KV Cache & MLA

假设输入 X 在 iter=1，则其长度为 $|X| = n-1$，在 iter=2，长度为 $|X'| = n$

我们可以认为：

$$X = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_{n-1} \end{bmatrix} \begin{cases} W^Q \\ W^K \\ W^V \end{cases}$$

$$X' = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_{n-1} \\ \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} X \\ \boldsymbol{x}_n \end{bmatrix}$$

我们考虑计算矩阵 XW^*,X 在前，因此是行成列，因此

$$XW^* = \begin{bmatrix} - & \boldsymbol{x}_1 & - \\ - & \boldsymbol{x}_2 & - \\ & \vdots & \\ - & \boldsymbol{x}_{n-1} & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_q \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & \boldsymbol{q}_1 & - \\ - & \boldsymbol{q}_2 & - \\ & \vdots & \\ - & \boldsymbol{q}_{n-1} & - \end{bmatrix}$$

$$X'^{W^*} = \begin{bmatrix} - & \boldsymbol{x}_1 & - \\ - & \boldsymbol{x}_2 & - \\ & \vdots & \\ - & \boldsymbol{x}_{n-1} & - \\ - & \boldsymbol{x}_n & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_q \\ | & | & & | \end{bmatrix} = \begin{bmatrix} - & \boldsymbol{q}_1 & - \\ - & \boldsymbol{q}_2 & - \\ & \vdots & \\ - & \boldsymbol{q}_{n-1} & - \\ - & \boldsymbol{q}_n & - \end{bmatrix}$$

$$= \begin{bmatrix} X \\ \boldsymbol{x}_n \end{bmatrix} \begin{bmatrix} | & | & & | \\ \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_q \\ | & | & & | \end{bmatrix} = \begin{bmatrix} Q \\ \boldsymbol{q}_n \end{bmatrix}$$

因此我们看上去只需要计算新的数据和权重相乘就可以获得 q_n, k_n, v_n，而不需要重复计算 $[*_{1:n-1}]$。

我们考虑 $QK^\top$ 矩阵 $(C \times C)$，由于 causal attention 即：

$$\Gamma = \begin{bmatrix} q_1 k_1^\top & -\infty & -\infty & -\infty & -\infty \\ q_2 k_1^\top & q_2 k_2^\top & -\infty & -\infty & -\infty \\ q_3 k_1^\top & q_3 k_2^\top & q_3 k_3^\top & -\infty & -\infty \\ \vdots & \vdots & \vdots & \ddots & -\infty \\ q_{n-1} k_1^\top & q_{n-1} k_2^\top & q_{n-1} k_3^\top & \cdots & q_{n-1} k_{n-1}^\top \end{bmatrix}$$

观察 iter=2 $QK^\top$

$$\begin{bmatrix} q_1k_1^\top & -\infty & -\infty & -\infty & -\infty & -\infty \\ q_2k_1^\top & q_2k_2^\top & -\infty & -\infty & -\infty & -\infty \\ q_3k_1^\top & q_3k_2^\top & q_3k_3^\top & -\infty & -\infty & -\infty \\ \vdots & \vdots & \vdots & \ddots & -\infty & -\infty \\ q_{n-1}k_1^\top & q_{n-1}k_2^\top & q_{n-1}k_3^\top & \cdots & q_{n-1}k_{n-1}^\top & -\infty \\ q_nk_1^\top & q_nk_2^\top & q_nk_3^\top & \cdots & q_nk_{n-1}^\top & q_nk_n^\top \end{bmatrix} = \begin{bmatrix} \Gamma & -\infty \\ q_nk_{1:n-1}^\top & q_nk_n^\top \end{bmatrix}$$

我们再去考虑 $O = a(QK^\top)V$

$$\begin{bmatrix} \Gamma & -\infty \\ q_nk_{1:n-1}^\top & q_nk_n^\top \end{bmatrix} \begin{bmatrix} V_{n-1} \\ v_n \end{bmatrix} = \begin{bmatrix} O_{n-1} \\ o_n \end{bmatrix}$$

考虑最后都是 $-\infty$，因此$a(-\infty)v_n = 0$, 因此O_{n-1} 和 iter=1 一致。而o_n 则会引入新数据。

因此最直觉的做法是直接缓存 Γ。但是这不是最优点。

考虑一个输入 $X = [x_{<n}, x_n]$ 输出$Y = [y_{<n}, y_n]$，而我们需要 y_n，因此我们不需要获得 $y_{<n}$。我们可以认为 Γ 是产生 $y_{<n}$ 的一个过程矩阵。

