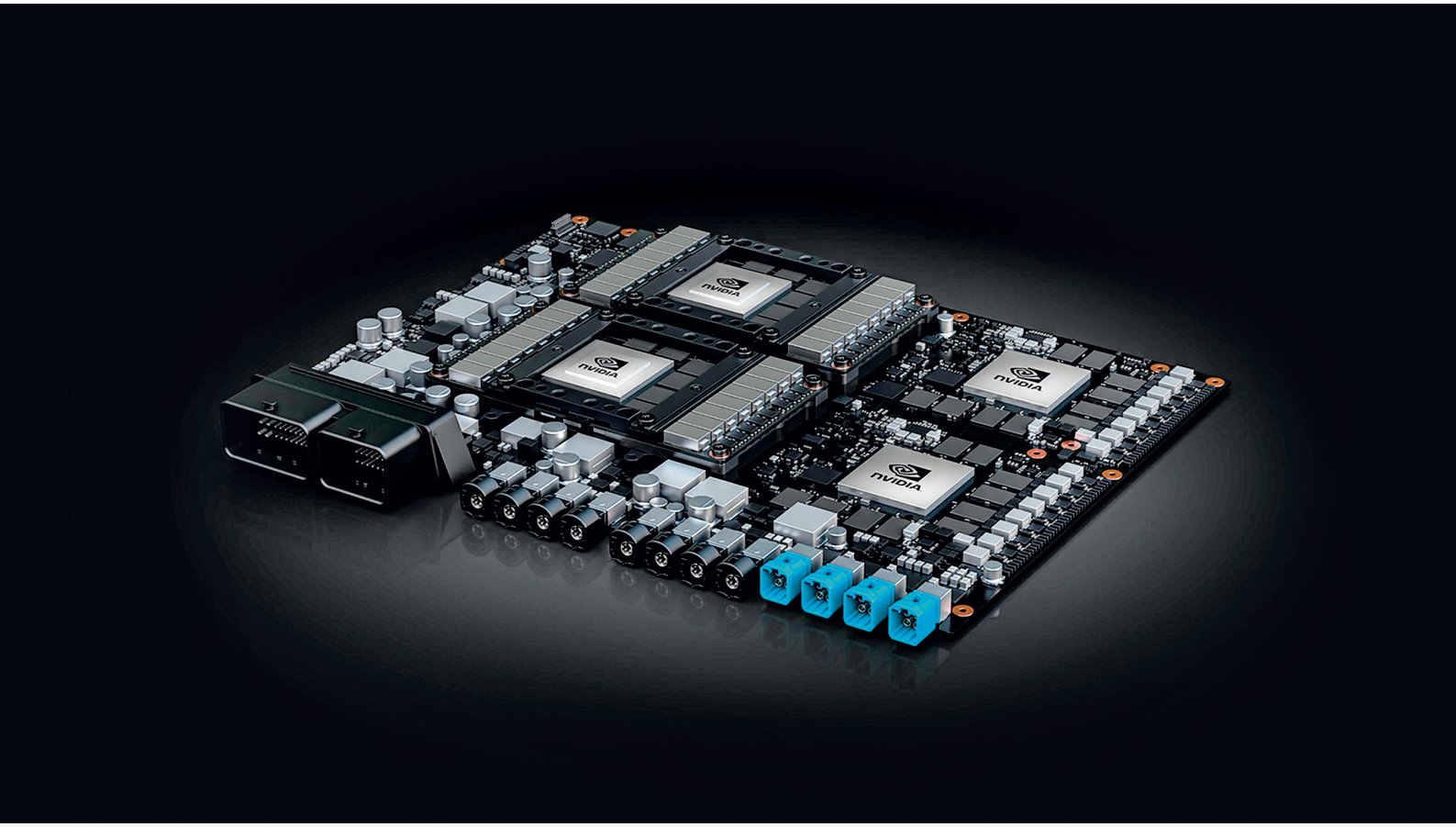

Acelerador para cálculos de Inteligencia Artificial

13/02/2020 Las funciones de conducción automatizadas y autónomas son imposibles de implementar sin Inteligencia Artificial (IA). La capacidad de cálculo requerida es proporcionada por chips especializados en computación paralela. Los investigadores también están trabajando en nuevas soluciones inspiradas en la biología, así como en ordenadores cuánticos que prometen una capacidad de cálculo aún mayor.

