Deep Learning

Dr. Sugata Ghosal
sugata.ghosal@pilani.bits-pilani.ac.in

BITS Pilani
Pilani Campus

Important Topics for Compre Exam (40 Marks)

Pre-Midsem (20-30%)

- Numerical / conceptual questions related to deep learning characteristics, Boolean function realization, universal classifiers, universal function approximation using Deep Networks
- Numerical / conceptual questions related to MLP construction, computational graphs
- Conceptual/Numerical questions related to regularization, optimization

Post-Midsem (70-80%)

- Numerical /conceptual questions on CNN
- Numerical /conceptual questions on RNN and applications
- Numerical / conceptual questions on Autoencoders (ordinary, CAE, sparse, contractive, denoising, KL divergence)
- Conceptual questions on VAE, GAN

- *Coding/Assignment related questions (15-20%)*

Question 1

- A network-in-network architecture is used to classify gray-scale images of size 64x64 into 100 classes. A micro-MLP with a single hidden layer of 10 nodes is used in the first layer instead of convolution filters of size 5x5 to generate an output feature map of depth (channel) 1. What will be the number of trainable parameters in the micro-MLP based convolution layer?

- Micro-MLP input size: 5*5
- Hidden nodes 10, output size 1
- Total # of trainable parameters = 5*5*10 (input-hidden weights) + 10 (bias @ hidden) + 10*1 (hidden-to-output nodes) +1 (bias @ output)

  = 271

Question 2

- Which of the following networks is (are) more suited architecturally for classifying images of varied size (e.g., the input image database has images of size 64x64, 96x96, 128x96, etc.)? No image rescaling is permitted.

- Networks that use global average pooling generate feature vectors whose size is independent of input data, after flattening.
  - Resnet, NiN, Inception all use global average pooling

- Max-Pooling function is
  - Differentiable: at the position of maximum, derivative value is 1, at all other positions it is 0.

Question 3

- Consider an LSTM network with one hidden layer of 20 nodes used for predicting the next word in one hot encoded representation in its output . No bias is used in any of the nodes. The corpus is of length 1000 words and there are 100 unique words. Assume a 10 dimensional word embedding module outside of the LSTM network, whose output is fed to the word predictor LSTM network. What will be the total number of trainable weights in the LSTM network?

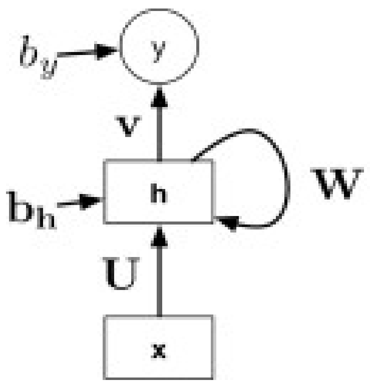
- Input size to the network: 10
- Number of hidden nodes: 20
- Number of output nodes: 100
- Total weights from input to LSTM hidden nodes = 4*(20+10)*20 = 2400
- Total weights from LSTM hidden nodes to output nodes = 20*100
- Total Weights = 4400

- $\mathbf{h}_t$ = step($\mathbf{W}\mathbf{h}_{t-1}$ + $\mathbf{U}\mathbf{x}$ - $\mathbf{b}_h$) and y_t = step($\mathbf{v}\mathbf{h}_t$ - b_y). **U**, **v** and **W,** b_y and are given by

  v = [1 -1] and b_y = 0.5 $\mathbf{b}_h^T$ = [-0.5 0,5]

  $$w = \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \quad U = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

  Left-to-right input string {0, 1, 0, 0}



- Input 0: h_1=step([0+0.5 0-0.5])=[1 0] y_1=step(1-0.5)=1
- Input 1: h_2=step([0+0.5 0-0.5])=[1 0] y_2=step(1-0.5)=1
- Input 0: h_3=step([1+0.5 1-0.5])=[1 1] y_3=step(0-0.5)=0
- Input 0: h_4=step([0+0.5 0-0.5])=[1 0] y_4=step(1-0.5)=1

Question 4

- In a network in network architecture, one hot output encoding is used to classify 100 classes of objects. The last convolution layer generates 7x7x128 features maps (depth=128). Assuming the fully connected subnetwork has one hidden layer with 50 nodes, what will be the total number of trainable weights (excluding biases) in the fully connected subnetwork (from the last convolution layer to the output layer)?

- NiN uses global averaging pooling
  - After flattening, feature vector size = 128 hitting the FC layer
  - So, total # of weights in FC layer = 128*50+50*100=11400

Question 5

- Consider a 1x3 convolution (as per Deep Learning terminology) kernel o =[o_1, o_2, o_3]= [-1 2 -1]. Output is generated by stacking 4 such operators on the 1-D input data (i.e., the operator is applied four times on the input). What is the width of the equivalent operator O that will give the same output, when applied only once on the input data?

- After stacking 2 filters, equiv filter size = 3+(3-1)
- After stacking 3 filters, equiv filter size = 3+(3-1)*2
- After stacking 4 filters, equiv filter size = 3+(3-1)*3 = 9

Question 6

- Consider a vanilla RNN network with one hidden layer of 20 nodes used for predicting the next word in a text corpus. No bias is used in any of the nodes. The corpus is of length 1000 words and there are 50 unique words. Assume a 10 dimensional word embedding module outside of the RNN network, whose output is fed to the word predictor RNN network. One hot encoding is used in the output. What will be the total number of trainable parameters in the RNN network?

- Input-to-hidden weights = (20+10)*20
- Hidden-to-output weights = 20*50
- Total weights = 1600

Question 7

- Design a recurrent neural network that outputs a parity bit for binary sequences of arbitrary length. The inputs are given as binary sequences from right to left and output of 1 is generated when number of ‘0’s in the string seen so far is even. For instance, the input string 1010010 would
  generate an output as follows:
  - Input: 0,1,0,0,1,0,1
  - Target output: 0,0,1,0,0,1,1
- All of the units use the hard threshold activation function, i.e., output is 1 if total weighted input is >=bias, else 0.