Practical No. 4

Tomasz Siłkowski
ts407106@students.mimuw.edu.pl

27 kwietnia 2023

2 Attention Exploration

a) Attention input for identity

For attention to approximately copy one of the value vectors v_i, $k_i^T q$ would have to significantly greater than $k_j^T q$ for any $j \neq i$.

b) Query input for attention to return average

$$q = C(k_a + k_b), \text{where } C \gg 0$$

c) Drawbacks of single-headed attention

1. $$q = C(\mu_a + \mu_b), \text{where } C \gg 0$$

   Because covariance matrices for all keys k are identity matrix multiplied by a vanishingly small constant, we know that keys (random variables) k_a and k_b are not correlated, therefore independent. That means μ_a and μ_b are their best estimators and are also perpendicular to each other, they are the best guess for requested value.

2. To restate the problem, we have $k_a \sim \mathcal{N}(\mu_a, \alpha I + \frac{1}{2}(\mu_a \mu_a^T))$ where α is vanishingly small.
   Let $q = C(\mu_a + \mu_b)$ (same as part 1) and let k_b be the key pointing in the same direction as k_a but with different norm. This means:

$$k_i^T q \approx \begin{cases} \varepsilon_a C & \text{for } i = a, \text{ where } \varepsilon_a \sim \mathcal{N}(1, \frac{1}{2}) \\ \varepsilon_b C & \text{for } i = b, \text{ where } \varepsilon_b \sim \mathcal{N}(1, \frac{1}{2}) \\ 0 & \text{otherwise} \end{cases}$$

Because of that, when calculating c:

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

Because ε_a and ε_b come from the same distribution, it is equally likely that c will be closer to v_a as to v_b. This means c will be closer to whichever v_i has bigger $|k_i|$ for $i \in \{a, b\}$.

d) Benefits of multi-headed attention

1.

$$
\begin{aligned}
q_1 &= C_1 \mu_a, \text{where } C_1 \gg 0 \\
q_2 &= C_2 \mu_b, \text{where } C_2 \gg 0
\end{aligned}
$$

2.

$$
\begin{aligned}
k_a^T q_1 &= C_1 \varepsilon_a \\
k_b^T q_2 &= C_2 \varepsilon_b
\end{aligned}
$$

Then:

$$
\begin{aligned}
c_1 &= v_a \\
c_2 &= v_b \\
c = \frac{1}{2}(c_1 + c_2) &\approx \frac{1}{2}(v_a + v_b)
\end{aligned}
$$

Basing my judgement on these calculations, I expect output c to be close to the average of v_a and v_b.