Natively Sparse Attention (NSA) : Comprendre la dernière innovation de DeepSeek

NSA offre une solution prometteuse pour les défis liés à la modélisation de contexte long. Il maintient les performances tout en réduisant les coûts computationnels.

Contexte et motivation de NSA

Défis des Modèles de Langage

Les modèles de langage à contexte long souffrent de coûts computationnels élevés.

Objectif de NSA

NSA vise à atténuer ces coûts tout en conservant les performances du modèle.

Innovations clés de NSA

1 **Stratégie hiérarchique dynamique**

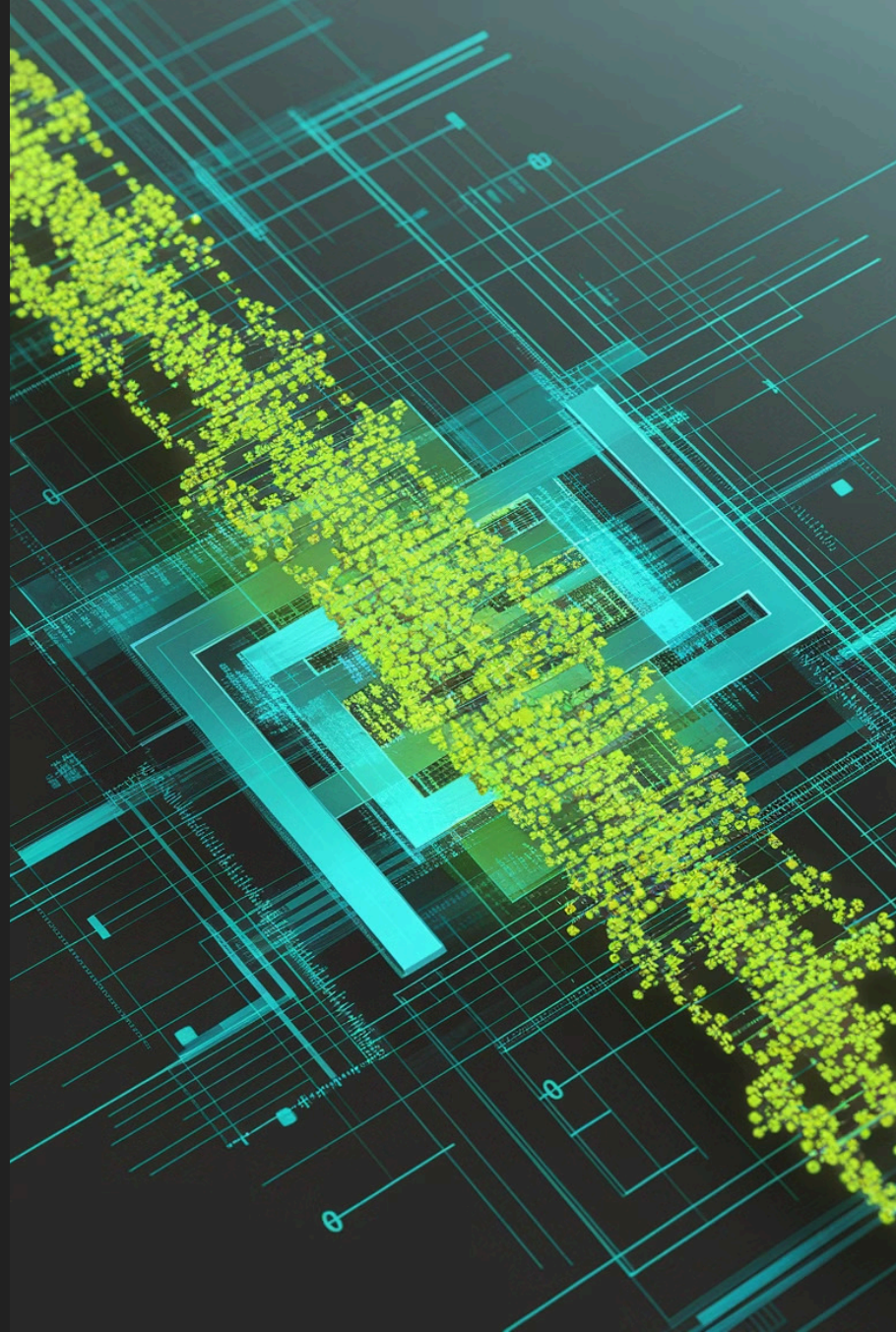
Compression grossière suivie d'une sélection fine des tokens.

2 **Optimisations matérielles**

Efficacité maximisée sur les matériels modernes.

3 **Entraînement de bout en bout**

Permet un entraînement complet avec des coûts réduits.



