quieren diseñar. A medida que esta tecnología se extienda, será esencial construir intencionalmente la experiencia del usuario para que el usuario controle esa tecnología y no al revés. Es, por tanto, una tendencia que se debe tener bajo el radar de los diseñadores de computadores.

Fig. 5. Ordenador actual, MAC.

III. Estado del arte

La aplicación del Maching Learning en el campo de la arquitectura informática se encuentra actualmente en sus etapas iniciales, con los pocos estudios exploratorios que muestran una promesa impresionante. Recientemente, se han realizado estudios pioneros sobre la aplicación de Maching Learning / deep learning a la programación de la CPU. En los trabajos, el programa utiliza predictores de rendimiento de la red neuronal artificial para mejorar el rendimiento del sistema en un programador basado en Linux en más del 30%. Otros enfoques para utilizar el machine / Deep Learning para la programación han sido clasificar las aplicaciones, así como identificar los atributos del proceso y el historial de ejecución de un programa. Este es el enfoque de que utilizó árboles de decisión para caracterizar programas completos y personalizar segmentos de tiempo de CPU para reducir el tiempo de respuesta de la aplicación al disminuir la cantidad de intercambios de contexto. El trabajo presentado en estudios utilizando valores de precisión de similitud estructural y máquinas de vectores de soporte y regresión lineal para predecir el rendimiento del hilo en diferentes tipos de núcleos a un nivel de granularidad alto (1 segundo).

En ese estudio, los tiempos de explosión de CPU de trabajos completos para cuadrículas computacionales se estiman utilizando un enfoque de Maching Learning. Realizan un enfoque que utiliza el Maching Learning para seleccionar si ejecutar una tarea en una CPU o GPU según el tamaño de los datos de entrada. Se propone un algoritmo que utiliza el aprendizaje por refuerzo para maximizar el IPC agregado normalizado. Demuestran la necesidad de una asignación central equilibrada, pero no proporcionan una implementación. Para la predicción de rama, se propuso utilizar un predictor basado en perceptrón para mejorar el rendimiento de la CPU. Varios estudios han aplicado el Maching Learning con el propósito de administrar la memoria caché. Algunos autores proponen el aprendizaje del perceptrón para la reutilización de predicciones y también presentan un método de predicción para la reutilización futura de bloques de caché utilizando diferentes tipos de parámetros. La predicción del comportamiento de la caché L2 se realiza mediante el Maching Learning con el fin de adaptar un planificador de procesos para reducir la contención de L2 compartida.

El resurgimiento reciente en la investigación de IA se atribuye, al menos en parte, a las capacidades mejoradas de procesamiento. Estas mejoras se ven reforzadas por optimizaciones de hardware que aprovechan el paralelismo disponible, la reutilización de datos, la dispersión, etc. en los algoritmos de ML existentes. En contraste, ha habido un trabajo relativamente limitado aplicando ML para mejorar el diseño arquitectónico, y la predicción de ramas es uno de los pocos ejemplos principales. Este trabajo incipiente, aunque limitado, presenta un enfoque auspicioso para el diseño arquitectónico. Este artículo presenta una visión general de ML aplicada al diseño y análisis arquitectónico.

IV. Desarrollo

La idea de machine learning esta fundamentada en hacer predicciones o decisiones sin una programación explicita, hace uso de esta idea por extracción de patrones en los datos.
La denominada ley de Moore donde se especifica la cantidad de transistores en un procesador a lo largo de los años se encuentra de cierto modo estancada, pues alcanzo un punto donde su crecimiento se ha disminuido esto da pie para que arquitecturas que reemplacen dicha ley emerjan, es de destacar que las capacidades y posibilidades de machine learning van en un aumento significativo, es aquí en esta colisión de información donde se halla la posibilidad de utilizar la arquitectura de computadores para mejorar el impacto de machine learning y de la misma manera se utilice machine learning para el desarrollo de la arquitectura de computadores.

El concepto de Machine Learning tiene funciones tanto en aplicaciones medicas como hasta autos que se manejan solos a continuación se planteara una serie de usos desarrollados con ML en la arquitectura de computadores se vera un poco los métodos empleados para llegar a ciertos resultados, y las mejoras en rendimiento en sus respectivos aspectos.
La exploración espacial de diseño de GPU a demostrado resultados favorables cuando se trata con ML, tomando variables como la frecuencia central o la frecuencia de memoria utilizando modelos de redes neuronales se posible obtener una mejora de 19.3 % en el rendimiento del sistema.
Dentro de la predicción multiplataforma se ha demostrado optimizaciones del 91 % utilizando algoritmos de regresión para predecir el rendimiento de GPU según el segmento del código con un error limitado al 11.6 %, posteriormente a este desarrollo se probó el bosque aleatorio (random forest) como algoritmo de clasificación que concluyo con un 94 % de precisión.
Para la predicción y clasificación de GPU's surgió una metodología donde se usan las generaciones anteriores para el desarrollo de generaciones futuras, se concluyó que entre más variables(aleatorias) sean las muestras entre ellas, mejores son los resultados, de esta manera lograron modelos con errores por debajo del 10 %, también se llevo a cabo el uso de redes neuronales para identificar patrones de trafico de la GPU, obteniendo igualmente resultados favorables con una confianza del 94 %.

