NLP

Homework 2

CSE4057/11687

2019040591

박진수

Self-Attention and Transformers

(a) Copying in attention

i. By definition, to be a categorical probability distribution, the following requirements must be satisfied.

If there are $K>0$ number of categories, given $p_1, \cdots, p_K$ as event probabilities, (1) $p_i \geq 0$, (2) $\sum p_i = 1$ must be satisfied.

With passing through the softmax function

$$\alpha_i = \frac{\exp(k_i^T q)}{\sum_{j=1}^{n} \exp(k_j^T q)},$$

they sum up to 1, and each α_i's have values $0 \leq \alpha_i \leq 1$.

ii. In order to make α_j close to 1 (almost all of its weight on α_j), the following condition should be satisfied

$k_j^T q \gg k_i^T q \quad \forall i \neq j$. In other words, the query vector q should be highly aligned with a specific key vector k_j.

(that particular one)

iii. $C = \sum_{i=1}^{n} v_i \alpha_i = \sum_{i=1}^{n} v_i \cdot \dfrac{\exp(k_i^T q)}{\sum_{j=1}^{n} \exp(k_j^T q)}$

Under the condition in (ii), most of the weights is on k_j, so the output C would be primarily determined by the corresponding value vector v_j compared to the other terms. And since the attention weight approximates to 1, C will approximate to v_j.

iv. If $k_j^T q$ is distinctly large compared to other $k_i^T q$ $\forall i \neq j$, which means they are highly aligned, the output C will approximate to v_j. And it resembles copying v_j.

(b) An average of two.

i. There exists some $c_1, c_2, \cdots, c_m$ such that $v_a = \overset{\text{scalar}}{c_1}\overset{\text{vector}}{a_1} + c_2 a_2 + \cdots + c_m a_m$

$= \overset{\text{Matrix}}{A}\overset{\text{vector}}{C}$

Also, there exists some $d_1, d_2, \cdots, d_p$ such that $v_b = d_1 b_1 + d_2 b_2 + \cdots + d_p b_p = Bd$

Since $MS = v_a$, $Mv_a + Mv_b = v_a$

$$\therefore \underbrace{Mv_a = v_a}_{(1)}, \quad \underbrace{Mv_b = 0}_{(2)} \quad (\because A^T B = 0)$$

$$A^TA = m\left\{\overbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}}^{d}\right. \; d\left\{\overbrace{\begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix}}^{m}\right. = \begin{bmatrix} a_1^T a & a_1^T a_2 & \cdots & a_1^T a_m \\ a_2^T a_1 & & & \vdots \\ \vdots & & \ddots & \\ a_m^T a_1 & \cdots & & a_m^T a_m \end{bmatrix}$$

$a_i^T a_i = 1$, $a_i^T a_j = 0 \quad \forall i \neq j$

Since all basis vectors have norm 1 and are orthogonal to each other.

$\therefore A^T A = I_{m \times m}$

For (1), $M v_a = v_a$

$MAc = v_a = Ac \rightarrow$ If $M = A^T A$, $\underbrace{MAc = AA^T Ac}$ $= Ac$

For (2), $MBd = 0$

$AA^T Bd = 0$

$\therefore A^T B = 0$

$\therefore M = AA^T$

ii. To satisfy $c \approx \frac{1}{2}(v_a + v_b)$,

$k_a^T q \approx k_b^T q \gg k_i^T q \quad \forall i$ where $i \neq a$, $i \neq b$

Thus, q can be represented as a vector containing k_a and k_b.

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

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

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

$k_i^T q = k_i^T (k_a + k_b)\beta = 0$

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

$(\beta >> 0)$

(c) Drawbacks of single headed attention

i. The elements of the covariance matrix will be vanishingly small, due to α. So k_i's will be very close to the means $\mu_i \in \mathbb{R}^d$.

$\therefore$ k_i can be represented as $k_i = \mu_i + \varepsilon_i$, where $\varepsilon_i \approx 0$.

Since the means μ_i are all perpendicular, $\mu_i^T \mu_j = 0$ if $i \neq j$, and unit norm $\|\mu_i\| = 1$, we can design a query q by taking the idea from the previous problem (b).

$\therefore$ $q = t(\mu_a + \mu_b)$, $t >> 0$

ii. For keys k_i where $i \neq a$, each vector is tightly clustered around its mean due to the small covariance $\Sigma_i = \alpha I$.

However, the key k_a has a covariance structure $\Sigma_a = \alpha I + \frac{1}{2}(\mu_a \mu_a^T)$.

It implies that k_a has significant variance in the direction of μ_a and negligible variance in orthogonal directions (due to the small αI term). Thus, k_a retains the direction of μ_a but its magnitude varies substantially across samples.

$\therefore$ The attention weight assigned to it will fluctuate significantly between samples (much smaller or larger than other keys)

This causes the output vector c to vary greatly across samples, as the attention may sometimes favor V_a, and other times not.

(d) Benefits of multi-headed attention.

i. According to the previous problem (c), the query q is expressed as

$q = t(\mu_a + \mu_b)$, $t \gg 0$, such that $c \approx \frac{1}{2}(v_a + v_b)$.

Since the final output of the multi-headed attention is their average,

$\frac{1}{2}(c_1 + c_2)$, we can set $c_1 = c_2 = \frac{1}{2}(v_a + v_b)$, $q_1 = q_2 = t(\mu_a + \mu_b)$

$\therefore\ c \approx \frac{1}{2}(c_1 + c_2) = \frac{1}{4}(v_a + v_b) + \frac{1}{4}(v_a + v_b) = \frac{1}{2}(v_a + v_b)$

ii.

When using multi-headed attention with two query vectors q_1 and q_2, each head produces its own context vector c_1 and c_2 based on the same set of keys and values. Due to the unique covariance of key k_a, as described in problem (c), the attention weights, and thus c_1 and c_2, may individually have high variance across samples. Since q_1 and q_2 are different, the fluctuations in c_1 and c_2 are not perfectly correlated. However, if we average as $c = \frac{1}{2}(c_1 + c_2)$, the variance tend to offset each other, making it more stable and consistent across samples.

$\therefore$ Multi-headed attention reduces the impact of fluctuations in any single head, enhancing robustness by aggregating information from multiple sources.

// Thank You