CS 224N: Assignment 5 (2021)

Zubin Gou

March 22, 2021

1 Attention exploration (21 points)

(a) Copying in attention

$$k_j^T q \gg k_i^T q, i \neq j$$

(b) An average of two

$$q = t(k_a + k_b), t \gg 0$$

(c) Drawbacks of single-headed attention

i.

$$q = t(u_a + u_b), t \gg 0$$

ii.

we got $k_a \sim \mathcal{N}\left(\mu_a, \alpha I + \frac{1}{2}(\mu_a \mu_a^\top)\right)$, and for vanishingly small α: $k_a \approx \varepsilon_a \mu_a$, $\varepsilon_a \sim \mathcal{N}(1, \frac{1}{2})$, when $q = t(u_a + u_b), t \gg 0$:

$$k_i^T q \approx 0 \text{ for } i \notin \{a, b\}$$
$$k_a^T q \approx \varepsilon_a t$$
$$k_b^T q \approx \varepsilon_b t$$

then:

$$\begin{aligned} c &\approx \frac{\exp(\varepsilon_a t)}{\exp(\varepsilon_a t) + \exp(\varepsilon_b t)} v_a + \frac{\exp(\varepsilon_b t)}{\exp(\varepsilon_a t) + \exp(\varepsilon_b t)} v_b \\ &= \frac{1}{\exp((\varepsilon_b - \varepsilon_a)t) + 1} v_a + \frac{1}{\exp((\varepsilon_a - \varepsilon_b)t) + 1} v_b \end{aligned}$$

since $\varepsilon_a, \varepsilon_b \sim \mathcal{N}(1, \frac{1}{2})$, when $\varepsilon_a > \varepsilon_b$, c will be closer to v_a, vice versa. (ie. c will be closer to those with larger $\|k\|$)

(d) Benefits of multi-headed attention

i.

$$q_a = t_1 \mu_a, t_1 \gg 0$$
$$q_b = t_2 \mu_b, t_2 \gg 0$$

ii.

$$k_a^T q = \varepsilon_a t_1$$
$$k_b^T q = \varepsilon_b t_2$$

then:

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

(e) Key-Query-Value self-attention in neural networks

i.

$$c_2 \approx u_a$$

It's impossible for c_2 to approximate u_b by adding either u_d or u_c to x_2. Say, if we add u_d, α_{21} increases, which means the weight of x_1 increases, but u_d and u_b will increase equally in c_2, that's why c_2 can never be approximated to u_b.

ii.

$$\begin{aligned} V &= u_b u_b^T \odot \frac{1}{\|u_b\|_2^2} - u_c u_c^T \odot \frac{1}{\|u_c\|_2^2} \\ &= \left(u_b u_b^T - u_c u_c^T\right) \odot \frac{1}{\beta^2} \end{aligned}$$
$$K = I$$
$$\begin{aligned} Q &= u_d u_a^T \odot \frac{1}{\|u_a\|_2^2} + u_c u_d^T \odot \frac{1}{\|u_d\|_2^2} \\ &= \left(u_d u_a^T + u_c u_d^T\right) \odot \frac{1}{\beta^2} \end{aligned}$$

Proof:

$$v_1 = u_b, v_2 = 0, v_3 = u_b - u_c$$
$$q_1 = u_c, q_2 = u_d, q_3 = 0$$
$$k_i = x_i, i \in \{1, 2, 3\}$$

so,

$$\alpha_1 \approx [0, 0, 1], \alpha_2 \approx [1, 0, 0]$$
$$c_1 \approx v_3 = u_b - u_c, c_2 \approx v_1 = u_b$$

2 Pretrained Transformer models and knowledge access (35 points)

(a)

None.

(b)

None.

(c)

None.

(d)

dev accuracy: *Correct: 7.0 out of 500.0: 1.4000000000000001%*

london baselone: *Correct: 25.0 out of 500.0: 5.0%*

(e) Define a span corruption function for pretraining.

None.

(f) Pretrain, finetune, and make predictions.

dev accuracy: *Correct: 115.0 out of 500.0: 23.0%*

(g) Research! Write and try out the synthesizer variant

i.

dev accuracy: *Correct: 72.0 out of 500.0: 14.40%*

ii.

synthesizer self-attention can't capture contextual information between different positions.

3 Considerations in pretrained knowledge (5 points)

(a)

The pretrained (vanilla) model contains extra knowledge trained by corrupted span strategy.

(b)

1. Misleading information: it made up an incorrect birth place that looks real.
2. Bias and stereotype.

(c)

It might generate the birthplace of some already known person with similar name. However, the similarity of the name has nothing to do with the birthplace in reality.