Notes : On expressive power of self attention matrices

Tam Le

March 2026

Notes de lecture basées sur le papier Likhosherstov, V., Choromanski, K., & Weller, A. (2021). On expressive power of self-attention matrices. arXiv preprint arXiv:2106.03764.

L'expressivité décrit les relations qu'un modèle prédictif peut approximer. C'est un aspect fondamental, car elle permet de comprendre dans quelle mesure un modèle peut représenter des relations complexes. Il est important de noter que cela ne concerne pas la procédure d'entraînement, mais uniquement ce que le modèle peut représenter en théorie si les paramètres sont bien choisis.

Un aspect important est d'étudier comment l'expressivité évolue avec la taille du modèle (nombre de paramètres, largeur, profondeur, etc.).

Exemples de questions d'expressivité

- Peut-on approximer les fonctions continues, avec un réseau de neurones ? Quelle taille de modèle le permet ?
- Peut-on approximer des algorithmes classique comme le tri ?
- Peut-on copier d'autres modèles prédictifs ;k-NN, arbres, réseaux de neurones plus "classiques".

C'est une question complexe, surtout pour les modèles sophistiqués. On peut déjà étudier l'expressivité de manière plus locale, en étudiant les briques élémentaires des réseaux de neurones.

Dans ce papier, on s'intéresse à la matrice de *self-attention*

$$\mathrm{SAN}[W_Q, W_K] : X \mapsto \mathrm{diag}\big(\exp(XW_QW_K^\top X^\top)\mathbf{1}\big)^{-1}\exp(XW_QW_K^\top X^\top).$$

avec

$$X \in \mathbb{R}^{L\times d}, \qquad W_Q \in \mathbb{R}^{d\times d}, \qquad W_K \in \mathbb{R}^{d\times d}$$

On simplifie en considérant la version non normalisée

$$\mathrm{SA}[W_Q, W_K] : X \;\mapsto\; \exp(X W_Q W_K^\top X^\top).$$

Question : Quelles matrices peut on représenter avec SA ? Peut-on fixer W_Q, W_K tels que pour toute matrice A, il existe X, satisfaisant

$$A \approx \mathrm{SA}[W_Q, W_K](X)\,?$$

Le papier considère des matrices A positives, parcimonieuses, avec beaucoup de zéros sur chaque ligne (c'est motivé par des observations empiriques).

Résultat du papier (informel) : si $d = O(k \log L)$ alors il existe W_Q, W_K tels que pour toute matrice A positive avec au plus k coefficients non nuls sur les lignes, il existe X tel que

$$A \approx \mathrm{SA}[W_Q, W_K](X).$$

On remarque que, sous hypothèses de bornitude inférieure / supérieure sur les coefficients, on se ramène à devoir montrer qu'il existe W_Q, W_K tels que pour toute matrice B, il existe X telle que

$$X W_Q W_K^\top X^\top \approx B$$

En effet, soit $A = (A_{ij})$ une matrice à coefficients positifs, avec au plus k coefficients non nuls par ligne. On suppose qu'il existe $0 < m \le M < \infty$ tels que

$$m \le A_{ij} \le M \qquad \text{pour tout } (i,j) \in S := \{(i,j) : A_{ij} > 0\}.$$

Fixons $\varepsilon > 0$ et définissons

$$B_{ij} := \log\Big(\frac{A_{ij}}{\varepsilon}\Big), \qquad (i,j) \in S.$$

et 0 sinon. Si $\|B - \hat{B}\|_\infty \le \delta$ alors, en posant $\hat{A} = \varepsilon e^{\hat{B}}$ pour tout $(i,j) \in S$, $|\hat{A}_{ij} - A_{ij}| = \varepsilon \big| e^{\hat{B}_{ij}} - e^{B_{ij}} \big|$. d'où $|\hat{A}_{ij} - A_{ij}| \le \varepsilon \cdot \frac{M}{\varepsilon} e^{\delta}\, \delta = M e^{\delta} \delta, \qquad (i,j) \in S.$
et on a une erreur de ε sinon sur les coeff nuls.

Compresser la matrice Si on réécrit

$$B = B^{(1)} B^{(2)\top},$$

alors on remarque que si on dispose d'une approximation

$$\hat{B}^{(1)} \hat{B}^{(2)\top} \approx B^{(1)} B^{(2)\top}$$

avec $\hat{B}^{(1)}, \hat{B}^{(2)} \in \mathbb{R}^{L\times d}$ avec d petit, alors on pourra reconstruire une écriture de la forme $XW_QW_K^\top X^\top$. En effet (quitte à changer d en $2d$), en prenant

$$X = \begin{pmatrix}\hat{B}^{(1)} & \hat{B}^{(2)}\end{pmatrix} \in \mathbb{R}^{L\times 2d},$$

et

$$W_Q = \begin{pmatrix} I_d & 0 \\ 0 & I_d \end{pmatrix}, \qquad W_K = \begin{pmatrix} 0 & 0 \\ I_d & 0 \end{pmatrix},$$

on obtient

$$XW_Q = \begin{pmatrix}\hat{B}^{(1)} & \hat{B}^{(2)}\end{pmatrix}, \qquad XW_K = \begin{pmatrix}\hat{B}^{(2)} & 0\end{pmatrix},$$

d'où

$$XW_QW_K^\top X^\top = (XW_Q)(XW_K)^\top = \hat{B}^{(1)}\hat{B}^{(2)\top}.$$

Donc il suffit de construire une approximation de rang petit de B. On voit que les coefficients de $B^{(1)}B^{(2)\top}$ sont les produits scalaires des lignes de chacune des matrices : $(B^{(1)}B^{(2)\top})_{ij} = \langle b_i^{(1)}, b_j^{(2)}\rangle$.

