The Mean and Variance of the Dot Product in Attention

In the Scaled Dot-Product Attention mechanism, we scale the dot product of the query (q) and key (k) vectors by $1/\sqrt{d_k}$. This is done to stabilize training by ensuring the inputs to the softmax function have a variance of 1. Here is the mathematical justification.

Assumptions

We start with two key assumptions about the components of the vectors q and k:

1. They are **independent** random variables.
2. Each component has a **mean of 0** and a **variance of 1**.

Let the vectors be $q=(q_1,q_2,\dots,q_{d_k})$ and $k=(k_1,k_2,\dots,k_{d_k})$.

For any component q_i or k_i:

- $E[q_i]=0$ and $E[k_i]=0$
- $Var(q_i)=1$ and $Var(k_i)=1$

---

1. Proving the Mean is 0

The dot product is defined as $q \cdot k=\sum_{i=1}^{d_k} q_i k_i$. Its expected value (mean) is:

$$E[q \cdot k]=E\left[\sum_{i=1}^{d_k} q_i k_i\right]$$

Due to the linearity of expectation, we can move the expectation inside the sum. Since q_i and k_i are independent, the expectation of their product is the product of their expectations:

$$E[q \cdot k]=\sum_{i=1}^{d_k} E[q_i k_i]=\sum_{i=1}^{d_k} E[q_i]E[k_i]$$

Substituting our assumed mean of 0:

$$E[q \cdot k]=\sum_{i=1}^{d_k} (0 \cdot 0)=0$$

Conclusion: The mean of the dot product is 0.

---

2. Proving the Variance is d_k

The formula for variance is $Var(X)=E[X^2]-(E[X])^2$. Since we just proved the mean is 0, this simplifies to $Var(q \cdot k)=E[(q \cdot k)^2]$.

Because all the terms $q_i k_i$ are independent, the variance of their sum is the sum of their variances:

$$\mathrm{Var}(q \cdot k)=\mathrm{Var}\left(\sum_{i=1}^{d_k} q_i k_i\right)=\sum_{i=1}^{d_k} \mathrm{Var}(q_i k_i)$$

Let's find the variance of a single term, $\mathrm{Var}(q_i k_i)$. Again, since its mean $E[q_i k_i]$ is 0, this simplifies to $\mathrm{Var}(q_i k_i)=E[(q_i k_i)^2]=E[q_i^2 k_i^2]$. Because q_i and k_i are independent, so are their squares:

$$\mathrm{Var}(q_i k_i)=E[q_i^2]E[k_i^2]$$

We find $E[q_i^2]$ using the variance formula for q_i itself: $\mathrm{Var}(q_i)=E[q_i^2]-(E[q_i])^2$. Given $\mathrm{Var}(q_i)=1$ and $E[q_i]=0$, we have $1=E[q_i^2]-0^2$, which means $E[q_i^2]=1$. Similarly, $E[k_i^2]=1$.

Plugging this back in, we find the variance of a single term is $\mathrm{Var}(q_i k_i)=1 \cdot 1=1$.

Finally, we substitute this into our sum:

$$\mathrm{Var}(q \cdot k)=\sum_{i=1}^{d_k} \mathrm{Var}(q_i k_i)=\sum_{i=1}^{d_k} 1=d_k$$

Conclusion: The variance of the dot product is d_k. This is why we divide by its standard deviation, 1/sqrt(dk), to re-normalize the variance to 1.