MHA（Multi-Head Attention）

$$
\boldsymbol{o}_t = \left[\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)}, \cdots, \boldsymbol{o}_t^{(h)}\right]
$$

$$
\boldsymbol{o}_t^{(s)} = Attention\left(\boldsymbol{q}_t^{(s)}, \boldsymbol{k}_{\leq t}^{(s)}, \boldsymbol{v}_{\leq t}^{(s)}\right) \triangleq \frac{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)\boldsymbol{v}_i^{(s)}}{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)} \tag{1}
$$

$$
\begin{aligned}
\boldsymbol{q}_i^{(s)} &= \boldsymbol{x}_i\boldsymbol{W}_q^{(s)} \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_q^{(s)} \in \mathbb{R}^{d\times d_k} \\
\boldsymbol{k}_i^{(s)} &= \boldsymbol{x}_i\boldsymbol{W}_k^{(s)} \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_k^{(s)} \in \mathbb{R}^{d\times d_k} \\
\boldsymbol{v}_i^{(s)} &= \boldsymbol{x}_i\boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d_v}, \quad \boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d\times d_v}
\end{aligned}
$$

MQA（Multi-Query Attention）

$$
\boldsymbol{o}_t = \left[\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)}, \cdots, \boldsymbol{o}_t^{(h)}\right]
$$

$$
\boldsymbol{o}_t^{(s)} = Attention\left(\boldsymbol{q}_t^{(s)}, \boldsymbol{k}_{\leq t}^{\cancel{(s)}}, \boldsymbol{v}_{\leq t}^{\cancel{(s)}}\right) \triangleq \frac{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{\cancel{(s)}\top}\right)\boldsymbol{v}_i^{\cancel{(s)}}}{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{\cancel{(s)}\top}\right)} \tag{2}
$$

$$
\begin{aligned}
\boldsymbol{q}_i^{(s)} &= \boldsymbol{x}_i\boldsymbol{W}_q^{(s)} \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_q^{(s)} \in \mathbb{R}^{d\times d_k} \\
\boldsymbol{k}_i^{\cancel{(s)}} &= \boldsymbol{x}_i\boldsymbol{W}_k^{\cancel{(s)}} \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_k^{\cancel{(s)}} \in \mathbb{R}^{d\times d_k} \\
\boldsymbol{v}_i^{\cancel{(s)}} &= \boldsymbol{x}_i\boldsymbol{W}_v^{\cancel{(s)}} \in \mathbb{R}^{d_v}, \quad \boldsymbol{W}_v^{\cancel{(s)}} \in \mathbb{R}^{d\times d_v}
\end{aligned}
$$

事后看来，GQA的思想也很朴素，它就是将所有Head分为g个组（g可以整除h），每组共享同一对K、V，用数学公式表示为

GQA（Group-Query Attention）

$$
\boldsymbol{o}_t = \left[\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)}, \cdots, \boldsymbol{o}_t^{(h)}\right]
$$

$$
\boldsymbol{o}_t^{(s)} = Attention\left(\boldsymbol{q}_t^{(s)}, \boldsymbol{k}_{\leq t}^{(\lceil sg/h\rceil)}, \boldsymbol{v}_{\leq t}^{(\lceil sg/h\rceil)}\right) \triangleq \frac{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(\lceil sg/h\rceil)\top}\right)\boldsymbol{v}_i^{(\lceil sg/h\rceil)}}{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(\lceil sg/h\rceil)\top}\right)} \tag{3}
$$

$$
\begin{aligned}
\boldsymbol{q}_i^{(s)} &= \boldsymbol{x}_i\boldsymbol{W}_q^{(s)} \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_q^{(s)} \in \mathbb{R}^{d\times d_k} \\
\boldsymbol{k}_i^{(\lceil sg/h\rceil)} &= \boldsymbol{x}_i\boldsymbol{W}_k^{(\lceil sg/h\rceil)} \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_k^{(\lceil sg/h\rceil)} \in \mathbb{R}^{d\times d_k} \\
\boldsymbol{v}_i^{(\lceil sg/h\rceil)} &= \boldsymbol{x}_i\boldsymbol{W}_v^{(\lceil sg/h\rceil)} \in \mathbb{R}^{d_v}, \quad \boldsymbol{W}_v^{(\lceil sg/h\rceil)} \in \mathbb{R}^{d\times d_v}
\end{aligned}
$$

这里的$\lceil\cdot\rceil$是上取整符号。GQA提供了MHA到MQA的自然过渡，当$g=h$时就是MHA，$g=1$时就是MQA，当$1<g<h$时，它只将KV Cache压缩到g/h，压缩率不如MQA，但同时也提供了更大的自由度，效果上更有保证。GQA最知名的使用者，大概是Meta开源的LLAMA2-70B，以及LLAMA3全

$$\boldsymbol{o}_t = \left[\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)}, \cdots, \boldsymbol{o}_t^{(h)}\right]$$

