Grokking Phenomenon

Un nouveau phénomène de généralisation retardée

MARREL Pierre-Emmanuel
ABIDA Youssef
VADUREL Benjamin
SEN Abdurrahman

1. Qu’est-ce que le grokking ?

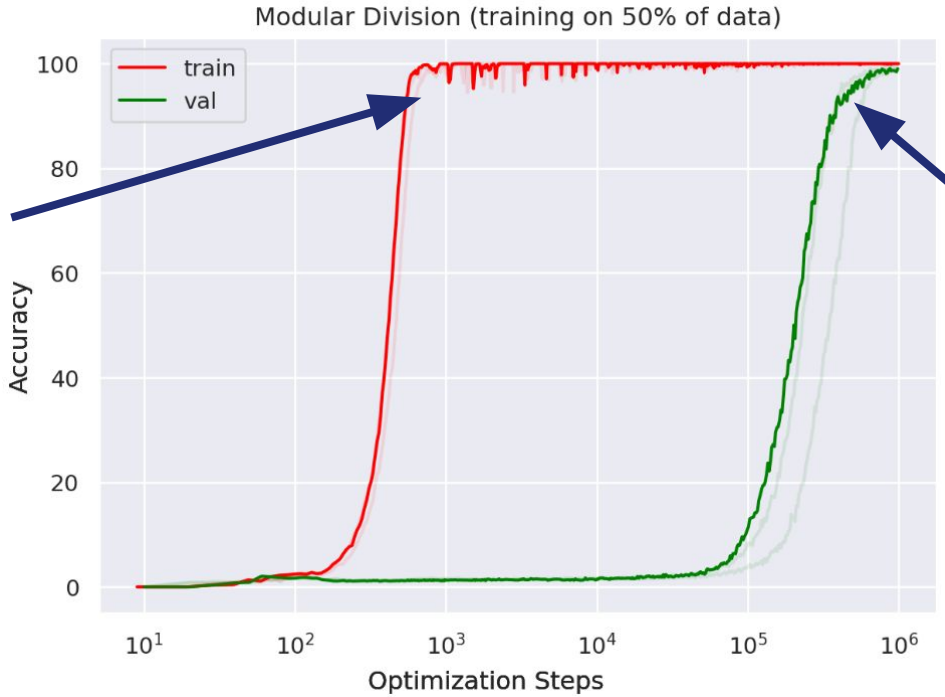
Qu’est-ce que le grokking ?

- Phénomène observé en Machine Learning
- Généralisation bien après le sur-apprentissage

Généralisation = Moment où le modèle approche 100% d’accuracy sur les tests

Qu’est-ce que le grokking ?

Sur-apprentissage



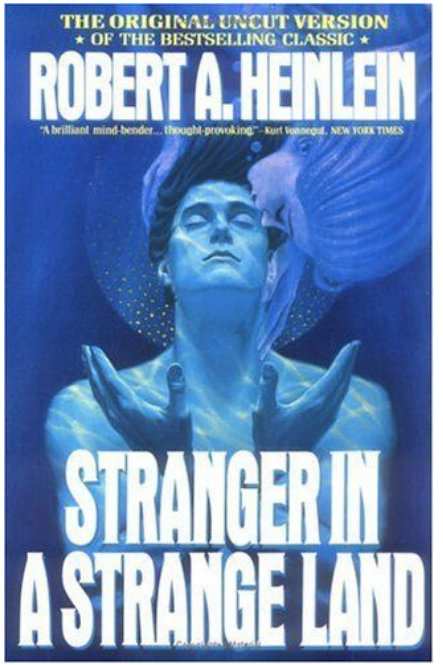
Généralisation

Power et al. (2022)
arXiv:2201.02177v

Évolution de l’accuracy sur un réseau de neurone qui grokke

Historique du Grokking

- Le terme provient de la nouvelle “Stranger in a strange land”
- Écrit en 1961
- Dans le livre :
  - To grok = Comprendre quelqu’un si profondément que l’on fait plus qu’un avec la personne.
  - “Un apprentissage que l’on ne comprend pas”



