Programming in Mathematics

Bo-Wen Shen, Ph.D.

Lecture #12d:
K–V Caching

Department of Mathematics and Statistics
San Diego State University
http://bwshen.sdsu.edu
DOI: https://doi.org/10.13140/RG.2.2.11127.84642

K-V Caching (Wikipedia, BWS)

- The dimensions of Q, K, and V are determined by the current number of tokens and the model's embedding size.
- Once the new token is generated, the autoregressive procedure appends it to the end of the input sequence, and the transformer layers repeat the matrix calculation for the next token.
- A mathematical analysis reveals that the new token introduces a new query, key, and value vector, appended to Q, K, and V, respectively.
- Appending these new vectors to the K and V matrices is sufficient for calculating the next token prediction.
- Consequently, storing the current K and V matrices in memory saves time by avoiding the recalculation of the attention matrix.
- This feature is known as K-V caching.
- This technique effectively reduces computational cost during inference.
- Please refer to Wikipedia article: DeepSeek to additional information.

K-V Caching

Let

$$W_Q,\ W_K \in \mathbb{R}^{d\times d}, \quad X \in \mathbb{R}^{T\times d},$$

and define

$$Q = X\,W_Q, \quad K = X\,W_K.$$

At time step t we have

$$Q^{(t)} \in \mathbb{R}^{t\times d}, \quad K^{(t)} \in \mathbb{R}^{t\times d}, \quad A^{(t)} = Q^{(t)}\,K^{(t)\top} \in \mathbb{R}^{t\times t}.$$

The above uses a superscript to indicate the number of iterations. In contrast, we use a subscript j to indicate the j-th row, which corresponds to a newly predicted word. For example, x_j q_j, k_j, and $\hat{a}_j$ indicate the j-th row of matrices X, Q, K, and $\hat{A}$.

Time step = 1

Time step $t = 1$

$$XW_Q = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \times \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = Q^{(1)}$$

$4 \times 3 \qquad 3 \times 3 \qquad 4 \times 3$

$$XW_K = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \times \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = K^{(1)}$$

$4 \times 3 \qquad 3 \times 3 \qquad 4 \times 3$

Raw Attention Scores

$$\hat{A}^{(1)} = Q^{(1)} K^{(1)\top} = \underset{4 \times 3}{\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}} \times \underset{3 \times 4}{\begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}} = \underset{4 \times 4}{\begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix}}.$$

Time step = 2

Time step $t = 2$

$$XW_Q = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \times \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = Q^{(2)}$$

$5 \times 3 \qquad 3 \times 3 \qquad 5 \times 3$

$$XW_K = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \times \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} = K^{(2)}$$

$5 \times 3 \qquad 3 \times 3 \qquad 5 \times 3$

Raw Attention Scores

Append the new token's embeddings (q_5, k_5). Then

$$Q^{(2)} = \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}, \quad K^{(2)\top} = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix}.$$

5×3 $\qquad$ 3×5

Hence

$$A^{(2)} = Q^{(2)} K^{(2)\top} = \begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{bmatrix},$$

