Topics
My posts
More
CATEGORIES
General
Site Feedback
All categories

2025 USA-NA-AIO Round 2, Problem 2, Part 12

USAAIO May 2025

MLA does not only enjoy its advantage of being more general than MHA and GQA, it is also computationally more efficient.

An intuitive approach of computing MLA.

1. Compute the key-projection matrix $\mathbf{W}^{\mathbf{UK},MLA}\mathbf{W}^{\mathbf{DKV},MLA} \in \mathbb{R}^{D\times D}$ and the value-projection matrix $\mathbf{W}^{\mathbf{UV},MLA}\mathbf{W}^{\mathbf{DKV},MLA} \in \mathbb{R}^{D\times D}$.
2. Follow the standard steps in MHA.

This approach is hereafter called a $\textbf{vanilla approach}$. This approach fails to enjoy the low-rank feature of $\mathbf{W}^{\mathbf{DKV},MLA}$, $\mathbf{W}^{\mathbf{UK},MLA}$, and $\mathbf{W}^{\mathbf{UV},MLA}$.

Part 12 (10 points, non-coding task)

In this part, you are asked to study an alternative approach to compute MLA.

1. Find a **head-independent reduced key-projection matrix** $\hat{\mathbf{W}}^{\mathbf{K},MLA} \in \mathbb{R}^{r\times D}$ and a **reduced query-projection matrix** $\hat{\mathbf{W}}^{\mathbf{Q},MLA} \in \mathbb{R}^{H\cdot r\times D}$, such that
   - The **reduced key** at position l_2 for head h in a being attended sequence is **head-independent** and is given by:

$$\hat{\mathbf{k}}_{l_2} = \hat{\mathbf{W}}^{\mathbf{K},MLA}\mathbf{y}_{l_2} \in \mathbb{R}^{r}$$

   - The **reduced query** at position l_1 for head h in an attending sequence is given by:

Skip to main content

$$\hat{\mathbf{q}}_{l_1,h} = \hat{\mathbf{W}}_h^{\mathbf{Q},MLA}\mathbf{x}_{l_1} \in \mathbb{R}^r$$

where

$$\hat{\mathbf{W}}^{\mathbf{Q},MLA} = \begin{bmatrix} \hat{\mathbf{W}}_0^{\mathbf{Q},MLA} \\ \hat{\mathbf{W}}_1^{\mathbf{Q},MLA} \\ \vdots \\ \hat{\mathbf{W}}_{H-1}^{\mathbf{Q},MLA} \end{bmatrix}$$

- The attention score (query-key similarity) is invariant in both the original and the reduced forms. That is

$$\frac{\mathbf{q}_{l_1,h}^\top \mathbf{v}_{l_2,h}}{\sqrt{D/H}} = \frac{\hat{\mathbf{q}}_{l_1,h}^\top \hat{\mathbf{v}}_{l_2}}{\sqrt{r}}. \quad (1)$$

2. Find a **head-independent reduced value-projection matrix** $\hat{\mathbf{W}}^{\mathbf{V},MLA} \in \mathbb{R}^{r\times D}$ and a **reduced out-projection matrix** $\hat{\mathbf{W}}^{O,MLA} \in \mathbb{R}^{D\times H\cdot r}$, such that

   - The **reduced value** with head h on position l_2 in a being attended sequence is head-independent and is given by:

$$\hat{\mathbf{v}}_{l_2} = \hat{\mathbf{W}}^{\mathbf{V},MLA}\mathbf{y}_{l_2} \in \mathbb{R}^r$$

   - Post-out-projection is invariant in both the original and the reduced forms.

     Let

$$\hat{\mathbf{W}}^{O,MLA} = \begin{bmatrix} \hat{\mathbf{W}}_0^{O,MLA} & \hat{\mathbf{W}}_1^{O,MLA} & \dots & \hat{\mathbf{W}}_{H-1}^{O,MLA} \end{bmatrix}$$

Skip to main content

Then we must have

$$\sum_{h=0}^{H-1} \mathbf{W}_h^O \sum_{l_2=0}^{L_2-1} \alpha_{h,l_1 l_2} \mathbf{v}_{l_2,h} = \sum_{h=0}^{H-1} \hat{\mathbf{W}}_h^{O,MLA} \sum_{l_2=0}^{L_2-1} \alpha_{h,l_1 l_2} \hat{\mathbf{v}}_{l_2,h}. \quad (2)$$

3. You answer of $\hat{\mathbf{W}}^{\mathbf{K},MLA}$, $\hat{\mathbf{W}}^{\mathbf{V},MLA}$, $\hat{\mathbf{W}}^{\mathbf{Q},MLA}$, and $\hat{\mathbf{W}}^{O,MLA}$ should be written in terms of $\mathbf{W}^{\mathbf{DKV}}$, $\mathbf{W}^{\mathbf{UK}}$, $\mathbf{W}^{\mathbf{UV}}$, $\mathbf{W}^{\mathbf{Q}}$, and $\mathbf{W}^O$.

USAAIO **May 2025**

Misplaced '#'

First, we study Equation (1).

