CS 6910 - Deep Learning - Endsem **Duration:**2 hours, **Marks: 75**

1. (a) **(2 Marks)** What is the technical name for the problem which occurs when all the weights of a neural network are initialized to 0 ?
   (b) **(2 Marks)** Which form of regularization ensures sparse weights ?
   (c) **(2 Marks)** Training an ensemble of large neural networks is obviously very hard. Suggest a method for doing this efficiently.
   (d) **(2 Marks)** In LSTMs what controls the flow of information and gradients ?
   (e) **(2 Marks)** Given the input dimensions W and H, filter size F, padding P, stride S used in a colvolution layer, give a formula for computing the dimensions of the output W' and H' of the convolution layer.

2. (a) **(2 Marks)** Write down the formula for the Gelu activation function.
   (b) **(2 Marks)** Pictorially explain any one learning rate used for training transformer based models.
   (c) **(2 Marks)** Draw the inception block as used in Inception Net (just one block, not the entire network).
   (d) **(2 Marks)** Write down the equations of the states and gates in GRU.
   (e) **(2 Marks)** Write down the equation for additive attention.

3. **(5 marks)** In class we discussed a technique (Word2Vec) to obtain word embeddings given a text corpus. We noted that these representations have interesting properties like syntactically similar words are "closer" to each other compared to dissimilar words. Say we wish to extend this idea to obtain representations for words across 2 languages (English and French). Further we wish to have similar words across language to have representations that are "close".

   We are given a corpus of English and French text. We are also given a dictionary $\mathcal{D}$ of words that maps words in English to equivalent words in French. Write down the objective function for training a Word2Vec model for the above task.

   A good language model assigns higher probability for "valid" sentences in a language and low probabilities for sentences which are incorrect (for instance - sentences which do not follow language grammar rules).

4. **(5 marks)** Convolutional neural networks consist of a series of convolutional layers and max pooling layers. The max pooling layers do not contain any weights. Say the output of one such convolution layer is a 4×4 feature map denoted by h (comprising of $h_{11}, h_{12}, ..., h_{44}$). On top of this we have a 2×2 max pooling layer whose output is denoted w.r.t. m (comprising of $m_{11}, m_{12}, m_{21}, m_{22}$). Say we have computed the derivative of the loss w.r.t. the output m of the max-pool layers. Express the derivative of the loss w.r.t. the output (h) of the convolution layer.

5. **(5 marks)** Consider a feedforward neural network fully connecting a m dimensional input layer to a n dimensional output layer. Represent this feedforward network as a convolutional neural network. Specifically write the number of filter(s) to be used and the sizes of these filters.

6. **(5 marks)** Consider an input volume of size $W \times H \times D$. Now consider two convolutional neural networks. The first one uses one $7 \times 7 \times D$ filter on the input image followed by a max pooling layer. The second one uses three $3 \times 3 \times D$ filters on the input image followed by a max pooling layer. Do you see any advantage of using one of these networks over the other. Explain your answer.

7. **(5 marks)** We saw that Residual Networks use identity connections to facilitate better flow of information (and gradients). On other hand, LSTMs use gates to facilitate better flow of information. One potential problem with Residual Networks is that we hardcode the manner in which the information should flow (for example, by always adding connections from layer $k-2$ to layer $k \quad \forall k$). Suggest a way to make this more adaptive (instead of hardcoding the connections).

8. **(5 Marks)** We saw that for a given decoder timestep t, the attention weights over the S encoder states (timesteps) are constrained to sum up to 1 (i.e., $\sum_{i=1}^{S} \alpha_{it} = 1$). However, there is no constraint on how the attention weights behave across timesteps. This could result in a situation where the attention weights are always concentrated on the same set of encoder states across different decoder timesteps. This may be undesirable (for example, in the case of translation this might mean that some words in the source sentence never get attention). Suggest a way to tweak the loss function to explicitly prevent this problem?

9. **(5 Marks)** Let y be the true output and $\hat{f}(x)$ be the output prediced by the models. Prove that $(E[(y-\hat{f}(x))^2] = Bias^2 + Variance + \sigma^2$ (irreducible error))

10. **(5 Marks)** What is the purpose of the feedforward layer in a transformer based encoder?

11. **(5 Marks)** Let $s_1, s_2, ..., s_M$ be the words in the source sentence and $t_1, t_2, ...t_N$ be the words in the target language. When translating between languages which have very similar language structure (e.g. Subject-Object-Verb ordering), the target word being decoded at timestep i (t_i depends only on source words lying in a small window of k words around t (e.g., s_{t-k} to s_{t+k}). How would you modify the encoder-decoder attention function to ensure that at time step t the attention on words outside the window is 0 (i.e., words other than s_{t-k} to s_{t+k} do not get any attention). I am looking for a solution which would be easy to implement in matrix form.

12. **(10 Marks)** Suggest a problem of social importance in the Indian context where AI/Deep Learning can help. For example, we can use a CNN based image classifier to predict the diseases that have affected a plant by looking at a real time image of the plant (say, taken by a farmer and uploaded on your app). Suggest other such applications of AI for social good outside the agriculture domain. Marks will be given based on (i) novelty of the problem (ii) availability of data or novel ideas for collecting data for training the model, (iii) potential social impact (iv) details of the solution and (v) feasibility of deployment.