Repo : Language Models with Context Re-Positioning

20 January 2026 21:42

Rope (Rotary pos Encod) : Linearly token pos assign e.g $(t_1, t_2, t_3 \dots t_n) \rightarrow$ n sequence

REPO (Content Reposition) : Dynamically token pos assign

Two Components :

① Pos Representation

$r_i = \text{Swish}(h_i W_g) \odot (h_i W_c)$

$r_i \rightarrow R^{d_p}$

Swish linear gate $(x) \rightarrow x \cdot \sigma(x)$

* $R^{d_p} = d/8$ for efficiency → 8 token pos at a time

$h_i \rightarrow$ hidden Emb

$W_g \rightarrow$ weight matrice for gating mechanism $[\in R^{d \times d_p}]$

$W_c \rightarrow$ " " " Content pathway $[\in R^{d \times d_p}]$

$\odot \rightarrow$ hadamard prod / element-wise multi

W_g # when a particular token gets reorganized in Content, how that token will influenced other tokens, that is learned by gating Weight matrice (W_g). It will score that token based on this, if score is high it will make it attend forward in pos, vice-versa.

$\text{gate}_i[j] = (h_i W_g)[j] \cdot \sigma((h_i W_g)[j])$

↳ Each dim $j \in [1, d_p]$ ↳ (Sigmoid) values stays in $[0, 1]$ when inputs are neg.

W_c # After the scoring by the gate, the Content matrix gives pos signal. if the words semantically similar they will group/cluster together even if seq distant. It might learn event causality rather than chronological order like repositioning the "reason" before the "Consequence".

$h_i W_c : R^d \rightarrow R^{d_p}$

Semantic → Pos state

② Pos Assignment

$z_i = r_i W_z$ ↳ $W_z \in R^{d_p \times 1}$

r_i gives a diff pos hypothesis for each d_p :

Dim 1 : How imp is this token ?

Dim 2 : Is this a sub or obj ?

and so on ....

W_z learns to combine these hypothesis into a single pos val :

$z_i = \sum_{j=1}^{d_p} r_i[j] \cdot W_z[j]$

Component ① + ② :

$f_\phi(h_i) = [\text{Swish}(h_i W_g) \odot (h_i W_c)] W_z$

REPO Attn :

$$A_{ij}^{\text{Repo}} = q_i^T g_\theta(f_\phi(h_j) - f_\phi(h_i)) k_j$$

$$= [q_i^T g_\theta(z_j - z_i) k_j]$$