Date: تاریخ:

Subject: a6 - 99521361 موضوع:

1) a) i) there are n α scores which represent the probabilities associated with each element in the sequence. these scores are normalized so they sum up to 1.

ii) $k_J \gg k_i \; \{i \in \{1, \ldots, n\}, i \neq J\}$

so the dot preduct of k_J and query will be large that means the softmax will put most of the probability on this.

iii) $\alpha_J \gg \alpha_i \rightarrow c = \sum_{i=1}^{n} v_i \alpha_i \sim v_J \alpha_J \sim v_J$

iv) if one of the key is similar to given query, the attention weight will be its value. so the output sense we copied the keys value

b) i)

$$\left.\begin{array}{l} v_a = c_1 a_1 + c_2 a_2 + \dots + c_m a_m = Ac \\ v_b = d_1 b_1 + d_2 b_2 + \dots + d_p b_p = Bd \\ Mv_a + Mv_b = v_a \end{array}\right\} MAc + MBd = v_a$$

we want $Mv_a = v_a$ and $Mv_b = 0$

$a_j^T b_k = 0$, $a_j^T a_i = 0 \to i \neq j$, $a_j^T a_i = 1 \to i = j$

if $M = A^T \to A^TAc + A^TBd = Ic + 0 = c$

and we know v_a is a collection of constancts c in terms of R^d

so $\rightarrow$ $M = A^T$

ii) $k_a^T q \sim k_b^T q \gg k_i^T q$ $(i \neq a, b)$

so the probability mass will be big on α_a and α_b

$q = \beta(k_a + k_b) \rightarrow \beta = k_a^T q = k_b^T q$

$(\beta \gg 0)$

$$\alpha_a = \alpha_b = \frac{\exp(\beta)}{n-2+2\exp(\beta)} \sim \frac{\exp(\beta)}{2\exp(\beta)} = \frac{1}{2} \quad (\beta \gg 0)$$

$$c = \sum_{i=1}^{n} v_i \alpha_i = 0 + \frac{1}{2} v_a + \frac{1}{2} v_b$$

c) i) variances are vanishingly small so $k_i \sim \mu_i$.
thus we can write equations like part b
and we have: $q = \beta(\mu_a + \mu_b)$ $(\beta \gg 0)$

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

$$\Rightarrow k_a \in [0{,}5\mu_a , 1{,}5\mu_a]$$

$$k_a \sim \gamma \mu_a \qquad (\gamma \sim N(\mu_a, \alpha I(\mu_a \mu_a^T)) \rightarrow \gamma \sim N(1, 1/2)$$

$$k_i \sim \mu_i \quad (i \neq a)$$

$$k_a^T q \sim \gamma \mu_a^T \beta(\mu_a + \mu_b) \qquad (\beta \gg 0)$$

$$k_b^T q \sim \mu_b^T \beta(\mu_a + \mu_b) \qquad (\beta \gg 0)$$

$$k_i^T q \sim \mu_i^T \beta(\mu_a + \mu_b) = 0 \qquad (\beta \gg 0)$$

$$\alpha_a \sim \frac{\exp(\gamma\beta)}{\exp(\gamma\beta) + \exp(\beta)} \sim \frac{1}{1 + \exp(\beta(1-\gamma))}$$

$$\alpha_b \sim \frac{\exp(\beta)}{\exp(\gamma\beta) + \exp(\beta)} \sim \frac{1}{1 + \exp(\beta(\gamma - 1))}$$

$$\rightarrow \begin{array}{l} \gamma = 0{,}5 \rightarrow \alpha_a \sim 0 \,, \alpha_b \sim 1 \\ \gamma = 1{,}5 \rightarrow \alpha_a \sim 1 \,, \alpha_b = 0 \end{array}$$

So $c = \alpha_a v_a + \alpha_b v_b$ oscillates between v_a and v_b

d) i) $q_1 = q_2 = \beta(v_a + v_b)$

$$\to c = \tfrac{1}{2}(c_1 + c_2) = \tfrac{1}{2}\left(\tfrac{1}{2}(v_a + v_b) + \tfrac{1}{2}(v_a + v_b)\right)$$

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

ii) As $q_1 = q_2 = \beta(v_a + v_b)$ so like part c(ii),

$$c = \frac{1}{1+\exp(\beta(1-\gamma))} v_a + \frac{1}{1+\exp(\beta(\gamma-1))} v_b .$$

$$\gamma \to 1 \Rightarrow c = \tfrac{1}{2} v_a + \tfrac{1}{2} v_b = \tfrac{1}{2}(v_a + v_b)$$

2) d) correct: 100 out of 500.0: 2.0%.
london: Correct: 25.0 out of 500.0: 5.0%.

f) Correct: 72.0 out of 500.0: 14,39%

g) i) Conect: 44.0out of 500.0 : 8,79%

ii) In the perceiver approach, the attention layer has important rule in which it reduces the complexity to $O(m \times d)$ (d is input dimensionality)
(cross-attention)
In Latent transformer blocks the complexity reduces to $O(m^2)$ (self-attention)
So, the complexity of multi-headed is $O(l^2 d + ld)$.
and the perceiver complexity is $O(dm + Lm^2)$

3) a) because the pretrained model had more knowledge about relation of words and patterns .so, it can generalize better on new examples.

b) 1) it can cause to the spread of misinformation so this leads to confusion.

2) it can cause users to use fals information in their works or some researches .soit's a legal impact.

c) the model try to find correlations and patterns based on other information of people with similar names. so, the prediction may be false and it cause concern as mentioned in 3b.