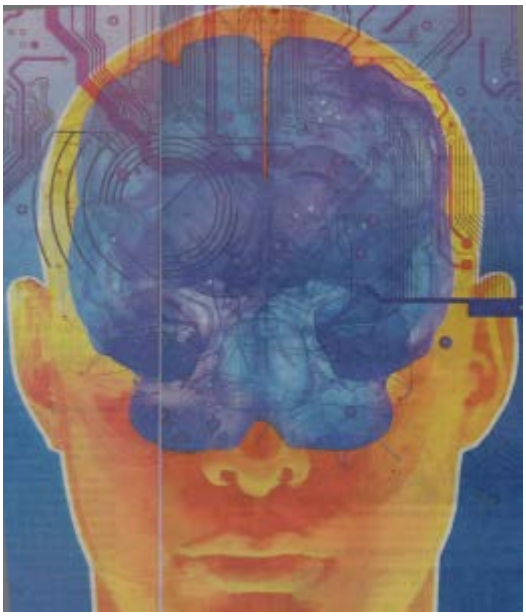
www.psicoadolescencia.com.ar

Computadoras inspiradas en el diseño del cerebro humano

Crece la demanda de que las máquinas aumenten su inteligencia y la complejidad de sus operaciones; los expertos buscan la respuesta en la naturaleza



Hoy todos esperamos demasiado de nuestras computadoras. Se supone que nos tendrían que hablar, que deberían reconocer desde rostros hasta flores y, dentro de poco, hasta manejarnos el auto. Toda esa inteligencia artificial exige una cantidad enorme de potencia informática, que fuerza los límites incluso de las máquinas más modernas. Para responder a esta creciente demanda, algunas de las empresas tecnológicas más grandes del mundo van tras la pista de la biología. Se pusieron a repensar la naturaleza de las computadoras y empezaron a fabricar máquinas más parecidas al cerebro humano, en el que un tronco encefálico supervisa el sistema nervioso y le delega determinadas tareas –ver, oír– a la corteza que lo rodea.

Después de años de estancamiento, la computadora vuelve a evolucionar, y esta migración hacia un nuevo tipo de máquina traerá variadas consecuencias a largo plazo. Es posible que esta transición también haga decrecer el poder de Intel, el gigante del diseño y de la manufactura de chips. Pero fundamentalmente es posible que termine reformulando una industria que genera 335 mil millones de dólares por año: la de los semiconductores que habitan en el corazón de todo lo que esté relacionado con la tecnología, desde los centros de datos que llevan Internet a su iPhone hasta los cascos para realidad virtual y los drones del futuro.

Diseño informático

"Es un cambio enorme –dice John Hennessy, el ex presidente de la Universidad de Stanford, quien a mediados de los años noventa escribió un libro pionero en el diseño informático y que hoy integra el equipo de Alphabet, la empresa matriz de Google.– El enfoque actual se agota y están tratando de rediseñar todo el sistema". Pero el enfoque actual tuvo un derrotero bastante satisfactorio. Por casi medio siglo, los fabricantes de computadoras han estado construyendo sistemas en torno a un chip único que hacía de todo –la unidad central de procesamiento– proveniente de una empresa como Intel, uno de los mayores fabricantes de semiconductores del mundo. Eso es lo que hay dentro de nuestras notebooks y nuestros celulares.

Ahora, los ingenieros en informática diseñan sistemas más complejos. En vez de canalizar todas las tareas a través de un chip fabricado por Intel, las máquinas nuevas dividen el trabajo en pequeñas partes que se distribuyen entre un gran número de terminales de chips más simples pero especializados, que consumen mucha menos energía. Los cambios dentro de las gigantescas centrales de datos de Google son un anuncio de lo que vendrá para el resto de la industria. Dentro de la mayoría de los servidores de Google, todavía existe un procesador central. Pero también hay enormes bancos de chips hechos a medida que trabajan a la par, ejecutando los algoritmos informáticos que controlan el reconocimiento de voz y las demás formas de inteligencia artificial.

Google llegó a este punto por necesidad. Durante años, la empresa había venido explotando la mayor red informática del mundo: un imperio de centros de datos y de cables tendidos desde California hasta Finlandia y Singapur. Pero para uno de sus investigadores, ya se estaba quedando corta. En 2011, Jeff Dean, uno de los ingenieros más célebres de la empresa, lideró un equipo de investigación para explorar el concepto de redes neuronales (en esencia, algoritmos informáticos capaces de aprender tareas por sí mismos). Estas pueden resultar útiles para una infinidad de cosas, como reconocer las palabras que se le dictan a un smartphone o las caras de una fotografía.

