#Attention & Transformer

1. Attention exploration

query: $q \in \mathbb{R}^d$ value: $\{v_1, \dots, v_n\}$, $v_i \in \mathbb{R}^d$ key: $\{k_1, \dots, k_n\}$ $k_i \in \mathbb{R}^d$

$$\Rightarrow \quad c = \sum_{i=1}^{n} \alpha_i v_i$$

$$\alpha_i = \frac{\exp(k_i^T q)}{\sum_{j=1}^{n} \exp(k_j^T q)} \quad \rightarrow \text{attention weight}$$

(a)

i. ① $0 \le \alpha_i \le 1$ ② $\sum_i \alpha_i = 1$ Then α could be interpreted as a distribution.

ii. $k_i^T q \ll k_j^T q$ or $k_i^T q = -\infty$ $(i \neq j)$

iii. under the condition ii, $c = v_j$

iv. The attention weight for the index $i \neq j$ is 0. Also, the similarity between q and k_j is very large then it would be a copy of v_j.

(b)

i. $M \rightarrow$ extract v_a from $s = v_a + v_b$; $Ms = v_a$ $(M \in \mathbb{R}^{d \times d})$

$$\begin{cases} v_a = c_1 a_1 + c_2 a_2 + \cdots + c_m a_m = Ac & (A \in \mathbb{R}^{d \times m} \quad c \in \mathbb{R}^d) \\ v_b = d_1 b_1 + d_2 b_2 + \cdots + d_p b_p = Bd & (B \in \mathbb{R}^{d \times p} \quad d \in \mathbb{R}^d) \end{cases}$$

$$Ms = Mv_a + Mv_b = MAc + MBd = Ac$$

if $M = A^T \Rightarrow A^T Ac + A^T Bd = Ic = c$

then $M = AA^T$

ii.

$$K = \begin{bmatrix} | & & | \\ k_1 & \cdots\cdots & k_n \\ | & & | \end{bmatrix} \qquad q = [- \; q \; -]$$

$k_a^T q = k_b^T q >> k_i^T q$

$\Rightarrow q = (k_a + k_b) \cdot \beta \quad (\beta >> 0)$

$\Rightarrow k_a^T q = \beta = k_b^T q \quad k_i^T q < 0$

$$\alpha_b = \alpha_a = \frac{\exp(\beta)}{(n-2) + \exp(\beta) \cdot 2} \approx \frac{1}{2}$$

(c) keys: $\{k_1, \dots, k_n\}$ $k_i \sim \mathcal{N}(\mu_i, \Sigma_i)$, $\mu_i \in \mathbb{R}^d$, $\mu_i^T \cdot \mu_j = 0$ $(i \neq j)$, $\|\mu_i\| = 1$

i. $\Sigma_i = \alpha I \quad \forall i \in \{1, \dots, n\}$

$k_a^T q = k_b^T q >> k_i^T q \quad (i \neq a, b)$

$q = (\mu_a + \mu_b) \cdot \beta$, Because $k_a = (\mu_a + \Sigma_a x) = \mu_a + \alpha \cdot I x_a \approx \mu_a$

$(\beta \gg 0)$ $k_b \approx \mu_b + \alpha I x_b \approx \mu_b$

$k_a^T q = \beta \cdot \mu_b \quad k_b^T q = \beta \mu_a \quad k_i^T q = 0$

note we have $\|\mu_i\| = 1$, $k_a^T q \approx k_b^T q \Rightarrow \alpha_a \approx \alpha_b \approx \frac{1}{2}$

ii. Under the high varience condition, $k_a^T q \approx k_b^T q$ doesn't hold true.

$$c = \alpha_a v_a + \alpha_b v_b + \sum_{i \neq a,b}^{1} \epsilon_i v_i .$$

where $\alpha_a + \alpha_b \approx 1$, ϵ_i is a small number (ideal situation)

if $k_a^T q = \beta \quad k_b^T q = \gamma \beta$

$$\Rightarrow \begin{cases} \alpha_a = \dfrac{\exp(\beta)}{\exp(\beta) + \exp(\gamma\beta)} = \dfrac{1}{1 + \exp((\gamma - 1)\beta)} & ① \\ \alpha_b = \dfrac{\exp(\gamma\beta)}{\exp(\gamma\beta) + \exp(\beta)} = \dfrac{1}{1 + \exp((1 - \gamma)\beta)} & ② \end{cases} \quad (\beta >> 0)$$

if β has a high varience (e.g. $0.1 \sim 2$), $\alpha_a, \alpha_b \approx 1.0$ or 0.1

(d) i.

$q_1 = k_a \quad q_2 = k_b \xrightarrow{\text{then}} q_1^T k_a = \|k_a\|_2^2 = 1 \qquad q_2^T k_b = \|k_b\|_2^2 = 1$

$c_1 = 1 \cdot v_a = v_a \qquad c_2 = 1 \cdot v_b = v_b$

ii. assume $\quad k_a^T q_1 = \beta \|k_a\|_2^2 = \beta \qquad \alpha_a = \dfrac{\beta}{\beta + 0 \cdots + 0} = 1$

$k_b^T q_2 = \|k_b\|_2^2 = 1 \qquad \alpha_b = \dfrac{1}{1 + 0 \cdots - 0} = 1$

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