Durante décadas, la electrónica se ha convertido en un elemento cada vez más frecuente en los vehículos. En la actualidad, múltiples dispositivos conectados en red controlan el motor, la transmisión o el sistema de información y entretenimiento, entre otras funciones. Los coches se han convertido hace ya tiempo en centros de cálculo rodantes, pero es ahora cuando van a experimentar un salto considerable, porque las funciones de conducción automatizada y la conducción autónoma requieren sistemas cada vez más potentes. Y debido a que el rendimiento necesario no se puede lograr con chips convencionales, le ha llegado el turno a los procesadores gráficos, las unidades de procesamiento tensorial (TPU) y otro tipo de hardware especialmente diseñado para los cálculos de redes neuronales.

Si bien las CPU convencionales (unidades centrales de procesamiento) se pueden usar universalmente, carecen de la arquitectura óptima para la IA, que requiere un tipo de cálculos concretos durante las fases de aprendizaje e inferencia con las redes neuronales. "Las multiplicaciones matriciales en estas redes son muy elaboradas", explica el Dr. Markus Götz del Centro Steinbuch para la Computación del Instituto de Tecnología de Karlsruhe (KIT). “Pero estos cálculos son muy susceptibles a la paralelización, particularmente con las tarjetas gráficas. Considerando que una CPU de gama alta con 24 núcleos y comandos vectoriales puede realizar 24 veces 4 cálculos por ciclo, con una tarjeta gráfica moderna se pueden superar los 5.000".

Los procesadores de gráficos (GPU, unidades de procesamiento de gráficos) están concebidos para realizar trabajos paralelos y tienen una arquitectura interna adaptada a ese fin: las GPU contienen cientos o miles de módulos de cómputo simples para operaciones de números enteros y de punto flotante, que pueden aplicar simultáneamente la misma operación a diferentes datos (instrucción única, datos múltiples). Por lo tanto, pueden ejecutar miles de operaciones informáticas por ciclo, por ejemplo, para calcular los píxeles de un paisaje virtual o las multiplicaciones matriciales para redes neuronales. Por lo tanto, no es de extrañar que los chips NVIDIA, fabricante de GPU, sean considerados actualmente como una herramienta esencial para la inteligencia artificial en general y la conducción autónoma en particular. Volkswagen, entre otros fabricantes, utiliza el hardware de esta compañía estadounidense. "Se necesita hardware especial para la conducción autónoma", dice Ralf Bauer, Gerente Senior de Desarrollo de Software en Porsche Engineering. “Las GPU son el punto de partida; más adelante, presumiblemente, llegarán chips para cada aplicación específica".

NVIDIA actualmente ofrece los procesos Xavier para la conducción autónoma gracias a un chip de silicio equipado con ocho CPU convencionales y una GPU específicamente optimizada para el aprendizaje robótico. Para la conducción automatizada de nivel 2+ (control longitudinal y lateral limitado, con mejores funcionalidades que el nivel 2 debido al uso de sensores estándar), está disponible la plataforma Drive AGX Xavier, que puede ejecutar un máximo de 30 trillones de operaciones informáticas por segundo (30 TOPS, Tera Operaciones por segundo). Para una conducción altamente automatizada y autónoma, NVIDIA dispone del Drive AGX Pegasus (320 TOPS), bajo cuyo control un vehículo de prueba ha realizado un trayecto de hasta 80 kilómetros sin intervención humana a través de Silicon Valley. Como sucesor de Xavier, NVIDIA ahora está desarrollando la GPU Orin, aunque actualmente se sabe poco sobre sus datos de rendimiento.