On va chercher $R \in \mathbb{R}^{d\times L}$ telle que

$$\langle Rb_i^{(1)}, Rb_j^{(2)}\rangle \approx \langle b_i^{(1)}, b_j^{(2)}\rangle.$$

De sorte à pouvoir poser

$$\hat{B}^{(1)} := B^{(1)}R^\top, \qquad \hat{B}^{(2)} := B^{(2)}R^\top.$$

Alors

$$(\hat{B}^{(1)}\hat{B}^{(2)\top})_{ij} = \langle Rb_i^{(1)}, Rb_j^{(2)}\rangle \approx \langle b_i^{(1)}, b_j^{(2)}\rangle = (B^{(1)}B^{(2)\top})_{ij}.$$

Idée du papier : Johnson–Lindenstrauss. Y a-t-il R linéaire qui envoie un nombre fini de points dans une dimension plus petite tout en préservant les distances / angles ? Surtout (par rapport aux autres techniques de réduction de dimension) : A-t-on une relation entre le nombre de points et la petite dimension ?

On commence par énoncer l'inégalité de concentration suivante.

Lemme. Soit $u \in \mathbb{R}^L$ fixé, et soit $P \in \mathbb{R}^{d\times L}$ une matrice de coefficients i.i.d gaussiens $\mathcal{N}(0,1)$. Alors, si $R = \frac{P}{\sqrt{d}}$, pour tout $\varepsilon \in (0,1)$,

$$\mathbb{P}\big(\,\big|\,\|Ru\|_2^2 - \|u\|_2^2\,\big| \geq \varepsilon\|u\|_2^2\big) \leq 2\exp(-\varepsilon^2 d/8),$$

Maintenant, si on veut borner la proba de déviation sur une famille finie de n points

$$u_1, \ldots, u_n \in \mathbb{R}^L,$$

on fait simplement une union :

$$\mathbb{P}\big(\exists k \in \{1,\dots,n\},\ \big|\, \|Ru_k\|_2^2 - \|u_k\|_2^2 \,\big| \geq \varepsilon \|u_k\|_2^2\big) \leq 2n \exp(-\varepsilon^2 d/8).$$

Donc si

$$\boxed{d = O\left(\frac{\log n}{\varepsilon^2}\right),}$$

alors cette probabilité est strictement < 1, et il existe donc au moins une réalisation de R telle que, pour tout k,

$$\boxed{(1-\varepsilon)\|u_k\|_2^2 \leq \|Ru_k\|_2^2 \leq (1+\varepsilon)\|u_k\|_2^2.}$$

On peut bien sûr remplacer les u_k par des différences $x_i - x_j$. C'est la forme usuelle du lemme de Johnson–Lindenstrauss.

Le résultat se convertit (par identités de polarisation) aisément en un contrôle sur les produits scalaires.

$$\boxed{|\langle Rx_i, Rx_j\rangle - \langle x_i, x_j\rangle| \leq \varepsilon \|x_i\| \|x_j\|.}$$

Application au problème. On veut contrôler simultanément

$$\langle Rb_i^{(1)}, Rb_j^{(2)}\rangle - \langle b_i^{(1)}, b_j^{(2)}\rangle$$

pour tous les couples $1 \leq i, j \leq L$. Par JL, il existe donc R de rang $d = O\Big(\frac{\log L}{\varepsilon^2}\Big)$ telle que

$$|\langle Rb_i^{(1)}, Rb_j^{(2)}\rangle - \langle b_i^{(1)}, b_j^{(2)}\rangle| \leq \varepsilon \|b_i^{(1)}\| \|b_j^{(2)}\| \qquad \forall\, i, j.$$

Là on voit que selon le choix de la représentation $B = B^{(1)} B^{(2)\top}$ on n'obtient pas la même borne. Le papier suppose une sparsité par ligne : si on a au plus k éléments non nuls par ligne, on peut choisir $B^{(1)} = B$, $B^{(2)} = I_L$ pour avoir un terme à gauche en $O(\sqrt{k}M)$ avec M un sup sur les coefficients, ce qui donnerait un d en

$$\boxed{d = O(\frac{M^2 k \log L}{\varepsilon'})}$$

. pour une garantie $\|B - \hat{B}\|_\infty \leq \varepsilon'$. On peut imaginer des choses analogues pour la sparsité colonne etc.

Remarque: Les auteurs propose une version de JL avec $R^{\text{ort}} = Q$ avec $R = Q\widetilde{R}$ décomposition QR (notations...); cela permet d'échantillonner uniformément parmi des matrices orthogonales. Ca permet d'avoir la borne

$$2(1 - \frac{1}{L+2}) \exp(-c\varepsilon^2 d).$$

sur la proba de déviation. Ca leur permet de fixer un nombre de tirage de L pour espérer avoir un bon R, pour les expériences...