Université Abdelmalek Essaadi
Faculté ses Sciences et techniques de Tanger
Département Génie Informatique
Master : MBD
Deep Learning
Pr . ELAACHAk LOTFI

Lab 3

Objective : The main purpose behind this lab is to get familiar with Pytorch, to build deep neural network architecture for Natural language process by using Sequence Models.

Work to do:

Part1 Classification Task:

1. By using scrapping libraries (Scrapy / BeautifulSoup), try to collect text data from several Arabic web site concerning one topic then prepare your Dataset as Below:

Text	Score
Text 1(Arabic Language)	6
Text 2(Arabic Language)	7.5

The score presents the relevance of each text (The score should be between 0 to 10).

2. Establish a preprocessing NLP pipeline (tokenization stemming lemmatization, Stop words, Discretization, etc) of the collected Dataset.

3. Train your models by using RNN, Bidirectional RNN GRU and LSTM Architectures and tuning hyper-parameters to get the best performance.

4. Evaluate the four languages models by using standards metrics and other metrics like blue score.

Part 2 Transformer (Text generation):

Install pytorch-transformers, then load the GPT2 Pre-trained model.

1. Fine tune the pre-trained model (GPT2) to a customized Dataset (You can generate your own DataSet).

2. Generate a new paragraph according to a given sentence.

You can follow this tutorial :
https://gist.github.com/mf1024/3df214d2f17f3dcc56450ddf0d5a4cd7

Notes :

- **At the end each student must give a brief synthesis about what he has learn during the proposed lab.**
- **Push the work in the Github repository and write a brief report in Github readme file.**

Tools:

Google colab or Kaggle, gitlab/github.