Transformer Layer: Notițe și Arhitectură

1. Definiții și Parametri

- n = lungimea secvenței în tokeni (ex. $n = 3$)
- d_{model} = dimensiunea embeddingurilor (ex. $d_{\text{model}} = 8$)
- h = număr de capete (heads) (ex. $h = 2$)
- $d = \dfrac{d_{\text{model}}}{h}$ dimensiune per head
- $X \in \mathbb{R}^{n \times d_{\text{model}}}$ (input)

Parametri antrenabili (Trainable params):

$$W_Q,\ W_K,\ W_V \in \mathbb{R}^{d_{\text{model}} \times d}, \qquad W_O \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$$

2. Exemplu numeric: Splitarea Matricei Q

Conform notițelor, matricea Q se sparge în head-uri distincte. Fie $Q \in \mathbb{R}^{n \times d_{\text{model}}}$ cu $n = 3, d_{\text{model}} = 8, h = 2$:

$$Q = \underbrace{\left[\begin{array}{cccc|cccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 1 & 0 & 1 & 3 & 3 & 2 & 2 \\ 0 & 1 & 1 & 0 & 9 & 8 & 7 & 6 \end{array}\right]}_{\begin{array}{cc} Q^{(1)} & Q^{(2)} \end{array}} \Bigg\} n = 3 \quad \in \mathbb{R}^{n \times d_{\text{model}}/h}$$

3. Algoritmul MHA (Pași)

1. Ai matricele W_Q, W_K, W_V (parametrii trainable).
2. Calculezi Q, K, V (prin înmulțirea X cu ponderile).
3. Spargi Q, K, V în head-uri $Q^{(1)}, Q^{(2)}, \ldots$ (vezi exemplul de mai sus).
4. Pentru fiecare head faci calculele din schemă ($Attention(Q, K, V)$).
5. Concatenezi matricele head $\Rightarrow$ MHA(X).

4. Schema Logică Detaliată: Multi-Head Attention

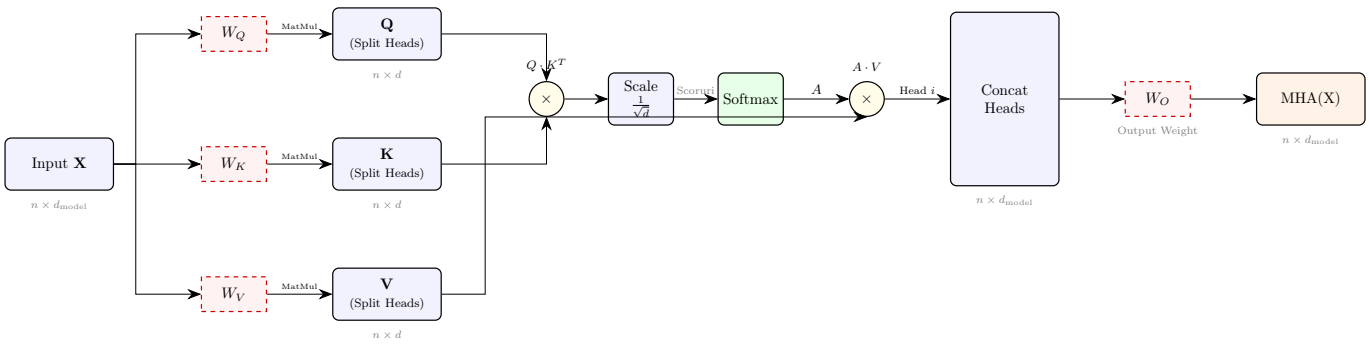

Figure 1: Fluxul de date în mecanismul Multi-Head Attention (Schema Explicită)

5. Formule Matematice

$$\mathrm{MHA}(X) = \mathrm{concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_h)\, W_O \in \mathbb{R}^{n \times d_{\mathrm{model}}}$$

Pentru fiecare head:

$$\mathrm{head}_i = A\,V \in \mathbb{R}^{n \times d}$$

$$A = \mathrm{softmax}(\tilde{S}) \in \mathbb{R}^{n \times n}, \qquad V = XW_V \in \mathbb{R}^{n \times d}$$

$$\tilde{S} = \frac{1}{\sqrt{d}}\, S, \qquad \text{unde} \qquad S = QK^\top \in \mathbb{R}^{n \times n}$$

$$Q = XW_Q, \qquad K = XW_K \qquad (\in \mathbb{R}^{n \times d})$$

Add & Norm (Layer Normalization)

După atenție, se aplică un strat rezidual și normalizare:

$$Y = \mathrm{LayerNorm}(X + \mathrm{MHA}(X))$$

Statistici per token: Se calculează media μ și varianța σ^2 pentru fiecare vector z (rând) din matrice.

$$\mu = \frac{1}{d_{\mathrm{model}}} \sum_{j=1}^{d_{\mathrm{model}}} z_j, \qquad \sigma^2 = \frac{1}{d_{\mathrm{model}}} \sum_{j=1}^{d_{\mathrm{model}}} (z_j - \mu)^2$$

Formula Layer Norm:

$$\mathrm{LN}(z) = \gamma \frac{z - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta$$

- **Intuiție:** Mutăm datele sub forma unei distribuții normale standard ($\mu = 0, \sigma = 1$).
- γ, β = parametri antrenabili.
- ϵ = termen de stabilizare numerică.

