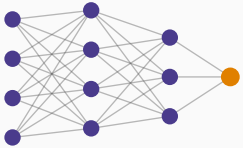
Apprentissage profond

Chapitre 10 : L'attention et les transformeurs



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Plan du chapitre

L’attention

Regarder directement, plutôt que se souvenir

Les deux défauts du récurrent

1. Le traitement est **séquentiel** : le pas t exige que le pas $t-1$ soit terminé. Impossible de paralléliser.
2. L'information doit **transiter** par tous les pas intermédiaires pour relier deux instants éloignés.

L'attention supprime les deux : chaque position peut consulter **directement** toutes les autres.

Requête, clé, valeur

Chaque position produit trois vecteurs :

Vecteur	Rôle
Requête q	ce que cette position cherche
Clé k	ce que cette position propose
Valeur v	ce qu'elle transmet si on la retient

Définition

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$

Le produit $QK^\top$ mesure la **compatibilité** entre chaque requête et chaque clé. Le softmax en fait des poids de somme 1. On en tire une moyenne pondérée des valeurs.

Pourquoi diviser par la racine de d

Intuition

En grande dimension, les produits scalaires prennent de grandes valeurs. Le softmax devient alors presque binaire — un poids proche de 1, tous les autres proches de 0.

Les gradients s'évanouissent, comme au chapitre 3. La division par $\sqrt{d}$ ramène les scores à une échelle où le softmax reste **doux** et différentiable utilement.

C'est un détail d'apparence, et sans lui rien ne s'entraîne.

Le transformeur

Les briques d’un bloc

1. **Attention multi-têtes** — plusieurs attentions en parallèle, chacune apprenant un type de relation.
2. **Connexion résiduelle** puis **normalisation par couche**.
3. Un petit réseau **entièrement connecté**, appliqué position par position.
4. Résiduelle et normalisation à nouveau.

On empile ces blocs.

Rien de neuf dans les briques

Les connexions résiduelles viennent du chapitre 8, la normalisation par couche du chapitre 7, le réseau dense du chapitre 2, le softmax du chapitre 2.

Seule l’attention est nouvelle. Le transformeur est un assemblage d’éléments déjà vus, autour d’une idée unique — et c’est ce qui le rend intelligible.

L’encodage de position

L’attention est **insensible à l’ordre** : elle traite la séquence comme un ensemble.

Il faut donc **injecter la position** dans les vecteurs d’entrée, par des fonctions sinusoïdales ou des vecteurs appris. Sans cela, une série chronologique et sa version mélangée seraient traitées

Ce que cela coûte

Une complexité quadratique

Le coût

Chaque position se compare à toutes : le coût croît comme le **carré** de la longueur de séquence.

Longueur	Comparaisons
100	10 000
1 000	1 000 000
10 000	100 000 000

L’arbitrage face au récurrent

Intuition

	Récurrent	Transformeur
Coût en longueur	**linéaire**	**quadratique**
Parallélisation	impossible	**totale**
Portée de la mémoire	limitée	**directe**
Données requises	modérées	**beaucoup**

Le transformeur coûte plus cher en calcul par séquence, mais s’entraîne bien plus vite en pratique — parce que tout se calcule **en parallèle** au lieu de pas à pas.

Une mise en garde pour les séries financières

Application en gestion

Les transformeurs dominent le texte parce que les corpus sont **immenses** — des milliards de mots.

Une série financière quotidienne sur vingt ans compte environ **cinq mille points**. C'est infime au regard de ce qu'un transformeur exige pour apprendre ses représentations.

Sur ces volumes, les modèles simples — GARCH, ARMA, boosting sur variables construites — restent le plus souvent supérieurs. **La taille des données commande le choix d'architecture, davantage que la nature du problème.**

Ce qu’il faut retenir

Le résumé du chapitre

Cinq points

1. L'**attention** permet à chaque position de consulter directement toutes les autres.
2. **Requête**, **clé**, **valeur** ; le softmax du produit $QK^\top$ donne les poids.
3. La division par $\sqrt{d}$ évite la saturation du softmax — indispensable.
4. Un bloc de transformeur assemble attention, **résiduelle**, **layer norm** et réseau dense.
5. Le coût est **quadratique** en longueur, mais tout se calcule en parallèle.

La suite

Toutes les architectures du cours sont en place. Le chapitre 11 les applique à la prévision financière, et les compare honnêtement aux modèles économétriques.