**CS 726: Samples questions on Deep Learning**

1. Consider a neural translation model using an encoder-decoder network. Suppose, we are considering a language pair where a word $x_i$ in source language maps to exactly one word $y_t$ but $t$ may not be equal to $i$. Thus, the source and target sentences are of equal length. How can you exploit this information to generate a better attention distribution at $t$-th decoding step?

   Assume the encoder RNN states are $\mathbf{h}^i$ for $i = 1, \ldots, n$, the decoder RNN states $\mathbf{z}^t$ for $t = 1, \ldots, n$. For your reference the standard method of computing attention is

   $$\text{Attn}(\mathbf{z}_{t-1}, \mathbf{h}_i) = A_{ti}, \qquad \alpha_{ti} = \frac{\exp(A_{ti})}{\sum_p \exp(A_{tp})} \tag{1}$$

   where $\alpha_{t1}, \ldots, \alpha_{tn}$ denotes the attention over encoder states at time $t$ and sum to 1. ..2

   $\alpha_{it} \propto \exp(A_{ti}) \prod_{t'<t}(1 - \alpha_{it'})$

2. Consider a second encoder-decoder network where decoder is arranged as a two-level tree. The first level tree generates a sequence of phrases and the second-level generates the words in each phrase. We want the attention for phrase to be over multiple tokens in the input [that is, the sum of attention values need not be 1].

   (a) We will first generate phrase attention at time $t$ as $b_{t1}, \ldots, b_{tn}$. Suggest how the logic to generate $b_{ti}$ could be made different so as to support multi-word attention? Write your equations using notations of Equation 2a. ..3

   We will allow the $b_{ti}$ to be binary variables. Thus, for each position we can independently choose to attend to that position or not. The modified equation for attention is now:

   $$\text{Attn}(\mathbf{z}_{t-1}, \mathbf{h}_i) = A_{ti}, \qquad \alpha_{ti} = \frac{1}{1 + \exp(-A_{ti})}$$

   (b) Using the above attention, we generate a hidden vector for the phrase at time $t$ (call it $B_t$). This vector $B_t$ is used to condition the generation of words in the phrase using another RNN. We use $\mathbf{s}^k$ to denote the state of this RNN at time $k$. At some time $k > t$ we generate a word $y_k$ in the phrase $B_t$ using attention like in normal encoder-decoder model. The difference is that we also have available to us the attention vector $b_{t1}, \ldots, b_{tn}$ of its parent phrase vector $B_t$. Using these, how can you design a better attention distribution for $y_k$? [The word level attention vectors $\alpha_{k1}, \ldots, \alpha_{kn}$ are required to sum to 1.] ..3

   We want the word-level attention to be non-zero only when phrase-level attention on that word is non-zero. One way to achieve this is by defining $\alpha_{ki} = \frac{\exp(A_{ki})b_{ti}}{\sum_p \exp(A_{kp})b_{tp}}$

   Other variants may be possible.

3. Answer the following questions on encoder-decoder models for translation.

   (a) The sequence of output tokens $y_1, \ldots, y_n$ for an input sequence $\mathbf{x}$ for which the probability $\Pr(y_1, \ldots, y_n|\mathbf{x})$ is maximum can be found exactly in polynomial time. True or False? Justify. ..2

False. In the encoder-decoder model the $\Pr(y_1, \ldots, y_n|\mathbf{x})$ is decomposed using chain rule where every token depends on all previous tokens. This makes the time to find the optimal sequence exponential in $n$.

*Some of you simply state that "the number of possible sequences is exponential in n". That is not a meaningful answer because if the tokens are independent of each other you can find the optimal sequence in linear time even though the set of possible sequences is exponential in n.*

(b) What is the main motivation behind the residual connection between the LSTM layers? ..2

The residual conections improve accuracy of the model since they permit better flow of the gradient in deep neural architectures.

(c) Why is the bidirectional LSTM used only in the first layer of the encoder, and not in the higher layers? ..2

With a unidirectional LSTM, the different layers can be processed in parallel. With a bi-directional model, each higher layer has to wait for the lower LSTM in both directions to terminate. This reduces parallelism.

(d) Why is the bidirectional LSTM not used in the decoder? ..2

Inference will become very expensive. During inference beam-search is used to find the most likely sequence. Beam search greedily generates the sequence token by token. Such generation is not possible with bi-directional LSTM.

4. Let $\mathbf{x} = [x_1, x_2, x_3]$ and $z$ be real random variables where $p(z) \sim N(0, 1)$. Also, we know that $[x_1, x_2, x_3] = [3z + 2, 4z + 1, z - 2]$. With this knowledge, we will plug in the optimal parameters of a VAE defined over them where encoder $q_\phi(z|\mathbf{x}) \sim N(\mu_{z|\mathbf{x}}, \sigma_{z|\mathbf{x}})$ and decoder $p_\theta(\mathbf{x}|z) \sim N(\mu_{\mathbf{x}|z}, \Sigma_{\mathbf{x}|z})$. Assume both encoder and decoder are linear layers where $\mu_{\mathbf{x}|z} = [\theta_1 z + \theta_2, \theta_3 z + \theta_4, \theta_5 z + \theta_6]$ and $\mu_{z|\mathbf{x}} = \phi_1 x_1 + \phi_2 x_2 + \phi_3 x_3 + \phi_4$.

