M20Temp17: Advanced NLP **Monsoon 2020**

Assignment 2

Deadline : 17-10-2020, 23:55 Hrs

Instructor: Dr. Manish Shrivastava *TA: Mounika Marreddy, Prashant Kodali*

1 General Instructions

1. Ensure that the submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors, and/or the internet. If any such attempt is caught then serious actions including an F grade in the course is possible.

2. Your grade will depend on the correctness of answers and output. In addition, due consideration will be given to the clarity and details of your answers and the legibility and structure of your code.

2 Problem Statement

We have seen different architectures for Neural Machine Translation. Starting with simple Seq2Seq network, and an improvement of this architecture by adding attention.

1. We had seen Seq2Seq architecture in the paper Sequence to Sequence Learning with Neural Networks, Stuskever et al.

2. Followed by how the concept of attention was introduced in Neural Machine translation in the paper Neural Machine Translation by Jointly Learning to Align and Translate, Bhadnau et al.

In this assignment, you have to code these two Neural Machine translation models and train them on a simple parallel corpus, from English to Hindi.

2.1 Dataset

The language pair we are using is English - Hindi. If you are not familiar with this language please reach out to use, we will share the dataset in a language that you are familiar with.

The dataset is attached as a zip file. Keeping in mind the compute, we are limiting to only 10000 sentences, of length 4-5. Randomly split these 10000 sentences into Train-Dev-Test in 70-10-20 ratio.

Zip file consists of individual text files for Hindi and English. Also consists of a pickled dataframe with english and hindi sentences as two columns. You can make use of either one. You can read the pickled file using the code snippet below:

```
import pandas as pd
df = pd.read_pickle('en_hi.pkl')
```

3 Deliverables

1. Code base for the two papers listed above. Seq2Seq and Bhadnau's attention for the given language pair.

2. Attention has to be coded ground up. Any pre-build methods/layers that implement attention are not be allowed.

3. Translated output for all the senteces in your test set.

4. Report in a seperate PDF file consisting of:

   (a) Analysis of 20 sentences from test set. Analysis should be qualitative and you are expected to draw insight into the model's working.

   (b) Performance graphs (loss during training, Bleu scores).

5. Trained models checkpoints.

4 Submission Format

All the following files zipped into a single file. File names should be your roll number, followed by "Assignment2". **Ex: 2018XXXXXX_Assignment2.**

- Allowed code submission formats: .py or .ipynb format. If you are submitting code in .py format then please mention the execution instruction in a README file.

- Translation of all test sentences in your test split. You can submit this as a csv or a PDF.

- Report should be a seperate pdf file. Visualisations/analysis in Jupyter notebook will be not be accepted.

- Trained models checkpoints. If the trained models can be shared in a external drive link. Mention the link in the PDF report.

5 Allowed Tools/libraries

1. This assignment shall be implemented in python.

2. You can make use of Deep learning frameworks like PyTorch, TensorFlow, Keras. All the associated functionality of Backprop, optimizers, loss functions,

3. Any pre-trained embeddings for the language pair.