EE538 Neural Networks

Homework 7

Due: 12:59 on June 2nd, 2021

1. Let's consider a Bidirectional Associative Memory in Lecture 9.
   (a) Generate *S* random binary vector pairs ($\boldsymbol{x}^s$,$\boldsymbol{y}^s$) (for s=1,..,*S*) with 1024 and 512 elements for $\boldsymbol{x}^s$ and $\boldsymbol{y}^s$, respectively. For simplicity, each element of $\boldsymbol{x}^s$ and $\boldsymbol{y}^s$ are generated either -1 or +1 with a probability of 0.5. Try 3 datasets D_{50}, D_{100}, and D_{200} with *S*=50, 100, and 200, respectively. (10 points)
   (b) Code a computer program to generate $\boldsymbol{y}$ from $\boldsymbol{x}$, and also $\boldsymbol{x}$ from $\boldsymbol{y}$ as
   $y_j = \mathrm{sgn}\left(\sum_{i=1}^{I} w_{ji} x_i\right)$, $x_i = \mathrm{sgn}\left(\sum_{j=1}^{J} w_{ji} y_j\right)$, $w_{ji} = \sum_{m=1}^{M} x_i^m y_j^m$. (10 points)
   (c) For the dataset D_{50}, use each $\boldsymbol{x}^s$ as an input vector $\boldsymbol{x}$ and generate an output vector $\boldsymbol{y}$ up to 10 iterations, i.e., 10 cycles of from $\boldsymbol{x}$ to $\boldsymbol{y}$, and then to $\boldsymbol{x}$ again. Plot the average of all output errors, i.e., the number of different elements between the generated $\boldsymbol{y}$ and true paired $\boldsymbol{y}$, versus the epoch. (20 points)
   (d) Repeat (c) for noisy input vector $\boldsymbol{x}$, which is m-elements different from one of the stored vector $\boldsymbol{x}^{10}$. Use *m* values of 0,1, 2, 3, 5, 7, 10, 15, 20, 30, 50, and 100, and generate 10 input vectors with random changes for each *m* value. Make all the curves in a single plot for easy comparison. (15 points)
   (e) Repeat (d) for the dataset D_{100} and D_{200}., and compare the results with those of (c). (15 points)

2. For the Transformer model, please answer the followings.
   (a) The Transformers usually use multi-heads, of which query and key embedding dimension is much smaller than that of the word imbedding. For example, $N_{mod}= N_{que} \times N_{head}$ and $N_{key} = N_{que}$. Discuss the advantages/disadvantages of this multi-head approach and the single-head approach with the same number of embedding dimensions. (10 points)
   (b) Please explain the idea behind the division by $\sqrt{N_{que}}$ before the Softmax operation. Hint. Read the footnotes in the Vaswani paper and judge the validity. Any other suggestion? (10 points)
   (c) Since $(\boldsymbol{q}_{hp})^T \boldsymbol{k}_{hp'} = (\boldsymbol{q}_p)^T \boldsymbol{W}_h^q (\boldsymbol{W}_h^k)^T \boldsymbol{k}_{p'}$ one may replace the two mapping matrices $\boldsymbol{W}_h^q$ and $\boldsymbol{W}_h^k$ with one matrix $\boldsymbol{W}_h^{qk} = \boldsymbol{W}_h^q (\boldsymbol{W}_h^k)^T$. Discuss the advantages/disadvantages of these two different approaches. (10 points)