TTIC 31230 Fundamentals of Deep Learning, 2020

Problems For Language Modeling, Translation and Attention.

In these problems, as in the lecture notes, capital letter indeces are used to indicate subtensors (slices) so that, for example, $M[I, J]$ denotes a matrix while $M[i, j]$ denotes one element of the matrix, $M[i, J]$ denotes the ith row, and $M[I, j]$ denotes the jth collumn.

Throughout these problems we assume a word embedding matrix $e[W, I]$ where $e[w, I]$ is the word vector for word w. We then have that $e[w, I]^\top h[t, I]$ is the inner product of the word vector $w[w, I]$ and the hidden state vector $h[t, I]$.

We will adopt the convention, similar to true Einstein notation, that repeated capital indeces in a product of tensors are implicitly summed. We can then write the inner product $e[w, I]^\top h[t, I]$ simply as $e[w, I]h[t, I]$ without the need for the (meaningless) transpose operation.

Problem 1. Consider an autoregressive RNN neural language model with $P_\Phi(w_{t+1}|w_1, \ldots, w_t)$ defined by

$$P_\Phi(w_t|w_1, \ldots, w_{t-1}) = \operatorname*{softmax}_{w_{t+1}} \; e[w_t, I]h[t-1, I]$$

Here $e[w, I]$ is the word vector for word w, $h[t, I]$ is the hidden state vector at time t of a left-to-right RNN, and as described above $e[w, I]h[t, I]$ is the inner prodcut of these two vectors where we have assumed that they have the same dimension. For the first word w_1 we have an externally provided initial hidden state $h[0, I]$ and $w_1, \ldots, w_0$ denotes the empty string. We train the model on the full loss

$$\begin{aligned}
\Phi^* &= \operatorname*{argmin}_\Phi E_{w_1,\ldots,w_T \sim \text{Train}} \; -\ln P_\Phi(w_1, \ldots, w_T) \\
&= \operatorname*{argmin}_\Phi E_{w_1,\ldots,w_T \sim \text{Train}} \; \sum_{t=1}^{T} -\ln P_\Phi(w_t|w_1, \ldots, w_{t-1})
\end{aligned}$$

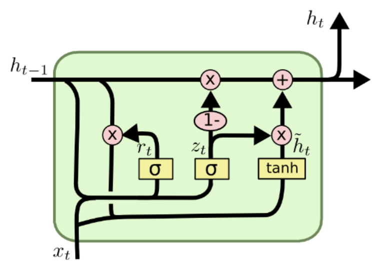
What is the order of run time as a function of sentence length T for the back-propagation for this model run on a sentence $w_1, \ldots, w_T$? Explain your answer.

Solution: The backprogation takes $O(T)$ time (not $O(T^2)$). The model consists of $O(T)$ objects each of which performs a single forward operation and a single backward operation. As the backpropagation procedes more of the loss terms in the sum over t get incorporated.

Problem. A UGRNN cell for computing $h[b, t, J]$ from $h[b, t-1, J]$ and $x[b, t, J]$ can be written as

$$
\begin{aligned}
G[b,t,j] &= \sigma\left(W^{h,G}[j,I]h[b,t-1,I] + W^{x,G}[j,K]x[b,t,K] - B^G[j]\right) \\
R[b,t,j] &= \tanh\left(W^{h,R}[j,I]h[b,t-1,I] + W^{x,R}[j,K]x[b,t,K] - B^R[j]\right) \\
h[b,t,j] &= G[b,t,j]h[b,t-1,j] + (1 - G[b,t,j])R[b,t,j]
\end{aligned}
$$

Modify the above equations so that they correspond to the following diagram for a Gated Recurent Unit (GRU).



[Christopher Olah]

Solution:

$$
\begin{aligned}
G_1[b,t,j] &= \sigma\left(W^{h,G_1}[j,I]h[b,t-1,I] + W^{x,G_1}[j,K]x[b,t,K] - B^G[j]\right) \\
\tilde{h}[b,t,j] &= G_1[b,t,j]x[b,t,j] \\
G_2[b,t,j] &= \sigma\left(W^{h,G_2}[j,I]h[b,t-1,I] + W^{x,G_2}[j,K]x[b,t,K] - B^G[j]\right) \\
R[b,t,j] &= \tanh\left(W^{h,R}[j,I]\tilde{h}[b,t-1,I] + W^{x,R}[j,K]x[b,t,K] - B^R[j]\right) \\
h[b,t,j] &= G_2[b,t,j]h[b,t-1,j] + (1 - G_2[b,t,j])R[b,t,j]
\end{aligned}
$$

Problem 2. This problem considers "blank language modeling" which is used in BERT. For blank language modelng we draw a sentence $w_1, \ldots, w_T$ from a corpose and blank out a word at random and ask the system to predict the blanked word. The cross-entropy loss for blank language modeling can be written as

$$
\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{w_1,\ldots,w_T \sim \text{Train}, t\sim\{1,\ldots,T\}} \quad -\ln P_\Phi(w_t | w_1, \ldots, w_{t-1}, w_{t+1}, \ldots, w_T)
$$

Consider a bidirectional RNN run on a sequence of words $w_1, \ldots, w_T$ such that for each time t we have a forward hidden state $\vec{h}[t, J]$ computed from $w_1, \ldots, w_t$ and a backward hidden state $\overleftarrow{h}[t, J]$ computed from $w_T,\ w_{T-1}, \ldots w_t$. Also assume that each word w has an associated word vector $e[w, J]$. Give a definition of $P(w_t \mid w_1, \ldots w_{t-1},\ w_{t+1}, \ldots, w_T)$ as a function of the vectors $\vec{h}[t-1, J]$ and $\overleftarrow{h}[t+1, J]$ and the word vectors $e[W, I]$. You can assume that $\vec{h}[T, J]$ and $\overleftarrow{[h]}[T, J]$ have the same shape (same dimensions) but do not make any assumptions about the dimension of the word vectors $e[W, I]$. You can assume whatever tensor parameters you want.

Solution: There are various acceptable solutions. A simple one is to assume the parameters include matrices $\vec{W}[I, J]$ and $\overleftarrow{W}[I, J]$. Using this convention and the standard convention for matrix-vector products we can then write a solution as

$$\begin{aligned} & P_\Phi(w_t | w_1, \ldots, w_{t-1}, w_{t+1}, \ldots, w_T) \\ = \quad & \operatorname*{softmax}_{w_t} \;\; e[w_t, I]\vec{W}[I, J]\vec{h}[t-1, J] \; + \; e[w_t, I]\overleftarrow{W}[I, J]\overleftarrow{h}[t+1, J] \end{aligned}$$