Pseudo code for Roformer

Xuhui Zhan

March 4, 2024

Algorithm with RoPE: $P \leftarrow DTransformerRoPE(x|\theta)$

/ GPT, a decoder-only transformer, applied RoPE, forward pass */*

Input: $x \in V^*$, a sequence of token IDs.

Output: $P \in (0,1)^{N_V \times \text{length}(x)}$, where the t-th column of P represents $\hat{P}_\theta(x[t+1]|x[1:t])$.

Hyperparameters:L, D, d_e, $d_{mlp} \in \mathbb{N}$, R_e the pre-defiend rotary matrix

Parameters: θ includes all of the following parameters:

- $W_e \in \mathbb{R}^{d_e \times N_V}$ the token embedding matrices.
- For $l \in [L]$:
  - W_l, multi-head attention parameters for layer l
  - $\gamma_l^1, \beta_l^1, \gamma_l^2, \beta_l^2 \in \mathbb{R}^{d_e}$, two sets of layer-norm parameters,
  - $W_{mlp1}^l \in \mathbb{R}^{d_{mlp} \times d_e}$, $b_{mlp1}^l \in \mathbb{R}^{d_{mlp}}$, $W_{mlp2}^l \in \mathbb{R}^{d_e \times d_{mlp}}$, $b_{mlp2}^l \in \mathbb{R}^{d_e}$, MLP parameters.
- $\gamma, \beta \in \mathbb{R}^{d_e}$, final layer-norm parameters.
- $W_u \in \mathbb{R}^{N_V \times d_e}$, the unembedding matrix.

```
$\ell \leftarrow \text{length}(x)$
**for** $t \in [\ell]$: $e_t \leftarrow R_e(t) W_e[:, x[t]]$
$X \leftarrow [e_1, e_2, \ldots, e_l]$
**for** $l = 1, 2, \ldots, L$ **do**
    **for** $t \in [\ell]$: $\tilde{X}[:, t] \leftarrow \text{layer\_norm}(X[:, t]|\gamma_l^1, \beta_l^1)$
    $X \leftarrow X + \text{MHAttention}(\tilde{X}|W_l, \text{Mask}[t, t'] = [[t \leq t']])$
    **for** $t \in [\ell]$: $\tilde{X}[:, t] \leftarrow \text{layer\_norm}(X[:, t]|\gamma_l^2, \beta_l^2)$
    $X \leftarrow X + \text{W}_{mlp2}^l \text{GELU}(\text{W}_{mlp1}^l \tilde{X} + b_{mlp1}^l 1^\top) + b_{mlp2} 1^\top$
**end**
**for** $t \in [\ell]$: $X[:, t] \leftarrow \text{layer\_norm}(X[:, t]|\gamma, \beta)$
**return** $P = \text{softmax}(W_u X)$
```