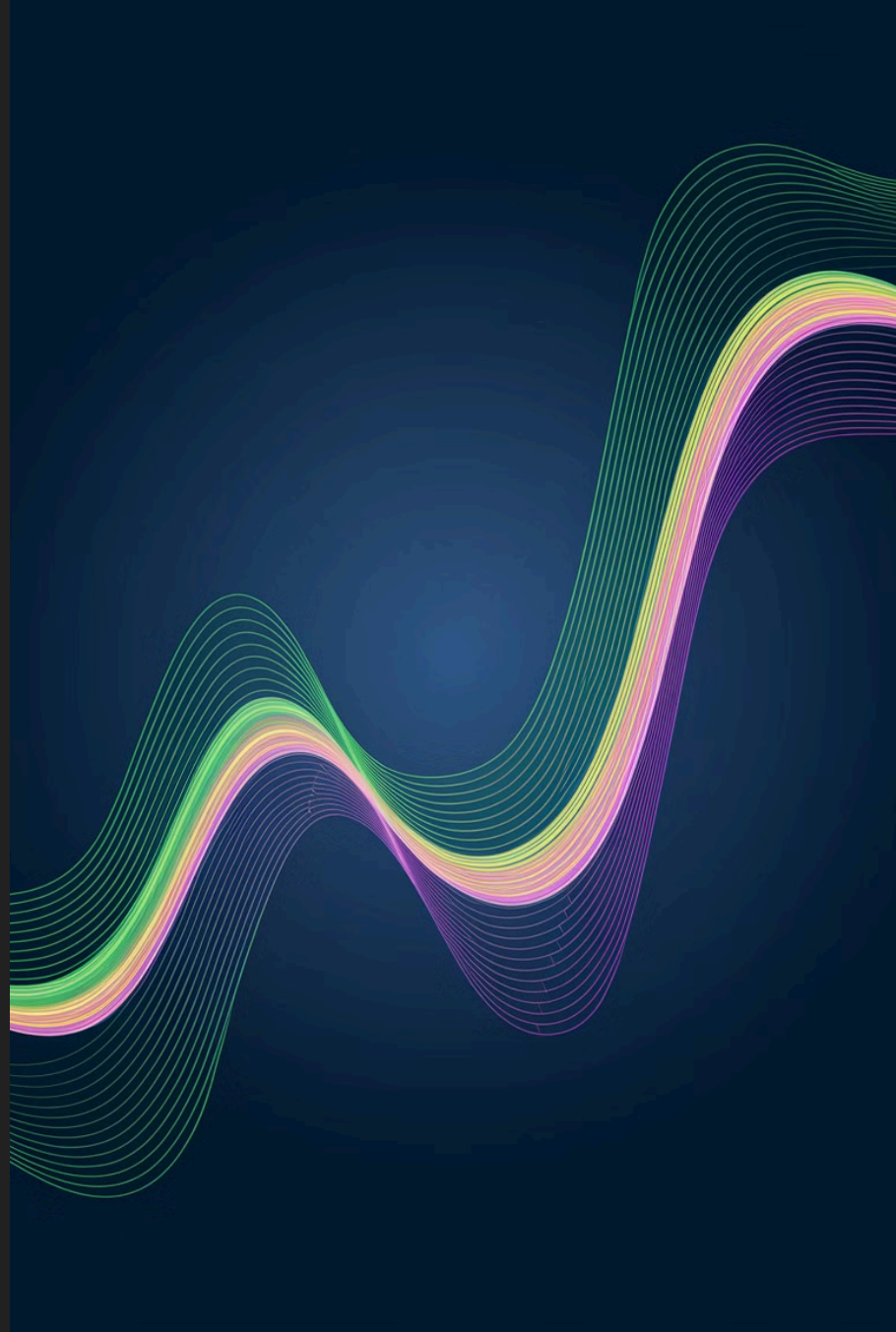
Résultats expérimentaux de NSA

Performances

NSA égale ou surpasse les modèles à attention complète.

Vitesse

Gains de vitesse significatifs pour les séquences longues.



Conception algorithmique : compression

1

Agrégation des tokens

Les tokens sont regroupés en blocs pour une gestion simplifiée.

2

Sélection basée sur l'importance

Réduction des calculs en se concentrant sur les tokens clés.

Capture du contexte proche

Fenêtre glissante

Exploitation d'une fenêtre glissante pour saisir les informations contextuelles.

Importance contextuelle

Assure une attention particulière aux relations locales.

Continuité

Maintien de la continuité sémantique au sein des séquences.

Optimisation matérielle avancée

Noyaux spécialisés

Utilisation de noyaux optimisés pour les GPU modernes.

Architectures GQA et MQA

Adaptation aux architectures GQA (Grouped-query attention) et MQA (Multi-query attention) pour une efficacité accrue.

Gains substantiels en vitesse

11.6

Facteur d'accélération

Amélioration significative de la vitesse de décodage.

5x

Entraînement

Accélération du processus d'entraînement.