

Hey Grok! I'm looking for a helpful tutor to walk through the solutions for a deep learning assignment. Based on the solutions we develop here, I'll share insights with the rest of the class that go alongside the official assignment solutions. Can you help me?

Sure, I'd be happy to help tutor you through your deep learning assignment! Deep learning is one of my favorite topics—it's like giving machines a brain, but with way more math and fewer existential crises (usually).

To get started, could you share the details of the assignment? Things like the problem statement, any provided code snippets, datasets, or specific questions you're working on would be super helpful. If it's something involving coding (e.g., PyTorch or TensorFlow implementations), we can iterate on solutions step by step, and I can even run code checks if needed.

What's the first part you'd like to tackle?

I'll begin by providing you with the entire file of the assignment, and then we can focus on the questions one-by-one. For now, just focus on problem 1, the very first question which takes a probabilistic view of the attention operation between key and value pairs within a transformer. What would be solutions to parts a, b, and c?

[PDF](#) Deep\_NN\_HW9.pdf

Great! Let's dive into **Problem 1** — this is a classic motivation question for why we scale the dot product by  $1/\sqrt{d_k}$  in the original Transformer paper.

We have two independent random vectors

$q, k \sim N(\mu, \sigma^2 I)$  i.i.d. in  $\mathbb{R}^d$

(each component is  $N(\mu, \sigma)$ ).

### Part (a)

Find  $E[q^T k]$  in terms of  $\mu, \sigma, d$ .

Since expectation is linear,

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

Because  $q_i$  and  $k_i$  are independent and identically distributed,

$$E[q_i k_i] = E[q_i] \cdot E[k_i] = \mu \cdot \mu = \mu^2$$

So

$$E[q^T k] = \sum_{i=1}^d \mu^2 = d \mu^2$$

Answer (a):  $d \mu^2$

### Part (b)

Now assume the practical case  $\mu = 0, \sigma = 1$ .

Find  $\text{Var}(q^T k)$  in terms of  $d$ .

Since  $q$  and  $k$  are independent and  $\mu = 0$ ,

$$\text{Var}(q^T k) = \text{Var}(\sum_{i=1}^d q_i k_i) = \sum_{i=1}^d \text{Var}(q_i k_i)$$

(the cross terms vanish because coordinates are independent)

For each term:  $q_i \sim N(0, 1)$ ,  $k_i \sim N(0, 1)$ , independent  $\rightarrow q_i k_i$  has variance

$$\text{Var}(q_i k_i) = E[(q_i k_i)^2] - (E[q_i k_i])^2 = E[q_i^2] E[k_i^2] - 0 = (1)(1) = 1$$

Thus

$$\text{Var}(q^T k) = \sum_{i=1}^d 1 = d$$

Answer (b):  $d$

### Part (c)

Still with  $\mu = 0, \sigma = 1$ .

We look at the scaled dot product:  $q^T k / s$

We want

$$E[q^T k / s] = 0 \text{ (already true because } E[q^T k] = 0)$$

and

$$\text{Var}(q^T k / s) = 1$$

From part (b),  $\text{Var}(q^T k) = d$ , so

$$\text{Var}(q^T k / s) = \text{Var}(q^T k) / s^2 = d / s^2$$

Set this equal to 1:

$$d / s^2 = 1$$

$$\Rightarrow s^2 = d$$

$$\Rightarrow s = \sqrt{d} \text{ (we take the positive root by convention)}$$

So the scaling factor should be  $\sqrt{d}$  (i.e., we divide by  $\sqrt{d}$ ).

**Answer (c):**  $s = \sqrt{d}$

(or equivalently, we compute the attention logits as  $q^T k / \sqrt{d}$ )

### Summary of answers for Problem 1

- (a)  $d \mu^2$
- (b)  $d$
- (c)  $\sqrt{d}$  (or  $s = \sqrt{d}$ )

This is exactly the reasoning from the “Attention Is All You Need” paper (Vaswani et al., 2017, Section 3.2.1): when queries and keys are roughly zero-mean and unit-variance in each dimension, the raw dot product has standard deviation  $\sqrt{d}$ , which grows with dimension and makes softmax produce extremely peaked (or very flat) distributions. Scaling by  $1/\sqrt{d}$  keeps the variance of the logits at 1, giving a much more stable softmax.

Let me know when you’re ready for Problem 2 (argmax attention) or any other part!

Your responses to parts b and c look good, but I notice that there’s an issue with part a.  
What kind of object is mu?

You’re absolutely right — I missed a crucial detail in part (a), and thank you for catching it!

Let’s look carefully at the exact wording in the PDF:

Suppose  $q, k \in \mathbb{R}^d$  are two random vectors with  $q, k$  iid  $\sim N(\mu, \sigma^2 I)$ , where  $\mu \in \mathbb{R}^d$  and  $\sigma \in \mathbb{R}^+$ .

So  $\mu$  is a vector in  $\mathbb{R}^d$ , not a scalar! That changes everything for part (a).

### Correct solution for Problem 1

**Part (a):  $E[q^T k]$  in terms of  $\mu$ ,  $\sigma$ , and  $d$**

Since  $q \sim N(\mu, \sigma^2 I)$  and  $k \sim N(\mu, \sigma^2 I)$  independently,

$$E[q^T k] = E[\sum_{i=1}^d q_i k_i] = \sum_{i=1}^d E[q_i] E[k_i] = \sum_{i=1}^d \mu_i \cdot \mu_i = \sum_{i=1}^d \mu_i^2 = \|\mu\|^2$$

Alternatively, in vector notation:

$$E[q^T k] = E[q]^T E[k] = \mu^T \mu = \|\mu\|^2$$

So the correct answer is:

**Answer (a):**  $\mu^T \mu$  (or equivalently  $\|\mu\|^2$  or  $\mu \cdot \mu$ )

(Definitely not  $d \mu^2$  — that would only be true if  $\mu$  were a scalar. My earlier answer was wrong because I misread  $\mu$  as scalar.)

**Part (b):** Still correct — when  $\mu = 0$  and  $\sigma = 1$ ,

$$\text{Var}(q^T k) = d$$

(because  $\text{Var}(q_i k_i) = 1$  for each  $i$ , and they’re independent)

**Part (c):** Also still correct — we want  $\text{Var}(q^T k / s) = 1$

$$\Rightarrow d / s^2 = 1$$

$$\Rightarrow s = \sqrt{d}$$

So scaling by  $1/\sqrt{d}$  keeps the variance of the attention logits at 1 when inputs are zero-mean and unit-variance per dimension.

**Final corrected answers for Problem 1:**

- (a)  $\mu^T \mu$  (or  $\|\mu\|^2$ , or  $\mu \cdot \mu$ )
- (b) d
- (c)  $\sqrt{d}$

Thanks again for the sharp catch — this is a really common trick in homework problems: they first give the general case with vector mean to make you think carefully, then specialize to  $\mu=0$  later. Easy to miss on a quick read!

Ready for Problem 2 (argmax attention) whenever you are!

↳ Explain vector mean in attention

↳ Multi-head attention intuition