(a) Provide the optimal parameters $\theta_1 \ldots \theta_6$ of the decoder mean. ..1 These will match the corresponding coefficients defined in $[x_1, x_2, x_3] = [3z + 2, 4z + 1, z - 2]$

(b) Provide the optimal parameters of the encoder means $\phi_1 \ldots, \phi_4$ ..3 Since the components of $\mathbf{x}$ are independent of each other given $z$, we can estimate $q(z|\mathbf{x})$ as $1/3[(x_1 - 2)/3 + (x_2 - 1)/4 + (x_3 + 2)]$ This when simplified will give us values for $\phi_1, \phi_2, \phi_3, \phi + 4$.

(c) What is the ideal NN to estimate $\Sigma_{\mathbf{x}|z}$ and $\sigma_{z|\mathbf{x}}$ from their respective inputs? ..3 The variance $\sigma_{z|\mathbf{x}}$ to be reliably estimated requires the sample itself and square of individual samples. Thus the $\sigma$ logic in the encoder must depend on $x_i$ and square of $x_i$. The variance of each $x_i$ is a constant function of its input and thus any linear layer with a bias parameter could suffice.

5. Consider a VAE for which prior $p(z)$ is Bernoulli distributed with parameter $\alpha$. The $\alpha$ now needs to be learned jointly with the encoder parameter $\phi$ and decoder parameters $\theta$. What will be the function at the last layer of the encoder? Let $e_i$ denote the output from the last encoder layer. Write the formulae for the $D_{KL}(q_\phi(z|x^i)\|p_\alpha(z))$ in terms of $\alpha$ and $e_i$. ..2
$e_i \log \frac{e_i}{\alpha} + (1 - e_i) \log \frac{1-e_i}{(1-\alpha)}$

6. Consider the attention-based encoder-decoder network for sequence prediction that at each time step $t$ outputs a probability distribution about possible output tokens: $\Pr(y_t|\mathbf{x}, y_1, \ldots, y_{t-1})$

Show an example where the sequence output by greedy inference has a smaller probability than the highest probability sequence of one longer length. You have to define two distributions: $\Pr(y_1|\mathbf{x})$, $\Pr(y_2|\mathbf{x}, y_1)$, and possibly others. Assume each $y$ can take 3 possible values: 1, 2, or eos. Fill in the tables below for your probabilities. Note the last token in each sequence always has to be eos.

$\Pr(y_1|\mathbf{x}) =$

| $y =$ | 1 | 2 | eos |
|---|---|---|---|
| | | | |

$\Pr(y_2|\mathbf{x}, y_1) =$

| $y_1 =$ | 1 | 2 |
|---|---|---|
| $y_2 = 1$ | | |
| $y_2 = 2$ | | |
| $y_2 =$eos | | |

Define other distributions if required here

The highest sequence probability is:

The greedy sequence probability is: ..3

| $y =$ | 1 | 2 | 3 |
|---|---|---|---|
| | 0.5 | 0.4 | 0.1 |

| $y_1 =$ | 1 | 2 |
|---|---|---|
| $y_2 = 1$ | 0.3 | 1.0 |
| $y_2 = 2$ | 0.3 | 0 |
| $y_2 =$eos | 0.4 | 0 |

$\Pr(y = eos|\mathbf{x}, y_1 = 2, y_2 = 1) = 1.$

7. Consider a speech to text task using an attention based encoder-decoder network. In standard attention, we assign a multinomial distribution over the importance $\alpha_{ti}$ at input position $i$ while outputing $y_t$ as follows:

$$\text{Attn}(\mathbf{z}_{t-1}, \mathbf{h}_i) = A_{ti}, \quad \alpha_{ti} = \frac{\exp(A_{ti})}{\sum_p \exp(A_{tp})}, \quad \mathbf{v}_{x,t} = \sum_i \alpha_{ti}\mathbf{h}_i$$

Now say we change the attention to be a Gaussian distribution over the input positions $i$. Write equations in terms of $\mathbf{h}_i$, and decoder state $\mathbf{z}_{t-1}$ to show how you could define $\mu_t, \sigma_t$ the mean and variance of the input attention position at time $t$. This is a design problem with no one correct answer. ..3 We treat the $\alpha_{ti}$ as fractional occurrence counts of Gaussian distribution positions. Then,

$\mu_t = \sum_i \alpha_{ti} i$ and $\sigma_t = \sum_i (i - \mu_t)^2 \alpha_{ti}$

**Total: 31**