Décryptage

Neural Network Pruning, ou l'art d'élaguer l'intelligence artificielle

Lucie Zhang, Analyste Zenon

Avril 2026

Les réseaux de neurones sont aujourd'hui largement déployés mais les modèles sont souvent de taille massive et sur-paramétrés, ce qui contribue à augmenter leur consommation énergétique et leur coût à l'entraînement. A titre d'exemple, GPT-4 possède 1760 milliards de paramètres et la consommation énergétique liée à l'utilisation de GPT-4o est estimée autour de 391 GWh en 2025. De plus, en phase d'inférence (utilisation), seulement une fraction des paramètres sont activés utilement lors des calculs. Dans ce contexte, l'IA frugale peut être un moyen de "faire mieux avec moins" : réduction des coûts d'infrastructure, diminution de la consommation énergétique, latence plus faible propice aux systèmes embarqués.

Le pruning (ou élagage) appliqué aux réseaux de neurones est une méthode algorithmique consistant à supprimer des paramètres jugés peu importants (poids, neurones, etc.) dans un modèle pour l'alléger sans altérer significativement sa performance. Elaguer un modèle peut permettre de réduire son temps d'inférence et sa consommation énergétique. Cela est donc adapté pour des utilisations en temps réel et facilite le déploiement de modèles complexes sur des appareils embarqués.

En fonction des paramètres supprimés, la stratégie de pruning peut être de deux types, avec leurs spécificités :

Dans le cas du pruning non structuré, des poids de connexions interneuronales voire des neurones sont supprimés de façon individuelle, indépendamment de leur position dans le modèle. Le pruning non structuré permet d'élaguer avec une forte granularité et d'optimiser précisément l'architecture du modèle. Le pruning non structuré s'adapte à diverses architectures de réseau et peut atteindre des taux de compression élevés.Le pruning non structuré crée des zéros dispersés de manière irrégulière et désordonnée dans les poids, ce qui nécessite un matériel informatique spécialisé pour traiter les calculs, sans quoi ce pruning ne raccourcirait pas automatiquement le temps d'inférence.

Avec le pruning structuré, des éléments structuraux entiers du modèle sont retirés, tels que des couches de neurones. Cette méthode permet de réduire effectivement la taille tout en conservant une structure ordonnée du réseau. Le pruning structuré est également compatible avec les matériels informatiques classiques. Il offre des taux de compression plus prévisibles et constants, car il cible des composants entiers. Les taux de compression sont généralement inférieurs à ceux permis pour le pruning non structuré et la granularité est aussi plus grossière.

Différentes méthodes existent pour déterminer les paramètres à supprimer : magnitude-based pruning, gradient-based pruning…

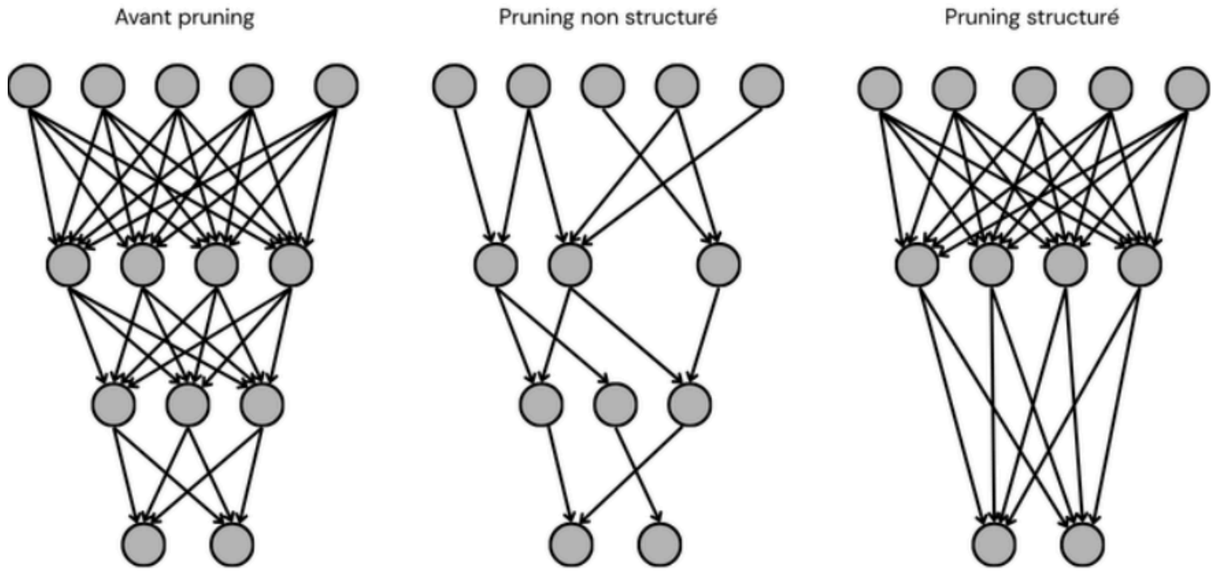

Figure - Différence entre pruning structuré et non-structuré (Zenon)

Quand utiliser le pruning et à quelle fréquence ?

Le pruning peut s'effectuer pendant la phase d'entraînement d'un modèle ou bien post-entraînement. Alléger le modèle alors qu'il est entraîné permet d'adapter ses paramètres au fur et à mesure sans altérer significativement sa performance. Sa complexité opérationnelle est cependant plus importante et peut allonger le temps d'entraînement. Le pruning post-entraînement est un processus plus simple à implémenter et compatible avec des modèles déjà entraînés, mais pouvant entraîner une plus forte perte de performance. Dans les deux cas, le pruning peut être appliqué en une seule passe (one-shot pruning) ou de manière itérative, en alternant plusieurs cycles de suppression et de ré-entraînement pour atteindre progressivement le niveau de compression souhaité. Le pruning peut aussi avoir lieu durant l'inférence d'un modèle : certains poids sont ignorés à l'inférence selon l'entrée reçue, sans être supprimés durablement. Le pruning dynamique améliore donc l'efficacité à l'exécution en réduisant le nombre de calculs effectués et les transferts de données, mais n'a aucun effet sur la taille du modèle en mémoire.

Tout l'enjeu du pruning réside dans la recherche d'un équilibre entre compression du modèle et conservation d'un certain degré de performances. Enfin, au-delà du pruning, d'autres méthodes algorithmiques pour rendre l'IA frugale ou plus performante existent et peuvent être mobilisées : quantization, distillation, fine-tuning.

Mention : Zhang L., *Neural Network Pruning, ou l'art d'élaguer l'intelligence artificielle.* Décryptage Zenon Reseach. Avril 2026

ISSN 2740-3963