3. Feed Forward position-wise (FFN)

Structură minim necesară:

- proiecție liniară (echivalent cu strat fully-connected)
- funcție neliniară (ReLU sau echivalenții)
- proiecție liniară

a) Proiecție liniară

Formula: $XW_1 + b_1$

Exemplu:

$$X = \begin{bmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Parametrii trainable:

$$W_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Calcul:

$$\begin{bmatrix} \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \end{bmatrix} (\text{rezultat parțial}) + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}^T = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 \\ 1 & 2 & 1 \end{bmatrix}$$

Intuiție: la col 2 adăugăm 1.

---

4. Backprop step-by-step

a) FFN

Forward:

1. $V = XW_1 + b_1 = \begin{bmatrix} 3 & 2 & 1 \\ 1 & 2 & 1 \end{bmatrix}$
2. $V = \text{ReLU}(V) = \begin{bmatrix} 3 & 2 & 1 \\ 1 & 2 & 1 \end{bmatrix}$
3. $Y = VW_2 + b_2 = \begin{bmatrix} 4 & 2 & 5 & 2 \\ 2 & 2 & 3 & 2 \end{bmatrix}$

Target: $T \rightarrow$ matrice cu 0 și 1. $T = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$.
Pierdere (Loss): $L = \frac{1}{2} \sum_{i,j} (Y_{ij} - T_{ij})^2$.
Gradientul către W_2, b_2, V:

$$\delta Y \triangleq \frac{\partial L}{\partial Y} \quad \text{(gradientul pierderii față de ieșirea stratului)}$$

$$\frac{\partial L}{\partial Y_{ij}} = \frac{1}{2} \cdot 2(Y_{ij} - T_{ij}) \implies \frac{\partial L}{\partial Y} = Y - T = \delta Y$$

$$\Rightarrow \delta Y = \begin{bmatrix} 3 & 2 & 5 & 1 \\ 2 & 2 & 2 & 1 \end{bmatrix} \quad \text{(gradientul pierderii față de ieșire)}$$

Gradientul față de W_2:

$$\frac{\partial L}{\partial W_2} = V^T \cdot \delta Y$$

$$\Rightarrow \frac{\partial L}{\partial W_2} = \begin{bmatrix} 3 & 1 \\ 2 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 & 5 & 1 \\ 2 & 2 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 8 & 17 & 4 \\ 10 & 8 & 14 & 4 \\ 5 & 4 & 7 & 2 \end{bmatrix}$$

Rule update: $W_2 \leftarrow W_2 - \eta \frac{\partial L}{\partial W_2}$.
Gradient față de b_2:

$$\frac{\partial L}{\partial b_2} = \sum_{i=1}^{2} \delta Y_i = [5 \quad 4 \quad 7 \quad 2]$$

Rule update: $b_2 \leftarrow b_2 - \eta \frac{\partial L}{\partial b_2}$.
Gradientul față de V:

$$\frac{\partial L}{\partial V} = \delta Y \cdot W_2^T$$

Intuiție:

$$Y = VW_2 + b_2$$

Considerăm L dependent de Y (practic loss-ul depinde de output-ul stratului).

$$\frac{\partial L}{\partial V} = \frac{\partial L}{\partial Y} \cdot \frac{\partial Y}{\partial V}$$

Regula lanțului se aplică în cazul ăsta pt că Loss-ul e influențat de V, dar prin intermediul lui Y.

$$\text{Gradientul } \delta Y = \frac{\partial L}{\partial Y} \Rightarrow \frac{\partial L}{\partial V} = \delta Y \cdot \frac{\partial Y}{\partial V}$$

Pt $Y = VW_2$ derivează $W_2^T \Rightarrow \frac{\partial L}{\partial V} = \delta Y \cdot W_2^T$.

---

5. Tiled Mat Mult

Matricele A și B (împărțite în blocuri 2×2):

$$A = \left[\begin{array}{cc|cc} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ \hline 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{array}\right], \quad B = \left[\begin{array}{cc|cc} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ \hline 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \end{array}\right]$$

Calculul pe blocuri : $C_{ij} = \sum A_{it} B_{tj}$:

$$A_{00} \cdot B_{00} = \begin{bmatrix} 1 & 2 \\ 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 5 & 5 \\ 17 & 17 \end{bmatrix}$$

$$A_{01} \cdot B_{10} = \begin{bmatrix} 3 & 4 \\ 7 & 8 \end{bmatrix} \begin{bmatrix} 3 & 3 \\ 4 & 4 \end{bmatrix} = \begin{bmatrix} 25 & 25 \\ 53 & 53 \end{bmatrix}$$

$$C_{00} = \begin{bmatrix} 30 & 30 \\ 70 & 70 \end{bmatrix}$$