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

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

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

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

MQA (Multi-Query Attention)
$$o_t^{(s)} = Attention\left(q_t^{(s)}, \mathbf{k}_{\leq t}, \mathbf{v}_{\leq t}\right) \triangleq \frac{\sum_{i \leq t} \exp\left(q_t^{(s)} \mathbf{k}_i^{\top}\right) \mathbf{v}_i^{\top}}{\sum_{i \leq t} \exp\left(q_t^{(s)} \mathbf{k}_i^{\top}\right)}$$
 (2)

$$egin{aligned} oldsymbol{q}_i^{(s)} &= oldsymbol{x}_i oldsymbol{W}_q^{(s)} \in \mathbb{R}^{d_k}, & oldsymbol{W}_q^{(s)} \in \mathbb{R}^{d imes d_k} \ oldsymbol{k}_i^{(s)} &= oldsymbol{x}_i oldsymbol{W}_i^{(s)} \in \mathbb{R}^{d_k}, & oldsymbol{W}_i^{(s)} \in \mathbb{R}^{d imes d_k} \ oldsymbol{v}_i^{(s)} &= oldsymbol{x}_i oldsymbol{W}_v^{(s)} \in \mathbb{R}^{d imes d_v} \end{aligned}$$

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

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

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

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

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

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

更换为一般的线性变换:

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

$$\boldsymbol{q}_{i}^{(s)} = \boldsymbol{x}_{i} \boldsymbol{W}_{q}^{(s)} \in \mathbb{R}^{d_{k}}, \quad \boldsymbol{W}_{q}^{(s)} \in \mathbb{R}^{d \times d_{k}}$$

$$\boldsymbol{k}_{i}^{(s)} = \boldsymbol{c}_{i} \boldsymbol{W}_{k}^{(s)} \in \mathbb{R}^{d_{k}}, \quad \boldsymbol{W}_{k}^{(s)} \in \mathbb{R}^{d_{c} \times d_{k}}$$

$$\boldsymbol{v}_{i}^{(s)} = \boldsymbol{c}_{i} \boldsymbol{W}_{v}^{(s)} \in \mathbb{R}^{d_{v}}, \quad \boldsymbol{W}_{v}^{(s)} \in \mathbb{R}^{d_{c} \times d_{v}}$$

$$\boldsymbol{c}_{i} = \boldsymbol{x}_{i} \boldsymbol{W}_{c} \in \mathbb{R}^{d_{c}}, \quad \boldsymbol{W}_{c} \in \mathbb{R}^{d \times d_{c}}$$

$$(5)$$

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

$$\boldsymbol{q}_{t}^{(s)}\boldsymbol{k}_{i}^{(s)\top} = \left(\boldsymbol{x}_{t}\boldsymbol{W}_{q}^{(s)}\right)\left(\boldsymbol{c}_{i}\boldsymbol{W}_{k}^{(s)}\right)^{\top} = \boldsymbol{x}_{t}\left(\boldsymbol{W}_{q}^{(s)}\boldsymbol{W}_{k}^{(s)\top}\right)\boldsymbol{c}_{i}^{\top}$$
(6)

这意味着推理阶段,我们可以将 $W_q^{(s)}W_k^{(s)}$ [—]合并起来作为Q的投影矩阵,那么 c_i 则取代了原本的 k_i ,同理,在 o_t 后面我们还有一个投影矩阵,于是 $v_i^{(s)}=c_iW_v^{(s)}$ 的 $W_v^{(s)}$ 也可以吸收到后面的投影矩阵中去,于是等效地 v_i 也可以用 c_i 代替,也就是说此时KV Cache只需要存下所有的 c_i 就行,而不至于存下所有的 $k_i^{(s)}$ 、 $v_i^{(s)}$ 。注意到 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。

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 $\boldsymbol{\mathcal{R}}_{t-i}$:

$$\mathbf{q}_{i}^{(s)} = \mathbf{x}_{i} \mathbf{W}_{q}^{(s)} \mathbf{\mathcal{R}}_{i} \quad , \quad \mathbf{k}_{i}^{(s)} = \mathbf{c}_{i} \mathbf{W}_{k}^{(s)} \mathbf{\mathcal{R}}_{i}$$

$$\mathbf{q}_{t}^{(s)} \mathbf{k}_{i}^{(s) \top} = \left(\mathbf{x}_{t} \mathbf{W}_{q}^{(s)} \mathbf{\mathcal{R}}_{t}\right) \left(\mathbf{c}_{i} \mathbf{W}_{k}^{(s)} \mathbf{\mathcal{R}}_{i}\right)^{\top} = \mathbf{x}_{t} \left(\mathbf{W}_{q}^{(s)} \mathbf{\mathcal{R}}_{t-i} \mathbf{W}_{k}^{(s) \top}\right) \mathbf{c}_{i}^{\top}$$

$$(7)$$

这里的 $W_q^{(s)}$ $\mathcal{R}_{t-i}W_k^{(s)\top}$ 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。

