

## TTIC 31230 Fundamentals of Deep Learning, 2020

### Problems For Language Modeling, Translation and Attention.

In these problems, as in the lecture notes, capital letter indices 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  $i$ th row, and  $M[I, j]$  denotes the  $j$ th column.

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 indices 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. RNN parallel run time.** Consider an autoregressive RNN neural language model with  $P_\Phi(w_{t+1}|w_1, \dots, w_t)$  defined by

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

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 product 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  $w_1, \dots, w_0$  denotes the empty string. We train the model on the full loss

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

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

**Solution:** The backpropagation 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 proceeds more of the loss terms in the sum over  $t$  get incorporated.

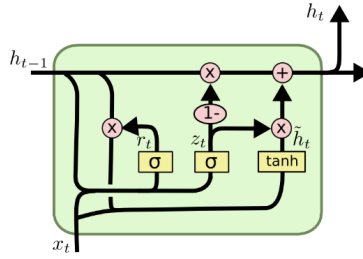
**Problem 2. Translating diagrams into equations.** A UGRNN cell for computing  $h[b, t, J]$  from  $h[b, t-1, J]$  and  $x[b, t, J]$  can be written as

$$G[b, t, j] = \sigma(W^{h,G}[j, I]h[b, t-1, I] + W^{x,G}[j, K]x[b, t, K] - B^G[j])$$

$$R[b, t, j] = \tanh(W^{h,R}[j, I]h[b, t-1, I] + W^{x,R}[j, K]x[b, t, K] - B^R[j])$$

$$h[b, t, j] = G[b, t, j]h[b, t-1, j] + (1 - G[b, t, j])R[b, t, j]$$

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



[Christopher Olah]

**Solution:**

$$G_1[b, t, j] = \sigma(W^{h,G_1}[j, I]h[b, t-1, I] + W^{x,G_1}[j, K]x[b, t, K] - B^{G_1}[j])$$

$$h'[b, t, j] = G_1[b, t, j]x[b, t, j]$$

$$G_2[b, t, j] = \sigma(W^{h,G_2}[j, I]h[b, t-1, I] + W^{x,G_2}[j, K]x[b, t, K] - B^{G_2}[j])$$

$$R[b, t, j] = \tanh(W^{h,R}[j, I]\tilde{h}[b, t-1, I] + W^{x,R}[j, K]h'[b, t, K] - B^R[j])$$

$$h[b, t, j] = (1 - G_2[b, t, j])h[b, t-1, j] + G_2[b, t, j]R[b, t, j]$$

**Problem 3. Blank Language modeling.** This problem considers “blank language modeling” which is used in BERT. For blank language modeling we draw a sentence  $w_1, \dots, w_T$  from a corpus 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

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

Consider a bidirectional RNN run on a sequence of words  $w_1, \dots, w_T$  such that for each time  $t$  we have a forward hidden state  $\vec{h}[t, J]$  computed from  $w_1, \dots, w_t$  and a backward hidden state  $\tilde{h}[t, J]$  computed from  $w_T, w_{T-1}, \dots, w_t$ . Also assume that each word  $w$  has an associated word vector  $e[w, J]$ . Give a definition of  $P(w_t | w_1, \dots, w_{t-1}, w_{t+1}, \dots, w_T)$  as a function of the vectors  $\vec{h}[t-1, J]$  and  $\tilde{h}[t+1, J]$  and the word vectors  $e[W, I]$ . You can assume that  $\vec{h}[T, J]$  and  $\tilde{h}[1, J]$  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  $\tilde{W}[I, J]$ . Using this convention and the standard convention for matrix-vector products we can then write a solution as

$$\begin{aligned} & P_{\Phi}(w_t | w_1, \dots, w_{t-1}, w_{t+1}, \dots, w_T) \\ &= \underset{w_t}{\text{softmax}} \quad e[w_t, I] \vec{W}[I, J] \vec{h}[t-1, J] + e[w_t, I] \tilde{W}[I, J] \tilde{h}[t+1, J] \end{aligned}$$

**Problem 4. Image Captioning as Translation with Attention.** In machine translation with attention the translation is generated from an autoregressive language model for the target language translation given the source language sentence. The source sentence is converted to an initial hidden vector  $h[0, J]$  for the decoding (usually the final hidden vector of a right-to-left RNN run on the input), plus the sequence  $M[T_{\text{in}}, J]$  of hidden vectors computed by the RNN on the source sentence where  $T$  is the length of the source sentence. We then define an autoregressive conditional language model

$$P_{\Phi}(w_1, \dots, w_{T_{\text{out}}} | h[0, J], M[T_{\text{in}}, J])$$

An autoregressive conditional language model with attention can be defined by

$$\begin{aligned} P(w_t | w_0, \dots, w_{t-1}) &= \underset{w_t}{\text{softmax}} \quad e[w_t, I] W^{\text{auto}}[I, J] h[t-1, J] \\ \alpha[t_{\text{in}}] &= \underset{t_{\text{in}}}{\text{softmax}} \quad h[t-1, J_1] W^{\text{key}}[J_1, J_2] M[t_{\text{in}}, J_2] \\ V[J] &= \sum_{t_{\text{in}}} \alpha[t_{\text{in}}] M[t_{\text{in}}, J] \\ h[t, J] &= \text{CELL}_{\Phi}(h[t-1, J], V[J], e[w_t, I]) \end{aligned}$$

Here CELL is some function taking (objects for) two vectors of dimensions  $J$  and one vector of dimension  $I$  and returning (an object for) a vector of dimension  $J$ .