En meses, Dean y su equipo crearon un servicio capaz de reconocer la palabra hablada con mucha mayor exactitud que el servicio existente de Google. Pero había una trampa: Dean se dio cuenta de que si los más de mil millones de teléfonos del mundo que operaban con Android usaban el servicio nuevo nada más que tres minutos por día, Google iba a tener que duplicar la capacidad de su central de datos con el fin de poder soportarlo. "Necesitamos otro Google", relata un testigo que le dijo Dean a Urs Hölzle, el científico informático nacido en Suiza encargado de supervisar el imperio de la central de datos de la empresa. Y le propuso una alternativa: Google podía construir su propio chip informático para ejecutar este tipo de inteligencia artificial. Pero lo que empezó en las centrales de datos ahora comienza a extenderse a otras zonas del panorama tecnológico. En los próximos años, empresas como Google, Apple y Samsung van a fabricar teléfonos con chips especializados para inteligencia artificial. Microsoft está diseñando un chip especial para unos auriculares de realidad aumentada. Y todos, desde Google hasta Toyota, están produciendo vehículos autónomos para los que van a hacer falta chips similares.

Según Gill Pratt, que fue gerente de programas en Darpa –una unidad de investigación del Departamento de Defensa de los Estados Unidos– y que ahora trabaja en vehículos autónomos para Toyota, esta tendencia hacia los chips especializados y hacia una nueva arquitectura informática podría llevar a una "explosión cámbrica" de la inteligencia artificial. A su entender, las máquinas que distribuyen los cálculos entre un gran número de chips diminutos de baja potencia pueden operar de forma más parecida a la de un cerebro humano, que utiliza eficientemente la energía de que dispone. "En el cerebro, la clave es la eficiencia energética", dice en una entrevista en el nuevo centro de investigación de Toyota en Silicon Valley.

Existen muchas clases de chips de silicio. Hay chips que almacenan información, hay chips que desempeñan tareas básicas en juguetes y televisores, y hay chips que ejecutan diversos procesos, desde las supercomputadoras que se emplean para crear los modelos de simulación de calentamiento global hasta las computadoras personales, los servidores de Internet y los smartphones. Por años, las unidades de procesamiento central, o CPU, que ejecutaban las PC y dispositivos similares estaban donde estaba el dinero. Y no había demasiada necesidad de cambiar nada.

De acuerdo con la ley de Moore, la máxima tan citada del cofundador de Intel Gordon Moore, el número de transistores que hay en un chip informático se duplica más o menos cada dos años. Esto produjo una mejora constante del desempeño, que se sostuvo durante décadas. Según otra ley del diseño de chips menos conocida, llamada “escala de Dennard” (en homenaje al investigador de IBM robert Dennard) o “escala Mosfet”, a pesar de que su desempeño mejoró, los chips siguieron consumiendo casi la misma cantidad de energía.

Pero para 2010 duplicar el número de transistores estaba tomando más tiempo de lo que la ley de Moore había predicho. La máxima de escala de Dennard también se había trastocado, mientras que los diseñadores de chips estaban alcanzando el límite de los materiales físicos que usaban para construir los procesadores. El resultado: si una empresa quería más potencia informática, no podía conformarse con actualizar los procesadores. Necesitaba más computadoras, más espacio y más electricidad.

Evolución constante

Los investigadores de la industria y del mundo académico trabajaron para extender la ley de Moore, explorando materiales y técnicas de diseño de chips completamente nuevos. Pero Doug Burger, un investigador de Microsoft, tuvo otra idea: en vez de depender de la evolución constante del procesador central, como la industria había venido haciendo desde la década del sesenta, ¿por qué no trasladar parte de la carga de trabajo a chips especializados? Durante las vacaciones de Navidad de 2010, Burger y un pequeño grupo de expertos en chips de Microsoft comenzaron a estudiar un nuevo hardware capaz de acelerar el desempeño de Bing, el motor de búsqueda de la empresa.

Para entonces, Microsoft recién empezaba a optimizar Bing a través de algoritmos de aprendizaje automático (las redes neuronales son una especie de máquina de aprender) capaces de mejorar los resultados de las búsquedas analizando la forma en que la gente usaba el servicio. Aunque estos algoritmos eran menos demandantes que las redes neuronales que más tarde iban a reconstruir Internet, a los chips de esa época les costaba seguirles el ritmo. Burger y su equipo exploraron varias opciones y finalmente se decidieron por el llamado arreglo de compuertas programables en campo, o FPGA (por su sigla en inglés): chips que se podían reprogramar sobre la marcha para desempeñar nuevas tareas. Microsoft fabrica software, como Windows, que se ejecuta en un CPU de Intel. Pero ese software no puede reprogramar el chip, ya que este fue diseñado para desempeñar únicamente ciertas tareas.

