Midterm

COMS W4995-010 Fall 2024
Mathematics of Machine Learning and Signal Recognition

Problem Statement

In MATLAB and only using basic MATLAB functions, you will be implementing a **selective-SSM**[1] with the following specifications:

Layers	1
Embedding Size	64
Sequence Length	4

Your task is to use the model for the Part-Of-Speech Tagging problem. We will use a standard dataset **CoNLL 2003**[2]. The dataset is split into training set, testing set and validation set and is provided to you in csv format for simplicity. In total, there are 46 Parts-of-Speech in the dataset. For our classification task, we will group them into 4 classes.

1. Noun: NN, NNS, NNP, NNPS, NN—SYM, PRP, PRP$
2. Verb: VB, VBD, VBG, VBN, VBP, VBZ
3. Adjective/Adverb: JJ, JJR, JJS, RB, RBR, RBS
4. Others: any remaining POS

To reduce the number of parameters, we will be using pre-trained word vectors from **word2vec**[3] of size 64 (Embedding Size). These vectors are also provided in csv format.

In the train_data.csv and valid_data.csv files, the first column ”tokens” contains the tokenized sentences, the second column ”pos_tags” contains the encoded POS tag for each token and the third column ”ner_tags” which contains the named entity recognition tags (this column can be ignored for this assignment).

The tags to be considered has been provided the assignment document.

Note: Please consider using ¡START¿ token to pad sentences when the current index in the sentence is less than 4, i.e., the sequence length of the model.

Results

Report the accuracy, precision and recall for each of the classes. Please also include a README file specifying how to run the project.

References

[1] Albert Gu, Tri Dao. “Mamba: Linear-Time Sequence Modeling with Selective State Spaces.” (2024).

[2] Erik F. Tjong Kim Sang and Fien De Meulder.“Introduction to the CoNLL- 2003 shared task: Language-independent named entity recognition.” In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pages 142– 147 (2003).

[3] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient estimation of word representations in vector space.” arXiv preprint arXiv:1301.3781 (2013).