更换为一般的线性变换：

$$\begin{gathered}
\boldsymbol{o}_t^{(s)} = Attention\left(\boldsymbol{q}_t^{(s)}, \boldsymbol{k}_{\leq t}^{(s)}, \boldsymbol{v}_{\leq t}^{(s)}\right) \triangleq \frac{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)\boldsymbol{v}_i^{(s)}}{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)} \\
\boldsymbol{q}_i^{(s)} = \boldsymbol{x}_i\boldsymbol{W}_q^{(s)} \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_q^{(s)} \in \mathbb{R}^{d\times d_k} \\
\boldsymbol{k}_i^{(s)} = \boldsymbol{c}_i\boldsymbol{W}_k^{(s)} \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_k^{(s)} \in \mathbb{R}^{d_c\times d_k} \\
\boldsymbol{v}_i^{(s)} = \boldsymbol{c}_i\boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d_v}, \quad \boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d_c\times d_v} \\
\boldsymbol{c}_i = \boldsymbol{x}_i\boldsymbol{W}_c \in \mathbb{R}^{d_c}, \quad \boldsymbol{W}_c \in \mathbb{R}^{d\times d_c}
\end{gathered} \tag{5}$$

矩阵吸收：

对此，MLA发现，我们可以结合Dot-Attention的具体形式，通过一个简单但不失巧妙的恒等变换来规避这个问题。首先，在训练阶段还是照常进行，此时优化空间不大；然后，在推理阶段，我们利用

$$\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top} = \left(\boldsymbol{x}_t\boldsymbol{W}_q^{(s)}\right)\left(\boldsymbol{c}_i\boldsymbol{W}_k^{(s)}\right)^{\top} = \boldsymbol{x}_t\left(\boldsymbol{W}_q^{(s)}\boldsymbol{W}_k^{(s)\top}\right)\boldsymbol{c}_i^{\top} \tag{6}$$

这意味着推理阶段，我们可以将$\boldsymbol{W}_q^{(s)}\boldsymbol{W}_k^{(s)\top}$合并起来作为Q的投影矩阵，那么$\boldsymbol{c}_i$则取代了原本的$\boldsymbol{k}_i$，同理，在$\boldsymbol{o}_t$后面我们还有一个投影矩阵，于是$\boldsymbol{v}_i^{(s)} = \boldsymbol{c}_i\boldsymbol{W}_v^{(s)}$的$\boldsymbol{W}_v^{(s)}$也可以吸收到后面的投影矩阵中去，于是等效地$\boldsymbol{v}_i$也可以用$\boldsymbol{c}_i$代替，也就是说此时KV Cache只需要存下所有的$\boldsymbol{c}_i$就行，而不至于存下所有的$\boldsymbol{k}_i^{(s)}$、$\boldsymbol{v}_i^{(s)}$。注意到$\boldsymbol{c}_i$跟$^{(s)}$无关，也就是说是所有头共享的，即MLA在推理阶段它可以恒等变换为一个MQA。

MLA 无法天然支持 RoPE：

刚才我们说了，MLA之所以能保持跟GQA一样大小的KV Cache，其关键一步是"将$\boldsymbol{W}_q^{(s)}\boldsymbol{W}_k^{(s)\top}$合并成一个（跟位置无关的）矩阵作为Q的投影矩阵"，但如果加了RoPE的话，这一步就无法实现了。这是因为RoPE是一个跟位置相关的、$d_k \times d_k$的分块对角矩阵$\boldsymbol{\mathcal{R}}_m$，满足$\boldsymbol{\mathcal{R}}_m\boldsymbol{\mathcal{R}}_n^{\top} = \boldsymbol{\mathcal{R}}_{m-n}$，MLA加入RoPE之后会让$\boldsymbol{W}_q^{(s)}\boldsymbol{W}_k^{(s)\top}$之间多插入了一项$\boldsymbol{\mathcal{R}}_{t-i}$：

$$\begin{gathered}
\boldsymbol{q}_i^{(s)} = \boldsymbol{x}_i\boldsymbol{W}_q^{(s)}\boldsymbol{\mathcal{R}}_i \quad , \quad \boldsymbol{k}_i^{(s)} = \boldsymbol{c}_i\boldsymbol{W}_k^{(s)}\boldsymbol{\mathcal{R}}_i \\
\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top} = \left(\boldsymbol{x}_t\boldsymbol{W}_q^{(s)}\boldsymbol{\mathcal{R}}_t\right)\left(\boldsymbol{c}_i\boldsymbol{W}_k^{(s)}\boldsymbol{\mathcal{R}}_i\right)^{\top} = \boldsymbol{x}_t\left(\boldsymbol{W}_q^{(s)}\boldsymbol{\mathcal{R}}_{t-i}\boldsymbol{W}_k^{(s)\top}\right)\boldsymbol{c}_i^{\top}
\end{gathered} \tag{7}$$