Historique du Grokking

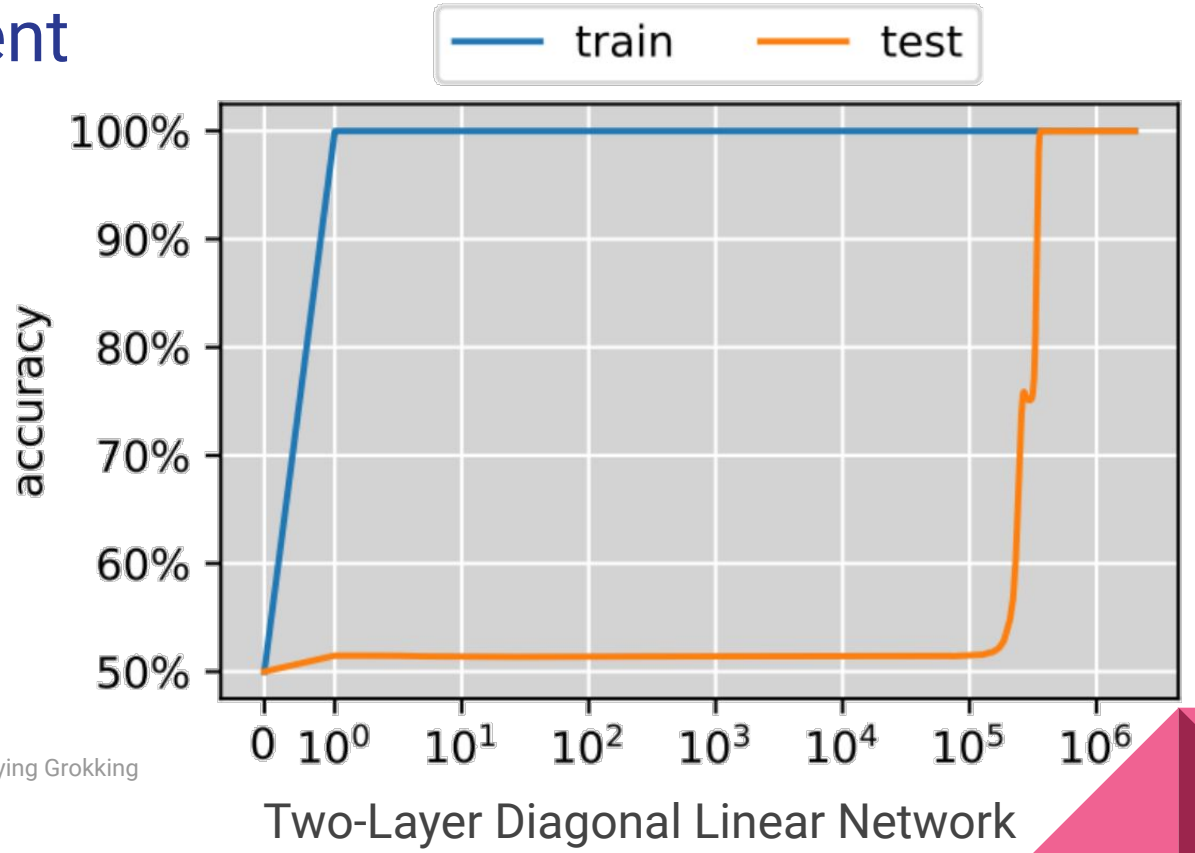
- Découvert par hasard en 2022 par des chercheurs chez OpenAI
  - Modèle entraîné pendant longtemps
  - Testé sur des modèles d'addition binaires modulaires
- Sujet de recherche actuel

2. Comment fonctionne le grokking?

Fonctionnement

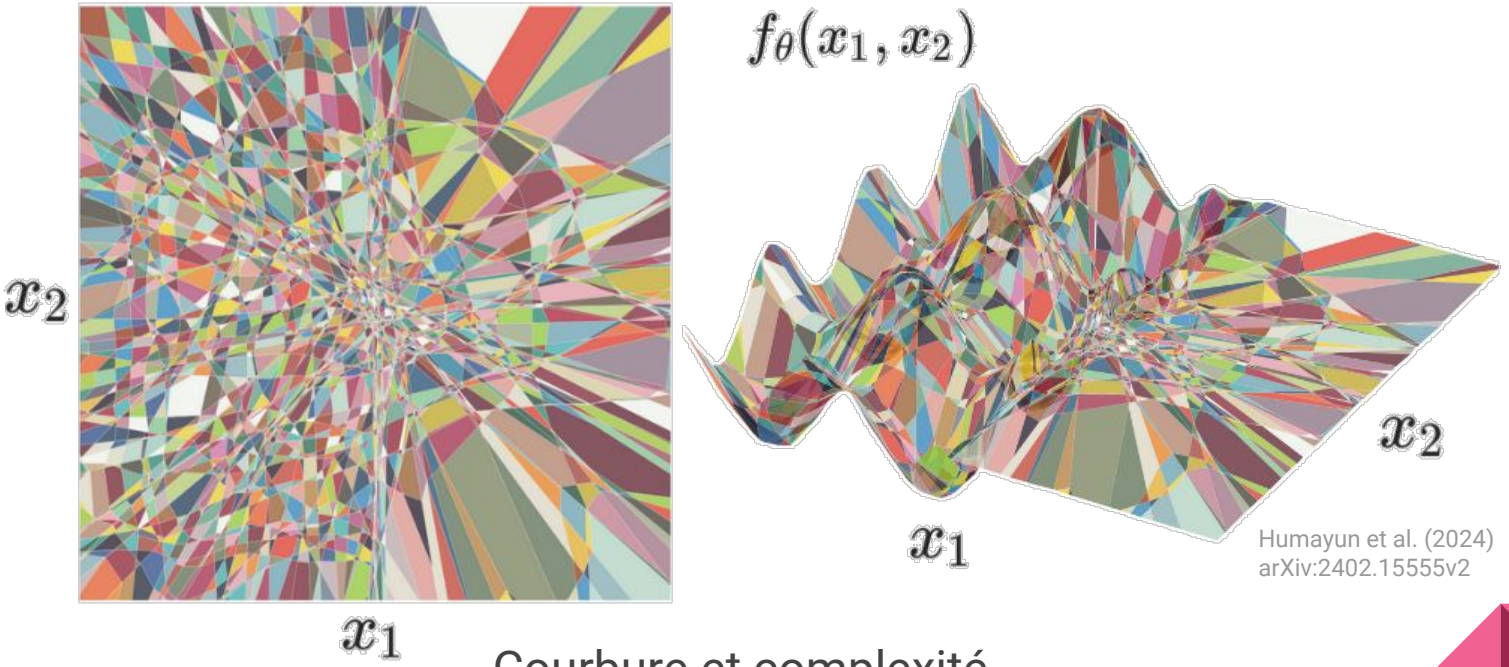
Plusieurs phases :

