Transformers - Mathematical derivation.

From the paper *Attention is all you need.*

Input: $x_i \in \mathbb{R}^d$. $(x_i)_{1 \le i \le n}$.

① Multi-head attention

Attention vectors are computed from each input x_i, $1 \le i \le n$, and independently for each "head" h, $1 \le h \le H$.

Keys : $k_h(x_i) = W_{h,k}^T x_i$

Queries : $q_h(x_i) = W_{h,q}^T x_i$

Values : $v_h(x_i) = W_{h,v}^T x_i$

② Attention weights

For all $1 \le i, j \le n$, $1 \le h \le H$,

$$\alpha_h(i,j) = \mathrm{Softmax}\left(\{q_h(x_i)^T k_h(x_j)\}_{1 \le j \le n}\right)$$

where $\mathrm{Softmax}(z_1, \dots, z_n) = \frac{1}{\sum_{1 \le i \le n} e^{z_i}} (e^{z_1}, \dots, e^{z_n})$.

③ Mixture of values

. Define for all $1 \le i \le n$ $\quad u_i = \sum_{h=1}^{H} W_{u,h}^T \Big(\underbrace{\sum_{j=1}^{n} \alpha_h(i,j) v_h(x_j)}_{\text{Mixture of values for each input on head } h.} \Big)$

. Layer normalization of the (u_i). $\longrightarrow$ $u_i \leftarrow \text{Layer norm}(u_i + x_i)$.

④ outputs.

. For all $1 \le i \le n$ $\quad z_i = W_{z,1}^T \, \sigma\left(W_{z,2}^T u_i\right)$

. Layer normalization of the (z_i). $\longrightarrow$ $z_i \leftarrow \text{Layer norm}(z_i + u_i)$.

Layer normalization of a vector $(v_1, \dots, v_n) = v$: $\quad v \leftarrow \beta_1 \underbrace{\sigma_v^{-1}}_{\text{empirical std}} (v - \underbrace{\mu_v}_{\text{empirical mean}}) + \beta_2$

Steps 1 to 4 provide a Regression function $T_\theta : (x_1, \dots, x_n) \longmapsto (z_1, \dots, z_n)$.

In practice, a Transformer network is given by: $T_{\theta_L} \circ \dots \circ T_{\theta_1}$!

⑤ Positional encoding.

Inputs are considered as unordered vectors to compute and assign attention weights. If input data are sequential (i.e i refers to a time index), several additional positional encodings have been considered.

one-hot : $X_i \leftarrow (X_i, e_i)^T$ e_i, i-th canonical vector of $\mathbb{R}^n$.

Sinusoidal: $p_{k,2i} = \sin\left(\frac{k}{\gamma^{2i/d}}\right)$; $p_{k,2i+1} = \cos\left(\frac{k}{\gamma^{2i/d}}\right)$

$$z = W^T \sigma(W^T u) + p.$$

(6) Connection to RNN - Time series.

Next session!