2025 USA-NA-AIO Round 2, Problem 2, Part 6

USAAIO **May 2025**

Next, let us study a variant of MHA: **Group Query Attention (GQA)**.

Recall that in MHA, the number of heads in queries, keys and values are the same, H. Thus, query $\mathbf{q}_{l_1,h}$ attends to key $\mathbf{k}_{l_2,h}$ with the same head index h.

In GQA, we relax this constraint by allowing keys and values to have G heads ($G \leq H$), where G is factor of H. For instance, if $H = 12$, then $G \in \{1, 2, 3, 4, 6, 12\}$.

In GQA, a query $\mathbf{q}_{l_1,h}$ with head h is permitted to attend to a key $\mathbf{k}_{l_2,g}$ and use value $\mathbf{v}_{l_2,g}$ in computing its output with head g if

$$h \equiv g \pmod{G}.$$

Thus, each head in keys and values is mapped to $\frac{H}{G} \geq 1$ heads in queries.

As an example, suppose $H = 12$ and $G = 3$. Then

- Head $g = 0$ in keys and values is associated with heads $h = 0, 3, 6, 9$ in queries.
- Head $g = 1$ in keys and values is associated with heads $h = 1, 4, 7, 10$ in queries.
- Head $g = 2$ in keys and values is associated with heads $h = 2, 5, 8, 11$ in queries.

Skip to main content

USAAIO **May 2025**

Part 6 (5 points, non-coding task)

Topics
My posts
More
CATEGORIES
General
Site Feedback
All categories

For $\mathbf{M} \in \{\mathbf{K}, \mathbf{V}\}$, Denote the $\mathbf{M}$-projection matrix as

$$\mathbf{W}^{\mathbf{M},GQA} = \begin{bmatrix} \mathbf{W}_0^{\mathbf{M},GQA} \\ \vdots \\ \mathbf{W}_{G-1}^{\mathbf{M},GQA} \end{bmatrix}$$

Now, we concatenate $\frac{H}{G}$ copies of the above matrix along axis 0:

$$\tilde{\mathbf{W}}^{\mathbf{M},GQA} = \begin{bmatrix} \mathbf{W}^{\mathbf{M},GQA} \\ \mathbf{W}^{\mathbf{M},GQA} \\ \vdots \\ \mathbf{W}^{\mathbf{M},GQA} \end{bmatrix}$$

What is the relationship between $\operatorname{rank}\left(\tilde{\mathbf{W}}^{\mathbf{M},GQA}\right)$ and $\operatorname{rank}\left(\mathbf{W}^{\mathbf{M},GQA}\right)$?

- Reasoning is required.

USAAIO **May 2025**

Misplaced '#'

Let $\{\mathbf{w}_i^* : i \in \{0, 1, \cdots, r-1\}\}$ be r linearly independent row vectors that span all row vectors $\mathbf{W}^{\mathbf{M},GQA}$.

Because each row vector in $\mathbf{W}^{\mathbf{M},GQA}$ has $\frac{H}{G}$ copies in $\tilde{\mathbf{W}}^{\mathbf{M},GQA}$, we must have that $\{\mathbf{w}_i^* : i \in \{0, 1, \cdots, r-1\}\}$ also spans $\tilde{\mathbf{W}}^{\mathbf{M},GQA}$.

Therefore,

Skip to main content

$$\operatorname{rank}\left(\tilde{\mathbf{W}}^{\mathrm{M},GQA}\right) = \operatorname{rank}\left(\mathbf{W}^{\mathrm{M},GQA}\right).$$

""" END OF THIS PART """

Related topics

Topic	Replies	Activity
2025 USA-NA-AIO Round 2, Problem 2, Part 9	1	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 12	1	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 1	2	Dec 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 2	2	Dec 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 7	1	May 2025