TTIC 31230 Fundamentals of Deep Learning

Transformer Problems.

Problem 1. A self-attention layer in the transformer takes a sequence of vectors $h_{\text{in}}[T, J]$ and computes a sequence of vectors $h_{\text{out}}[T, J]$ using the following equations where k ranges over "heads". Heads are intended to allow for different relationship between words such as "coreference" or "subject of" for a verb. But the actual meaning emerges during training and is typically difficult or impossible to interpret. In the following equations we typically hve $U < J$ and we require $I = J/K$ so that the concatenation of K vectors of dimension I is a vector of dimension J.

$$\begin{aligned}
\text{Query}[k, t, U] &= W^Q[k, U, J] h_{\text{in}}[t, J] \\
\text{Key}[k, t, U] &= W^K[k, U, J] h_{\text{in}}[t, J] \\
\alpha[k, t_1, t_2] &= \underset{t_2}{\text{softmax}}\ \text{Query}[k, t_1, U]\text{Key}[k, t_2, U] \\
\text{Value}[k, t, I] &= W^V[k, I, J] h_{\text{in}}[t, J] \\
\text{Out}[k, t, I] &= \sum_{t'} \alpha[k, t, t'] \text{Value}[k, t', I] \\
h_{\text{out}}[t, J] &= \text{Out}[1, t, I]; \cdots ; \text{Out}[K, t, I]
\end{aligned}$$

A summation over N terms can be done in parallel in $O(\log N)$ time.

(a) For a given head k and position t_1 what is the parallel running time of the above softmax operation, as a function of T and U where we first compute the scores to be used in the softmax and then compute the normalizing constant Z.

Solution: The scores can be computed in parallel in $\ln U$ time and then Z can be computed in $\ln T$ time. We then get $O(\ln T + \ln U)$. In practice the inner product used in computing the scores would be done in $O(U)$ time giving $O(U + \ln T)$.

(b) What is the order of running time of the self-attention layer as a function of T, J and K (we have I and U are both less than J.)

Solution: $O(\ln T + \ln J)$. In practice the inner products would be done serially which would give $O(J + \ln T)$.

Problem 2. Just as CNNs can be done in two dimensions for vision and in one dimension for language, the Transformer can be done in two dimensions for vision — the so-called spatial transformer.

(a) Rewrite the equations from problem 1 so that the time index t is replaced by spatial dimensions x and y.

Solution:

$$
\begin{aligned}
\text{Query}[k, x, y, U] &= W^Q[k, U, J] h_{\text{in}}[x, y, J] \\
\text{Key}[k, x, y, U] &= W^K[k, U, J] h_{\text{in}}[x, y, J] \\
\alpha[k, x_1, y_1, x_2, y_2] &= \underset{x_2, y_2}{\text{softmax}}\ \text{Query}[k, x_1, y_1, U] \text{Key}[k, x_2, y_2, U] \\
\text{Value}[k, x, y, I] &= W^V[k, I, J] h_{\text{in}}[x, y, J] \\
\text{Out}[k, x, y, I] &= \sum_{x', y'} \alpha[k, x, y, x', y'] \text{Value}[k, x', y', I] \\
h_{\text{out}}[x, y, J] &= \text{Out}[1, x, y, I]; \cdots ; \text{Out}[K, x, y, I]
\end{aligned}
$$

(b) Assuming that summations take logarithmic parallel time, give the parallel order of run time for the spatial self-attention layer as a function of X, Y, J and K (we have that I and U are both less than J).

Solution: $O(\ln XY + \ln J)$