November 27, 2024

Explanation

The equation $Z = QK^T$

The attention scores A are computed as:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right),$$

The matrix Z, which represents the attention scores before applying softmax:

$$Z = QK^T,$$

In which, we have:

- Q has dimension (num_tokens $\times$ dim_head)
- K has dimension (num_tokens $\times$ dim_head)
- Z has dimension (num_tokens $\times$ num_tokens)

Gradient of Z with respect to Q

To compute the gradient of the loss $\mathcal{L}$ with respect to Q, we apply the chain rule of differentiation. First, we need the gradient of Z with respect to Q. Since $Z = QK^T$, we have:

$$\frac{\partial Z}{\partial Q} = K.$$

Because $Z = QK^T$:
- Each row of Q influences the corresponding rows of Z
- Therefore, the derivative of Z with respect to Q is just K (not K^T).

Backpropagating the Gradient to Q

Using the chain rule, the gradient of the loss $\mathcal{L}$ with respect to Q is:

$$\frac{\partial \mathcal{L}}{\partial Q} = \frac{\partial \mathcal{L}}{\partial Z} \cdot \frac{\partial Z}{\partial Q}.$$

Since $\frac{\partial Z}{\partial Q} = K$, we get:

$$\frac{\partial \mathcal{L}}{\partial Q} = \frac{\partial \mathcal{L}}{\partial A} \cdot K,$$

where $\frac{\partial \mathcal{L}}{\partial A}$ is the gradient of the loss with respect to the attention scores A, and K is the key matrix.

Gradient of Z with respect to K

To compute the gradient with respect to K, take the derivative of $Z = QK^T$ with respect to K:

$$\frac{\partial Z}{\partial K} = Q^T.$$

Thus, the gradient of the loss with respect to K is:

$$\frac{\partial \mathcal{L}}{\partial K} = \left(\frac{\partial \mathcal{L}}{\partial A}\right)^T \cdot Q.$$

Conclusion

Gradient of Z with respect to Q:

$$\frac{\partial \mathcal{L}}{\partial Q} = \frac{\partial \mathcal{L}}{\partial A} \cdot K.$$

Gradient of Z with respect to K:

$$\frac{\partial \mathcal{L}}{\partial K} = \left(\frac{\partial \mathcal{L}}{\partial A}\right)^T \cdot Q.$$