Compulsory Assignment #3

Data Mining, Machine Learning and Deep Learning [KAN-CDSCO1004U]

Somnath Mazumdar
sma.digi@cbs.dk
Copenhagen Business School, Denmark

Deadline: {Refer digital exam for exact date and time}

Instructions

1. This compulsory assignment contains three main parts. Some parts may have more than one question. Please answer all the questions.
2. You must upload your solutions before the deadline to the digital exam http://exam.cbs.dk/.
3. If your answers involve any math, please feel free to use whatever format you like [Latex, Word, plain paper and pen], but make sure that it is readable.
4. It is always a good practice to use comments extensively in your code so that it will be easy for other people to understand it.
5. The code for answering the questions should be submitted as **one single *jupyter* notebook.**
6. For answering Question 3, please consult books, peer-reviewed research articles, and technical reports (and white papers to a certain extent). **Citing websites are highly discouraged**.

Assignments

Question 1 : **Bi-Directional LSTM**

Download two attached datasets (Fake.csv and True.csv), which contain real and fake news. In the given dataset, there is a total of five columns. Now build a Bi-LSTM model to detect fake news using TensorFlow and other available libraries. Some hints are as follows:

1. You might pre-process both datasets.
2. You should also use Natural Language Toolkit (NLTK) library to handle "stopwords" in the dataset.
3. For a better model, remove stopwords and words with two or fewer characters and then split data into test and train sets.
4. You might also need to create a tokenizer to tokenize the words and sequences of tokenized words.
5. You can call Bi-Directional LSTM (from Keras library) and where fitting set Sigmoid and ReLU as activation function, Adam as an optimizer and Binary cross-entropy loss function should be used.
6. During training, set the batch size to 64 and the number of epochs to 2.
7. Finally, print the model accuracy. If the predicted value is greater than 0.5, then it is a piece of real news.

Question 2 : **Convolutional Neural Network (CNN)**

Build a Convolutional Neural Network (CNN) to classify images from the CIFAR-10 dataset. The network architecture should have the following layers:

- Convolution layer: kernel_size = [5 x 5]
- Convolution layer: kernel_size = [5 x 5]
- Batch normalization layer
- Convolution layer: kernel_size = [5 x 5]
- Max pooling: pool size = [2 x 2]
- Convolution layer: kernel_size = [5 x 5]
- Convolution layer: kernel_size = [5 x 5]
- Batch normalization layer
- Max pooling: pool size = [2 x 2]
- Convolution layer: kernel_size = [5 x 5]
- Convolution layer: kernel_size = [5 x 5]
- Convolution layer: kernel_size = [5 x 5]
- Max pooling: pool size = [2 x 2]
- Dropout layer
- Dense layer: units = [512]
- Dense layer: units = [256]
- Dropout layer
- Dense layer: units = [10]

The test accuracy should be more than 80%. The CNN model architecture implemented loosely mirrors the AlexNet [1].

Data: CIFAR-10 is another standard image classification dataset to classify a coloured 32 x 32-pixel image data into ten image classes, namely, airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. You can import dataset using the command *tf.keras.datasets.cifar10.load_data()* [2].

Question 3 : **Written Assignment**

Write an approximately two-page extended abstract to discuss the applications of **one** of the following techniques.

1. Federated Learning
2. Reinforcement Learning

To answer the above question, you can also do a tiny literature review to find out how your chosen techniques are used across various application domains. Feel free to choose whatever direction you want to explore. Write your reflections and critical comments rather than just reproducing what you found in the literature review. Your report should conform to academic standards and use APA referencing/citation style.

References

[1] Wikipedia, "Alex net." https://en.wikipedia.org/wiki/AlexNet.

[2] Keras, "Cifar10 small images classification dataset." https://keras.io/api/datasets/cifar10/.