CS114B (Spring 2021) Written Assignment 4
Attention Exploration*

Due May 14, 2021

Introduction

Multi-headed self-attention is the core modeling component of Transformers. In this assignment, we'll get some practice working with the self-attention equations, and motivate why multi-headed self-attention can be preferable to single-headed self-attention.

Assignment

(a) Recall that attention can be viewed as an operation on a query $q \in \mathbb{R}^d$, a set of value vectors $\{v_1, \ldots, v_n\}$, $v_i \in \mathbb{R}^d$, and a set of key vectors $\{k_1, \ldots, k_n\}$, $k_i \in \mathbb{R}^d$, specified as follows:

$$c = \sum_{i=1}^{n} v_i \alpha_i \tag{1}$$

$$\alpha_i = \frac{\exp(k_i^\top q)}{\sum_{j=1}^{n} \exp(k_j^\top q)}, \tag{2}$$

where α_i are frequently called the "attention weights", and the output $c \in \mathbb{R}^d$ is a correspondingly weighted average over the value vectors.

We'll first show that it's particularly simple for attention to "copy" a value vector to the output c. Describe (in one sentence or so) what properties of the inputs to the attention operation would result in the output c being approximately equal to v_j for some $j \in \{1, \ldots, n\}$. Specifically, what must be true about the query q, the values $\{v_1, \ldots, v_n\}$, and/or the keys $\{k_1, \ldots, k_n\}$?

*This assignment is adapted from the CS 224N course at Stanford.

(b) Consider a set of key vectors $\{k_1, \ldots, k_n\}$ where all key vectors are perpendicular, that is, $k_i \perp k_j$ for all $i \neq j$. Let $||k_i|| = 1$ for all i. Let $\{v_1, \ldots, v_n\}$ be a set of arbitrary value vectors. Let $v_a, v_b \in \{v_1, \ldots, v_n\}$ be two of the value vectors. Give an expression for a query vector q such that the output c is approximately equal to the average of v_a and v_b, that is, $\frac{1}{2}(v_a + v_b)$[1]. Note that you can reference the corresponding key vectors of v_a and v_b as k_a and k_b.

(c) In the previous part, we saw how it was *possible* for a single-headed attention to focus equally on two values. The same concept could easily be extended to any subset of values. In this question we'll see why it's not a *practical* solution. Consider a set of key vectors $\{k_1, \ldots, k_n\}$ that are now randomly sampled, $k_i \sim \mathcal{N}(\mu_i, \Sigma_i)$, where the means μ_i are known to you, but the covariances Σ_i are unknown. Further, assume that the means μ_i are all perpendicular; $\mu_i^\top \mu_j = 0$ if $i \neq j$, and unit norm, $||\mu_i|| = 1$.

i. Assume that the covariance matrices are $\Sigma_i = \alpha I$, for vanishingly small α. Design a query q in terms of the μ_i such that as before, $c \approx \frac{1}{2}(v_a + v_b)$, and provide a brief argument as to why it works.

ii. Though single-headed attention is resistant to small perturbations in the keys, some types of larger perturbations may pose a bigger issue. Specifically, in some cases, one key vector k_a may be larger or smaller in norm than the others, while still pointing in the same direction as μ_a. As an example, let us consider a covariance for item a as $\Sigma_a = \alpha I + \frac{1}{2}(\mu_a \mu_a^\top)$ for vanishingly small α (as shown in figure 1). Further, let $\Sigma_i = \alpha I$ for all $i \neq a$.

When you sample $\{k_1, \ldots, k_n\}$ multiple times, and use the q vector that you defined in part i, what qualitatively do you expect the vector c will look like for different samples?

[1] Hint: while the softmax function will never *exactly* average the two vectors, you can get close by using a large scalar multiple in the expression.

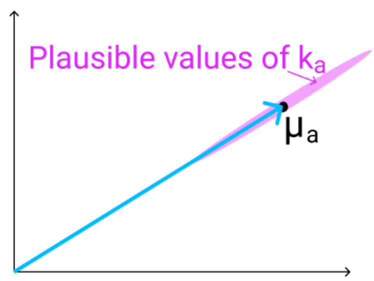

Figure 1: The vector μ_a (shown here in 2D as an example), with the range of possible values of k_a shown in purple. As mentioned previously, k_a points in roughly the same direction as μ_a, but may have larger or smaller magnitude.