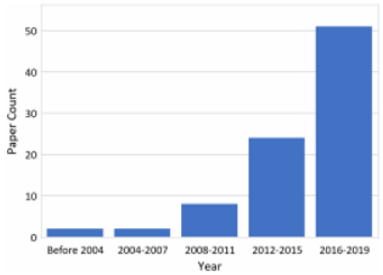


Fig. 6. Publicaciones sobre la aplicación de ML a la arquitectura de los computadores.

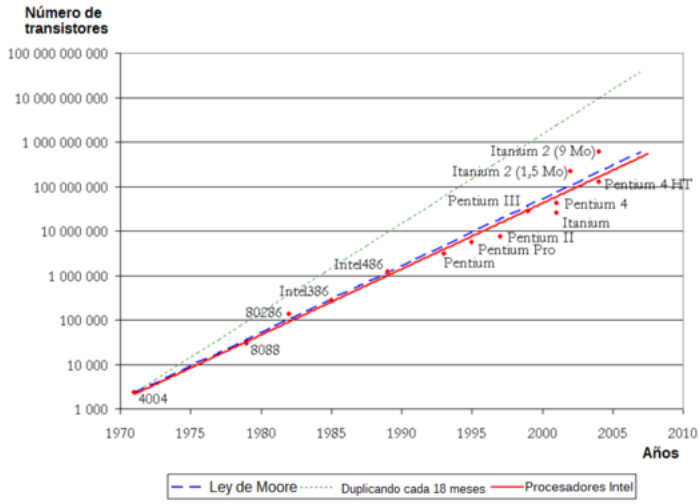


Fig. 7. Ley de Moore.

Mejorar la memoria cache entra dentro de los parámetros del machine learning, dicho método logra “aprender” las complejidades propuestas por el rendimiento del cache y mejorar el mismo.
Observando ventajas de rendimiento sobre el enrutamiento de la ruta mas corta, intensidad, estimaciones de nodos vecinos, redes con tiempo de entrega, topologías, con la ayuda de implementación de métodos como el Q-learning es posible maximizar el rendimiento de múltiples saltos y mejorar la equidad al reducir imprevistos. El uso de redes neuronales redujo el tiempo de ejecución un 17.8 % utilizando modelos con precisiones superiores al 90 %.
Asimismo, gracias al machine learning es posible entrenar un árbol de decisión para predecir fallas en un centro de operación de redes, utilizando parámetros como pueden ser; la temperatura, utilización y desgaste de dispositivos; gracias a esto se obtuvo mejoras sustanciales en la latencia que mejoro un 32%, la eficiencia energética con mejoras del 67% y la fiabilidad cuyo tiempo de medio de falla incremento un 77% en comparación a como se llevaba trabajando regularmente.

Mencionemos ahora dos aplicaciones más de machine learning inicialmente ML en línea, abordemos en que sucede cuando los tiempos de ejecución deben ser “reales”, aquí encontramos problemas como puede ser la sobrecarga de procesamiento de energía, área y el ya nombrado tiempo real, también es de destacar ML para la implementación arquitectónica, es decir mejorar tareas como el diseño o la simulación o dicho de otro modo ML fuera de línea.

Gracias al concepto de ML en línea es posible beneficiarse de la adaptabilidad a las características de carga de trabajo en tiempo de ejecución sin embargo con un modelo sin una precisión confiable el rendimiento del sistema se ve afectado negativamente es por ello por lo que se pueden considerar adaptaciones al control y al aprendizaje para evitar estos impactos perjudiciales.
Es así como surge la idea de “shadow operations” u operaciones sombra al español, el cual retroalimentación sobre la acción sin afectar negativamente al sistema, esto se logra gracias a la aplicación de redes neuronales que desarrolla una tarea de control basado en el aprendizaje supervisado, en la mayoría de los trabajos ML reemplaza los enfoques existentes, sin embargo se ha demostrado que combinando el enfoque tradicional con la implementación de ML se obtienen ventajas significativas esto debido a las capacidades de predicción y toma de decisiones de los dos enfoques, lo que permite una sincronización de

mejoras en el rendimiento, es así entonces como se reduce el costo de implementación de ML.