这里的$\boldsymbol{W}_q^{(s)}\boldsymbol{\mathcal{R}}_{t-i}\boldsymbol{W}_k^{(s)\top}$就无法合并为一个固定的投影矩阵了（跟位置差$t-i$相关），从而MLA的想法无法结合RoPE实现。

解耦的 RoPE：

最后发布的MLA，采取了一种混合的方法——每个Attention Head的Q、K新增d_r个维度用来添加RoPE，其中K新增的维度每个Head共享：

$$\boldsymbol{o}_t = \left[\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)}, \cdots, \boldsymbol{o}_t^{(h)}\right]$$

$$\begin{gathered}
\boldsymbol{o}_t^{(s)} = Attention\left(\boldsymbol{q}_t^{(s)}, \boldsymbol{k}_{\leq t}^{(s)}, \boldsymbol{v}_{\leq t}^{(s)}\right) \triangleq \frac{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)\boldsymbol{v}_i^{(s)}}{\sum_{i\leq t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)} \\
\boldsymbol{q}_i^{(s)} = \left[\boldsymbol{x}_i\boldsymbol{W}_{qc}^{(s)}, \boldsymbol{x}_i\boldsymbol{W}_{qr}^{(s)}\boldsymbol{\mathcal{R}}_i\right] \in \mathbb{R}^{d_k+d_r}, \quad \boldsymbol{W}_{qc}^{(s)} \in \mathbb{R}^{d\times d_k}, \boldsymbol{W}_{qr}^{(s)} \in \mathbb{R}^{d\times d_r} \\
\boldsymbol{k}_i^{(s)} = \left[\boldsymbol{c}_i\boldsymbol{W}_{kc}^{(s)}, \boldsymbol{x}_i\boldsymbol{W}_{kr}^{\cancel{(s)}}\boldsymbol{\mathcal{R}}_i\right] \in \mathbb{R}^{d_k+d_r}, \quad \boldsymbol{W}_{kc}^{(s)} \in \mathbb{R}^{d_c\times d_k}, \boldsymbol{W}_{kr}^{\cancel{(s)}} \in \mathbb{R}^{d\times d_r} \\
\boldsymbol{v}_i^{(s)} = \boldsymbol{c}_i\boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d_v}, \quad \boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d_c\times d_v} \\
\boldsymbol{c}_i = \boldsymbol{x}_i\boldsymbol{W}_c \in \mathbb{R}^{d_c}, \quad \boldsymbol{W}_c \in \mathbb{R}^{d\times d_c}
\end{gathered} \tag{9}$$

这样一来，没有RoPE的维度就可以重复"Part 1"的操作，在推理时KV Cache只需要存$\boldsymbol{c}_i$，新增的带RoPE的维度就可以用来补充位置信息，并且由于所有Head共享，所以也就只有在K Cache这里增加了d_r个维度，原论文取了$d_r = d_k/2 = 64$，相比原本的$d_c = 512$，增加的幅度不大。

训练阶段的 MLA：

$$
\begin{gathered}
\boldsymbol{o}_t = \left[\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)}, \cdots, \boldsymbol{o}_t^{(h)}\right] \\
\boldsymbol{o}_t^{(s)} = Attention\left(\boldsymbol{q}_t^{(s)}, \boldsymbol{k}_{\le t}^{(s)}, \boldsymbol{v}_{\le t}^{(s)}\right) \triangleq \frac{\sum_{i\le t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)\boldsymbol{v}_i^{(s)}}{\sum_{i\le t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)} \\
\boldsymbol{q}_i^{(s)} = \left[\boldsymbol{c}_i'\boldsymbol{W}_{qc}^{(s)}, \boldsymbol{c}_i'\boldsymbol{W}_{qr}^{(s)}\mathcal{R}_i\right] \in \mathbb{R}^{d_k+d_r}, \quad \boldsymbol{W}_{qc}^{(s)} \in \mathbb{R}^{d_c'\times d_k}, \boldsymbol{W}_{qr}^{(s)} \in \mathbb{R}^{d_c'\times d_r} \\
\boldsymbol{k}_i^{(s)} = \left[\boldsymbol{c}_i\boldsymbol{W}_{kc}^{(s)}, \boldsymbol{x}_i\boldsymbol{W}_{kr}^{\cancel{(s)}}\mathcal{R}_i\right] \in \mathbb{R}^{d_k+d_r}, \quad \boldsymbol{W}_{kc}^{(s)} \in \mathbb{R}^{d_c\times d_k}, \boldsymbol{W}_{kr}^{\cancel{(s)}} \in \mathbb{R}^{d\times d_r} \\
\boldsymbol{v}_i^{(s)} = \boldsymbol{c}_i\boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d_v}, \quad \boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d_c\times d_v} \\
\boldsymbol{c}_i' = \boldsymbol{x}_i\boldsymbol{W}_c' \in \mathbb{R}^{d_c'}, \quad \boldsymbol{W}_c' \in \mathbb{R}^{d\times d_c'} \\
\boldsymbol{c}_i = \boldsymbol{x}_i\boldsymbol{W}_c \in \mathbb{R}^{d_c}, \quad \boldsymbol{W}_c \in \mathbb{R}^{d\times d_c}
\end{gathered}
\tag{10}
$$

