Pseudocode for Multi-Query Attention (MQA)

Siyu Yang

Algorithm 1 Uptraining Multi-Head to Multi-Query Attention

Require: M_{MHA}: Multi-Head Attention model
Require: D_{train}: Training dataset
Require: $\alpha = 0.05$: Proportion of original training compute for uptraining

$M_{\text{MQA}} \leftarrow \text{ConvertToMQA}(M_{\text{MHA}})$
function ConvertToMQA(M_{MHA})
    Initialize M_{MQA} with M_{MHA}'s architecture
    for each attention layer in M_{MHA} **do**
        $K_{\text{pooled}} \leftarrow \text{mean}(\{K_h | h \in \text{Heads}\})$
        $V_{\text{pooled}} \leftarrow \text{mean}(\{V_h | h \in \text{Heads}\})$
        Assign $K_{\text{pooled}}, V_{\text{pooled}}$ to corresponding layer in M_{MQA}
    end for
    return M_{MQA}
end function
$steps_{\text{total}} \leftarrow$ Number of steps in M_{MHA}'s original pre-training
$steps_{\text{uptrain}} \leftarrow \lceil steps_{\text{total}} \times \alpha \rceil$
for $step = 1$ to $steps_{\text{uptrain}}$ **do**
    $batch \leftarrow$ Sample from D_{train}
    Update M_{MQA} on $batch$ ▷ Using pooled K and V
end for
return M_{MQA}