Rewrite these equations for image captioning where instead of  $M[t_{\text{in}}, J]$  we are given an image feature tensor  $M[x, y, K]$

**Solution:**

$$P(w_t \mid w_0, \dots, w_{t-1}) = \underset{w_t}{\text{softmax}} e[w_t, I] W^{\text{auto}}[I, J] h[t-1, J]$$

$$\alpha[x, y] = \underset{x, y}{\text{softmax}} h[t-1, J] W^{\text{key}}[J, K] M[x, y, K]$$

$$V[K] = \sum_{x, y} \alpha[x, y] M[x, y, K]$$

$$h[t, J] = \text{CELL}_{\Phi}(h[t-1, J], V[K], e[w_t, I])$$

**Problem 5. Gated CNNs**

A UGRNN is defined by the following equations.

$$\tilde{R}_t[b, j] = \left( \sum_i W^{h, R}[j, i] h_{t-1}[b, i] \right) + \left( \sum_k W^{x, R}[j, k] x_t[b, k] \right) - B^R[j]$$

$$R_t[b, j] = \tanh(\tilde{R}_t[b, j])$$

$$\tilde{G}_t[b, j] = \left( \sum_i W^{h, G}[j, i] h_{t-1}[b, i] \right) + \left( \sum_k W^{x, G}[j, k] x_t[b, k] \right) - B^G[j]$$

$$G_t[b, j] = \sigma(\tilde{G}_t[b, j])$$

$$h_t[b, j] = G_t[b, j] h_{t-1}[b, j] + (1 - G_t[b, j]) R_t[b, j]$$

Modify these to form a data-dependent data-flow CNN for vision — an Update-Gate CNN (UGCNN). More specifically, give equations analogous to those for UGRNN for computing a CNN “box”  $L_{\ell+1}[b, x, y, j]$  from  $L_{\ell}[b, x, y, i]$  (stride 1) using a computed “gate box”  $G_{\ell+1}[b, x, y, j]$  and an “update box”  $R_{\ell+1}[b, x, y, j]$ .

$$\Phi = (W_{\ell+1}^{L, R}[\Delta x, \Delta y, j, j'], B_{\ell+1}^R[j], W_{\ell+1}^{L, G}[\Delta x, \Delta y, j, j'], B_{\ell+1}^G[j])$$

**Solution:**

$$R_{\ell+1}[b, x, y, j] = \tanh \left( \left( \sum_{\Delta x, \Delta y, j'} W_{\ell+1}^{L, R}[\Delta x, \Delta y, j', j] L_{\ell}[b, x + \Delta x, y + \Delta y, j'] \right) - B_{\ell+1}^R[j] \right)$$

$$G_{\ell+1}[b, x, y, j] = \sigma \left( \left( \sum_{\Delta x, \Delta y, i} W_{\ell+1}^{L, G}[\Delta x, \Delta y, i, j] L_{\ell}[b, x + \Delta x, y + \Delta y, i] \right) - B_{\ell+1}^G[j] \right)$$

$$L_{\ell+1}[b, x, y, j] = G_{\ell+1}[b, x, y, j] L_{\ell}[b, x, y, j] + (1 - G_{\ell+1}[b, x, y, j]) R_{\ell+1}[b, x, y, j]$$

**Problem 6. Language CNNs.** This problem is on CNNs for sentences. We consider a model with parameters

$$\Phi = (e[w, i], W_1[\Delta t, i, i'], B_1[i], \dots, W_L[\Delta t, i, i'], B_L[i])$$

The matrix  $e$  is the word embedding matrix where  $e[w, I]$  is the vector embedding of word  $w$ .

(a) Give an equation for the convolution layer  $L_0[b, t, i]$  as a function of the word embeddings and the input sentence  $w_1, \dots, w_T$ .

**Solution:**

$$L_0[b, t, i] = e[w[t], i]$$

(b) Give an equation for  $L_{\ell+1}[b, t, i]$  as a function of  $L_\ell[b, t, i]$  and the parameters  $W_{\ell+1}[\Delta t, i', i]$  and  $B_{\ell+1}[i]$  and where  $L_{\ell+1}$  is computed stride 2.

**Solution:**

$$L_{\ell+1}[b, t, i] = \sigma \left( \left( \sum_{\Delta t, i'} W_{\ell+1}[\Delta t, i', i] L_\ell[2t + \Delta t, i'] \right) - B_{\ell+1}[i] \right)$$

(c) Assuming all computations can be done in parallel as soon the inputs have been computed, what is the **parallel** order of run time for this convolutional model as a function of the input length  $T$  and the number of layers  $L$  (assume all parameter tensors of size  $O(1)$ ). Compare this with the parallel run time of an RNN.

**Solution:** The CNN has  $O(L)$  parallel run time while the RNN is  $O(T)$  or  $O(T + L)$  with  $L$  layers of RNN.