- Mémorisation
- Généralisation
- Cleanup



Wei Hu
Toward Demystifying Grokking

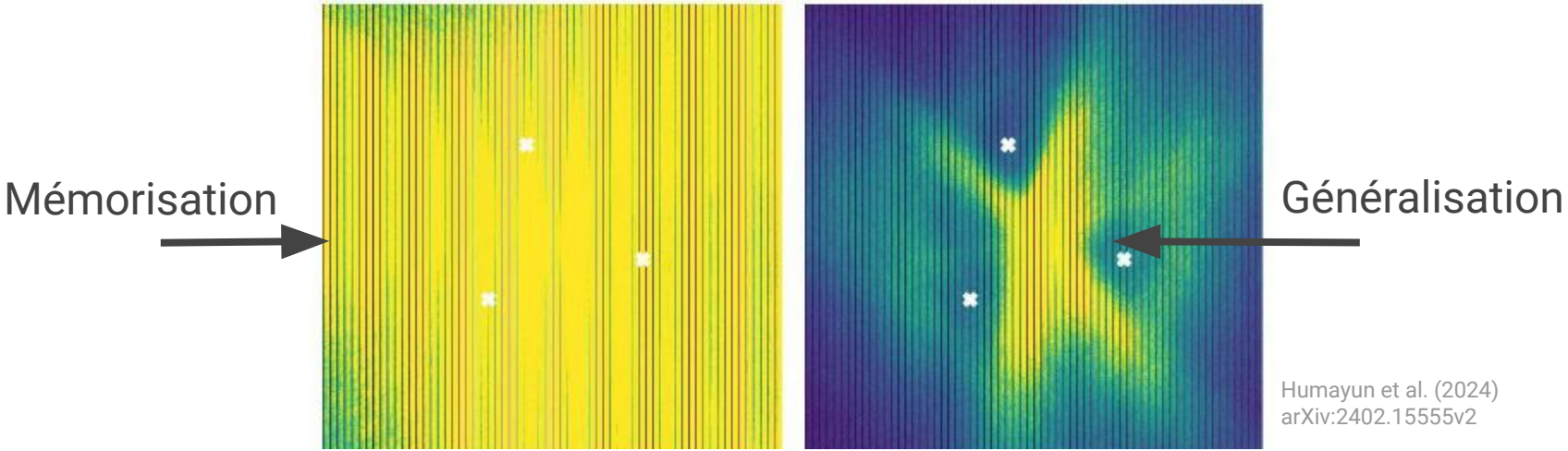
Complexité locale



Humayun et al. (2024)
arXiv:2402.15555v2

Courbure et complexité.

