< NLP Homework 2 >

Self - Attention & Transformers

Computer Software Engineering
20200814121
KIM SUNG HYUK (김성혁)

(a) - i) Each α_i $(i = 1 \cdots n)$ are probability,

$1 \geq \alpha_i \geq 0$, $\sum_{i=1}^{n} \alpha_i = 1$. # category $= n$ $(n > 0)$

$\therefore$ α can be interpleted as a categorical probability distribution.

(a) - ii) $K_j^T q \gg K_i^T q$ (for all $i = 1, \cdots, n$ except $i = j$)

(a) - iii) ~~$K_j^T q \gg$~~ $\alpha_j \gg \sum_{i \neq j} \alpha_i$. ~~$\alpha_j$~~ and $\sum_{i=1}^{n} \alpha_i = 1$

$\therefore$ $\alpha_j \approx 1$. $\therefore$ $c = \sum_{i=1}^{n} v_i \alpha_i \approx v_j \alpha_j \approx v_j$

(a) - iv) If q aligns very closely with one of all key vectors (k_i),

attention weight significantly larger than other $k_i^T q$.

As a result, c becomes almost identical to v_j, in that situation we ~~copi~~copy a value vector (v_j) to the output c //

(b) - i)

$V_n = c_1 a_1 + \cdots + c_m a_m$, $A = [a_1, \cdots, a_m]$, ~~$C = [c_1, \cdots, c_m]$~~

$C = \begin{bmatrix} c_1 \\ \vdots \\ c_d \end{bmatrix}$

$MS = r_a \iff M(V_a + V_b) = V_a \iff \underline{MV_a + MV_b = V_a}$

∴ We will construct M that satisfy ① $MV_a = V_a$ ② $MV_b = 0$.

① $MV_a = M(c_1 a_1 + \cdots + c_m a_m) = MAC = r_a$.

let, $M = A^T$, $A^T A C = c_1 a_1^T a_1 + c_2 a_2^T a_2 + \cdots + c_m a_m^T a_m = C = r_a$

$(\because a_i^T a_i = 1)$

and $C = r_a \in \mathbb{R}^d$ //

② $MV_b = M(d_1 b_1 + \cdots + d_p b_p) = 0$ $(\because ②)$

let, $M = A^T$, $MV_b = d_1 a_1^T b_1 + \cdots + d_p a_p^T b_p = 0$ $(\because a_i^T b_j = 0$ for $i, j)$

∴ Choose $M = A^T$ then , satisfy ① and ②.

∴ $MS = r_a$ //.

(b) - ii)

Let $k_a^T q = k_b^T q = C \gg 0$

then, $\alpha_a = \alpha_b = \dfrac{\exp(C)}{\sum_{i=1}^{n} \exp(k_i^T q)} = \dfrac{\exp(C)}{n-2 + 2\exp(C)}$, and $\lim_{C \to \infty} \dfrac{\exp(C)}{n-2+2\exp(C)} = \dfrac{1}{2}$.

∴ $q = C(k_a + k_b)$ where $C \gg 0$ //

(C) - i)

α is vanishingly small so diagonal of Σ matrix is also vanishingly small.

$\therefore$ $k_i \sim N(\mu_i, \Sigma_w) \Rightarrow k_i \approx \mu_i$.

$k_a \sim N(\mu_a, \Sigma_a) \Rightarrow \underline{k_a \approx \mu_a}$, $k_b \sim N(\mu_b, \Sigma_b) \Rightarrow \underline{k_b \approx \mu_b}$

and $\mu_i^T \mu_j = 0$ if $i \neq j$ $\therefore$ $q = C(\mu_a + \mu_b)$ where $C >> 0$.

then output $c \approx \frac{1}{2}(r_a + r_b)$ //

(C) - ii)

~~let, $k_a \in [0.5\mu_a, 1.5\mu_a]$ and $k_i = \mu_i$ $(i \neq a)$.~~

~~Then, $k_a = f\mu_a$ where $f \sim N(1, 0.5)$~~

let, $k_a = f\mu_a$ where $f \sim N(1, 0.5)$. $k_i = \mu_i$ $(i \neq a)$.

$q = C(\mu_a + \mu_b)$,

$k_a^T q = f\mu_a^T \cdot C \cdot (\mu_a + \mu_b) = f \cdot C$ $(\because \mu_a^T \mu_a = 1)$

$k_b^T q = \mu_b^T (\mu_a + \mu_b) = C$ $(\because \mu_b^T \mu_b = 1)$ where $C >> 0$.

$k_i^T q = 0$ $(i \neq a$ and $i \neq b)$

$$\alpha_a = \frac{\exp(fC)}{\exp(C) + \exp(fC)}$$

$$\alpha_b = \frac{\exp(C)}{\exp(C) + \exp(fC)}.$$

① $f \to 0.5$: $\alpha_a \approx 0$, $\alpha_b \approx 1$

$\therefore$ $c = r_b$

② $f \to 1.5$: $\alpha_a \approx 1$, $\alpha_b \approx 0$

$\therefore$ $c = r_a$.

In (a)-i, we know $c = \frac{1}{2}(r_a + r_b)$.

But now, c is fluctating between r_a and r_b //.

So, output c will vary more significantly due to the fluctuating magnitude of k_a.

(d) – 9)

let, $q_1 = \mu_a$. then, $k_a^T q \approx \mu_a^T \mu_a \approx 0$ (if $i \neq a$).

$\therefore \quad C_1 = \sum_{i=1}^{n} v_i a_i = v_a$

Like same manner, let $q_2 = \mu_b$ then $C_2 = v_b$

$\therefore \quad C = \frac{1}{2}(C_1 + C_2) = \frac{1}{2}(v_a + v_b)$ //

(d) – 10)

$q_1 = \mu_a$, $q_2 = \mu_b$.

C_1, C_2 are derived from different query vector so ~~there~~ their variances are independent.

C_1 is fluctuating due to large variances in magnitude.

But C_2 will remain stable ($\Leftrightarrow C_2 \approx v_b$)

And averaging C_1 and C_2 will reduce variance of final output.

This means, multi-headed attention is much more stable for handling large variance than single-headed attention. Because as we seen in (c), single-headed attention is sensitive to magnitude fluctuations in individual keys. //