因此我们关注最后需要产生的一个数据

$$\begin{bmatrix} q_nk_1^\top & q_nk_2^\top & q_nk_3^\top & \cdots & q_nk_{n-1}^\top & q_nk_n^\top \end{bmatrix} \begin{bmatrix} V_{n-1} \\ v_n \end{bmatrix}$$

会发现这里需要 $q_n, k_{1:n}, v_{1:n}$，而$k_{1:n-1}$ 和 $v_{1:n-1}$已被计算，所以直接缓存即可。因此这里我们只会去缓存 K, V 而不去缓存 Q（因为无需缓存）。

这里会发现 K, V 会随着长度增加，而越来越大。

Grouped Query Attention (GQA), Multi Query Attention (MQA)

计算 q_nK 时，随着 sequence length 增加，K在增大。因此通过 HBM 和 SRAM 的通信成为短板。（类似于 DRAM 和 L1）。

因此我们需要降低 K, V 的大小。最简单的做法是降低KV数量，因此提出了 MQA 和 GQA 以分别表示如下：

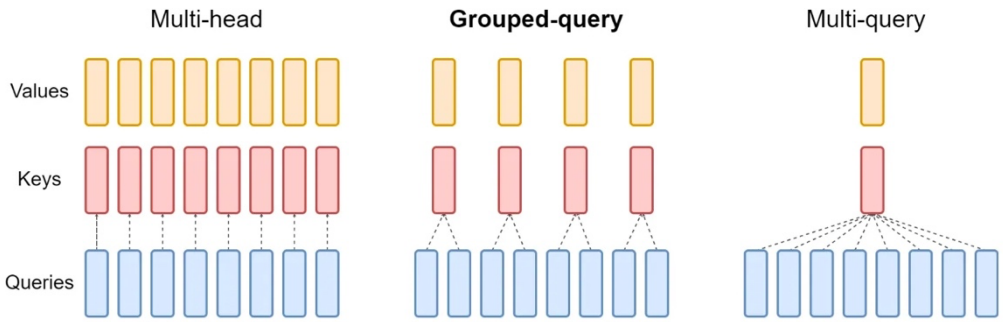
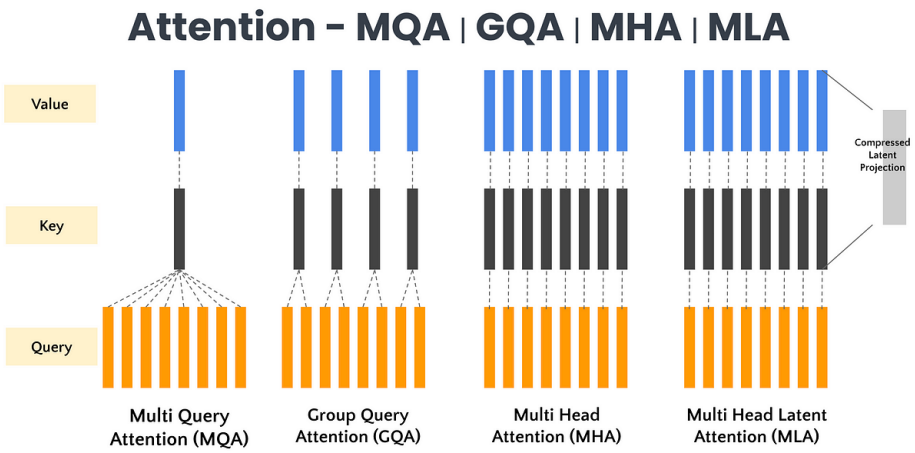


Figure 2: Overview of grouped-query method. Multi-head attention has H query, key, and value heads. Multi-query attention shares single key and value heads across all query heads. Grouped-query attention instead shares single key and value heads for each *group* of query heads, interpolating between multi-head and multi-query attention.



而 MLA 则使用 Latent Space 表示。假设隐层为 h_t

$$
\begin{aligned}
c_t^Q &= W^{\downarrow Q} h_t \\
q_t^C &= W^{\uparrow Q} c_t^Q, q_t^R = RoPE\left(W^{QR} c_t^Q\right) \\
q_t &= [q_t^C, q_t^R]
\end{aligned}
$$

$$
\begin{aligned}
c_t^{KV} &= W^{\downarrow KV} h_t \\
k_t^C &= W^{\uparrow K} c_t^{KV}, k_t^R = RePE(W^{KR} h_t) \\
k_t &= [k_t^C, k_t^R] \\
v_t^C &= W^{\uparrow V} c_t^{KV}
\end{aligned}
$$