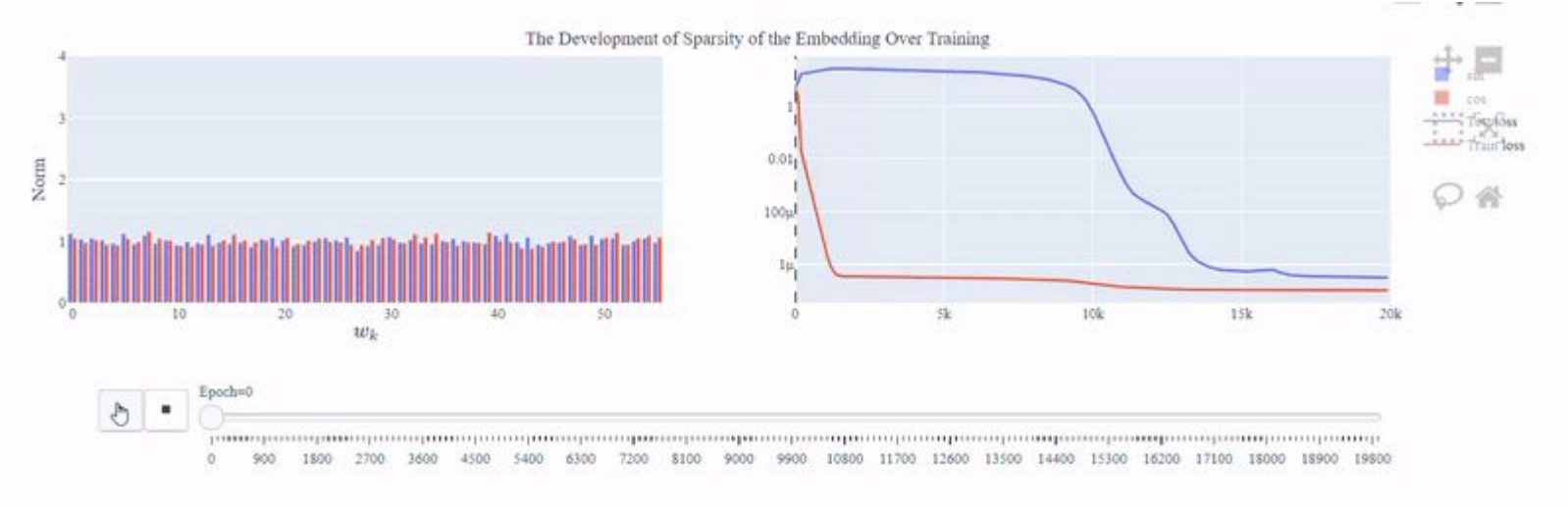
Mémorisation



Humayun et al. (2024)
arXiv:2402.15555v2

Représentation de la complexité locale et trois échantillons aléatoires

Généralisation et Cleanup



Évolution des poids normalisés

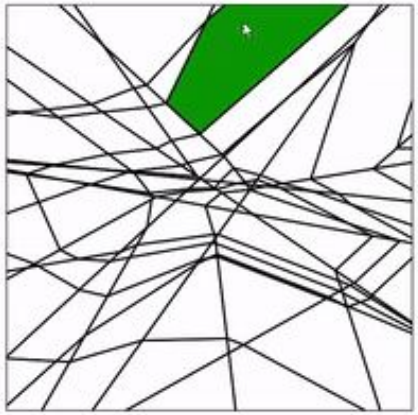
Source : https://www.neelnanda.io/grokking-paper

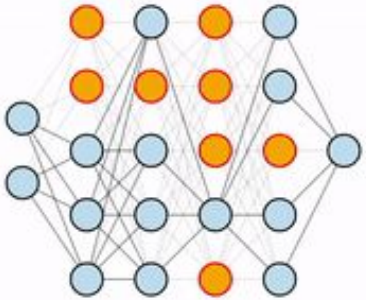
3. Mesure de progression

Mesure de progression

- Des mesures pour évaluer dans quelle phase se trouve le réseau de neurone
- Interprétabilité mécanistique

Connecting Spline Theory and Mechanistic Interpretability

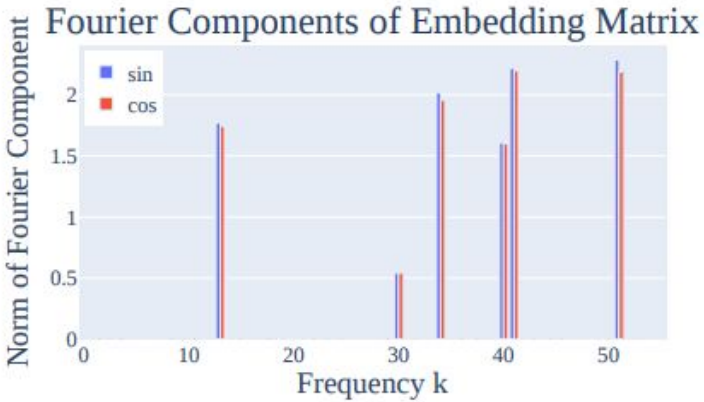


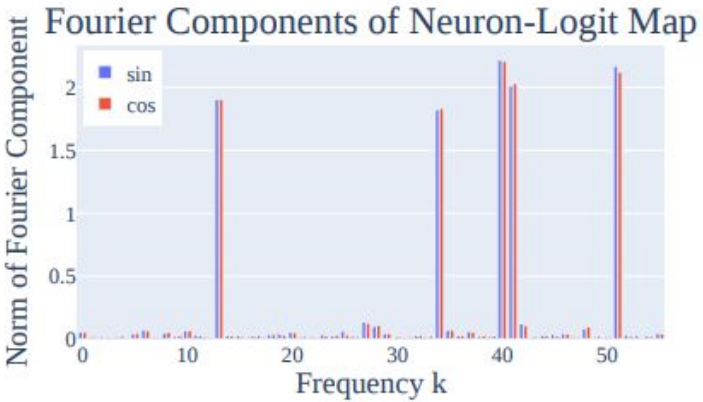


Humayun et al. (2024)
arXiv:2402.15555v2

Mesure de progression

- Transformée de Fourier pour extraire les neurones clés.

