E1246 - Natural Language Understanding
Assignment2 : LSTM Language Models

Divya Agarwal(14722)
divyaagarwal@iisc.ac.in

Abstract

This assignment is about designing a language model on **Gutenberg** corpus using LSTM Model with two different approaches one is word level LSTM model and the other is using character level LSTM based model.and the second part of assignment is about generating sentences with better model.

1 Introduction

In this assignment the tasks were to implement LSTM model for language generation. Because of the limitation of computational resources i implemented this language model on limited files of gutenberg corpus.

- **Task 1:**First task was to design a language model i.e word level LSTM model. i used only three files of gutenberg corpus for this model because the limitation of computational resources. file names are: austen-emma.txt, austen-sence.txt, austen-persuasion.
- **Task 2:** In this i designed an another language model i.e character level LSTM model. i used only one files of gutenberg corpus for this model. file name: austen-sence.txt.
- **Task 3:** In this task we suppose to generate sentences using better model. i used word level LSTM model as it was generating better sentences.

2 Word Level LSTM Model

- **Prepossessing Tokens:** Firstly i filtered raw data i.e removing punctuation etc. then i generate sequences with filtered tokens here i took sequence size to 50. then i split the corpus into training(80 percentage) and testing(20 percentage) part.
- **Model Parameters:** I used adam optimizer of keras and cross entropy loss function. i used three files as previously mentioned. and batch size i took 128 with 50 epochs.
- **Generating Sentences:** I generated sentences with help of this this model by taking seed text of 50 words and predicts 51st word. i generated sentence of 10 words like this.

3 Character Level LSTM Model

- **Prepossessing:** Preprocessing phase is similar like word level lstm model. Firstly i filtered raw data i.e removing punctuation etc. then i generate sequences using characters, here i took sequence size to 50. then i split the corpus into training(80 percentage) and testing(20 percentage) part.
- **Model Parameters:** I used RMSprop optimizer in keras and cross entropy loss function. i used one files as previously mentioned. and batch size i took 128 with 80 epochs.
- **Generating Sentences:** I generated sentences with help of this this model by taking seed text of 80 characters and predicts 51st character. i generated sentence of 10 words like this.

4 Result

- **Results for word level LSTM model:**
  Loss: 7.3125
  Perplexity: 158.9577
- **Results for character level LSTM model:**
  Loss: 8.0143
  Perplexity:258.5500
- **Generated sentence using word level LSTM model:**
  1. observation his wife and the whole of the house and the whole
  2. to convince her that she had been so much as possible to
  3. comforts and the whole of the house and the whole of the
  4. the hero of a man of the house and the whole of
  5. even that she had been so much as possible to say that

5 Git Hub Link

https://github.com/Divyagarwalfeb/NLU-project.git