PHDOpen Assigment2

Maciej Domaradzki

1 Single-head attention

1.1

Attention weights can be interpreted as categorical probability distribution, because they are between 0 and 1, they sum up to 1 and higher weight means more important item.

1.2

$k_i \cdot q$ must be significantly greater then $k_j \cdot q$. In such a case $\exp(k_i \cdot q) >> \exp(k_j \cdot q)$, so $\alpha_i >> \alpha_j$

1.3

Let $q = t(k_i + k_j)$, where $t >> 0$. Then $k_k \cdot q = 0$ for $k \notin \{i, j\}$ and $k_k \cdot q = t$ otherwise, so for large t $\alpha_k \approx 0$ for $k \notin \{i, k\}$ and $\frac{1}{2}$ otherwise. Therefore:

$$c = \sum_{k=1}^{n} \alpha_k \cdot v_k = \sum_{k=1, k \notin \{i,j\}}^{n} 0 \cdot v_k + \alpha_i \cdot v_i + \alpha_j \cdot v_j = \frac{1}{2} \cdot v_i + \frac{1}{2} \cdot v_j = \frac{1}{2} \cdot (v_i + v_j)$$

1.4

Let $A = [s_1 s_2 \ldots s_S]$. According to this page $P = A(A^T A)^{-1} A^T$ gives an orthogonal projection to the subspace formed by vectors $(s_1, s_2, \ldots, s_S)$. Therefore $c^T \cdot P = \frac{1}{2} v_i$ (because v_i belongs to subspace formed by vectors s_k and v_j belongs to subspace formed by vectors t_k, which is orthogonal to the first one), so, if $M = 2 \cdot P = A(A^T A)^{-1} A^T$, then $c^T M = 2 \cdot c^T \cdot P = v_i$.

1.5

Let $q = t(\mu_i + \mu_j)$, where $t >> 0$. $\epsilon << 1$, so $k_k \approx \mu_k$. Therefore $k_k \cdot q \approx 0$ for $k \notin \{i, j\}$ and $k_k \cdot q \approx t$ otherwise, so for large t $\alpha_k \approx 0$ for $k \notin \{i, k\}$ and $\frac{1}{2}$ otherwise. Therefore:

$$c = \sum_{k=1}^{n} \alpha_k \cdot v_k = \sum_{k=1, k \notin \{i,j\}}^{n} 0 \cdot v_k + \alpha_i \cdot v_i + \alpha_j \cdot v_j = \frac{1}{2} \cdot v_i + \frac{1}{2} \cdot v_j = \frac{1}{2} \cdot (v_i + v_j)$$