No todos los fabricantes de automóviles utilizan GPU. En 2016, Tesla comenzó a desarrollar sus propios procesadores para redes neuronales. En lugar de procesadores gráficos de NVIDIA, la compañía con sede en EE. UU. ha estado instalando su chip FSD (Full Self-Driving) en sus vehículos desde principios de 2019. Además de dos unidades de procesamiento neuronal (NPU) con 72 TOPS cada una, también contiene doce núcleos de CPU convencionales para cálculos generales y una GPU para el procesamiento posterior de datos de imagen y vídeo. Tanto las NPU, como las GPU, están especializadas en las operaciones en paralelo y, por lo tanto, en la ejecución rápida de operaciones de suma y multiplicación.

Chip de Google para aplicaciones de Inteligencia Artificial

Google es un recién llegado al negocio de los chips: desde 2015, la compañía de tecnología ha estado utilizando unidades de procesamiento tensorial (TPU) de desarrollo propio en sus centros de datos. El nombre proviene del término matemático "tensor", que abarca vectores y matrices, entre otros elementos. Es por ello que la biblioteca de software de Google ampliamente utilizada para la inteligencia artificial se llama TensorFlow. En 2018, Google presentó la tercera generación de sus TPU, que contienen cuatro "unidades de multiplicación de matriz" y se dice que son capaces de 90 TFLOPS (Tera operaciones de punto flotante por segundo). Waymo, filial de Google, utiliza TPU para entrenar redes neuronales para la conducción autónoma.

Los chips específicos de la aplicación como el FSD de Tesla o los TPU de Google solo abaratan sus precios con su fabricación en masa. Una alternativa son los FPGA (matrices de puertas programables en campo). Estos chips digitales de uso universal contienen innumerables bloques informáticos y de memoria que se pueden combinar entre sí a través de la programación y con los que es posible introducir algoritmos en el hardware, como con un chip específico, pero mucho más económico. Los FPGA se pueden adaptar fácilmente a los requisitos específicos de una aplicación de IA, lo que genera beneficios en términos de rendimiento y consumo de energía. La start-up Kortiq, con sede en Múnich, ha desarrollado su arquitectura AIScale para FPGA, que simplifica las redes neuronales para el reconocimiento de imágenes y optimiza los cálculos para que los requisitos del hardware disminuyan significativamente y los resultados estén disponibles hasta diez veces más rápido.

Algunos investigadores están buscando una relación aún más estrecha entre las células nerviosas de los chips específicos empleados en Inteligencia Artificial. Investigadores de la Universidad de Heidelberg han desarrollado el sistema neuromórfico BrainScaleS, cuyas neuronas artificiales se implementan como interruptores analógicos en chips de silicio: el cuerpo celular consta de unos 1.000 transistores y dos condensadores, y las sinapsis requieren aproximadamente 150 transistores. Los cuerpos celulares individuales se pueden combinar como módulos para formar varios tipos de neuronas artificiales. Estas sinapsis pueden, como en la naturaleza, formar fuertes conexiones, y también hay tipos excitadores e inhibidores. La salida de las neuronas consiste en "picos", pulsos de voltaje corto que duran unos pocos microsegundos, que funcionan como entradas para las otras neuronas artificiales.

Neurochips energéticamente eficientes

Pero BrainScaleS no solo se usa para investigar el cerebro humano. Las neuronas técnicas también se pueden utilizar para resolver problemas técnicos, como la detección de objetos para la conducción autónoma. Por un lado, ofrecen una alta capacidad informática de aproximadamente un cuatrillón de operaciones informáticas por módulo con 200.000 neuronas. Por otro lado, la solución analógica también consume muy poca energía. "En los circuitos digitales, por ejemplo, se utilizan unos 10.000 transistores para cada operación", explica Johannes Schemmel, de la Universidad de Heidelberg. "Nos las arreglamos con mucho menos, lo que nos permite lograr aproximadamente 100 TOPS por vatio".

