Topics
My posts
More
CATEGORIES
General
Site Feedback
All categories

2025 USA-NA-AIO Round 2, Problem 2, Part 9

USAAIO May 2025

Now, let us study another variant of MHA: **Multi-head Latent Attention (MLA)**. MLA was introduced by **DeepSeek**. It is a core component of DeepSeek's large language model (LLM).

The key intuition of MLA is as follows. In MHA, the key and value projection matrices

$$\mathbf{W}^{\mathbf{K},MHA} \in \mathbb{R}^{H \cdot D_{qk} \times D_2}, \quad \mathbf{W}^{\mathbf{V},MHA} \in \mathbb{R}^{H \cdot D_v \times D_2}$$

may be high dimensional.

For instance, suppose $H \cdot D_{qk} = H \cdot D_v = D_2 = 4096$.

However, it is not necessarily the case that these matrices are with high ranks (such as 4096). Their actual ranks (or top few ranks that make their truncated singular value decomposition (SVD) to be close to the actual matrices) may be much lower than that.

To capture the low-rank feature, MLA proposed the following model:

$$\begin{aligned}\mathbf{W}^{\mathbf{K},MHA} &= \mathbf{W}^{\mathbf{UK},MLA}\mathbf{W}^{\mathbf{DKV},MLA} \\ \mathbf{W}^{\mathbf{V},MHA} &= \mathbf{W}^{\mathbf{UV},MLA}\mathbf{W}^{\mathbf{DKV},MLA}\end{aligned}$$

where

- $\mathbf{W}^{\mathbf{DKV},MLA} \in \mathbb{R}^{r \times D_2}$: down-projection matrix for computing keys and values.
- $\mathbf{W}^{\mathbf{UK},MLA} \in \mathbb{R}^{H \cdot D_{qk} \times r}$: up-projection matrix for computing keys.
- $\mathbf{W}^{\mathbf{UV},MLA} \in \mathbb{R}^{H \cdot D_v \times r}$: up-projection matrix for computing values.

Skip to main content

In practice, rank r is typically much smaller than $\min\{H \cdot D_{qk}, H \cdot D_v, D_2\}$.

In all remaining parts of this problem, to simplify your analysis and highlight the relationships of MHA, GQA and MLA, we make the following assumptions:

- $D_1 = D_2 = D$.
- $D_{qk} = D_v = d$.
- d is a factor of D.

Under these assumptions, the number heads H satisfies

$$H = \frac{D}{d}.$$

Part 9 (10 points, non-coding task)

In this part, you are asked to prove that GQA can be equivalently represented by MLA.

In your solution, it is sufficient for you to prove that for $\mathbf{M} \in \{\mathbf{K}, \mathbf{V}\}$, for matrix

$$\tilde{\mathbf{W}}^{\mathbf{M},GQA} = \begin{bmatrix} \mathbf{W}^{\mathbf{M},GQA} \\ \mathbf{W}^{\mathbf{M},GQA} \\ \vdots \\ \mathbf{W}^{\mathbf{M},GQA} \end{bmatrix} \in \mathbb{R}^{D \times D}$$

(defined in Part 6) who is the concatenation of $\frac{H}{G}$ copies of

Skip to main content

$$\mathbf{W}^{\mathrm{M},GQA} = \begin{bmatrix} \mathbf{W}_0^{\mathrm{M},GQA} \\ \vdots \\ \mathbf{W}_{G-1}^{\mathrm{M},GQA} \end{bmatrix} \in \mathbb{R}^{G\cdot d\times D}$$

matrix $\tilde{\mathbf{W}}^{\mathrm{M},GQA}$ can be decomposed as

$$\tilde{\mathbf{W}}^{\mathrm{M},GQA} = \mathbf{W}^{\mathbf{UM},MLA}\mathbf{W}^{\mathbf{DKV},MLA}$$

where

- $\mathbf{W}^{\mathbf{DKV},MLA} \in \mathbb{R}^{r\times D}$: down-projection matrix for computing keys and values.
- $\mathbf{W}^{\mathbf{UM},MLA} \in \mathbb{R}^{D\times r}$: up-projection matrix for computing M (keys or values).
- $r = G \cdot d$.

USAAIO **May 2025**

Misplaced '#'

We have

$$\begin{aligned} \operatorname{rank}\left(\tilde{\mathbf{W}}^{\mathrm{M},GQA}\right) &= \operatorname{rank}\left(\mathbf{W}^{\mathrm{M},GQA}\right) \\ &\leq \min\{G\cdot d, D\} \\ &= \min\{r, D\} \\ &= r, \end{aligned}$$

Skip to main content

where the first equality follows from Part 6.

Therefore, SVD implies

$$\tilde{\mathbf{W}}^{\mathrm{M},GQA} = \sum_{i=0}^{r-1} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$$

$$= \underbrace{\begin{bmatrix} \mathbf{u}_0 & \mathbf{u}_1 & \cdots & \mathbf{u}_{r-1} \end{bmatrix}}_{\mathbf{W}^{\mathrm{UM},MLA}} \underbrace{\begin{bmatrix} \sigma_0 \mathbf{v}_0^\top \\ \sigma_1 \mathbf{v}_1^\top \\ \vdots \\ \sigma_{r-1} \mathbf{v}_{r-1}^\top \end{bmatrix}}_{\mathbf{W}^{\mathrm{DKV},MLA}}.$$

""" END OF THIS PART """

✦ Related topics

Topic	Replies	Activity
2025 USA-NA-AIO Round 2, Problem 2, Part 12	1	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 6	2	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 11	1	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 1	2	Dec 2025

Skip to main content

Topic	Replies	Activity
2025 USA-NA-AIO Round 2, Problem 2, Part 10	1	May 2025