Natural Language Processing: Homework 2

Taewook Kang

October 23, 2024

Problem (a.i)

Explain why α can be interpreted as a categorical probability distribution.

Solution. Since $\sum_{i=1}^{n} \alpha_i = 1$ and each score α_i lies between 0 and 1, we can interpret α_i as the probability of selecting the category with index i. Thus, α represents the categorical probability distribution.

Problem (a.ii)

Describe (in one sentence) under what conditions the categorical distribution α puts almost all of its weight on some α_j, where $j \in \{1, \cdots, n\}$ (i.e. $\alpha_j \gg \sum_{i \neq j} \alpha_i$). What must be true about the query q and/or the keys $\{k_1, \cdots, k_n\}$?

Solution. If the query q and a specific key k_j have high similarity in the vector space, while other keys do not, the dot product $q^\top k_j$ becomes significantly larger than the dot products between q and the other keys. Consequently, the attention weight associated with $q^\top k_j$, i.e. α_j, becomes much larger than the others, leading to $\alpha_j \gg \sum_{i \neq j} \alpha_i$.

Problem (a.iii)

Under the conditions you gave in (ii), describe the output c.

Solution. The result will be $c \approx v_j$, as α_j becomes much larger than all other attention weights, concentrating most of the probability mass on α_j.

Problem (a.iv)

Explain (in two sentences or fewer) what your answer to (ii) and (iii) means intuitively.

Solution. If the dot product between the query q and the key k_j of the j-th word is significantly higher compared to the other words, the attention weight α_j will approach 1. Therefore, the attention mechanism will focus almost exclusively on the j^{th} word, resulting in $c \approx v_j$, effectively "copying" the value v_j to the output c.

Problem (b.i)

Construct M such that for any $v_a, v_b, Ms = v_a$.

Solution. Given that v_a lies in the subspace $A \in \mathbb{R}^{d\times m}$ and v_b lies in the subspace $B \in \mathbb{R}^{d\times p}$, we can express v_a and v_b as follows:

$$v_a = \sum_{i=1}^{m} c_m a_m = Ac, \quad v_b = \sum_{i=1}^{p} c'_p b_p = Bc', \tag{1}$$

where $c \in \mathbb{R}^m$ and $c' \in \mathbb{R}^p$ are coefficient vectors corresponding to the subspaces A and B, respectively.
We are given that the matrix $M \in \mathbb{R}^{d\times d}$ should satisfy the following equation:

$$Ms = Mv_a + Mv_b = v_a, \quad s, v_a, v_b \in \mathbb{R}^d. \tag{2}$$

Since the basis vectors of A are orthonormal, i.e. $a_j^\top a_k = 1$ if $j = k$ and $a_j^\top a_k = 0$ if $j \neq k$, we have:

$$A^\top A = \mathbf{I} \in \mathbb{R}^{m\times m}. \tag{3}$$

Additionally, because the subspaces A and B are orthogonal, i.e. $a_j^\top b_k = 0$ for all j, k, the following equation is true:

$$A^\top B = \mathbf{0} \in \mathbb{R}^{m\times p}. \tag{4}$$

If we set $M = AA^\top$, we can substitute M into $AA^\top$ in equation (2), using equation (1):

$$(AA^\top)Ac + (AA^\top)Bc' = A(A^\top A)c + A(A^\top B)c' = A\mathbf{I}c + A\mathbf{0}c' = Ac = v_a. \tag{5}$$

Thus, from equation (5), we can conclude that $M = AA^\top \in \mathbb{R}^{d\times d}$ allows us to extract v_a from the sum $v_a + v_b$.

Problem (b.ii)

Find an expression for a query vector q such that $c \approx \frac{1}{2}(v_a + v_b)$, and justify your answer.

Solution. Since all key vectors are orthonormal, the following equation holds:

$$k_i^\top k_i = 1, \quad k_i^\top k_j = 0 \text{ where } i \neq j. \tag{6}$$

Given that $c = \sum_{i=1}^{n} v_i \alpha_i \approx \frac{v_a + v_b}{2}$, we can infer that $\alpha_a = \alpha_b \approx \frac{1}{2}$ and $\alpha_{i \neq a,b} \approx 0$. If we set the query $q = \beta(k_a + k_b)$, where $0 \ll \beta$ to ensure that the term $\exp(0) = 1$ is negligible, we observe that α_a and α_b approach $\frac{1}{2}$:

$$k_a^\top q = k_a^\top \beta(k_a + k_b) = \beta k_a^\top k_a + \beta \times 0 = \beta. \tag{7}$$

$$k_a^\top q = k_b^\top \beta(k_a + k_b) = \beta \times 0 + \beta k_b^\top k_b = \beta. \tag{8}$$

For $i \neq a, b$, we have:

$$k_i^\top q = k_i^\top \beta(k_a + k_b) = \beta(k_i^\top k_a + k_i^\top k_b) = \beta \times (0 + 0) = 0 \quad \text{for all } i \neq a, b. \tag{9}$$

