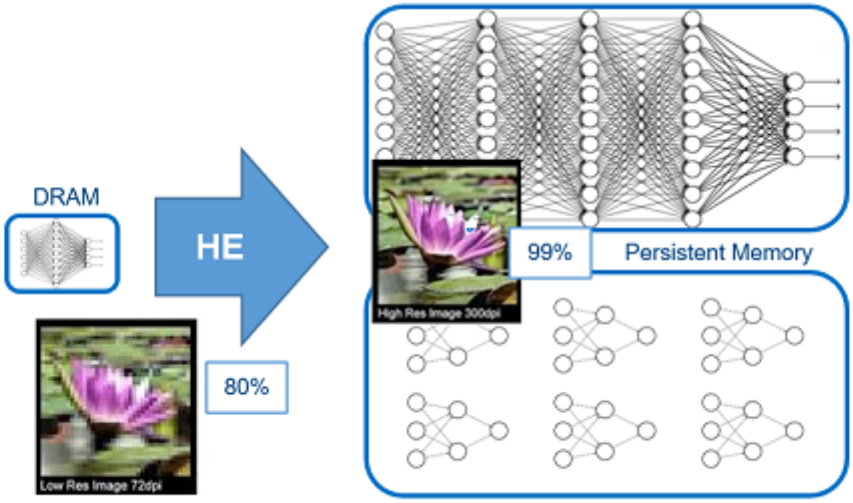
Published on *BSC-CNS* (https://www.bsc.es)

El BSC ejecuta, por primera vez, grandes redes neuronales encriptadas utilizando la memoria persistente Intel Optane y los procesadores escalables Intel Xeon

Hasta el momento, el uso de la encriptación homomórfica había limitado a modelos de redes neuronales para dispositivos móviles.



El Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC) conjuntamente con Intel han hecho posible, por primera vez, la ejecución encriptada de grandes redes neuronales de forma eficiente, gracias a la memoria persistente Intel Optane (PMem) y a los procesadores escalables Intel Xeon con aceleración de IA incorporada. Hasta el momento, el tamaño de memoria principal soportada por la tecnología actual había limitado el uso de la encriptación homomórfica a modelos de redes neuronales pequeñas (hasta 1,7 millones de parámetros), diseñados para dispositivos móviles. Por lo tanto, el cifrado de grandes redes neuronales es un avance tecnológico importante.

Este tipo de encriptación, Homomorphic Encryption, que no se puede romper incluso con computadores cuánticos, posibilita operaciones directamente sobre datos encriptados, de manera que quien opera con los datos no tiene acceso a su contenido. Dado que esta encriptación no necesita ser desencriptada para operar, se garantiza la privacidad en entornos no seguros (como el cloud).

El principal reto de la encriptación homomórfica es su "sobrecoste" al incrementar el tamaño de los datos, que puede llegar a multiplicarlos por un factor de hasta 10.000. La memoria persistente Intel Optane ofrece capacidades muy superiores a las de DRAM y tiempo de acceso mucho más rápido que otras memorias no volátiles. Si bien no es tan rápido como la tecnología de memoria principal, la combinación de ambos con un patrón de acceso eficiente ofrece atractivos beneficios precio / rendimiento.
Esta nueva tecnología tiene aplicación en la ejecución privada de redes neuronales en entornos remotos no confiables, como la nube e incluye, tanto la protección de la propiedad intelectual relacionada con el propio modelo de red neuronal, como los datos utilizados, que lo haría compatible con la ley de protección de datos.

Estos datos pueden incluir información personal, médicos, secretos industriales o de estado, etc.

La investigación ha sido realizada por un equipo de investigadores del BSC, junto con un equipo internacional de Intel, con miembros tanto en Europa como en Estados Unidos, liderados por el Investigador del BSC Antonio J. Peña.
Peña lidera el equipo de Aceleradores y comunicaciones para computación de altas prestaciones en el departamento de Ciencias de la Computación del BSC. Su investigación se centra en la heterogeneidad de recursos de hardware y comunicaciones sobre redes de alto rendimiento.

Según Peña, “esta nueva tecnología debe permitir el uso generalizado de redes neuronales en entornos de nube, incluyendo, por primera vez, allí donde se requiera confidencialidad indiscutible para los datos o el propio modelo de red neuronal”.
Para Fabian Boemer, responsable técnico de Intel que participa en esta investigación, “el cálculo de cifrado homomórfico es tanto computacional como intensivo en cuanto a memoria. Para paliar el cuello de botella que se puede generar en el acceso a la memoria, estamos investigando diferentes arquitecturas que permitan una computación más eficiente. Este trabajo es un primer paso importante para resolver este desafío que a menudo se pasa por alto. Entre otras tecnologías, estamos investigando el uso de la memoria persistente Intel Optane para mantener los datos a los que se accede constantemente cerca del procesador durante el cómputo mediante encriptacion homomorfica".

El artículo científico relacionado con esta investigación está aceptado para publicación en la revista IEEE Transactions on Computers, donde se analiza la ejecución del popular modelo ResNet-50, que incorpora 25 millones de parámetros, llegando a consumir cerca de 1 TB de memoria, más del doble del disponible a un nodo de cómputo del superordenador MareNostrum 4.

En este artículo también se menciona una arquitectura de computador eficiente para esta tarea con sólo 1/3 de la memoria RAM habitual, que consume alrededor de 10 veces más energía por byte que la memoria persistente Intel Optane, posibilitando así configuraciones con una eficiencia energética muy mejorada y sostenibilidad de la solución.

También está disponible públicamente la versión de los autores en la plataforma arXiv:
https://arxiv.org/abs/2103.16139

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---