CS 224N- Assignment 5 (2021)

Attention exploration

(a) Copying in attention: Describe (in one sentence) what properties of the inputs to the attention operation would result in the output c being approximately equal to vj for some j ∈ {1, . . . , n}. Specifically, what must be true about the query q, the values {v1, . . . , vn} and/or the keys {k1, . . . , kn}?

Sol: To achieve the goal, we must have $k_j^T q >> k_i^T q,\ i \neq j$.

(b) An average of two: Give an expression for a query vector q such that the output c is approximately equal to the average of va and vb, that is, 1/2(va + vb).

Sol: $q = t(u_a + u_b),\ t >> 0.$

(c) Drawbacks of single-headed attention:

i. Design a query q in terms of the μi such that as before, c ≈ 1/2(va + vb), and provide a brief argument as to why it works.

Sol: $q = t(u_a + u_b),\ t >> 0.$

ii. Though single-headed attention is resistant to small perturbations in the keys, some types of larger perturbations may pose a bigger issue. When you sample {k1, . . . , kn} multiple times, and use the q vector that you defined in part i., what qualitatively do you expect the vector c will look like for different samples?

Sol: it can be shown that $k_a \sim N(\mu_a, \alpha I + 1/2(\mu_a \mu_a^T))$, and for vanishingly small α:
$k_a \approx \epsilon_a \mu_a, \epsilon_a \sim N(1, 1/2)$, when $q = t(u_a + u_b), t >> 0$, we have $k_i^T q \approx 0\ for\ i \notin \{a, b\}$, $k_a^T q \approx \epsilon_a t$, $k_b^T q \approx \epsilon_b t$. Thus we have $c \rightarrow v_a$.

(d) Benefits of multi-headed attention:

i.

Sol: $q_a = t_1 \mu_a, t_1 >> 0, q_b = t_2 \mu_b, t_2 >> 0.$

ii.

Sol: $c \approx \frac{1}{2}(v_a + v_b).$

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

i. $c_2 \approx u_a$. It is impossible for c_2 to approximate u_b by just adding either u_d or u_c to x_2. Since u_d and u_b will increase equally in c_2.

ii. Let

$$V = (u_b u_b^T - u_c u_c^T) \cdot \frac{1}{\beta^2}$$
$$K = I$$
$$Q = (u_d u_a^T - u_c u_d^T) \cdot \frac{1}{\beta^2}$$

It can be showed that we can have the desired results in this way.

2. Pretrained Transformer models and knowledge access

(g)

ii. What might the synthesizer self-attention not be able to do, in a single layer, what the key-query-value self-attention can do?

Sol: the synthesizer self-attention cannot capture the similarity between the embeddings, i.e. the context information.

3. Considerations in pretrained knowledge

(a) Succinctly explain why the pretrained (vanilla) model was able to achieve an accuracy of above 10%, whereas the non-pretrained model was not.

Sol: The pretrained model contains extra information from the extra corpus, which can be transfered.

(b) Come up with two reasons why this indeterminacy of model behavior may cause concern for suc applications.

Sol: 1. Bias and stereotype; 2. It can generate some results that seem to be realistic but actually are totally wrong!!

(c)

Sol: For example, it can generate the birthplace of some already known person with similar name. Whereas the similarity of name has nothing to do with the birthplace.