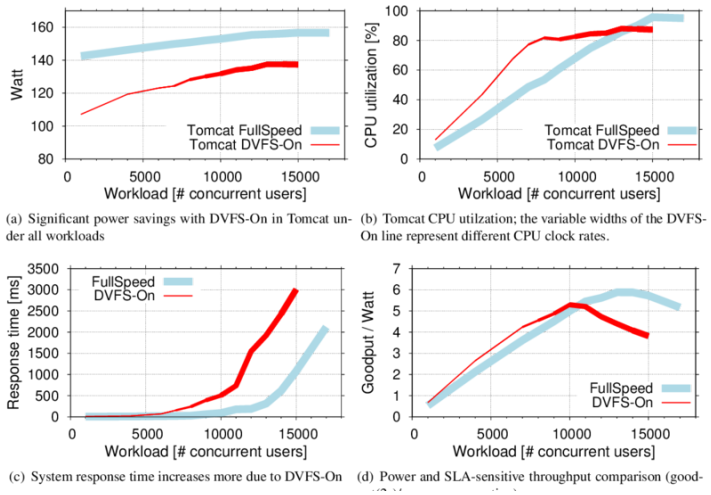
Ahora la utilidad de los modelos fuera de línea de ML están en gran medida ligados a los enfoques tradicionales de diseño es por ello que su enfoque principal es mejorar la eficiencia de los datos y la precisión del modelo, gracias al desarrollo de redes neuronales han surgido optimizaciones para mejorar esta eficiencia, una propuesta indica que al filtrar valores de rendimiento o predicciones de potencia se destaca que algunos modelos pueden ser muy útiles en ciertas aplicaciones pero muy débiles en otras, esto se soluciono tomando una muestra del 60% de los modelos mas cercanos a la media.

Dada la naturaleza de las aplicaciones de ML muchas veces es posible pasar por alto las características que tiene la tarea y centrarse únicamente los datos, pudiendo llegar a perder el conocimiento de dominio adicional que podría mejorar la interpretabilidad o el rendimiento general del modelo, en algunas aplicaciones, este conocimiento puede ayudar a identificar comportamientos fallidos y, de nuevo, mejorar la utilidad general del modelo.

Dicho todo esto es correcto plantearnos el futuro del área, hacia donde nos esta llevando las nuevas metodologías, las estrategias de implementación o la capacidad de aprendizaje y hacia dónde queremos dirigirla.

El análisis de aplicaciones utilizando bloques básicos ha sido durante mucho tiempo un método útil para la simulación, la predicción a nivel de fase ofrece beneficios análogos para el ML aplicado a la arquitectura, concretamente, el trabajo futuro podría explorar predicciones para el control y la reconfiguración del sistema en función del comportamiento a nivel de fase, en lugar de ventanas estáticas o conducta a nivel de aplicación.

ML para la mejora del rendimiento de la energía (DVFS específicamente) demuestra cambios muy rápidos en el consumo de energía, del orden de las instrucciones de 1000 para algunas aplicaciones, el inconveniente viene en la explotación de estos intervalos pues requieren de una cuidadosa consideración tanto en el modelo como en el algoritmo, es por eso que se plante que a futuro los modelos y algoritmos pueden ser optimizados o considerar enfoques mas adecuados para nuevos modelos.
Modelos cada vez más y más complejos requieren de nuevas técnicas para reducir gastos y costos; dos técnicas efectivas con



(a) Significant power savings with DVFS-On in Tomcat under all workloads

(b) Tomcat CPU utilzation; the variable widths of the DVFS-On line represent different CPU clock rates.

(c) System response time increases more due to DVFS-On

(d) Power and SLA-sensitive throughput comparison (good-

Fig. 8. Comportamiento DVFS en ordenador.

beneficios demostrados en aceleradores, la poda de modelos (pruning model) y el peso de cuantización (weight quantization). Aunque también se están explorando muchos otros enfoques. Si tenemos un modelo con una complejidad elevada en ML puede representar un problema en las aplicaciones de este, en el caso de las redes neuronales se requiere almacenamiento de peso de red y capacidades de procesamiento adicionales, es así como estudios con redes neuronales nos permiten observar las mejoras en la complejidad por el método de la poda, la idea general de este método es “podar” las conexiones que puedan tratarse como innecesarias, este método demuestra poca dispersión en los datos y una pequeña cantidad en su error.

Cuando hablamos del peso de cuantización nos referimos a los valores de estado en Q-learning (algoritmo de aprendizaje reforzado) que permiten la implementación practica de Q-table es así como, las redes neuronales se benefician de la reducción potencial en el tiempo de ejecución, la potencia y el área al reducir la precisión del acumulador múltiple dicho esto en un futuro en la aplicación de ML se puede aprovechar el funcionamiento de hardware similares mientras se buscan niveles de

cuantificación óptimos para diferentes tareas y esquemas de control.

