ABIDA Youssef p2024398 | SEN Abdurrahman p1510824 | MARREL Pierre-Emmanuel p2006524 | VADUREL Benjamin p2000154

Grokking Phenomenon

Synthèse - IAC

I. Introduction

Le phénomène de *grokking* désigne, en machine learning, la capacité d'un réseau de neurones à généraliser bien après le sur-apprentissage. Cette émergence tardive, observée après le sur-apprentissage (ou *overfitting*) soulève des questions sur la façon dont les modèles apprennent réellement et remet en cause certaines idées de conception de l'apprentissage.

Ce phénomène a été récemment découvert en 2022 par des chercheurs de Google et OpenAI (Power et al).

Cette synthèse explore le fonctionnement du *grokking*, ses implications sur l'apprentissage des réseaux de neurones et ses liens avec la cognition.

II. Fonctionnement du Grokking

L'entraînement d'un réseau de neurones qui finit par grokker se divise en plusieurs phases : la mémorisation, la généralisation et le cleanup.

Mémorisation

Lorsqu'on entraîne un réseau de neurones, l'objectif est d'atteindre pour ses prédictions une précision proche de 100%. Une fois le réseau entraîné, il a mémorisé les données d'entraînement et arrive toujours à trouver le bon labelling sur ces données. Cependant, si le modèle est trop entraîné, il va perdre sa capacité à généraliser et à faire des prédictions correctes sur des données de test. C'est ce qu'on appelle le surentraînement. Avant, sur les données de test, la précision augmentait. Puis, lors du surentraînement, cette précision chute énormément.

Généralisation et Cleanup

Lors du surapprentissage, les poids des neurones se répartissent de manière homogène. Ce qui signifie que le modèle ne concentre pas ses calculs sur certains points. Il ne raisonne pas, il apprend juste à reconnaître les données sur lesquelles il s'entraîne.

Au bout d'un certain nombre d'étapes d'entraînement, cette complexité se déplace vers les frontières de décision du modèle, c'est-à-dire les limites qui séparent des classes. Grâce au Weight Decay, les valeurs des poids des neurones diminuent globalement excepté pour certains neurones spécifiques. Les régions de décision vont alors migrer subitement pendant la phase de Cleanup (Nanda et al). C'est à ce moment que l'on observe le grokking. La précision mesurée sur des données de tests augmente fortement et finit par dépasser l'ancienne précision avant le surentraînement.

En bref, les neurones qui ne se spécialisent pas ont fini par créer des frontières de décisions et permet une généralisation du choix de prédiction. Le modèle se base uniquement sur certains neurones clés lui permettant de prendre sa décision.

Mesures de progression

Comme beaucoup de temps est nécessaire pour qu'un réseau grok, des chercheurs ont proposé des *"mesures de progression"*, pouvant être calculées pendant l'entraînement du réseau en enlevant des ensembles de neurones et qui permettent de savoir dans quelle phase du grokking on se trouve *(Mémorisation, Généralisation ou Cleanup)* :

- Excluded Loss : en neutralisant les neurones clés, la précision devient hasardeuse;
- Restricted Loss : en neutralisant tous les autres neurones, la précision augmente un petit peu.

Importance de la quantité de données et poids

Comme expliqué précédemment, pour que le grokking ait lieu, il faut continuer d'entraîner le modèle bien au-delà de l'overfitting, il faut donc logiquement de grande quantité de données. Pour l'instant, on observe le grokking principalement sur des jeux de données algorithmiques (par exemple l'addition modulaire).

Il a été observé que l'initialisation des poids dans le réseau de neurone avait aussi un impact sur le grokking, généralement si on les initialise avec de grandes valeurs, le réseau sur-apprendra plus vite et arrive donc au grokking plus vite aussi.

III. Lien avec la cognition

Les différentes phases qui mènent au grokking peuvent rappeler le processus d'apprentissage observable chez les êtres humains : la mémorisation et la généralisation correspondent directement aux points de départ et d'arrivée dans ce processus. Le mécanisme de Weight Decay ressemble très fortement au Synaptic Pruning (élagage synaptique) qui élimine les synapses (connexions) inutiles ou redondantes du cerveau afin de le rendre plus rapide et efficace.

Le phénomène de grokking dans son ensemble peut être assimilé à un apprentissage inconscient qui, sur le long terme, permet de construire un raisonnement rapide et fiable basé sur la détection de motifs récurrents.

L'interprétation mécaniste de Neel Nanda et al soutient cette hypothèse et montre même que le grokking apparaît après le début de la phase de cleanup (i.e le pruning).

Pour illustrer l'apprentissage inconscient chez les humains, prenons l'exemple de l'apprentissage d'une langue : avec le temps, un apprenant peut déduire le sens d'un mot nouveau grâce au contexte qui l'entoure. Ce raisonnement s'apparente au grokking.

IV. Discussions sur le grokking

La littérature montre que sur certains jeux de données, un réseau qui a grok donne de bons résultats. Cependant, nous pouvons nous interroger sur la pertinence de l'utilisation du grokking. En effet, le grokking ne propose pas de bonne performances sur tous les jeux de données et il est difficile de prévoir à l'avance si un modèle va grokker.

L'entraînement et l'utilisation des modèles de machine learning constituent déjà une grande consommation d'énergie de nos jours. Faire grokker un modèle, nécessitant beaucoup plus de temps, va davantage augmenter cette consommation.

De plus, l'obtention de grand volume de données est indispensable pour observer le grokking. L'obtention de ces données pose un problème et demande un travail de labellisation conséquent.

Pour pallier les lacunes d'un modèle, il ne faut pas oublier les autres techniques de l'intelligence artificielle comme un système à base de règles (par exemple, ChatGPT qui utilise des modules, autre que de la prédiction de texte, pour les opérations arithmétiques).

Si OpenAI dit s'intéresser au grokking, ce serait pour mieux comprendre ce qu'il se passe dans les boîtes noires de modèles de machine learning, c'est-à-dire comprendre la représentation que se fait un modèle d'un problème. Cette technique est dans les recherches actuelles grâce aux promesses de bonnes performances observées.