IN3310/IN4310 Transformers Exercise

Dhananjay Tomar

To illustrate why the dot products get large, assume that the components of q and k are independent random variables with mean 0 and variance 1. Then their dot product, $q.k = \sum_{i=1}^{d_k} q_i k_i$ has mean 0 and variance d_k.

$$E(q.k) = E(q_0 k_0) + E(q_1 k_1) + \ldots + E(q_{d_k} k_{d_k})$$

$$E(q.k) = E(q_0)E(k_0) + E(q_1)E(k_1) + \ldots + E(q_{d_k})E(k_{d_k})$$

We can do $E(q_i k_i) = E(q_i)E(k_i)$ because q_i and k_i are independent.

$E(q.k) = 0 * 0 + 0 * 0 + \ldots + 0 * 0 = 0$ because q_i and k_i have mean 0.

$$Var(q.k) = Var(q_0 k_0) + Var(q_1 k_1) + \ldots + Var(q_{d_k} k_{d_k})$$

(The covariance terms are equal to zero since the variables are independent. You can try to prove that on your own)

First, let's calculate variance of $q_i * k_i$ using the basic variance formula: $Var(X) = E(X^2) - [E(X)]^2$

$$Var(q_i k_i) = E(q_i^2 k_i^2) - E(q_i k_i)^2$$

$$Var(q_i k_i) = E(q_i^2)E(k_i^2) - E(q_i k_i)^2 = E(q_i^2)E(k_i^2) - 0^2 = E(q_i^2)E(k_i^2)$$

Now let's calculate $E(q_i^2)$ or equivalently $E(k_i^2)$

$$Var(q_i) = E(q_i^2) - E(q_i)^2 = E(q_i^2) - 0^2 = E(q_i^2)$$

Therefore $E(q_i^2) = Var(q_i) = 1$

By substituting it in the equation above we get

$$Var(q_i k_i) = E(q_i^2)E(k_i^2) = 1 * 1 = 1$$

Hence $Var(q.k) = 1 + 1 + \ldots + 1 = d_k$