Los investigadores acaban de desarrollar la segunda generación de sus circuitos y hablan con socios de la industria para llegar a posibles acuerdos de colaboración.

Poder cuántico desde la nube

En el futuro, incluso los ordenadores cuánticos podrían usarse en el campo de la IA. Su unidad fundamental no es el bit binario, sino el qubit, con un número infinito de valores posibles. Gracias a las leyes de la mecánica cuántica, los cálculos pueden ser altamente paralelizados y, por lo tanto, acelerados. Al mismo tiempo, las computadoras cuánticas son difíciles de implementar porque los qubits están representados por sistemas físicos sensibles como electrones, fotones e iones. Esto se demostró, por ejemplo, con el IBM Q System One, que la compañía presentó en la feria de electrónica CES 2019 en Las Vegas. El interior del ordenador cuántico debe estar rigurosamente protegido contra vibraciones, campos eléctricos y fluctuaciones de temperatura.

Células nerviosas y neuronas artificiales

Las células nerviosas del cerebro humano reciben sus señales de otras neuronas a través de sinapsis que se encuentran en las dendritas o directamente en el cuerpo celular. Las sinapsis pueden tener un efecto excitador o inhibidor. Todas las entradas se suman en el axón y, si se excede un umbral en el proceso, la célula nerviosa dispara una señal de aproximadamente un milisegundo que se propaga a lo largo del axón y llega a otras neuronas.

Las neuronas artificiales imitan este comportamiento de una manera bastante fiel. En las redes neuronales convencionales con múltiples capas, cada "célula nerviosa" recibe una suma ponderada como entrada. Esta suma consiste en las salidas de las neuronas de la capa anterior y el factor de ponderación wi, en el que se almacena la experiencia de aprendizaje de la red neuronal. Estos factores de ponderación corresponden a las sinapsis y también pueden ser excitadores o inhibidores. Un valor de umbral configurable determina, como en una célula nerviosa, cuándo se activa la neurona artificial.

Aprendizaje e inferencia con redes neuronales

Las redes neuronales naturales y artificiales aprenden de los cambios en la fuerza de las conexiones sinápticas y los factores de ponderación. Usando métodos matemáticos, el factor de ponderación wi se reajusta continuamente hasta que la red neuronal pueda ubicar imágenes de manera confiable, por ejemplo, en categorías específicas. Mediante inferencia, los datos se envían a la entrada, y la salida se usa, por ejemplo, para tomar decisiones.

Tanto en el aprendizaje como en la inferencia en las redes neuronales profundas (redes con múltiples capas de neuronas artificiales), las mismas operaciones matemáticas ocurren repetidamente. Si se suman tanto las salidas de las neuronas de la capa 1 como las entradas de las neuronas de la capa 2

como vectores de columna, todos los cálculos se pueden representar como multiplicaciones matriciales. En el proceso, se producen numerosas multiplicaciones y sumas independientes entre sí que se pueden ejecutar en paralelo. Las CPU convencionales no están diseñadas para eso, y es por ello que los procesadores gráficos, TPU y otros aceleradores de IA son muy superiores a ellos.

En resumen

Los chips convencionales alcanzan sus límites cuando se trata de cálculos para redes neuronales. Los procesadores gráficos y el hardware especial para IA desarrollados por compañías como NVIDIA y Google son mucho más potentes. Los chips neuromórficos son sustancialmente similares a las neuronas reales y funcionan de manera muy eficiente. Los ordenadores cuánticos también podrían aumentar enormemente la capacidad informática.

Información

Texto: Christian Buck
Con la contribución de: Ralf Bauer, Dr. Christian Koelen

Texto publicado previamente en la revista Porsche Engineering Magazine, número 2/2019

Link Collection

Link to this article
https://newsroom.porsche.com/es_ES/tecnologia/2020/es-porsche-engineering-acelerador-calculos-inteligencia-artificial-conduccion-autonoma-19928.html

External Links
https://www.porscheengineering.com/peg/en/