ECE 537 Fundamentals of Speech Processing
Problem Set 9

UNIVERSITY OF ILLINOIS
Department of Electrical and Computer Engineering

Assigned: Sunday, 11/6/2022; Due: Friday, 11/11/2022
Reading: Vaswani et al., "Attention is All You Need," 2017

1. In the Transformer, positional embeddings enable a query and a key to match one another based on their relative position, rather than based on their content. The positional embedding is a d_{model}-dimensional vector whose $(2i)^{\text{th}}$ and $(2i)^{\text{st}}$ elements, at time t, are:

$$e^t_{2i} = \sin\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right)$$
$$e^t_{2i+1} = \cos\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right)$$

These are then added to the query and key prior to computing attention, thus

$$e^t = [e^t_0, e^t_1, \cdots, e^t_{d_{\text{model}}-1}]$$
$$E = \begin{bmatrix} e^0 \\ \vdots \\ e^{n-1} \end{bmatrix}$$
$$\text{head}_i = \text{Attention}\left((Q+E)W_i^Q, (K+E)W_i^K, (V+E)W_i^V\right),$$

where the matrices W_i^Q and W_k^K each are of dimension $d_{\text{model}} \times d_k$.

Consider the matrix $A = W_i^K W_i^{Q,T}$. All parts of this problem will ask you to calculate input-output relations of the Attention operation by thinking about the values of the following submatrix:

$$A_{(2i:2i+1),(2j:2j+1)} = \begin{bmatrix} A_{2i,2j} & A_{2i,2j+1} \\ A_{2i+1,2j} & A_{2i+1,2j+1} \end{bmatrix},$$

(a) (1 point) What should be the values of $A_{(2i:2i+1),(2j:2j+1)}$ so that the attention, $\alpha_{\tau,t}$, is maximized when the time alignment of the key (the time index τ of vector k^τ) precedes the time index of the query (time index t of vector q^t) by exactly T time steps ($\tau = t - T$)?

(b) (1 point) Suppose that you require a less-precise time alignment. Suppose that you want the attention to be maximized when $t - \tau \in \left[T - \frac{B}{2}, T + \frac{B}{2}\right]$, i.e., the separation between τ and t should be $T \pm \frac{B}{2}$. This can be accomplished by starting with the A matrix you computed in part (a), and then zeroing some of its elements. Which elements of A should be zeroed so that $\alpha_{\tau,t}$ is maximum for an $t - \tau \in \left[T - \frac{B}{2}, T + \frac{B}{2}\right]$?

(c) (1 point) The previous parts have considered a head that cares about relative position. In this part, instead, consider a head that only cares whether or not a query and key have the same content. Suppose, for example, that for this head, A is the identity matrix. Then, assuming that $|k^\tau| = 1$ and $|q^t| = 1$, the inner product $k^\tau A q^{t,T} = k^\tau q^{t,T}$ is maximized if $k^\tau = q^t$.

Suppose that $k^\tau = q^t$ and A is the identity matrix, but we also have to consider the contribution of the positional embeddings. Let's define $\tilde{q}$ and $\tilde{k}$ to be the position-enhanced embeddings, thus

$$\tilde{q}_{2i}^t = q_{2i}^t + e_{2i}^t$$
$$\tilde{k}_{2i}^\tau = k_{2i}^\tau + e_{2i}^\tau$$

What are the mean and variance of the inner product $\tilde{k}^\tau \tilde{q}^{t,T}$, assuming that $k^\tau = q^t$, $|k^\tau| = 1$ and $|q^t| = 1$?

Hint: what are the mean and variance of $\sin\alpha\sin\beta$ if α and β are each independent random variables uniformly distributed between 0 and 2π? What are the mean and variance of $k_{2i}^\tau e_{2i}^t$ if k_{2i}^τ is a zero-mean, unit-variance random variable independent of e_{21}^t? How many such terms are added together in order to compute the inner product $\tilde{k}^\tau \tilde{q}^{t,T}$?