

Série d'exercices S2

(AI et LLMs)

Université Ibn Tofail

Exercice 1 : Mathématiques de l'Espace Sémantique

Sujet : Calcul de similarité et interprétation géométrique.

Dans le cours, la similarité entre deux concepts est mesurée par la *Similarité Cosinus*. Considérons un espace d'embedding très simplifié à **2 dimensions** pour faciliter le calcul. Soient les vecteurs représentant les tokens "Roi" et "Reine" :

$$\vec{v}(\text{"Roi"}) = [3, 4] \quad \text{et} \quad \vec{v}(\text{"Reine"}) = [4, 3]$$

1. Calculez la **norme** (la longueur euclidienne) de chaque vecteur ($||\vec{v}||$).
2. Calculez le **produit scalaire** (dot product) des deux vecteurs : $\vec{v}(\text{"Roi"}) \cdot \vec{v}(\text{"Reine"})$.
3. En utilisant la formule du cours, calculez la **similarité cosinus** entre ces deux mots :

$$\text{Cosinus}(\theta) = \frac{A \cdot B}{||A|| \times ||B||}$$

Le résultat est-il proche de 1 ? Interprétez ce résultat en termes de proximité sémantique.

4. Si un troisième vecteur $\vec{v}(\text{"Camion"}) = [-3, 4]$ existait, calculez sa similarité avec "Roi". Que cela signifie-t-il géométriquement par rapport à l'angle entre les vecteurs (angle aigu, droit ou obtus) ?

Exercice 2 : Au cœur de l'Attention (Self-Attention)

Sujet : Dimensions matricielles et mécanisme Q, K, V.

Dans un bloc Transformer, le mécanisme d'attention projette les embeddings d'entrée vers trois matrices : **Query (Q)**, **Key (K)** et **Value (V)**. Supposons les dimensions suivantes pour un modèle réduit :

- Dimension de l'embedding (d_{model}) = **64**
 - Longueur de la séquence (nombre de tokens) = **10**
 - Nombre de têtes d'attention (h) = **4**
1. Pour une seule tête d'attention, quelle est la dimension des matrices de poids W^Q , W^K et W^V si l'on suit la règle standard $d_v = d_k = d_{model}/h$?
 2. Si la matrice Q représente "ce que je cherche" et la matrice K représente "ce que je contiens", expliquez pourquoi nous effectuons le produit scalaire $Q \cdot K^T$. Que représente le résultat brut (avant softmax) de cette opération ?
 3. Après avoir calculé les scores d'attention et appliqué la fonction Softmax, nous multiplions le résultat par la matrice **V (Values)**. Pourquoi ?
 4. *Mise en situation* : Si l'attention (poids softmax) pour le mot "Le" sur le mot "Chat" est de 0.9 et sur le mot "Chien" est de 0.1, comment est construit le nouveau vecteur contextuel du mot "Le" à partir des vecteurs V_{Chat} et V_{Chien} ?

Exercice 3 : Architecture et Nombre de Paramètres

Sujet : Comprendre le coût mémoire d'un LLM.

La série S1 demandait la taille de la matrice d'embedding. Allons plus loin en analysant un **Bloc Transformer** complet. Un bloc est composé de deux sous-couches principales : l'Attention Multi-têtes et le réseau Feed-Forward (FFN).

Données : Dimension du modèle $d_{model} = 512$.

1. **Feed-Forward Network (FFN) :** Le cours explique que le FFN projette le vecteur dans une dimension plus grande (expansion) avant de le réduire (contraction). Le facteur d'expansion standard est $\times 4$.
 - (a) Combien de poids contient la première couche linéaire du FFN (matrice de taille $512 \rightarrow 2048$) ?
 - (b) Combien de poids contient la seconde couche (retour vers $2048 \rightarrow 512$) ?
 2. **Attention vs FFN :** En comparant vos résultats, quelle partie du bloc consomme le plus de mémoire pour stocker les paramètres : le mécanisme d'attention ou le réseau FFN ?
-

Exercice 4 : Descente de Gradient (Calcul manuel)

Sujet : Mise à jour des poids et "Manager Prudent".

Le cours utilise l'analogie du "Manager Prudent" pour le taux d'apprentissage (Learning Rate). Appliquons cela mathématiquement. Soit un poids unique $w = 1.5$ dans le modèle. Après une "Forward Pass", nous calculons une erreur (Loss). Lors de la "Backward Pass", le gradient calculé pour ce poids est :

$$\frac{\partial E}{\partial w} = +2.0$$

1. Que signifie un gradient de $+2.0$? Si nous augmentons la valeur de w , l'erreur va-t-elle augmenter ou diminuer ?
 2. Nous appliquons l'algorithme d'optimisation avec un **Learning Rate (η) de 0.01**. Calculez la nouvelle valeur du poids $w_{nouveau}$.
 3. Si le Learning Rate avait été de **100** (absurdement haut), quelle aurait été la nouvelle valeur ? Pourquoi cela illustre-t-il le risque de "divergence" ou d'instabilité mentionné dans le cours ?
-

Exercice 5 (Bonus de réflexion) : Le Masque

Sujet : Différence Encodeur vs Décodeur.

Dans le schéma de l'architecture Transformer (Figure 1 du cours), le bloc Décodeur utilise une couche appelée "**Masked Multi-Head Attention**".

Question : Pourquoi ce "masque" est-il crucial lors de la phase d'entraînement ? Si le modèle pouvait "voir" les tokens à droite de sa position actuelle (le futur), pourquoi l'apprentissage de la prédiction du mot suivant deviendrait-il trivial et inutile ?