Transformer with GLU Variant Pseudocode

Zheling Zhang

March 5, 2024

Algorithm 1 Transformer model with Gated Linear Unit (GLU) variant

Require: x, a sequence of token IDs.

Ensure: y, the output sequence after passing through the Transformer model with a GLU variant.

Hyperparameters: N, the number of layers; d_{model}, the dimensionality of token embeddings; d_{ff}, the dimensionality of the feedforward layer; h, the number of attention heads.

Parameters θ include all following parameters:

  $E \in R^{d_{\text{vocab}} \times d_{\text{model}}}$, the token embedding matrix.

  $PE \in R^{\text{max\_position} \times d_{\text{model}}}$, the positional embedding matrix.

  For each layer $l \in [1...N]$:

    $W_Q^l, W_K^l, W_V^l \in R^{d_{\text{model}} \times (d_{\text{model}}/h)}$, attention parameter matrices for each head.

    $W_O^l \in R^{(d_{\text{model}}/h) \times d_{\text{model}}}$, output projection matrix for multi-head attention.

    $\gamma^l, \beta^l \in R^{d_{\text{model}}}$, parameters for layer normalization before and after the multi-head attention, respectively.

    $W_1^l \in R^{d_{\text{model}} \times d_{\text{ff}}}$, $W_2^l \in R^{d_{\text{ff}} \times d_{\text{model}}}$, weights for the feedforward network.

    $W_g^l \in R^{d_{\text{ff}} \times d_{\text{model}}}$, weights for the gating mechanism in the GLU variant.

    $b_1^l, b_2^l, b_g^l \in R^{d_{\text{model}}}$, biases for the feedforward network and the gating mechanism.

  $W_u \in R^{d_{\text{model}} \times d_{\text{vocab}}}$, the unembedding matrix.

Initialize the output sequence y to an empty list.

Compute the embedded input sequence $E_x = E[x] + PE[\text{pos}]$, where pos is the position sequence.

for each layer $l \in [1...N]$ **do**

  Apply layer normalization: $X_{\text{norm}} = \text{LayerNorm}(E_x, \gamma^l, \beta^l)$.

  Calculate self-attention: $Z = \text{MultiHeadAttention}(X_{\text{norm}}, W_Q^l, W_K^l, W_V^l, W_O^l)$.

  Apply residual connection and layer normalization: $X_{\text{norm}} = \text{LayerNorm}(X_{\text{norm}} + Z, \gamma^l, \beta^l)$.

  Apply the first feedforward projection: $F1 = W_1^l X_{\text{norm}} + b_1^l$.

  Apply the GLU variant: $G = \text{GLU\_Variant}(F1, W_g^l, b_g^l)$.

  Apply the second feedforward projection: $F2 = W_2^l G + b_2^l$.

  Apply residual connection: $E_x = X_{\text{norm}} + F2$.

  Append the result to the output sequence y.

Compute the unembedded output: $P = \text{softmax}(W_u y)$.

return P

function GLU_VARIANT(X, W_g, b_g)

  Split the input matrix into two equal parts: $A, B = \text{split}(X, 2, \text{axis} = -1)$.

  Apply a non-linearity to the first part: $A = \tanh(A)$.

  Apply the gating mechanism to the second part: $B = \text{sigmoid}(W_g B + b_g)$.

  Element-wise multiplication of the results: $G = A \times B$.

  return G