Assignment 2 Report

Rupali Dakhane
Computer Science and Automation
Indian Institute Of Science
rupalid@iisc.ac.in

Abstract

Build a language model on **Gutenberg**(D2) corpus after dividing the dataset into train, dev and test. The following tasks are to be performed:

- **Task 1:** Build the best token level LSTM-based language model and compare this model with best classical Language Model.
- **Task 2:** Build the best character level LSTM-based language model and compare this model with best classical Language Model.
- **Task 3:** Generate a sentence of 10 tokens useing the best of above models.

1 Preprocessing

I have divided the datasets in train ,dev and test in the ratio of 8:1:1. The preprocessing includes:

- **Removal of punctuations.**
- **Handling of Unknown words:** I have fixed the size of Vocabulary which includes the most common words of train data. All the out of vocabulary words in train and test data have been replaced by token UNK.

2 Implementation

I have implemented the above two LSTM-based language models using different datasets.

2.1 Token-level LSTM-based Language Model

The model is developed for dataset consisting of five files of Gutenberg Corpus and has following specifications:

- **Embedding:** The word embedding layer expects input sequences to be comprised of integers. Each word in our vocabulary is mapped to a unique integer and encodes input sequences. The encoding is done using Tokenizer and the mappings are saved to map the test data later. The length of input sequence is taken as 50.
- **Layers:** Two LSTM hidden layers with 100 LSTM cells are used in the model with ReLU activation function.
- **Number of Epochs:** The model is run for 20 epochs until the accuracy increases.

2.2 Character-level LSTM-based Language Model

The model is developed for dataset consisting of one file of Gutenberg Corpus and has following specifications:

- **Embedding:** The sequences of characters must be encoded as integers.Each unique character will be assigned a specific integer value and each sequence of characters will be encoded as a sequence of integers. Mappings are created using a sorted set of unique characters in the raw train data. The mapping is a dictionary of character values to integer values. The model takes one-hot character embedding of size 75.
- **Layers:** One LSTM hidden layer with 75 memory cells are used in the model.
- **Number of Epochs:** The model is run for 15 epochs until the accuracy increases.

2.3 Sentence Generation

The sentences are generated by giving random seed text of 50 words as input and target words are predicted using token level Language Model.

3 Result

3.1 Comparison of Language Models

Perplexities	
Token-level LSTM-based Language Model	89.74
Character-level LSTM-based Language Model	3.62
Classical Language Model	121.036

3.2 Generated Sentences

Some examples of sentences generated from Language Model:
Token-level Language Model

- that i am not afraid of the lord and i will bring forth the people
- he said unto him i will not be in the land of egypt
- a man that is in the house of israel shall be the king s house

Character-level Language Model

- satisfied that he was a little and who was a great
- what she had said and the same the connect

4 Acuuracy/Measures

Perplexity is used as the measure for this task.
(All values in Result section table contain perplexity value)

5 Github Link

https://github.com/Rupali0408/NLU_Assignment2