Natalia Rutecka

NLP - report of practical 4

Exercise 2

a) $\exists_i \forall_{j \neq i} k_i^T q \gg k_j^T q$ (if $k_i^T q \geq 0$ it is sufficient for $k_j^T q$ to be just a little smaller than $k_i^T q$, since we use softmax to calculate weights)

b) We need $k_a^T q = k_b^T q \gg k_i^T$ for $i \neq a, b$. Since the dot product of perpendicular vectors is equal to 0 we can take: $q = c_1(k_a + k_b) + c_2 \sum_{i=1..n, i \neq a,b} k_i + x$, where $c_1 \gg c_2, x \in R^d, x \perp \{k_1, \ldots, k_n\}$ (in particular x can be a zero vector).

c)

1. Since for every i $\Sigma_i = \alpha I$ for vanishingly small α we can assume $\mu_i \approx k_i$. Therefore using the formula from b) we get $q = c_1(\mu_a + \mu_b) + c_2 \sum_{i=1..n, i \neq a,b} \mu_i + x$, where $c_1 \gg c_2, x \in R^d, x \perp \{\mu_1, \ldots, \mu_n\}$
2. Here we still have $\alpha_a \gg \alpha_i$ and $\alpha_b \gg \alpha_i$ for $i \neq a, b$. However the equality $\alpha_a = \alpha_b$ doesn't hold anymore. Moreover, we have $\Sigma_a \approx 0.5 \mu_a \mu_a^T$ and we can notice that since μ_a has norm 1, $\mu_a \mu_a^T$ is a projection onto the subspace spanned by μ_a. Therefore, $k_a \approx s\mu_a$, where $s \sim N(1, 0.5)$. We can notice that the mean of the normal distribution corresponds to the case $c \approx 0.5(v_a + v_b)$, however the value of c in a given iteration will be closer to the vector v with its corresponding vector k having the bigger norm.

d) 1. We can take: $q_1 = a_1 \mu_a + a_2 \sum_{i=1..n, i \neq a} \mu_i + x$ and $q_2 = a_3 \mu_b + a_4 \sum_{i=1..n, i \neq b} \mu_i + x$ where $a_1 \gg a_2, a_3 \gg a_4, x \in R^d, x \perp \{\mu_1, \ldots, \mu_n\}$ Then $c_1 \approx v_a$ and $c_2 \approx v_b$, therefore $c \approx 0.5(v_a + v_b)$

1. Despite the oscillations in the norm of k_a we expect $k_a^T q_1 \gg k_i^T q_1$ for $i \neq a$ and $k_b^T q_2 \gg k_i^T q_2$ for $i \neq b$. Therefore $c_1 \approx v_a$ and $c_2 \approx v_b$ hold and we have $c \approx 0.5(v_a + v_b)$

Exercise 3

d) The accuracy of the finetuned model on the dev set was equal to 0.4% (correct: 2/500), whereas the accuracy of always predicting "London" was equal to 5%.

f) The accuracy of the model that was pretrained and then finetuned was equal to 12% (correct: 60/500).

g) Linformer: Context mapping matrix is low-rank and most of the information it contains can be expressed using its few largest singular values. Therefore, we can use projection to obtain k-dimensional keys and values for a hyperparameter k and compute attention as usual. If $k \ll n$ we can expect a significant speedup.

BigBird: We can view the attention as a graph, with edges between tokens that attend to each other, and try to sparsify the graph in a way that keeps it well-connected. In BigBird have three types of tokens that we attend to:

- tokens that are close in text (in a window of a fixed size)
- a small number of global tokens, such that they attend to every token and every token attends to them
- a small number of random tokens.

BigBird is a universal approximator of sequence-to-sequence functions, meaning that it is able to learn any function given enough data and parameters. Similarly to regular transformers, BigBird is also Turing-complete, which means that it is able to simulate any Turing machine and perform complex computations.

Exercise 4

a) During pretraining, the model was able to learn general structure of the language due to higher versatility of the dataset and gather information about people mentioned in the text.

b) 1. Misleading information - since the language in the correct and incorrect answers is very similar and all the answers seem equally believable it might get really hard to fact-check the information provided by the model (significantly harder than the information provided by humans, where the reliable information is often different in terms of the used language).

1. Bias - the model might produce biased information that is hard to filter

c) The model might try to predict the birthplace based on the similarity of the name to the names in the training set. Such a prediction will not be accurate since in reality the name of a person has little to do with their birthplace.