

The diagram illustrates the Softmax operation applied to the product of two matrices,  $\mathbf{Y}$  and  $\mathbf{V}$ , resulting in the matrix  $\mathbf{QK}^T$ .

On the left, a large light blue rectangle represents matrix  $\mathbf{Y}$  with dimensions  $N \times D_v$ . A small red square in the top-left corner indicates a specific element or a small subset of the matrix.

In the center, the word "Softmax" is written, indicating the operation applied to the product.

On the right, a large light blue rectangle represents matrix  $\mathbf{QK}^T$  with dimensions  $N \times N$ . A red horizontal bar at the top of this matrix indicates a specific row or a subset of rows.

To the right of  $\mathbf{QK}^T$ , a multiplication symbol  $\times$  is shown, followed by a vertical light blue rectangle representing matrix  $\mathbf{V}$  with dimensions  $N \times D_v$ . A red vertical bar in the center of  $\mathbf{V}$  indicates a specific column or a subset of columns.

$$N \times D_v \quad = \quad \text{Softmax} \quad \left\{ \begin{array}{c} N \times N \end{array} \right\} \quad \times \quad N \times D_v$$