Con un FPGA, Microsoft podía cambiar el tipo de funcionamiento del chip. Podía programarlo para que fuera eficaz con determinados algoritmos de aprendizaje automático. Luego, podría reprogramar el chip para optimizarlo en la ejecución de la lógica que envía millones de paquetes

de datos a través de su red informática. Era el mismo chip, pero reaccionaba de manera diferente. En 2015, Microsoft empezó a instalar esos chips masivamente. Ahora, casi todos los servidores nuevos conectados a un centro de datos de Microsoft incluyen uno de estos chips programables. Estos ayudan a elegir los resultados cuando se busca con Bing y también colaboran con Azure (el servicio de computación en la nube) en el transporte de la información a través de su red de máquinas subyacentes.

Hace un año, otro equipo de investigadores de Microsoft construyó una red neuronal –en espejo con el trabajo realizado por Jeff Dean en Google– que, al menos para una de las mediciones, podía reconocer las palabras habladas con mayor precisión que el ser humano promedio. La iniciativa fue liderada por Xuedong Huang, especialista en reconocimiento de voz nacido en China. Poco después de que el equipo publicó el artículo que describe su trabajo, Huang fue a cenar a las colinas de Palo Alto, California, con su viejo amigo Jen-Hsun Huang, director ejecutivo de la fábrica de chips Nvidia. Ambos tenían motivos para celebrar, y brindaron con champagne.

Xuedong Huang y sus colegas investigadores de Microsoft habían entrenado su servicio de reconocimiento de voz utilizando un gran número de chips especializados provistos por Nvidia, en lugar de confiar en los chips genéricos de Intel. De no haberse producido ese cambio, el avance no habría sido posible. “En casi un año cerramos la brecha con los humanos –dice Huang, de Microsoft–. Si no hubiéramos tenido la infraestructura nos hubiese llevado al menos cinco años más”. Como los sistemas basados en redes neuronales pueden aprender por su cuenta, son capaces de evolucionar con mayor rapidez que los servicios tradicionales. No dependen tanto de que los ingenieros escriban líneas interminables de códigos para explicarles cómo se tienen que comportar.

Pero entrenar redes neuronales de esta forma exige un trabajo intensivo de ensayo y error. Para crear una red que pueda reconocer las palabras tan bien como un ser humano, los investigadores tienen que instruirla repetidamente, ajustando una y otra vez los algoritmos para mejorar el entrenamiento. Este proceso involucra miles de algoritmos, lo que exige una potencia informática enorme. Y si para hacerlo las empresas como Microsoft usan chips comunes y corrientes, el proceso demanda muchísimo más tiempo, ya que esos chips no son capaces de manejar la carga. De este modo se consume demasiada energía eléctrica.

De manera que las empresas líderes de Internet están preparando sus redes neuronales con el auxilio de otra clase de chip, conocido como unidad procesadora de gráficos o GPU (por su sigla en inglés). Estos chips de baja potencia –por lo general, fabricados por Nvidia– fueron diseñados originalmente para procesar las imágenes de juegos y otros programas, y trabajaban mano a mano con el chip –por lo general, fabricado por Intel– en el corazón de una computadora. Los GPU son capaces de procesar los cálculos que exigen las redes neuronales con una eficiencia mucho mayor que los CPU. Como resultado, Nvidia está prosperando, ahora les vende gran número de GPU a los gigantes de Internet de Estados Unidos y a las empresas en línea más importantes del globo, sobre todo en China. La recaudación trimestral por las ventas del centro de datos se triplicó, superando en 409 millones de dólares las del año pasado. “Esto es un poco como haber estado en el principio de Internet”, dice Jen-Hsun Huang. En otras palabras, el panorama tecnológico cambia con rapidez, y Nvidia está en el corazón de ese cambio.

Crear chips especializados

Los GPU son el principal instrumento que las empresas usan para enseñarles una tarea determinada a sus redes neuronales, pero esa es sólo una parte del proceso. Una vez que una red neuronal fue entrenada para una tarea, tiene que ejecutarla, y esto requiere una potencia informática de otro orden.