En ML aun hay muchos campos por explotar, su enfoque es impredecible en cuanto a las arquitecturas emergentes, actualmente como se ha llevado desarrollando ML tiene la capacidad de mejorar aspectos como energía, rendimiento al igual que la vida útil y la latencia, estos últimos particularmente problemáticos en las tecnologías emergentes, ya que estas tecnologías no pueden encontrar fácilmente una implementación generalizada sin alguna base de confianza es por ello que se espera a futuro modelos que proporcionen métodos de trabajo para equilibrar de forma eficiente estrategias de control de manera dinámica. El desarrollo de arquitecturas con machine learning permite una aceleración en el desarrollo sin un conocimiento previo de las practicas tradicionales en las arquitecturas emergentes, en el momento de asignación de tareas o predicción de ramas se pueden utilizar practicas tradicionales o de ML, en las arquitecturas emergentes la ventana de que practica utilizar puede ser difusa, por eso como recomendaciones futuras se deja explorar la aplicación de ML a preocupaciones novedosas, como la conectividad y la reconfigurabilidad en intercaladores y aceleradores específicos de dominio.

V. CONCLUSIONES

- Desde sus primeros y a lo largo del tiempo el machine learning, que viene siendo uno de los subcampos mas importantes de la inteligencia artificial, se ha venido implementado paso a paso para ayudar a mejorar y facilitar el alcance funcional de las computadoras modernas, maquinas y/o robots para que éstas aprendan del entorno y desarrollen tareas, basicamente que actue de igual forma a como lo harían los seres humanos.
- Para el libre desarrollo del machine learning se es necesario contar con un ambiente y con unas condiciones ideales, gracias al exponencial desarrollo la arquitectura de computadores con la creación de nuevos microprocesadores y el avance que hemos tenido en la parte de software, este campo se ha visto enormemente beneficiado ya que estos algoritmos cuentan con una complejidad muy alta y no cualquier computador cuenta con los recursos necesarios para su ejecucion.
- Dentro de los pro del Machine Learning existen elementos que todavía cuentan con áreas de oportunidad. Sin embargo, también se han encontrado diversas alternativas para mitigar los posibles riesgos. Es importante tener en cuenta que mientras más robusto sea el diseño de datos, las posibilidades de que haya un modelo predictivo mal formulado son menores.

VI. AGRADECIMIENTOS

Agradecimientos para Carlos Jaime Barrios, por la suministracion de material para la realizacion de este trabajo junto a las clases y diferentes prácticas para el conocimineto de la arquitectura de los computadores y poder hacer la relación entre el ML y su aplicaión de en las computadores del futuro para mejorar o aplicar la inteligencia artificial de una manera más innovadora en cuanto al mercado.

REFERENCES

[1] G. Kent, "¿Qué es machine learning? [Guía completa para principiantes]" Junio 26 2017 [online] Disponible en:https://blog.adext.com/machine-learning-guia-completa/.
[2] Video de "Hecho en Alemania", "Machine learning": "computadoras que piensan", Mayo 2 2018 [online] Disponible en: https://www.dw.com/es/machine-learning-computadoras-que-piensan/av-42176441.
[3] J. Calvo, M. A. Gúzman, D. Ramos, Manegement Solutions, Machine Learning, una pieza clave en la transformación de los modelos de negocio, 2018 [texto, online] Disponible en: https://www.managementsolutions.com/sites/default/files/publicaciones/esp/machine-learning.pdf
[4] J. Russell, Study Examines Efforts (and Prospects) for ML Use in Computer Architecture Design, Enero 9, 2020, [online] Disponible en: https://www.hpcwire.com/2020/01/09/study-examines-efforts-and-prospects-for-ml-use-in-computer-architecture-design/
[5] D. Nemirovsky, T Arkose, N. Markovic, M. Nemirovsky, O. Unsal, A. Cristal1, M. Valero, art "A General Guide to Applying Machine Learning to Computer Architecture" vol. 5, 2018 [online] Disponible en: https://pdfs.semanticscholar.org/e385/b179a48061b122789321e6b3760882c87787.pdf
[6] Foro Quora, "How could machine learning techniques be applied to Computer Architecture?", [online] Disponible en: https://www.quora.com/How-could-machine-learning-techniques-be-applied-to-Computer-Architecture
[7] J. Reda, Ing. Computación, Universidad Autonoma del Carmen, "Arquitectura de las Computadoras", Mayo 12 2008, [online] Disponible en:https://arquitecturaico.blogspot.com/2008/05/historia-y-evolucion.html
[8] Andrew Ng, conferencia Chief Scientist at Baidu, [video, online] Disponible en: https://www.youtube.com/watch?v=O0VN0pGgBZM&feature=youtu.be