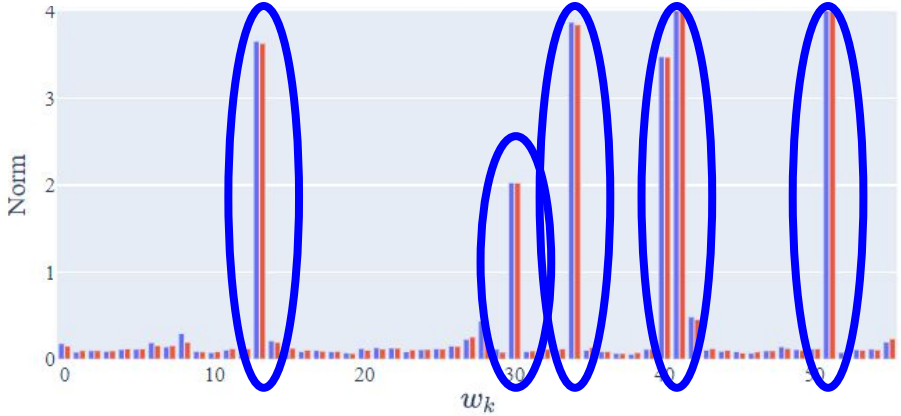


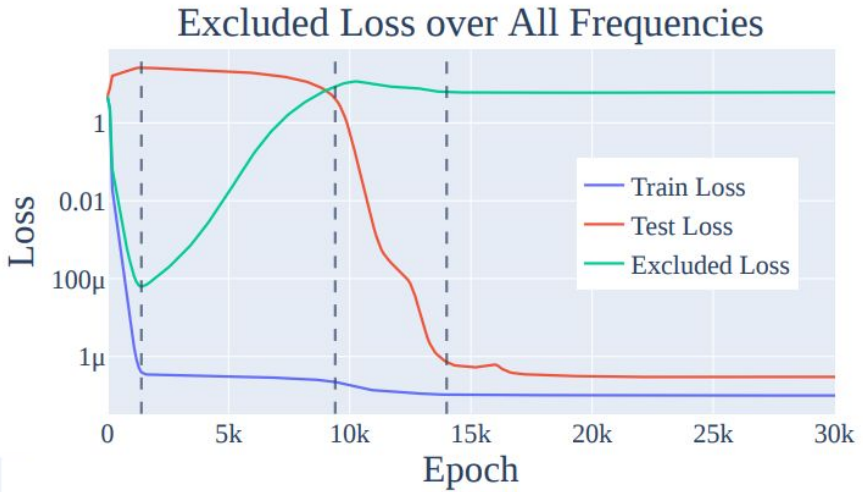


Nanda et al. (2023)
arXiv:2301.05217v3

Excluded loss

On enlève les poids importants

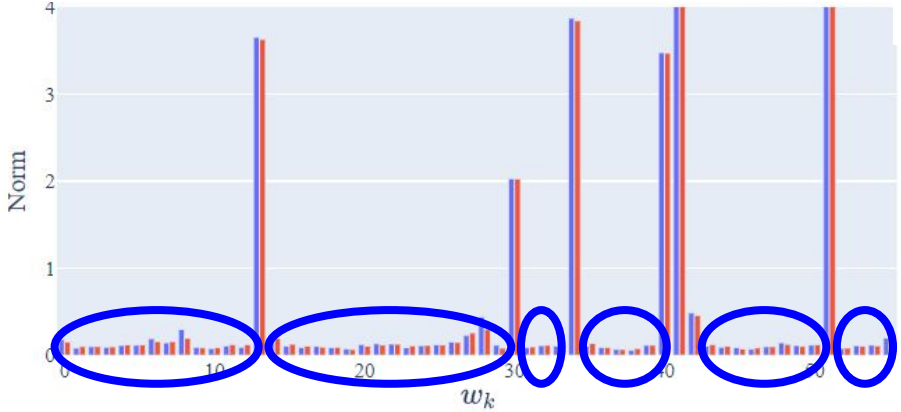


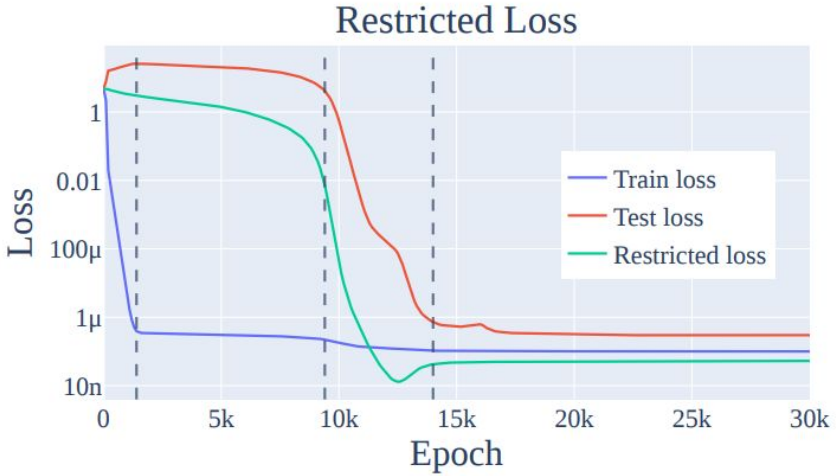


Nanda et al. (2023)
arXiv:2301.05217v3

Restricted loss

On enlève les poids faibles





Nanda et al. (2023)
arXiv:2301.05217v3

4. Différents points de vue sur les phases

Vision complexité locale

- **The first descent** (Mémorisation)
- **The ascent phase** (Généralisation)