5×5

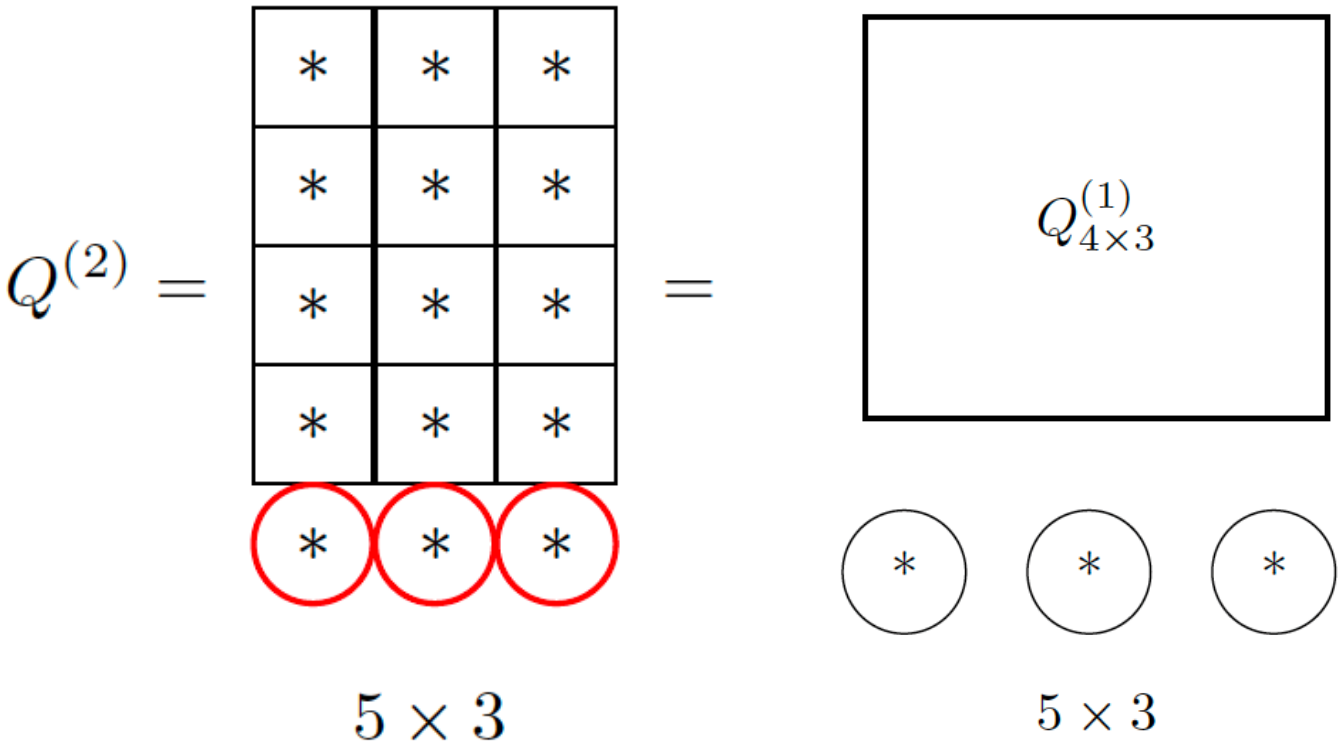
Revisit: Time step = 2

Revisit: Time step $t = 2$

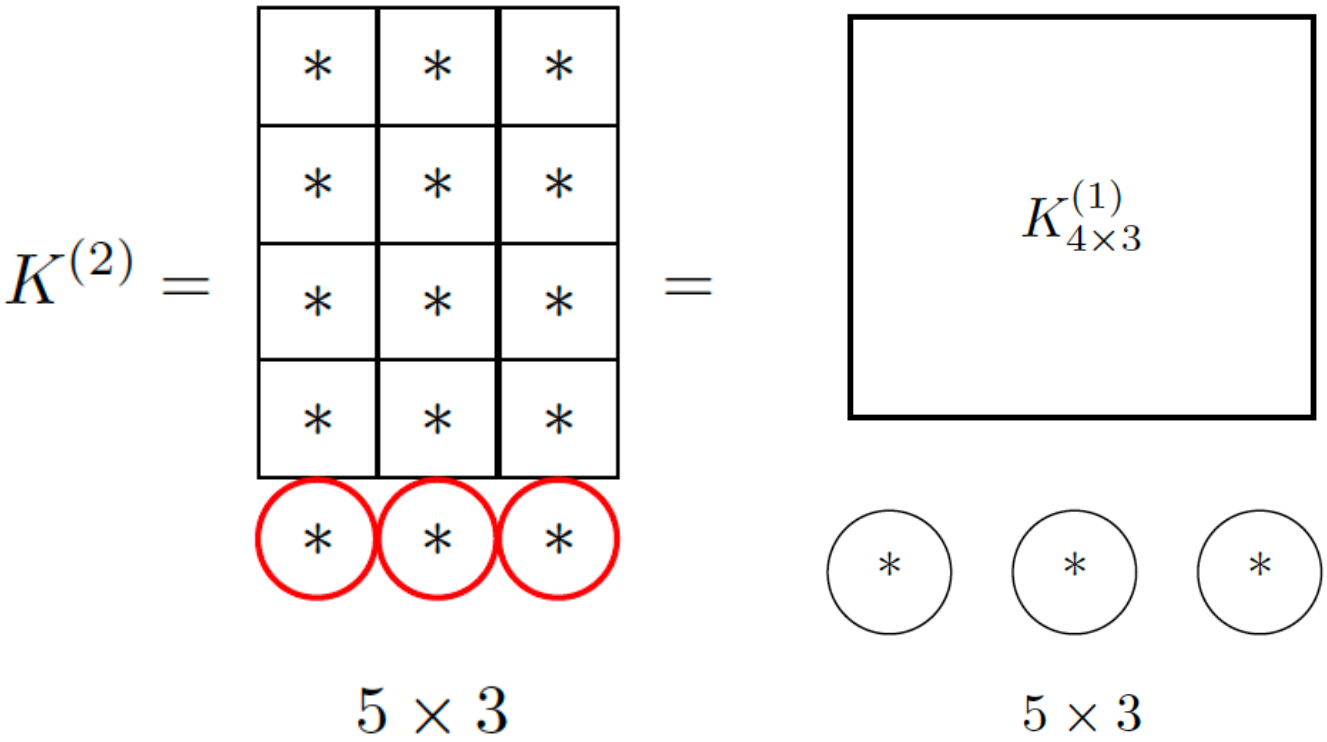
$$x_5 W_Q = \underset{1\times 3}{\bigcirc\bigcirc\bigcirc} \times \underset{3\times 3}{\begin{array}{|c|c|c|}\hline * & * & * \\ \hline * & * & * \\ \hline * & * & * \\ \hline \end{array}} = \underset{1\times 3}{\bigcirc\bigcirc\bigcirc} = q_5$$