Using these dot products, the attention weights α_a, α_b, and $\alpha_{i \neq a,b}$ are computed as:

$$\alpha_a = \frac{\exp(k_a^\top q)}{\sum_{i=1}^{n} \exp(k_i^\top q)} \approx \frac{\exp(\beta)}{2 \exp \beta} = \frac{1}{2}, \quad \alpha_b \approx \frac{1}{2}. \tag{10}$$

$$\alpha_i \approx \frac{\exp(0)}{2 \exp(\beta)} \approx 0 \quad \text{for all } i \neq a, b. \tag{11}$$

With these values of α_a and α_b, the output c is expressed as:

$$c = \sum_{i=1}^{n} v_i \alpha_i \approx \frac{1}{2} v_a + \frac{1}{2} v_b = \frac{1}{2}(v_a + v_b). \tag{12}$$

Therefore, the the query vector q can be appropriately expressed as:

$$q = \beta(k_a + k_b), \text{where } 0 \ll \beta. \tag{13}$$

Problem (c.i)

Design a query q in terms of the μ_i such that as before, $c \approx \frac{1}{2}(v_a + v_b)$, and provide a brief argument as to why it works.

Solution. Since $\Sigma_i = \alpha\mathbf{I}$ and α is vanishingly small, the key vector $k_i \sim \mathcal{N}(\mu_i, \Sigma_i)$ can be approximated as $k_i \approx \mu_i$. Therefore, if we set the query q as $\beta(\mu_a + \mu_b)$ (similar to the solution of Problem (b.ii)), we can drive the output c to approach $\frac{1}{2}(v_a + v_b)$.

$$k_a^\top q \approx \mu_a^\top \beta(\mu_a + \mu_b) = \beta\mu_a^\top \mu_a + \beta\mu_b^\top \mu_a = \beta \quad (\mu_a^\top \mu_a = 1, \mu_a^\top \mu_b = 0). \tag{14}$$

$$k_b^\top q \approx \mu_b^\top \beta(\mu_a + \mu_b) = \beta\mu_b^\top \mu_a + \beta\mu_b^\top \mu_b = \beta \quad (\mu_b^\top \mu_a = 0, \mu_b^\top \mu_b = 1). \tag{15}$$

$$k_i^\top q \approx \mu_i^\top \beta(\mu_a + \mu_b) = 0, \text{for all } i \neq a, b \quad (\mu_i^\top \mu_j = 0). \tag{16}$$

$$\therefore \alpha_a \approx \frac{1}{2}, \alpha_b \approx \frac{1}{2}, \alpha_{i\neq a,b} \approx 0. \tag{17}$$

With these values of α_a and α_b, the output c is given by:

$$c = \sum_{i=1}^{n} v_i \alpha_i \approx \frac{1}{2}v_a + \frac{1}{2}v_b = \frac{1}{2}(v_a + v_b). \tag{18}$$

Therefore, the appropriate query for this condition is:

$$q = \beta(\mu_a + \mu_b), \text{where } 0 \ll \beta. \tag{19}$$

Problem (c.ii)

When you sample $\{k_1, \cdots, k_n\}$ multiple times, and use the q vector that you defined in part i., what do you expect the vector c will look like qualitatively for different samples?

Solution. Given that $\mu_i^\top \mu_i = 1, \mu_i^\top \mu_k = 0$ and α is vanishingly small, the covariance matrix $\Sigma_a = \alpha\mathbf{I} + \frac{1}{2}(\mu_a \mu_a^\top)$ means that k_a will vary between $0.5\mu_a$ (i.e. $\mu_a - \frac{1}{2}\mu_a$) and $1.5\mu_a$ (i.e. $\mu_a + \frac{1}{2}\mu_a$). As a result, k_a will follow the distribution:

$$k_a \approx \epsilon\mu_a, \quad \epsilon \sim \mathcal{N}(1, 0.5). \tag{20}$$

For the remaining k_i for all $i \neq a$, each k_i will be approximately:

$$k_i \approx \mu_i \quad \text{for all } i \neq a. \tag{21}$$

When $q = \beta(\mu_a + \mu_b)$, the dot product $k_a^\top q, k_b^\top q$, and $k_{i\neq a,b}^\top q$ are computed as follows:

$$k_a^\top q \approx \epsilon\mu_a^\top \{\beta(\mu_a + \mu_b)\} = \beta\epsilon \quad \text{where } \epsilon \sim \mathcal{N}(1, 0.5). \tag{22}$$

$$k_b^\top q \approx \mu_b^\top \{\beta(\mu_a + \mu_b)\} = \beta. \tag{23}$$

$$k_i^\top q \approx \mu_i^\top \{\beta(\mu_a + \mu_b)\} = 0 \quad \text{for all } i \neq a, b. \tag{24}$$