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RoP E,其中K新增的维度每个Head共享:

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

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

$$\boldsymbol{q}_{i}^{(s)} = \left[\boldsymbol{x}_{i} \boldsymbol{W}_{qc}^{(s)}, \boldsymbol{x}_{i} \boldsymbol{W}_{qr}^{(s)} \boldsymbol{\mathcal{R}}_{i}\right] \in \mathbb{R}^{d_{k} + d_{r}}, \quad \boldsymbol{W}_{qc}^{(s)} \in \mathbb{R}^{d \times d_{k}}, \boldsymbol{W}_{qr}^{(s)} \in \mathbb{R}^{d \times d_{r}}$$

$$\boldsymbol{k}_{i}^{(s)} = \left[\boldsymbol{c}_{i} \boldsymbol{W}_{kc}^{(s)}, \boldsymbol{x}_{i} \boldsymbol{W}_{kr}^{(s)} \boldsymbol{\mathcal{R}}_{i}\right] \in \mathbb{R}^{d_{k} + d_{r}}, \quad \boldsymbol{W}_{kc}^{(s)} \in \mathbb{R}^{d_{c} \times d_{k}}, \boldsymbol{W}_{kr}^{(s)} \in \mathbb{R}^{d \times d_{r}}$$

$$\boldsymbol{v}_{i}^{(s)} = \boldsymbol{c}_{i} \boldsymbol{W}_{v}^{(s)} \in \mathbb{R}^{d_{v}}, \quad \boldsymbol{W}_{v}^{(s)} \in \mathbb{R}^{d_{c} \times d_{v}}$$

$$(9)$$

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 c_i ,新增的带R oPE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。

 $oldsymbol{c}_i = oldsymbol{x}_i oldsymbol{W}_c \in \mathbb{R}^{d_c}, \quad oldsymbol{W}_c \in \mathbb{R}^{d imes d_c}$

起中败收:

MLA 无法天然支持 RoPE:

解耦的 RoPE:

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

$$oldsymbol{o}_t^{(s)} = Attention\left(oldsymbol{q}_t^{(s)}, oldsymbol{k}_{\leq t}^{(s)}, oldsymbol{v}_{\leq t}^{(s)}
ight) riangleq rac{\sum_{i \leq t} \exp\left(oldsymbol{q}_t^{(s)} oldsymbol{k}_i^{(s) op}
ight) oldsymbol{v}_i^{(s)}}{\sum_{i \leq t} \exp\left(oldsymbol{q}_t^{(s)} oldsymbol{k}_i^{(s) op}
ight)}$$

训练阶段的 MLA:

$$q_{i}^{(s)} = \begin{bmatrix} \boldsymbol{c}_{i}' \boldsymbol{W}_{qc}^{(s)}, \boldsymbol{c}_{i}' \boldsymbol{W}_{qr}^{(s)} \boldsymbol{\mathcal{R}}_{i} \end{bmatrix} \in \mathbb{R}^{d_{k}+d_{r}}, \quad \boldsymbol{W}_{qc}^{(s)} \in \mathbb{R}^{d_{c}' \times d_{k}}, \boldsymbol{W}_{qr}^{(s)} \in \mathbb{R}^{d_{c}' \times d_{r}}$$

$$\boldsymbol{k}_{i}^{(s)} = \begin{bmatrix} \boldsymbol{c}_{i} \boldsymbol{W}_{kc}^{(s)}, \boldsymbol{x}_{i} \boldsymbol{W}_{kr}^{(s)} \boldsymbol{\mathcal{R}}_{i} \end{bmatrix} \in \mathbb{R}^{d_{k}+d_{r}}, \quad \boldsymbol{W}_{kc}^{(s)} \in \mathbb{R}^{d_{c} \times d_{k}}, \boldsymbol{W}_{kr}^{(s)} \in \mathbb{R}^{d \times d_{r}}$$

$$\boldsymbol{v}_{i}^{(s)} = \boldsymbol{c}_{i} \boldsymbol{W}_{v}^{(s)} \in \mathbb{R}^{d_{v}}, \quad \boldsymbol{W}_{v}^{(s)} \in \mathbb{R}^{d_{c} \times d_{v}}$$

$$\boldsymbol{c}_{i}' = \boldsymbol{x}_{i} \boldsymbol{W}_{c}' \in \mathbb{R}^{d_{c}'}, \quad \boldsymbol{W}_{c}' \in \mathbb{R}^{d \times d_{c}'}$$

$$\boldsymbol{c}_{i} = \boldsymbol{x}_{i} \boldsymbol{W}_{c} \in \mathbb{R}^{d_{c}}, \quad \boldsymbol{W}_{c} \in \mathbb{R}^{d \times d_{c}}$$

$$\boldsymbol{W}_{c} \in \mathbb{R}^{d \times d_{c}}$$

$$oldsymbol{o}_t = \left[oldsymbol{o}_t^{(1)} oldsymbol{W}_v^{(1)}, oldsymbol{o}_t^{(2)} oldsymbol{W}_v^{(2)}, \cdots, oldsymbol{o}_t^{(h)} oldsymbol{W}_v^{(h)}
ight]$$

