A5 Written

Alok Regmi[*]

May 26, 2021

Contents

Q.1.a.

Copying in attention. Describe (in one sentence) what properties of the inputs to the attention operation would result in the output c being approximately equal to vj for some j {1, . . . , n}. Specifically, what must be true about the query q, the values {v1 , . . . , vn } and/or the keys {k1 , . . . , kn }?

answer

Since our softmax function never gives output that's exactly 0 to all the elements, we will copy our v_j into the attention output only if our value vector is represented as one-hot vector.

Q.1.b.

Assume key vectors as perpendicular vectors and values be arbitrary. Let two values from value vectors be v_a and v_b . Give expression for query vector q such that the output c is approximately equal to average of the two.

answer

- This has to be related to our keys. Keys are independent of each other.

[*]sagar.r.alok@gmail.com

- We need not scale $[k_a, k_b]$ since it's already assumed that $||k_I|| = 1$.

$$q = \frac{k_a + k_b}{2}$$

Then,

$$qk^T = [k_a.q, k_b.q, ..., k_i.q]$$

Since q is linear combination of two vectors k_a and k_b, all the dot products except for k_a and k_b will be 0. Thus,

$$qk^T = [\frac{k_a.k_a}{2}, \frac{k_b.k_b}{2}, 0, 0, ..., 0]$$

Now alpha will be almost non-negligible for all the values that are 0. We can scale up the vector by scalar s if required so that the probabilities get close to 0.5.

Q.1.c.i

Now assuming key vectors are randomly sampled $k_i \sim \mathcal{N}(\mu_i, \sum_i)$ with means μ_i known but covariances $\sum_i$ unknown. Further, all means μ_i are perpendicular and unit norm. $||\mu_i|| = 1$.

Further assume, covariance matrices $\sum_i = \alpha I$, for vanishingly small α.

Q.1.c.ii