Midterm Exam

ECE 09495/09595

Advanced Emerging Topics in Computational Intelligence, Machine Learning and Data Mining
Emerging Topics in Computational Intelligence, Machine Learning and Data Mining

Due: April 21 before end of the day

Question 1 (20%):

The CIFAR-10 dataset (https://www.cs.toronto.edu/~kriz/cifar.html) consists of 60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. We will train a CNN for the classification. We have a working model (**Q1.py**) and the validation accuracy of this model is around 79.70%.

1. **Task**:   Please use **data augmentation** and see if you can improve validation accuracy. Please feel free to tune any hyperparameters you like including number of convolution layers, convolution kernel size, max pooling layers, optimizers type, leaning rate, number of epochs, batch size, etc. You may want to explore (https://keras.io/preprocessing/image/#imagedatagenerator) for data augmentation.
2. **Result Reporting**:   Please send me a graph of epochs vs training/validation accuracy for your model that achieved highest validation accuracy. Please do not include your code.

Question 2 (20%):
We are using the CIFAR-10 dataset and we are using ResNet (**Q2.py**). The model is working and produces reasonable validation accuracy. Please see network architecture (for ResNet20) in the attached figure.
1. **Task**:   Please run the model for different depths (20, 32, 44, 56). Please modify the code to remove skip connections and train the model for different depths (20, 32, 44, 56).
2. **Result Reporting**: Please fill the following table:

Model Depth	With skip connection		Without skip connection	
	Training accuracy	Validation accuracy	Training accuracy	Validation accuracy
20				
32				
44				
56				

Question 3 (20%):

We are using the same CIFAR-10 data with an Inception module (Q3.py) this time. The network consists of a single inception module and does not seem to perform well (training accuracy > 99.9% , and validation accuracy ≈ 60%).

1. **Task**:   Please modify the network to improve its validation accuracy.

2. **Result Reporting**:   Please include your code and report the maximum validation accuracy.

Question 4 (20%):

We will train a character level LSTM on a small text, few paragraphs from Chapter 1 of Bertrand Russell's book "Conquest of Happiness".

1. **Task**:   Train the model (*Q4_Train.py*) for various input size, i.e., the number of character given to the LSTM for making a prediction about the next character (try 5/10/20/30/40/50). Test your model using Q4_Test.py for each trained model and generate 10/50/100/500 characters for each model.
2. **Result Reporting**:   Looking at the generated text, please comment on:
   a. What is the effect of number of characters given to the LSTM during training?
   b. What is the effect of number of characters generated by the LSTM during testing?

Question 5 (20%):

We want to extend the same LSTM network for a big dataset (A Tale of Two Cities, a novel by Charles Dickens). The model is available *Q5_Train.py*.

1. **Task**:   Please make appropriate choices for the:
   a. Number of LSTMs
   b. Size of each LSTM
   c. Dropout
   d. Number of epochs
   e. The batch size
2. **Result Reporting**:   Please provide your code and a paragraph of almost 500 characters generated by your trained LSTM.

Extra Credit (20%):

Extend the character level LSTM to word level LSTM for the same input (A Tale of Two Cities, a novel by Charles Dickens). Train the model and provide the code. (Hint: try https://keras.io/preprocessing/text/#tokenizer for text preprocessing)