推理阶段的MLA:

$$oldsymbol{o}_t^{(s)} = Attention\left(oldsymbol{q}_t^{(s)}, oldsymbol{k}_{\leq t}^{(s)}, oldsymbol{c}_{\leq t}
ight) riangleq rac{\sum_{i \leq t} \exp\left(oldsymbol{q}_t^{(s)} oldsymbol{k}_i^{(s) op}
ight) oldsymbol{c}_i}{\sum_{i \leq t} \exp\left(oldsymbol{q}_t^{(s)} oldsymbol{k}_i^{(s) op}
ight)}$$

$$q_i^{(s)} = \begin{bmatrix} \boldsymbol{c}_i' \boldsymbol{W}_{qc}^{(s)} \boldsymbol{W}_{kc}^{(s)\top}, \boldsymbol{c}_i' \boldsymbol{W}_{qr}^{(s)} \boldsymbol{\mathcal{R}}_i \end{bmatrix} \in \mathbb{R}^{d_c + d_r}$$

$$\boldsymbol{k}_i^{(s)} = \begin{bmatrix} \boldsymbol{c}_i, \boldsymbol{x}_i \boldsymbol{W}_{kr}^{(s)} \boldsymbol{\mathcal{R}}_i \end{bmatrix} \in \mathbb{R}^{d_c + d_r}$$

$$\boldsymbol{W}_{qc}^{(s)} \in \mathbb{R}^{d_c' \times d_k}, \boldsymbol{W}_{kc}^{(s)} \in \mathbb{R}^{d_c \times d_k}, \boldsymbol{W}_{qr}^{(s)} \in \mathbb{R}^{d_c' \times d_r}, \boldsymbol{W}_{kr}^{(s)} \in \mathbb{R}^{d \times d_r}$$

$$\boldsymbol{c}_i' = \boldsymbol{x}_i \boldsymbol{W}_c' \in \mathbb{R}^{d_c'}, \quad \boldsymbol{W}_c' \in \mathbb{R}^{d \times d_c'}$$

$$(12)$$

$$egin{aligned} oldsymbol{c}_i' &= oldsymbol{x}_i oldsymbol{W}_c' \in \mathbb{R}^{d_c}, & oldsymbol{W}_c' \in \mathbb{R}^{d imes d_c} \ oldsymbol{c}_i &= oldsymbol{x}_i oldsymbol{W}_c \in \mathbb{R}^{d imes d_c}, & oldsymbol{W}_c \in \mathbb{R}^{d imes d_c} \end{aligned}$$

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

$$\boldsymbol{o}_{t}^{(s)} = Attention\left(\boldsymbol{q}_{t}^{(s)}, \boldsymbol{k}_{\leq t}^{(s)}, \boldsymbol{v}_{\leq t}^{(s)}\right) \triangleq \frac{\sum_{i \leq t} \exp\left(\boldsymbol{q}_{t}^{(s)} \boldsymbol{k}_{i}^{(s)\top}\right) \boldsymbol{v}_{i}^{(s)}}{\sum_{i \leq t} \exp\left(\boldsymbol{q}_{t}^{(s)} \boldsymbol{k}_{i}^{(s)\top}\right)}$$
(11)

带 RoPE 的 MHA:

$$egin{aligned} oldsymbol{q}_i^{(s)} &= oldsymbol{x}_i oldsymbol{W}_q^{(s)} oldsymbol{\mathcal{R}}_i \in \mathbb{R}^{d_k}, & oldsymbol{W}_q^{(s)} \in \mathbb{R}^{d imes d_k} \ oldsymbol{k}_i^{(s)} &= oldsymbol{x}_i oldsymbol{W}_k^{(s)} oldsymbol{\mathcal{R}}_i \in \mathbb{R}^{d_k}, & oldsymbol{W}_k^{(s)} \in \mathbb{R}^{d imes d_k} \ oldsymbol{v}_i^{(s)} &= oldsymbol{x}_i oldsymbol{W}_v^{(s)} \in \mathbb{R}^{d imes d_v} \end{aligned}$$

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。