COL 865: Special Topics in Computer Applications - Deep Learning.
Major Exam. Part (A), Section II. Weightage: 15%

November 23, 2017

Notes:

- **Time: 4:45 pm to 5:45 pm. Max Score: 20 points.**
- **There is a total of 5 questions. Each question carries 4 points.**
- **Start each answer on a new page of your answer sheet.**
- **This section is open handwritten notes only. Laptop/internet connectivity is not allowed.**

1. **Backpropagation in Recurrent Networks**
   Consider the recurrent architecture as shown in Figure 1. Let $x^{(t)}$ denote the input at time step t. Let each $x^{(t)} \in \mathcal{R}^d$. Let $h^{(t)}$ denote the hidden state at time t. Further, input to hidden connections are parameterized by a weight matrix U while hidden to hidden connections are parameterized by weight matrix W. Let each hidden state consist of r units. Therefore, $W \in \mathcal{R}^{r \times r}$ and $U \in \mathcal{R}^{d \times r}$. In our example, $t \in \{1,2\}$. Let $h^{(0)}$ be a r-sized vector of all $1's$. Let $o = h^{(2)}$ denote the output of the final hidden layer which is then fed through a softmax layer to get the output $\hat{y}$. Let $\mathcal{L}(\hat{y}, y)$ denote the loss function. Assume ReLUs as the activation units and that they are operating in the active range (i.e., input to the unit > 0). You can ignore the bias terms for simplification.

   Derive the expression for $\nabla_U \mathcal{L}(\hat{y}, y)$ in terms of $\nabla_o \mathcal{L}(\hat{y}, y)$. Note that you need account for the fact that the U's are tied with each other. You can use the following facts. Let b and c be column vectors such that $c = A^T b$. Let b affect $\mathcal{L}$ through c. Then,

   - $\nabla_b \mathcal{L} = A \nabla_c \mathcal{L}$
   - $\nabla_A \mathcal{L} = b(\nabla_c \mathcal{L})^T$

2. **L2-regularization and Early Stopping**

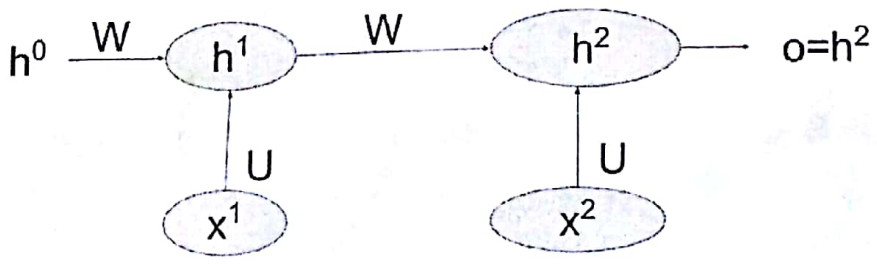

Figure 1: Recurrent Network

(a) Recall that the expression for the $L2$-regularized weights $\tilde{w}$ can be written as:

$$\tilde{w} = Q(\Lambda + \alpha I)^{-1}\Lambda Q^T w^* \quad (1)$$

Here, $H = Q\Lambda Q^T$ is the eigenvalue decomposition of the Hessian matrix H of the error function. Λ is the diagonal matrix of eigenvalues of H with i^{th} eigenvalue given by λ_i. α is the regularization parameter. w^* denotes the vector of optimum unregularized weights. Describe $\tilde{w}$ in terms of stretching/shrinking of the original parameter space as a function of w. Clearly describe the relationship of this shrinking/stretching with the eigenvalues of the matrix H.

(b) Under certain assumptions, when doing early stopping the weights $w^{(\tau)}$ obtained after τ iterations can be described as:

$$w^{(\tau)} = Q(I - (I - \epsilon\Lambda)^{\tau})Q^T w^* \quad (2)$$

Here, ϵ is a very small constant and I denotes the identity matrix. Other symbols are as before. Show that in the above setting, early stopping can be seen as a form of $L2$-regularization for appropriate choice of the hyperparameters α, τ and ϵ. In particular, you should derive an expression equating two diagonal matrices, each of whose entries are a function of the hyperparameters above. Assuming $\epsilon\lambda_i << 1$ and $\lambda_i/\alpha << 1$, show that $\tau \approx \frac{1}{\epsilon\alpha}$. Hint: Use the expansion $log(1 + x) \approx x$ when $|x|$ is very small.

3. Maxout Units

Maxout units(Goodfellow et al., 2013) are used to introduce non-linearity as follows. Instead of applying an element-wise function $g(z)$, maxout units divide the input vector $\mathbf{z}$ into groups of k values. Each maxout unit then outputs the maximum element of one of these groups:

$$g(\mathbf{z})_i = \max_{j \in G(i)} z_j$$

where $G(i)$ is the set of indices into the inputs for group i, i.e., $\{(i-1)k + 1, \ldots, ik\}$.

(a) What kind of functions can a maxout unit learn? Be as precise as possible.

(b) Are these functions convex? Argue.

(c) Describe the relationship between an maxout unit and the standard ReLU?

4. Pictorial Understanding

(a) Figure 2 plots the performance of three different architectures (a) 3-layered convolutional network (b) 11-layered convolutional network (c) 3-layered fully connected network. x-axis plots number of parameters and y-axis plots the test accuracy.

i. Match the type of the architecture to the corresponding curve in the plot.

ii. Justify the performance of each architecture in absolute terms as well as relative to other curves in the graph.

(b) Figure 3 presents an intuitive understanding of how using an absolute value rectifier can help carve out linear regions whose number grows exponentially with the depth of the network. Explain the figure to articulate this understanding (you should explain each of the three pictures and also the relationship between them).

5. Short Explanations

(a) Describe the concept of teacher forcing training in recurrent networks. Explain how teacher forcing allows you to train the weights in each time step in parallel. You should argue in mathematical terms. Feel free to take the help of a system architecture drawing.

(b) We have studied the concept of zero-padding in CNNs.

i. What is the advantage of zero padding in a CNN? Explain with the help of a 1-dimensional example.

ii. In a 1-dimensional CNN with kernel size k and depth d, how many zero pads (in each layer) would be required so that the output size is preserved to be equal to the input? Justify your answer.