推理阶段的MLA：

$$
\begin{gathered}
\boldsymbol{o}_t = \left[\boldsymbol{o}_t^{(1)}\boldsymbol{W}_v^{(1)}, \boldsymbol{o}_t^{(2)}\boldsymbol{W}_v^{(2)}, \cdots, \boldsymbol{o}_t^{(h)}\boldsymbol{W}_v^{(h)}\right] \\
\boldsymbol{o}_t^{(s)} = Attention\left(\boldsymbol{q}_t^{(s)}, \boldsymbol{k}_{\le t}^{(s)}, \boldsymbol{c}_{\le t}\right) \triangleq \frac{\sum_{i\le t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)\boldsymbol{c}_i}{\sum_{i\le t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)} \\
\boldsymbol{q}_i^{(s)} = \left[\boldsymbol{c}_i'\boldsymbol{W}_{qc}^{(s)}\boldsymbol{W}_{kc}^{(s)\top}, \boldsymbol{c}_i'\boldsymbol{W}_{qr}^{(s)}\mathcal{R}_i\right] \in \mathbb{R}^{d_c+d_r} \\
\boldsymbol{k}_i^{(s)} = \left[\boldsymbol{c}_i, \boldsymbol{x}_i\boldsymbol{W}_{kr}^{\cancel{(s)}}\mathcal{R}_i\right] \in \mathbb{R}^{d_c+d_r} \\
\boldsymbol{W}_{qc}^{(s)} \in \mathbb{R}^{d_c'\times d_k}, \boldsymbol{W}_{kc}^{(s)} \in \mathbb{R}^{d_c\times d_k}, \boldsymbol{W}_{qr}^{(s)} \in \mathbb{R}^{d_c'\times d_r}, \boldsymbol{W}_{kr}^{\cancel{(s)}} \in \mathbb{R}^{d\times d_r} \\
\boldsymbol{c}_i' = \boldsymbol{x}_i\boldsymbol{W}_c' \in \mathbb{R}^{d_c'}, \quad \boldsymbol{W}_c' \in \mathbb{R}^{d\times d_c'} \\
\boldsymbol{c}_i = \boldsymbol{x}_i\boldsymbol{W}_c \in \mathbb{R}^{d_c}, \quad \boldsymbol{W}_c \in \mathbb{R}^{d\times d_c}
\end{gathered}
\tag{12}
$$

带 RoPE 的 MHA：

$$
\begin{gathered}
\boldsymbol{o}_t = \left[\boldsymbol{o}_t^{(1)}, \boldsymbol{o}_t^{(2)}, \cdots, \boldsymbol{o}_t^{(h)}\right] \\
\boldsymbol{o}_t^{(s)} = Attention\left(\boldsymbol{q}_t^{(s)}, \boldsymbol{k}_{\le t}^{(s)}, \boldsymbol{v}_{\le t}^{(s)}\right) \triangleq \frac{\sum_{i\le t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)\boldsymbol{v}_i^{(s)}}{\sum_{i\le t}\exp\left(\boldsymbol{q}_t^{(s)}\boldsymbol{k}_i^{(s)\top}\right)} \\
\boldsymbol{q}_i^{(s)} = \boldsymbol{x}_i\boldsymbol{W}_q^{(s)}\mathcal{R}_i \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_q^{(s)} \in \mathbb{R}^{d\times d_k} \\
\boldsymbol{k}_i^{(s)} = \boldsymbol{x}_i\boldsymbol{W}_k^{(s)}\mathcal{R}_i \in \mathbb{R}^{d_k}, \quad \boldsymbol{W}_k^{(s)} \in \mathbb{R}^{d\times d_k} \\
\boldsymbol{v}_i^{(s)} = \boldsymbol{x}_i\boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d_v}, \quad \boldsymbol{W}_v^{(s)} \in \mathbb{R}^{d\times d_v}
\end{gathered}
\tag{11}
$$

可以发现，其实在训练阶段，除了多了一步低秩投影以及只在部分维度加RoPE外，MLA与Q、K的Head Size由d_k换成$d_k + d_r$的MHA基本无异。