Por ejemplo, después de haber entrenado un algoritmo de reconocimiento de voz, Microsoft lo ofrece como servicio en línea, y en realidad recién entonces empieza a identificar las órdenes que la gente les imparte a sus smartphones. Los GPU no son tan eficientes durante esa etapa del proceso. Por eso, muchas empresas están fabricando chips especializados para hacer lo que otros chips aprenden. Google creó su propio chip especializado, una unidad de procesamiento de tensor, o TPU (por su sigla en inglés). Nvidia está fabricando un chip similar. Y Microsoft reprogramó chips especializados, como Altera, que fue adquirido por Intel, para ejecutar redes neuronales con mayor facilidad.

Otras empresas siguen sus pasos. Qualcomm, que se especializa en chips para smartphones, y un buen número de empresas emergentes también están trabajando con los chips AI, esperando llevarse su parte de un mercado en rápida expansión. La firma de investigaciones tecnológicas IDC predice que para 2021 los ingresos de los servidores equipados con chips alternativos van a alcanzar los 6800 millones de dólares, casi el 10 por ciento del mercado global de servidores. Burger señaló que, dentro de la red global de máquinas de Microsoft, los chips alternativos todavía representan una parte modesta de la operación. Y Bart Sano, el vicepresidente de Ingeniería que dirige el desarrollo de software y hardware para la red de Google, dice lo mismo acerca de los chips emplazados en sus centros de datos.

Mike Mayberry, que preside Intel Labs, minimizó el cambio de orientación hacia los procesadores alternativos, tal vez porque Intel controla más del 90 por ciento del mercado de los centros de procesamiento de datos, lo que lo hace el mayor vendedor de chips tradicionales. Mayberry sostiene que si los procesadores centrales se modificaran de la manera correcta, podrían desempeñar tareas nuevas sin necesidad de ayuda adicional. Pero esta nueva raza de silicio se propaga con rapidez y, cada vez más, Intel es una empresa en conflicto consigo misma. De alguna manera niega que el mercado cambia, pero al mismo tiempo modifica su actividad para mantenerse al día con el cambio.

Hace dos años, Intel gastó 16.700 millones de dólares en comprar Altera, el fabricante de los chips programables que usa Microsoft. Fue la mayor adquisición en la historia de Intel. Y el año pasado la empresa pagó 408 millones por Nervana, una compañía que estaba explorando un chip especializado en ejecutar redes neuronales. Ahora, liderada por el equipo de Nervana, Intel está desarrollando un chip dedicado para entrenar y ejecutar redes neuronales. “Intel tiene el problema típico de las grandes empresas –dice Bill Coughran, uno de los socios de Sequoia Capital, la empresa de capital de riesgo de Silicon Valley que se dedicó durante casi una década a supervisar la infraestructura en línea de Google–. Necesita encontrar cómo pasar a las áreas nuevas y en expansión sin perjudicar el negocio tradicional”.

El conflicto interno de Intel es más evidente cuando los directivos de la empresa hablan de la decadencia de la ley de Moore. En una entrevista con The New York Times, Naveen Rao, fundador

de Nervana y ahora ejecutivo de Intel, dijo que Intel podría “exprimir” la ley de Moore unos años más. Oficialmente, la posición de la empresa es que las mejoras en los chips tradicionales van a continuar hasta bien entrada la próxima década. Mayberry, de Intel, también argumentó que el uso de chips adicionales no es nuevo. Según él, en el pasado los fabricantes de computadoras usaban chips independientes para tareas como procesar audio.

Pero ahora el alcance de esta tendencia es mucho mayor. Y está cambiando el mercado en nuevas formas. Intel no solo compite con los fabricantes de chips como Nvidia y Qualcomm, sino también con las empresas como Google y Microsoft. Google está diseñando la segunda generación de sus chips TPU. La empresa afirma que, para finales de este año, cualquier comerciante o desarrollador que sea cliente de su servicio informático en la nube podrá usar los nuevos chips para ejecutar su software. Mientras se lleva a cabo esta transición en el interior de los enormes centros de datos en los que se sustenta Internet, será cuestión de tiempo hasta que se extienda al conjunto de la industria.

La esperanza es que esta nueva casta de chips portátiles contribuya para que los dispositivos soporten por su cuenta cada vez más, y más complejas, tareas sin tener que recurrir a los centros de datos remotos: teléfonos que reconozcan los comandos de voz sin acceder a Internet y autos que reconozcan el mundo que los rodea con una velocidad y precisión que hasta hoy no son posibles. En otras palabras, un vehículo autónomo necesita cámaras, radar y láseres. Pero además le hace falta un cerebro.

LA NACION 30 Nov 2017
Publicado originalmente por The New York Times
Texto Cade Metz
Traducción de Jaime Arrambide
http://www.pressreader.com/argentina/la-nacion/20171130/281925953334490