(a) i All the values of α_i are non-negative ($\alpha_i \geq 0$)

and their summation is 1. ($\sum_{i=1}^{n} \alpha_i = 1$)

These two features make α able to be considered as probability distribution

ii When the query vector q is significantly more aligned with the key vector k_j

than any other key vector k_i ($i \neq j$), resulting in a much larger $k_j^T q$

the distribution of α concentrates almost all of its weight on the single component α_j

iii When $\alpha_j \simeq 1$ and $\alpha_i \simeq 0$ ($i \neq j$), $c = \sum_{i=1}^{n} v_i \alpha_i \simeq v_j \alpha_j$

iv If the query is highly similar to a specific key.

the attention mechanism acts as a copying, or lookup system.

It allows the model to dynamically focus on a single information

from a larger set of inputs and pass its value directly to the output

(b) i $A = [a_1 \ a_2 \cdots a_m]$. $M = A \cdot A^T$ $Ms = M(v_a + v_b) = Mv_a + Mv_b$

① $Mv_a = (AA^T)(Ac) = A(A^TA)c = Ac = v_a$

$$A^TA = \begin{bmatrix} -a_1^T- \\ \vdots \\ -a_m^T- \end{bmatrix} \begin{bmatrix} | & & | \\ a_1 & \cdots & a_m \\ | & & | \end{bmatrix} = \begin{bmatrix} a_1^Ta_1 & \cdots & a_1^Ta_m \\ \vdots & & \vdots \\ a_m^Ta_1 & \cdots & a_m^Ta_m \end{bmatrix} = I \quad (\because a_i^Ta_j = \begin{cases} 1 & (i=j) \\ 0 & (i \neq j) \end{cases})$$

$$v_a = c_1a_1 + c_2a_2 + \cdots + c_ma_m = A \cdot \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix} = A \cdot c$$

② $Mv_b = AA^TBc' = 0$

$$A^TB = \begin{bmatrix} -a_1^T- \\ \vdots \\ -a_m^T- \end{bmatrix} \begin{bmatrix} | & & | \\ b_1 & \cdots & b_p \\ | & & | \end{bmatrix} = \begin{bmatrix} a_1^Tb_1 & \cdots & a_1^Tb_p \\ \vdots & & \vdots \\ a_m^Tb_1 & \cdots & a_m^Tb_p \end{bmatrix} = 0 \quad (\because a_i^Tb_j = 0 \text{ for all } i, j)$$

by ①,②, $Ms = M(v_a + v_b) = Mv_a + Mv_b = v_a + 0 = v_a$.

ii $c \simeq \frac{1}{2}(v_a + v_b) \rightarrow \alpha_a \simeq 0.5, \ \alpha_b \simeq 0.5, \ \alpha_i (i \neq a,b) \simeq 0$

let $q = k_a + k_b$

$$\left(\begin{aligned} k_a^T q &= k_a^T(k_a + k_b) = 1 + 0 = 1. \\ k_b^T q &= k_b^T(k_a + k_b) = 0 + 1 = 1. \\ k_i^T q &= k_i^T(k_a + k_b) = 0 + 0 = 0. \end{aligned} \right) \quad \alpha_a = \frac{e^1}{2e^1 + \sum_{i \neq a,b} e^0} \neq \frac{1}{2}$$

using hint, let $q = C \cdot (k_a + k_b)$ which C is extremely large scalar value

$$\alpha_a = \alpha_b = \frac{e^C}{2e^C + (n-2)e^0} \simeq \frac{1}{2}.$$

(c) i $\Sigma_i = \alpha I$ (α: vanishinglly small) $\Rightarrow k_i \simeq \mu_i$

So, let $q = c(\mu_a + \mu_b)$, $k_i^T q \simeq \mu_i^T q$. $\therefore c \simeq \frac{1}{2}(\mu_a + \mu_b)$

ii let's rewrite score $s_a = k_a^T q = c(k_a^T \mu_a + k_a^T \mu_b)$

$$\mathrm{Var}[s_a] = \mathrm{Var}[k_a^T q] = q^T \mathrm{cov}[k_a^T] q = c^2(\mu_a + \mu_b)^T (\alpha I + \tfrac{1}{2}\mu_a\mu_a^T)(\mu_a+\mu_b)$$

$$= c^2(\alpha \mu_a^T\mu_a + \tfrac{1}{2}\mu_a^T\mu_a\mu_a^T\mu_a + \alpha\mu_a^T\mu_b + \tfrac{1}{2}\mu_a^T\mu_a\mu_a^T\mu_b$$

$$+ \alpha\mu_b^T\mu_a + \tfrac{1}{2}\mu_b^T\mu_a\mu_a^T\mu_a + \alpha\mu_b^T\mu_b + \tfrac{1}{2}\mu_b^T\mu_a\mu_a^T\mu_b)$$

($\mu_a^T\mu_a = \mu_b^T\mu_b = 1$, $\mu_a^T\mu_b = \mu_b^T\mu_a = 0$)

$$= c^2(2\alpha + \tfrac{1}{2}) \simeq \frac{c^2}{2}$$

$$\mathrm{Var}[s_b] = q^T \mathrm{Cov}[k_b^T] q = c^2(\mu_a+\mu_b)^T \alpha I (\mu_a+\mu_b) = 2\alpha c^2 \simeq 0$$

$\therefore c = v_1 a_1 + v_2 a_2 + \cdots + v_m a_m$ isn't a stable mean

and single-head attention is not robust for this kind of noise.

(d) i let $q_1 = \mu_a$, $q_2 = \mu_b$

in head 1, $s_{1,a} = k_a^T \mu_a \simeq \mu_a^T\mu_a = 1$, $s_{1,i} = k_i^T \mu_a \simeq \mu_i^T \mu_a = 0$

$\alpha_{1,a} \simeq 1$, $\alpha_{1,i} \simeq 0$ $\therefore c_1 = \Sigma\, \alpha_{1,i} v_i \simeq v_a$

similarly, $\alpha_{2,b} \simeq 1$, $\alpha_{2,i} \simeq 0$ $\therefore c_2 = \Sigma\, \alpha_{2,i} v_i \simeq v_b$

$c = \frac{1}{2}(c_a + c_b) \simeq \frac{1}{2}(v_a + v_b)$

ii $\Sigma_a = \alpha I + \frac{1}{2}(\mu_a\mu_a^T)$

In head 1, k_a has the same direction with μ_a and different magnitude.

So, c_1 has a high variance and c_1 becomes unstable.

While in head 2, k_a and μ_b are orthogonal : the noise of k_a is negligible to c_2.

So c_2 keeps close to v_b with low variance.

Finally, $c = \frac{1}{2}(c_1 + c_2)$, it makes its variance smaller than single-head attention