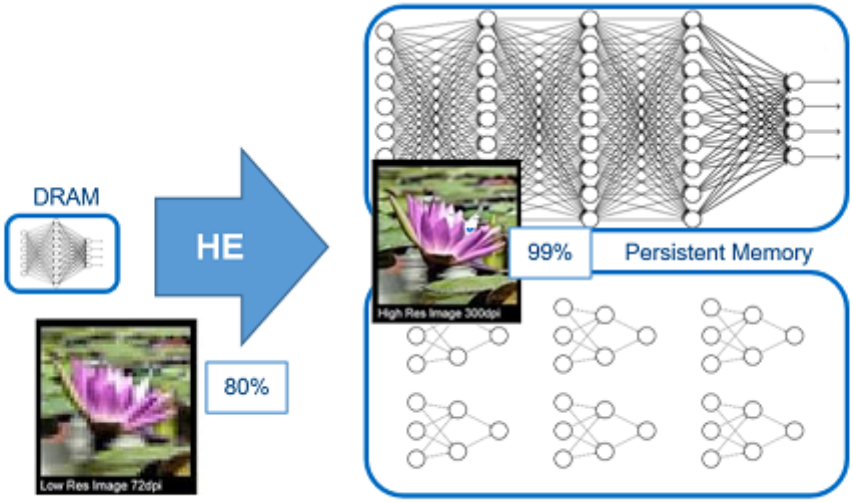
Published on *BSC-CNS* (https://www.bsc.es)

El BSC executa, per primera vegada, grans xarxes neuronals encriptades utilitzant la memòria persistent Intel Optane i els processadors escalables Intel Xeon

Fins al moment, l'ús de l'encriptació homomòrfica s'havia limitat a models de xarxes neuronals per a dispositius mòbils.



El Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC) conjuntament amb Intel han fet possible, per primer cop, l'execució encriptada de grans xarxes neuronals de forma eficient, gràcies a la memòria persistent Intel Optane (PMem) i als processadors escalables Intel Xeon amb acceleració d'IA incorporada. Fins al moment, la mida de memòria principal suportada per la tecnologia actual havia limitat l'ús de la encriptació homomòrfica a models de xarxes neuronals petites (fins a 1,7 milions de paràmetres), dissenyats per a dispositius mòbils. Per tant, el xifrat de grans xarxes neuronals és un avanç tecnològic important.

Aquest tipus d'encriptació, Homomorphic Encryption, que no es pot trencar fins i tot amb computadors quàntics, possibilita operacions directament sobre dades encriptades, de manera que qui opera amb les dades no té accés al seu contingut. Atès que aquesta encriptació no necessita ser desencriptada per operar, es garanteix la privacitat en entorns no segurs (com el cloud).

El principal repte de l'encriptació homomòrfica és el seu "sobrecost" a l'incrementar la grandària de les dades, que pot arribar a multiplicar per un factor de fins a 10.000. La memòria persistent Intel Optane ofereix capacitats molt superiors a les de DRAM i temps d'accés molt més ràpid que altres memòries no volàtils. Si bé no és tan ràpid com la tecnologia de memòria principal, la combinació de tots dos amb un patró d'accés eficient ofereix atractius beneficis preu / rendiment.
Aquesta nova tecnologia té aplicació en l'execució privada de xarxes neuronals en entorns remots no fiables, com el núvol i inclou, tant la protecció de la propietat intel·lectual relacionada amb el propi model de xarxa

neuronal, com les dades utilitzades, que ho faria compatible amb la llei de protecció de dades. Aquestes dades poden incloure informació personal, mèdiques, secrets industrials o d'estat, etc.

La recerca ha estat realitzada per un equip d'investigadors del BSC, juntament amb un equip internacional d'Intel, amb membres tant a Europa com als Estats Units, liderats per l'investigador del BSC Antonio J. Peña. Peña lidera l'equip d'Acceleradors i comunicacions per a computació d'altes prestacions en el Departament de Ciències de la computació del BSC. La seva recerca se centra en l'heterogeneïtat de recursos de hardware i comunicacions sobre xarxes d'alt rendiment.

Segons Peña, "aquesta nova tecnologia ha de permetre l'ús generalitzat de xarxes neuronals en entorns de núvol, incloent, per primera vegada, allà on es requereixi confidencialitat indiscutible per a les dades o el propi model de xarxa neuronal".
Per a Fabian Boemer, responsable tècnic d'Intel que participa en aquesta investigació, "el càlcul de xifrat homomòrfic és tant computacional com intensiu pel que fa a la memòria. Per pal·liar el coll d'ampolla que es pot generar en l'accés a la memòria, estem investigant diferents arquitectures que permetin una computació més eficient. Aquest treball és un primer pas important per resoldre aquest desafiament que sovint es passa per alt. Entre altres tecnologies, estem investigant l'ús de la memòria persistent Intel Optane per mantenir les dades a les quals s'accedeix constantment a prop de el processador durant el còmput mitjançant encriptació homomòrfica".

L'article científic relacionat amb aquesta recerca està acceptat per a la seva publicació a la revista IEEE Transactions on Computers, on s'analitza l'execució del popular model ResNet-50, que incorpora 25 milions de paràmetres, arribant a consumir prop d'1 TB de memòria, més del doble del disponible a un node de còmput del superordinador MareNostrum 4.

En aquest article també s'esmenta una arquitectura de computador eficient per a aquesta tasca amb només 1/3 de la memòria RAM habitual, que consumeix al voltant de 10 vegades més energia per byte que la memòria persistent Intel Optane, possibilitant així configuracions amb una eficiència energètica molt millorada i sostenibilitat de la solució.

També està disponible públicament la versió dels autors a la plataforma arXiv:
https://arxiv.org/abs/2103.16139

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---