$$x_5 W_K = \underset{1\times 3}{\bigcirc\bigcirc\bigcirc} \times \underset{3\times 3}{\begin{array}{|c|c|c|}\hline * & * & * \\ \hline * & * & * \\ \hline * & * & * \\ \hline \end{array}} = \underset{1\times 3}{\bigcirc\bigcirc\bigcirc} = k_5$$

K-V Caching

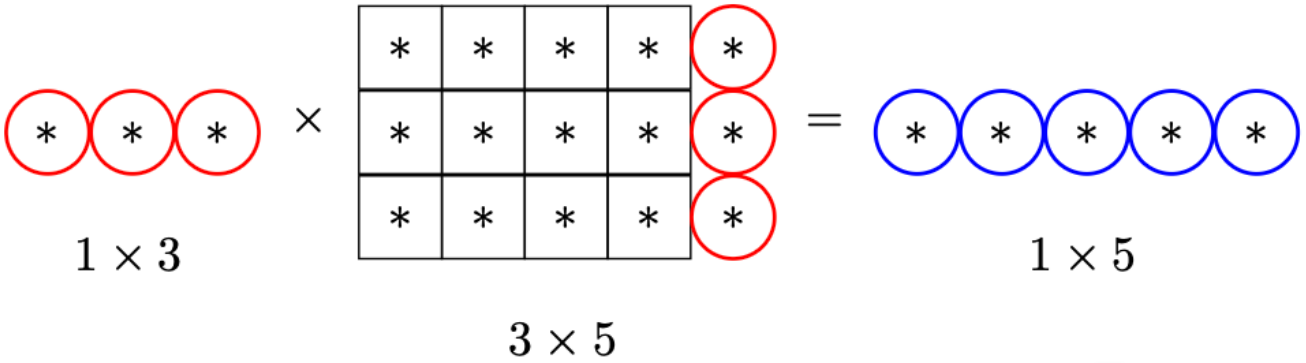


K-V Caching



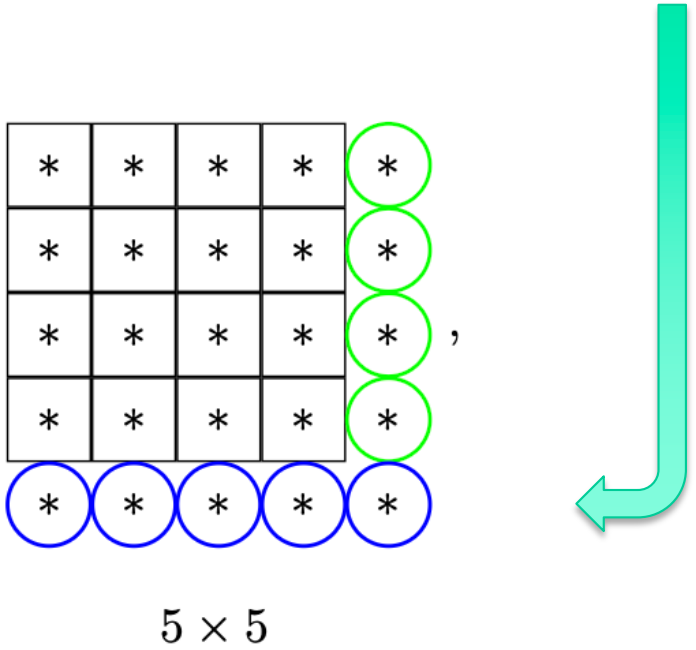
K-V Caching

$$\hat{a}_5 = q_5 \times K^{(2)\top} =$$



Hence

$$\hat{A}^{(2)} = Q^{(2)}\, K^{(2)\top} =$$



where only the blue-circled last row and green-circled last column are newly computed, and the 4 × 4 block at the upper-left is reused from $\hat{A}^{(1)}$.

K-V Caching

- In autoregressive next-token prediction, we apply a causal mask to the attention matrix $\hat{A}$, zeroing out (by assigning $-\infty$ logits) every entry above the main diagonal—that is, the "upper" triangle that would correspond to attending to future tokens.
- What remains after masking is the lower-triangular portion
- (including the diagonal), so there's no need to compute the green-circled last column.