- **The second descent phase or region migration phase** (Cleanup)

source : https://arxiv.org/pdf/2402.15555

Vision valeurs des poids

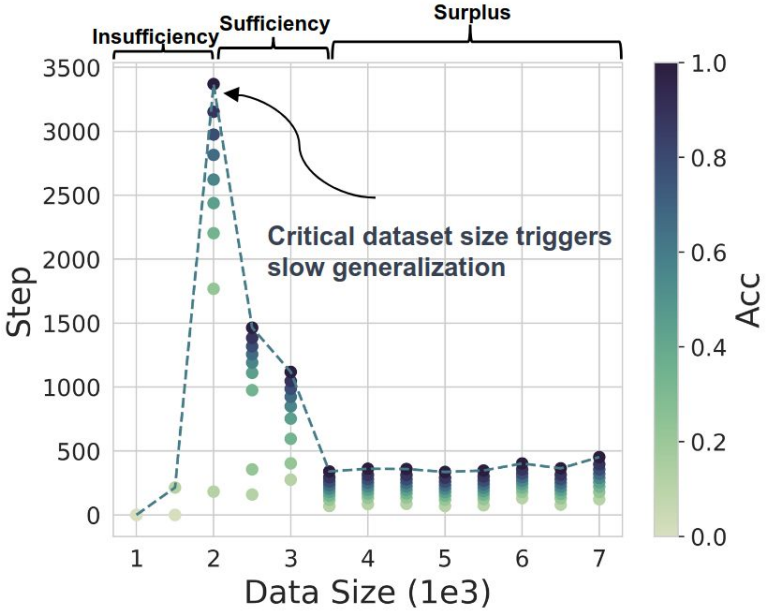
- **Memorization** → Mémorisation
- **Circuit formation** → Généralisation
- **Cleanup** → Cleanup

source : https://arxiv.org/pdf/2301.05217

5. Résultats

Importance de la quantité de données et poids

- Grokking accéléré



Zhu et al. (2024)
arXiv:2401.10463v3

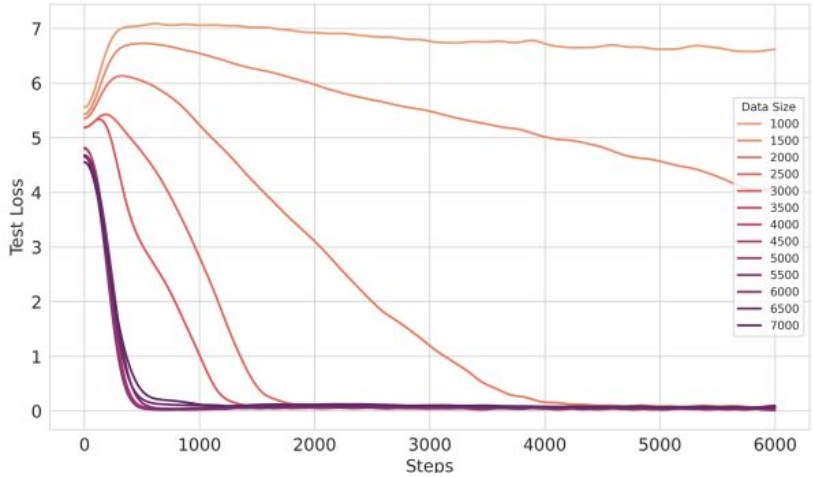
Temps nécessaire au grokking par rapport à la taille du jeu d'entraînement

