Neural Network and Deep Learning

Solutions to Assignment 3

Author: Shashank Gupta

Problem 1a.

There is a teacher forcing ratio used in the code. Explain how that ratio is used in the code (in English description) and why do you think it is used (read online for why).

Solution:

In the code, the teacher forcing ratio is fixed at 0.5 and compared with a randomly generated number between 0 and 1 indicating randomly training the seq-to-seq model with either teacher forcing method or the normal method (passing the decoder's guess as the next input). Teacher forcing is a concept of using ground truth target labels to correct the mistakes of model during training and this helps the model to converge to the minima at a faster rate but when the trained network is exploited it may show instability so a general rule on when to use teacher forcing cannot be given.

Problem 2a.

What are the dimensions of the matrices K and V, given Q is dimension $N \times d_k$?

Solution:

The attention matrix uses Q and K to produce the attention logit matrix - L of dimension $N \times N$. Which is normalized to obtain attention weights using the Softmax function.

From the text given in question, we know each Key has dimension (d_k) and it has to be (d_k) otherwise $Q.K^T$ matrix multiplication wouldn't be possible. Now, adding other information, that is there are N such queries and N such keys indicate that Q and K matrix has same dimensions that is $N \times d_k$.

This implies the dimension of K matrix is $N \times d_k$.

Secondly, the output has a dimension dv indicating V matrix has d_v columns by the law of simple matrix multiplication. The logit matrix has dimension $N \times N$ which indicates the matrix V must have N rows.

The above two statement implies, the dimension of V matrix must be $N \times d_v$.

Problem 2b.

An attention layer has fixed size of input (say *512*). But all input sequences may not be of the same length *512*, thus, they need to be padded. But there should not be any attention paid to the padding (i.e., attention weight should be *0* for padding). This can be accomplished by a mask M. Suppose the padding is post-padding, for example, for a sequence *X1, X2* of length *2*, the padded sequences

with $N - 2$ pads are *X1, X2, pad, pad, . . . pad* of length N. For this example, we make a mask M *= [0, 0, 1, . . ., 1]* with *0* in place of valid input and 1 in place of padded entry. Using M we want to create a matrix of the attention logits L' such that $L'_{i,j} = L_{i,j}$ for j being a valid key and $L'_{i,j} = -\infty$ for key j corresponding to a pad. Write a differentiable equation in *Q, KT, M* to produce such a matrix L'. Differentiable means you should be easily able to get the partial derivative of $L'_{i,j}$ with respect to any entry in *Q or K*.

Solution:

Following is an equation to produce the desired attention logits (*L'*):

$$L' = (1-M) \odot L + M \odot (-inf)$$

$$Where,\ L = QK^T/\sqrt{d_k}$$

The Hadamard product ensures element-wise multiplication. Therefore, if the mask matrix (*M*) has an element *1* which corresponds to a padding then, *-inf* will fill the position of an invalid key else the position will be filled by a valid key.

In the above equation *L'* is differentiable with respect to *Q* and *K*. *M* is a constant matrix containing zeroes and ones with *0* as valid input and *1* as padded input. Hence, *L'* is differentiable as long as *L* is differentiable.