(d) Now we'll see some of the power of multi-headed attention. We'll consider a simple version of multi-headed attention which is identical to single-headed self-attention as we've presented it in this homework, except two query vectors (q_1 and q_2) are defined, which leads to a pair of vectors (c_1 and c_2), each the output of single-headed attention given its respective query vector. The final output of the multi-headed attention is their average, $\frac{1}{2}(c_1 + c_2)$. As in question (c), consider a set of key vectors $\{k_1, \ldots, k_n\}$ that are randomly sampled, $k_i \sim \mathcal{N}(\mu_i, \Sigma_i)$, where the means μ_i are known to you, but the covariances Σ_i are unknown. Also as before, assume that the means μ_i are mutually orthogonal; $\mu_i^\top \mu_j = 0$ if $i \neq j$, and unit norm, $||\mu_i|| = 1$.

i. Assume that the covariance matrices are $\Sigma_i = \alpha I$, for vanishingly small α. Design q_1 and q_2 such that c is approximately equal to $\frac{1}{2}(v_a + v_b)$.

ii. Assume that the covariance matrices are $\Sigma_a = \alpha I + \frac{1}{2}(\mu_a \mu_a^\top)$ for vanishingly small α, and $\Sigma_i = \alpha I$ for all $i \neq a$. Take the query vectors q_1 and q_2 that you designed in part i. What, qualitatively, do you expect the output c to look like across different samples of the key vectors? Please briefly explain why. You can ignore cases in which $q_i^\top k_a < 0$.

(e) So far, we've discussed attention as a function on a set of key vectors, a set of value vectors, and a query vector. In Transformers, we perform *self-attention*, which roughly means that we draw the keys, values, and queries from the same data. More precisely, let $\{x_1, \ldots, x_n\}$ be a sequence of vectors in $\mathbb{R}^d$. Think of each x_i as representing word i in a sentence. One form of self-attention defines keys, queries, and values as follows. Let $V, K, Q \in \mathbb{R}^{d \times d}$ be parameter matrices. Then

$$v_i = Vx_i, i \in \{1, \ldots, n\} \tag{3}$$

$$k_i = Kx_i, i \in \{1, \ldots, n\} \tag{4}$$

$$q_i = Qx_i, i \in \{1, \ldots, n\}. \tag{5}$$

Then we get a context vector for each input i; we have $c_i = \Sigma_{j=1}^{n} \alpha_{ij} v_j$, where α_{ij} is defined as $\alpha_{ij} = \frac{\exp(k_j^\top q_i)}{\sum_{\ell=1}^{n} \exp(k_\ell^\top q_i)}$. Note that this is single-headed self-attention.

In this question, we'll show how key-value-query attention like this allows the network to use different aspects of the input vectors x_i in how it defines keys, queries, and values. Intuitively, this allows networks to choose different aspects of x_i to be the "content" (value vector) versus what it uses to determine "where to look" for content (keys and queries).

i. First, consider if we didn't have key-query-value attention. For keys, queries, and values we'll just use x_i; that is, $v_i = q_i = k_i = x_i$. We'll consider a specific set of x_i. In particular, let u_a, u_b, u_c, u_d be mutually orthogonal vectors in $\mathbb{R}^d$, each with equal norm $||u_a|| = ||u_b|| = ||u_c|| = ||u_d|| = \beta$, where β is very large. Now, let our x_i be:

$$x_1 = u_d + u_b \tag{6}$$

$$x_2 = u_a \tag{7}$$

$$x_3 = u_c + u_b. \tag{8}$$

If we perform self-attention with these vectors, what vector does c_2 approximate? Would it be possible for c_2 to approximate u_b by adding either u_d or u_c to x_2? Explain why or why not (either math or English is fine).

ii. Now consider using key-query-value attention as we've defined it originally. Using the same definitions of x_1, x_2, and x_3 as in part i, specify matrices K, Q, V such that $c_2 \approx u_b$, and $c_1 \approx u_b - u_c$. There are many solutions to this problem, so it will be easier for you (and the graders), if you first find V such that $v_1 = u_b$ and $v_3 = u_b - u_c$, then work on Q and K. Some outer product properties may be helpful (as summarized in this footnote)[2].

Submission Instructions

Please submit your solutions (in PDF format) to LATTE.

Important: There are two assignment activity modules on LATTE, marked "HW4 (Graduating This Semester)" and "HW4 (Not Graduating This Semester)". Both contain the same assignment with the same due date, and you only have to submit your assignment once. If you are graduating this semester, please be sure to use the "HW4 (Graduating This Semester)" submission box, so we know to grade your assignment in time for you to graduate!

[2]For orthogonal vectors $u, v, w \in \mathbb{R}^d$, the outer product $uv^\top$ is a matrix in $\mathbb{R}^{d\times d}$, and $(uv^\top)v = u(v^\top v) = u||v||_2^2$, and $(uv^\top)w = u(v^\top w) = u * 0$ (The last equality is because v and w are orthogonal).