Résultats

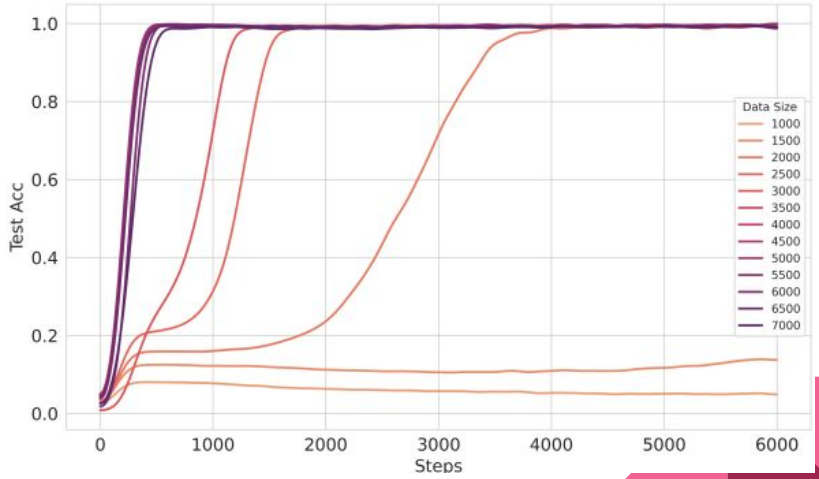
Algorithmique :

- Addition modulaire ((*a* + *b*) % *c*)

Test loss



Test accuracy

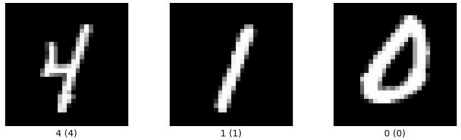


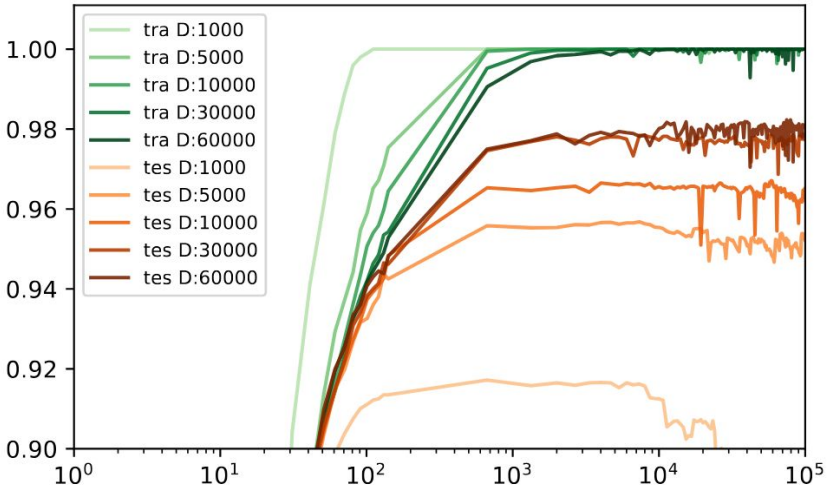
Zhu et al. (2024)
arXiv:2401.10463v3

Test loss et accuracy pour addition modulaire avec grokking

Résultats

Image
- MNIST





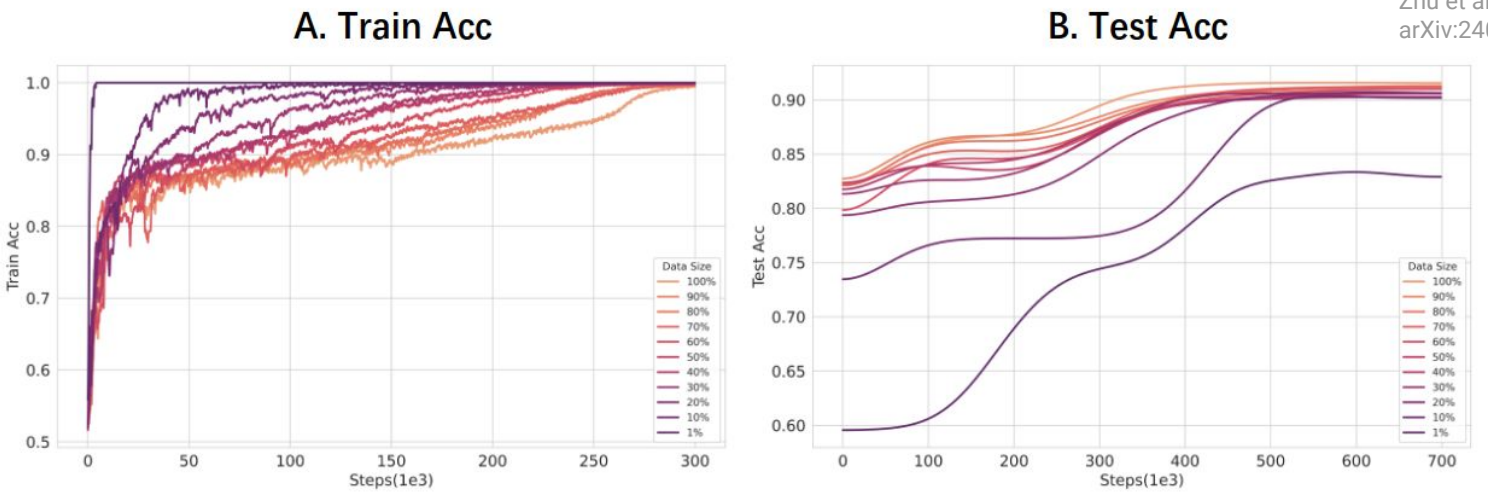
Humayun et al. (2024)
arXiv:2402.15555v2