Given the collection of $k_i^\top q$, and noting that $\alpha_{i\neq a,b} \approx 0$, the attention weights α_a and α_b are calculated as:

$$\alpha_a \approx \frac{\exp(\beta\epsilon)}{\exp(\beta\epsilon) + \exp(\beta)} = \frac{1}{1 + \exp(\beta(1-\epsilon))}. \tag{25}$$

$$\alpha_b \approx \frac{\exp(\beta)}{\exp(\beta\epsilon) + \exp(\beta)} = \frac{1}{1 + \exp(\beta(\epsilon-1))}. \tag{26}$$

When β is large ($0 \ll \beta$) and ϵ ranges from 0.5 to 1.5, the behavior of α_a and α_b is as follows:

$$\alpha_a \approx \frac{1}{1+\infty} = 0, \quad \alpha_b \approx \frac{1}{1+0} = 1 \quad \text{if } \epsilon = 0.5. \tag{27}$$

$$\alpha_a \approx \frac{1}{1+0} = 1, \quad \alpha_b \approx \frac{1}{1+\infty} = 0 \quad \text{if } \epsilon = 1.5. \tag{28}$$

Therefore, given that $c \approx \alpha_a v_a + \alpha_b v_b$ when β is large enough to make other terms negligible, the output c will fluctuate between v_a and v_b. This causes the output c to become unstable, as it varies depending on the sampled key vectors.

Problem (d.i)

Design q_1 and q_2 such that c is approximately equal to $\frac{1}{2}(v_a + v_b)$.

Solution. In the equation $c = \frac{1}{2}(c_1 + c_2) \approx \frac{1}{2}(v_a + v_b)$, let's assume that $c_1 \approx v_a$ and $c_2 \approx v_b$. To ensure the attention weights $\alpha_a = 1$ and $\alpha_b = 1$ for c_1 and c_2, we can set the queries q_1 and q_2 as follows:

$$q_1 = \beta\mu_a \quad \text{where } 0 \ll \beta. \tag{29}$$

$$q_2 = \beta\mu_b \quad \text{where } 0 \ll \beta. \tag{30}$$

For example, for c_1, we can demonstrate that c_1 approximates v_a when using the query $q_1 = \beta k_a$. Given that $k_i \sim \mathcal{N}(\mu_i, \Sigma_i) \approx \mu_i$, due to the vanishingly small α in $\Sigma_i = \alpha\mathbf{I}$, this results in:

$$k_a^\top q_1 = \beta(\mu_a^\top \mu_a) = \beta. \tag{31}$$

$$k_i^\top q_1 = \beta(\mu_i^\top \mu_a) = 0, \quad \text{for all } i \neq a. \tag{32}$$

Thus, the attention weights become:

$$\alpha_a \approx \frac{\exp(\beta)}{\exp(\beta)} = 1, \quad \alpha_i \approx \frac{\exp(0)}{\exp(\beta)} \approx 0 \quad \text{for all } i \neq a. \tag{33}$$

Finally, we approximate c_1 as:

$$c_1 = \sum_{i=1}^{n} \alpha_i v_i = v_a. \tag{34}$$

A similar procedure applies to c_2 with the query $q_2 = \beta\mu_b$, leading to $c_2 \approx v_b$. Therefore, we conclude that $c \approx \frac{1}{2}(v_a + v_b)$, and we can define the queries as $q_1 = \beta\mu_a$ and $q_2 = \beta\mu_b$.

Problem (d.ii)

What, qualitatively, do you expect the output c to look like across different samples of the key vectors? Explain briefly in terms of variance in c_1 and c_2.

Solution. If we use the query vectors $q_1 = \beta k_a$, and $q_2 = \beta k_b$, the dot product between the keys and queries will be as follows:

$$k_a^\top q_1 \approx \beta\epsilon(\mu_a^\top \mu_a) = \beta\epsilon \quad \text{where } \epsilon \sim \mathcal{N}(1, 0.5). \tag{35}$$

$$k_b^\top q_2 \approx \beta(\mu_b^\top \mu_b) = \beta. \tag{36}$$

Using these results, we can compute the attention weights for a with q_1 (since all other key-query dot products will be negligible when β is sufficiently large):

$$\alpha_a = \frac{\exp(\beta\epsilon)}{\exp(\beta\epsilon)} = 1, \quad \alpha_i = 0 \text{ for all } i \neq a. \tag{37}$$

Similarly, we can compute the attention weights for b with q_2:

$$\alpha_b = \frac{\exp(\beta)}{\exp(\beta)} = 1, \quad \alpha_i = 0 \text{ for all } i \neq b. \tag{38}$$

As a result, we can conclue:

$$c_1 = \sum_{i=1}^{n} \alpha_i v_i \approx v_a, \quad c_2 = \sum_{i=1}^{n} \alpha_i v_i \approx v_b. \tag{39}$$

This means that $c \approx \frac{1}{2}(v_a + v_b)$. Therefore, regardless of the sampled key vectors k, we can compute a stable output c due to the multiple query vectors used in multi-headed attention, ensuring stability and consistency in the result.