Topics
My posts
More
CATEGORIES
General
Site Feedback
All categories

2025 USA-NA-AIO Round 2, Problem 2, Part 14

USAAIO May 2025

Part 14 (5 points, non-coding task)

In generative AI, such as GPT, we autoprogressively generate tokens. For a given position l, the keys and values on this position $\mathbf{k}_l$ and $\mathbf{v}_l$ are repeatly used in generating tokens for positions $l' > l$.

Therefore, the values of $\mathbf{k}_l$ and $\mathbf{v}_l$ are typically stored in cache (no need to revise your code in earlier parts if your code does not support this). We call such storage as kv**-cache**.

Do the following tasks to compute kv-cache in different models while doing autoregressive inference: (reasoning is required)

1. In MHA, the kv-cache at each position is $2D$. Explain why.
2. In MLA, what is the kv-cache at each position?

USAAIO May 2025

Misplaced '#'

1. In MHA, $\mathbf{k}_l, \mathbf{v}_l \in \mathbb{R}^D$. Therefore, the kv-cache at each position is $\boxed{2D}$.
2. In MLA, because $\mathbf{W}^{\mathbf{DKV}} \in \mathbb{R}^{r \times D}$, we have $\hat{\mathbf{k}}_l, \hat{\mathbf{v}}_l \in \mathbb{R}^r$.

   In addition, because $\hat{\mathbf{k}}_l = \hat{\mathbf{v}}_l$, the kv-cache at each position is $\boxed{r}$.

Skip to main content

""" 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 9	1	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 5	1	May 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 2	2	Dec 2025
2025 USA-NA-AIO Round 2, Problem 2, Part 1	2	Dec 2025