For the LHS in (1), we have

$$\frac{\mathbf{q}_{l_1,h}^\top \mathbf{v}_{l_2,h}}{\sqrt{D/H}} = \frac{1}{\sqrt{D/H}} \left(\mathbf{W}_h^{\mathbf{Q}} \mathbf{x}_{l_1}\right)^\top \left(\mathbf{W}_h^{\mathbf{UK}} \mathbf{W}^{\mathbf{DKV}} \mathbf{y}_{l_2}\right)$$
$$= \frac{1}{\sqrt{D/H}} \mathbf{x}_{l_1}^\top \mathbf{W}_h^{\mathbf{Q},\top} \mathbf{W}_h^{\mathbf{UK}} \mathbf{W}^{\mathbf{DKV}} \mathbf{y}_{l_2} \quad (1.1)$$

For the RHS in (1), we have

$$\frac{\hat{\mathbf{q}}_{l_1,h}^\top \hat{\mathbf{v}}_{l_2}}{\sqrt{r}} = \frac{1}{\sqrt{r}} \left(\hat{\mathbf{W}}_h^{\mathbf{Q},MLA} \mathbf{x}_{l_1}\right)^\top \left(\hat{\mathbf{W}}^{\mathbf{K},MLA} \mathbf{y}_{l_2}\right)$$
$$= \frac{1}{\sqrt{r}} \mathbf{x}_{l_1}^\top \hat{\mathbf{W}}_h^{\mathbf{Q},MLA,\top} \hat{\mathbf{W}}^{\mathbf{K},MLA} \mathbf{y}_{l_2} \quad (1.2)$$

Skip to main content

By equating (1.1) and (1.2), we can set

$$\hat{\mathbf{W}}^{\mathbf{K},MLA} = \boxed{\mathbf{W}^{\mathbf{DKV}}}$$

and

$$\hat{\mathbf{W}}_h^{\mathbf{Q},MLA} = \frac{\sqrt{r}}{\sqrt{D/H}} \mathbf{W}_h^{\mathbf{UK},\top} \mathbf{W}_h^{\mathbf{Q}}$$

Therefore,

$$\hat{\mathbf{W}}^{\mathbf{Q},MLA} = \begin{bmatrix} \hat{\mathbf{W}}_0^{\mathbf{Q},MLA} \\ \hat{\mathbf{W}}_1^{\mathbf{Q},MLA} \\ \vdots \\ \hat{\mathbf{W}}_{H-1}^{\mathbf{Q},MLA} \end{bmatrix} = \boxed{\frac{\sqrt{r}}{\sqrt{D/H}} \begin{bmatrix} \mathbf{W}_0^{\mathbf{UK},\top} \mathbf{W}_0^{\mathbf{Q}} \\ \mathbf{W}_1^{\mathbf{UK},\top} \mathbf{W}_1^{\mathbf{Q}} \\ \vdots \\ \mathbf{W}_{H-1}^{\mathbf{UK},\top} \mathbf{W}_{H-1}^{\mathbf{Q}} \end{bmatrix}}$$

Second, we study Equation (2).

For the LHS in (2), we have

$$\sum_{h=0}^{H-1} \mathbf{W}_h^{O,MLA} \sum_{l_2=0}^{L_2-1} \alpha_{h,l_1 l_2} \mathbf{v}_{l_2,h} = \sum_{h=0}^{H-1} \sum_{l_2=0}^{L_2-1} \alpha_{h,l_1 l_2} \mathbf{W}_h^{O} \mathbf{W}_h^{\mathbf{UV}} \mathbf{W}^{\mathbf{DKV}} \mathbf{y}_{l_2}$$

For the RHS in (2), we have

Skip to main content

$$\sum_{h=0}^{H-1} \hat{\mathbf{W}}_h^{O,MLA} \sum_{l_2=0}^{L_2-1} \alpha_{h,l_1 l_2} \hat{\mathbf{v}}_{l_2,h} = \sum_{h=0}^{H-1} \sum_{l_2=0}^{L_2-1} \alpha_{h,l_1 l_2} \hat{\mathbf{W}}_h^{O,MLA} \hat{\mathbf{W}}^{\mathbf{V},MLA} \mathbf{y}_{l_2}$$

By equating (2.1) and (2.2), we can set

$$\hat{\mathbf{W}}^{\mathbf{V},MLA} = \boxed{\mathbf{W}^{\mathbf{DKV}}}$$

and

$$\hat{\mathbf{W}}_h^{O,MLA} = \mathbf{W}_h^O \mathbf{W}_h^{\mathbf{UV}}$$

Therefore

$$\begin{aligned} \hat{\mathbf{W}}^{O,MLA} &= \begin{bmatrix} \hat{\mathbf{W}}_0^{O,MLA} & \hat{\mathbf{W}}_1^{O,MLA} & \cdots & \hat{\mathbf{W}}_{H-1}^{O,MLA} \end{bmatrix} \\ &= \boxed{\begin{bmatrix} \mathbf{W}_0^O \mathbf{W}_0^{\mathbf{UV}} & \mathbf{W}_1^O \mathbf{W}_1^{\mathbf{UV}} & \cdots & \mathbf{W}_{H-1}^O \mathbf{W}_{H-1}^{\mathbf{UV}} \end{bmatrix}} \end{aligned}$$

""" END OF THIS PART """

✨ Related topics

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

Skip to main content

Topic	Replies	Activity
2025 USA-NA-AIO Round 2, Problem 2, Part 11	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 13	1	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 2	2	Dec 2025