ELMo

Advanced NLP Assignment 2

Implementation of the ELMo Language Model from the paper "Deep contextualized word representations"

T H Arjun, CSD, 2019111012

All the assignment requirements were implemented including:

- Implementation of the ELMo Language Model as described by the Paper with a characted level embedding with BiLM with Source Code.
- The model was implemented from scratch in pytorch and trained on the given corpus.
- Visualisation of Cosine Similarity and Euclidean Distance in matrices for 5 different pairs of words as asked in the assignment.
- Qualitative Analysis of the embeddings produced in the Report
- Analysis of the contextual Embeddings Produced by the Model
- Report

Report is included as Report.pdf

Reproducing the Results

To reproduce the results you can download the ELMo Model Checkpoints and Data Folder

- The requirements can be installed by `pip install -r requirements.txt`. I recommend doing so in a conda environment with `python 3.7`
- `preprocess.ipynb` generates the corpus file from the given files of the assignment.
- `ELMo.ipynb` is the source code of the ELMo Model and has the classes and code related to the model definition, and training.
- `inference.ipynb` has example on how to load and use the trained checkpoint and tokenizer. It also has code related to inference and observations in report.
- `run.sh` is the sample script to run the code on HPC Cluster (takes about 1 hour / epoch on 2080Ti)
- Training the model takes about 2 days for 50 epochs and hence I recommend running as sbatch script as in `run.sh` using papermill.
- Images folder has the images included in the report.