Grading
No rigorous proofs are required, brief explanations will suffice.

- 30 points - 3
- 50 points - 4
- 70 points - 5

Submit your solutions to julek@straszynski.pl by the end of June 2022.

Q 1. Single-head attention

In its basic form, attention can be viewed as a function on:

- a query vector $q \in \mathbb{R}^d$,
- n key vectors $\{k_1, \ldots, k_n\} : \forall_i k_i \in \mathbb{R}^d$,
- n value vectors $\{v_1, \ldots, v_n\} : \forall_i v_i \in \mathbb{R}^d$,

returning output vector $c = \sum_{i=1}^{n} \alpha_i v_i$,
where attention weights α_i are defined as follows:

$$\alpha_i = \frac{\exp(k_i \cdot q)}{\sum_{j=1}^{n} \exp(k_j \cdot q)}$$

where $\cdot$ denotes a dot product: $v \cdot u = v^\intercal u$

(a) (5 points) Why attention weights can be interpreted as categorical probability distribution?

(b) (5 points) What are the necessary conditions on $q, k_1, \ldots, k_n$ so that there would exist α_i overwhelmigly larger than all other α_j?

(c) (10 points) Let's assume that all key vectors are orthogonal, i.e. $k_i \cdot k_j = 0$ for $i \neq j$ and have norm 1. What should query q look like for output c to be equal to $(v_i + v_j)/2$ for some $i \neq j$?

(d) (10 points) Let's assume that we're given output $c = (v_i + v_j)/2$ for some unknown i, j. What we know is that v_i lies in a subspace formed by S basis vectors $s_1, \ldots, s_S$, i.e.

$$v_1 = a_1 s_1 + \ldots a_t s_S$$

for some $a_1, \ldots a_t$. Similarly, v_j lies in a subspace formed by basis vectors $t_1, \ldots, t_T$. Let's further assume that those subspaces are orthogonal and basis vectors have norm 1. Find such matrix M such that $cM = v_i$

(e) (10 points) Let's assume now that our keys are randomly sampled: k_i from multivariate normal distribution with mean vector μ_i and covariance matrix Σ_i. Let's assume that $\forall_i \Sigma_i = \epsilon I$, where $\epsilon \ll 1$. Also, let all mean vectors be pairwise orthogonal and have norm 1. What should the query q look like for output c to be roughly equal to $(v_i + v_j)/2$ for some $i \neq j$?

(f) (10 points) Let's modify the setting in (e) a bit. Let us fix some $x \in \{1, \ldots, n\}$ and change covariance matrix to $\Sigma_x = \epsilon I + (\mu_a \cdot \mu_a)/2$. After sampling $\{k_1, \ldots k_n\}$ multiple times, what will output c look like for your q from (e)?

Q 2. Programming

(a) (30 points) Complete Lab 3 at `http://mimuw.edu.pl/~jks`