Practical No. 4

Jan Dziuba

April 30, 2023

1 Part 2

a)

There are many combinations of q, $\{v_1, \dots v_n\}$ and $\{k_1, \dots k_n\}$ such that $c \approx v_j$.
If we want a solution that works for any set of vectors $\{v_1, \dots v_n\}$ then the following deduction is true.
If $c \approx v_j$, then $\alpha_j \approx 1$ and $\alpha_i \approx 0$, for all $i \neq j$. Then $k_j^T q \gg k_i^T q$ for all $i \neq j$.

b)

We want a solution that works for any set of vectors $\{v_1, \dots v_n\}$.
So the following deduction is true.
If $c \approx \frac{1}{2}(v_a + v_b)$ then $\alpha_a \approx 1/2$, $\alpha_b \approx 1/2$ and $\alpha_i \approx 0$ for all $i \neq (a \vee b)$.
So $k_a^T q \approx k_b^T q$, $k_a^T q \gg k_i^T q$ for all $i \neq (a \vee b)$.
For that to happen We could have $q = (100 + n) * (k_a + k_b)$.
Then

$$\alpha_a = \alpha_b = \exp(100 + n)/(2 * \exp(100 + n) + n - 2) \approx 1/2$$

$$\alpha_i = 1/(2 * \exp(100 + n) + n - 2) \approx 0 \quad \text{for all} \quad i \neq (a \vee b)$$

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

c)

1. α is vanishingly small, so $k_i \approx \mu_i$. If we take $q = (100 + n) * (k_a + k_b)$
   then as in subsection b) $\alpha_a \approx \alpha_b \approx 1/2$, $\alpha_i \approx 0$ for all $i \neq (a \vee b)$. So $c \approx \frac{1}{2}(v_a + v_b)$

2. As in point 1. $\alpha_i \approx 0$ for all $i \neq (a \vee b)$. α is vanishingly small, so $k_a \approx \mathcal{N}(\mu_a, 1/2\mu_a)$
   So c will still be weighted average of v_a, v_b.

d)

1. Just take $q = (100 + n) * (k_a + k_b)$ as before, and we get the same result.

2. As in subsection c) c will be weighted average of v_a, v_b.

e)

1.

$$k_1^T \cdot q_2 = x_1^T \cdot x_2 = (u_d + u_b)^T \cdot u_a = 0$$
$$k_2^T \cdot q_2 = x_2^T \cdot x_2 = u_a^T \cdot u_a = \beta$$
$$k_3^T \cdot q_2 = x_3^T \cdot x_2 = (u_c + u_b)^T \cdot u_a = 0$$

$$\alpha_{21} = \frac{1}{2 + \exp(\beta)}$$
$$\alpha_{22} = \frac{\beta}{2 + \exp(\beta)}$$
$$\alpha_{23} = \frac{1}{2 + \exp(\beta)}$$

$$c_2 = \frac{v_1 + v_3 + \exp(\beta) \cdot v_2}{2 + \exp(\beta)} \approx v_2 = x_2$$

It would not be possible for c_2 to approximate μ_b by adding either μ_d or μ_c to x_2 because μ vectors are mutually orthogonal, so all terms containing μ_b will reduce, and no combination of μ_i where $i \neq b$ will approximate μ_b

2 Part 3

d)

model's accuracy on the dev set: Correct: 3.0 out of 500.0: 0.6%

London accuracy: Correct: 25.0 out of 500.0: 5.0%

f)

Correct: 108.0 out of 500.0: 21.6%

g)

1. `https://paperswithcode.com/method/linformer`

   Linformer is a linear Transformer that utilises a linear self-attention mechanism to tackle the self-attention bottleneck with Transformer models. The original scaled dot-product attention is decomposed into multiple smaller attentions through linear projections, such that the combination of these operations forms a low-rank factorization of the original attention.

2. `https://paperswithcode.com/method/bigbird`

   BigBird is a Transformer with a sparse attention mechanism that reduces the quadratic dependency of self-attention to linear in the number of tokens. BigBird is a universal approximator of sequence functions and is Turing complete, thereby preserving these properties of the quadratic, full attention model. In particular, BigBird consists of three main parts:

   - A set of global tokens attending on all parts of the sequence.
   - All tokens attending to a set of local neighboring tokens.
   - All tokens attending to a set of random tokens.

   This leads to a high performing attention mechanism scaling to much longer sequence lengths (8x).

3. As said in the point above BigBird is a universal approximator of sequence functions and is Turing complete, thereby preserving these properties of the quadratic, full attention model.

3 Part 4

a)

Pretraining allows model to create deeper representation of the data.

b)

If the model is good at lying, then user won't know when model makes an error. It would also make it harder to create mechanisms, that would catch those errors.

c)

The model may look for similar names or birthplaces. This should cause concern because model could be clueless but still confident in its answer.