Accuracy sur différentes tailles de dataset MNIST

Résultats

- Sur plusieurs autres dataset…
  - IMDB Yelp Langage naturel

Zhu et al. (2024)
arXiv:2401.10463v3



Accuracy sur différentes tailles de dataset Yelp

6. Cognition

Lien avec la cognition

- Apprentissage inconscient chez les humains
  - Raisonnement rapide et fiable
- Cleanup ≈ Synaptic Pruning
- Exemple : Apprentissage d’une langue

Références

- Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). **Grokking: Generalization beyond overfitting on small algorithmic datasets**. arXiv:2201.02177.
- Humayun, A. I., Balestriero, R., & Baraniuk, R. (2024). **Deep networks always grok and here is why**. arXiv:2402.15555.
- Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). **Progress measures for grokking via mechanistic interpretability**. arXiv:2301.05217.

Références

- Zhu, X., Fu, Y., Zhou, B., & Lin, Z. (2024). **Critical data size of language models from a grokking perspective**. arXiv:2401.10463.
- Liu, Z., Kitouni, O., Nolte, N. S., Michaud, E., Tegmark, M., & Williams, M. (2022). **Towards understanding grokking: An effective theory of representation learning**. arXiv:2205.10343.
- Liu, Z., Michaud, E. J., & Tegmark, M. (2022). **Omnigrok: Grokking beyond algorithmic data**. arXiv:2210.01117.

Merci pour votre attention.
Des questions ?

Débat

Qu’est-ce que l’apprentissage ?

Qu’est-ce-que l’apprentissage ?

Qu’est-ce-que l’apprentissage ?

- 4 phases de l’apprentissage
- Apprentissage par pratique supervisée
- L'apprentissage par essais et erreurs

Du point de vue écologique ?

Est ce que cela en vaut la peine ?

- Quelques chiffres :
  - Grokking : 10^5 époque
  - Entrainement de GPT-3 : 1300 MWh (Mégawatt par heure)
  - Consommation pour regarder une vidéo en ligne : 0.8 kWh (Kilowatt par heure)
  - Soit 1 625 000x plus.

Est-ce que les problématiques environnementales doivent ralentir le progrès technologique ?

Est-ce réalisable ?

Est-ce réalisable ?

- Temps d'entraînement ?
- Assez de données ?

Table 1: Statistics of the datasets used in our experiments.

Tasks	Datasets	Train	Test
Modular Arithmetic	Addition	9,576	3,193
Sentiment Classification	IMDB Yelp	21,072 352,232	7,025 117,411

Zhu et al. (2024)
arXiv:2401.10463v3

Tailles des jeux de données pour les expériences de Zhu et al.

Quelques citations

Learning Deep Architecture for AI, Y. Bengio

Humans often describe such concepts in hierarchical ways, with multiple levels of abstraction. The brain also appears to process information through multiple stages of transformation and representation. [...] Inspired by the architectural depth of the brain, neural network researchers had wanted for decades to train deep multi-layer neural networks.

Minsky, The Society of Mind, 1986

Yet, though all grown-up persons know how to do such things, no one understands how we learn to do them! And that is what will concern us here. To pile up blocks into heaps and rows: these are skills each of us learned so long ago that we can't remember learning them at all. Now they seem mere common sense-and that's what makes psychology hard. This forgetfulness,the amnesia of infancy, makes us assume that all our wonderful abilities were always there inside our minds,and we never stop to ask ourselves how they began and grew.

Références (débat)

- Luccioni, Viguier, Ligozat (2022). **Estimating The Carbon Footprint Of Bloom, A 176b Parameter Language Model**
  arXiv:2211.02001v1