Project Final Presentation

Semi-supervised Sequence Learning

Sabiha Sultana[1], Piyakorn Munegan[2],
Sanjay Kanakkot Viswanathan[3],
Mohammed Rizwan Amanullah[4]

Department of Computing,
Macquarie University

November 2, 2021

Github Repository: https://github.com/sanjay-kv/Semi-supervised-sequence-learning-Project.

Recap - Semi-supervised Sequence Learning Project

- Goal: Replicating project which aims on unlabelled data to improve sequence learning with Recurrent network
- (LM-LSTM) Language modelling. Predict what comes next in sequence.
- (SA-LSTM) Reads input sequence into vectors and predict the input sequence again.
- Requirements: AWS(EC2) with Python 3, tensorflow v 1.15.5
- Source: https://arxiv.org/pdf/1511.01432v1.pdf
- Source Repo: https://github.com/tensorflow/models/tree/master/research/adversarial
- Dataset:
  - IMDB movie review dataset
  - DBpedia dataset (DBpedia abstracts and its categories)

Recap - Replication Process

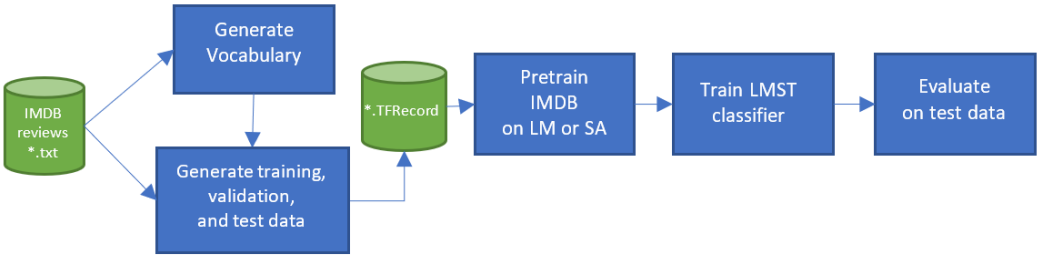

Figure 1: Explains the flow of replication for Sentiment Analysis.

IMDB - Data Generation

The Process

1. Created Python Script.
2. Scrapped the data in 7 categories.
3. BeautifulSoup python package.

Used python script to scrap the recent reviews from IMDB Website. Scrapped 15% of original data-set, latter manually annotated 2000 file, among them as positive and negative sentiment category and 1000 treated as unlabelled data.

Table 1: A summary of total data scrapped using Beautifulsoup from IMDB Website.

Dataset	Labelled	Unlabelled
IMDB	2000	1000

- Repository: https://github.com/sanjay-kv/Semi-supervised-sequence-learning-Project/tree/main/imdb-review-scrapping

Amazon New Data Generation

- Scrapped amazon.in website using selenium.
- We have extracted product listing and product information for ten categories.
  - After extraction, we are manually tagging the product listing.

Table 2: A summary of total data scrapped using Beautifulsoup from IMDB Website.

Dataset	**Scrapped Listings**
Amazon	3154

- Repository: https://github.com/sanjay-kv/Semi-supervised-sequence-learning-Project/tree/main/amazon-scrapping

Dataset for the sentiment classification task

Dataset

1. Original IMDB movie review dataset.
   - Training set: 25,000 labeled and 50,000 unlabeled reviews.
   - Test set: 25,000 labeled reviews.

2. Replication of 5% the size of the original IMDB dataset.
   - Training set: 1,111 labeled and 1,111 unlabeled reviews.
   - Test set: 1,111 labeled reviews.

3. Replication of new IMDB dataset.
   - Training set: 1,000 labeled and 1,000 unlabeled reviews.
   - Test set: 1,000 labeled reviews.

Dataset for the product classification task

Dataset

1. Dataset: DBpedia dataset.
   - Training set: 560,000 labeled examples.
   - Test set: 70,000 examples.

2. Replication of 2% the size of the original DBpedia dataset.
   - Training set: 11,200 labeled examples.
   - Test set: 1,400 examples.

3. Replication of new Amazon dataset
   - Training set: 2,795 examples.
   - Test set: 395 examples.

Pre-training Process

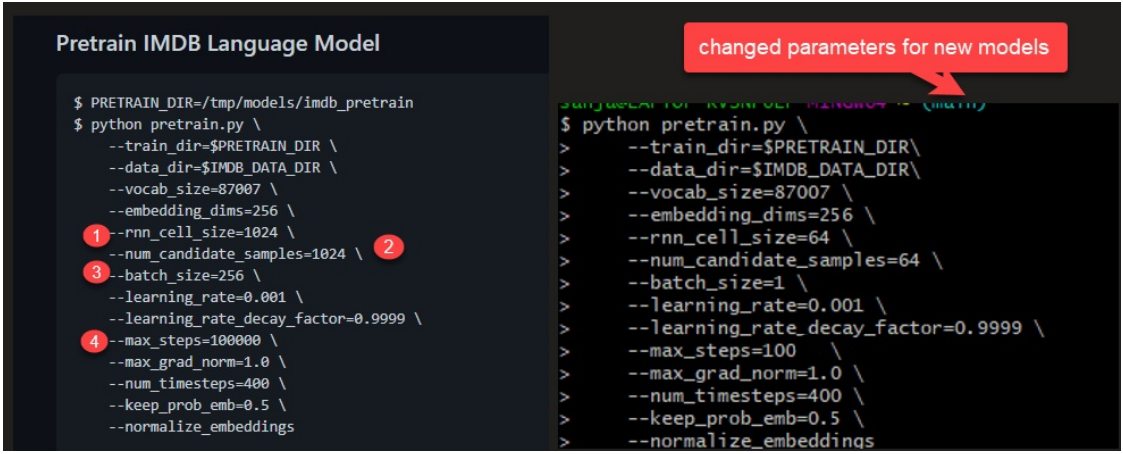

Figure 2: On left the Default parameter, On right Parameters we used for replication

Evaluation Results

Table 3: Summary of the accuracy scores of models on the IMDB sentiment classification task

Dataset	**LM-LSTM**	**SA-LSTM**	**LSTM**
Original IMDB dataset (original work)	92.36	92.76	86.5
5 % of the original IMDB dataset	57.6	56.8	-
New IMDB dataset	50	50	50

Table 4: Summary of the accuracy scores of models on the product classification task

Dataset	**LM-LSTM**	**SA-LSTM**	**LSTM**	**SVM**
Original DBpedia dataset (original work)	98.5	97.66	86.36	-
2 % of the original DBpedia dataset	7.1	7.1	7.1	-
New Amazon dataset	9.4	9.4	9.4	4.54

Pretraining: LM = Language Model, and SA = Sequence Autoencoder

Github Repository: https://github.com/sanjay-kv/Semi-supervised-sequence-learning-Project.

Findings/ Conclusion

The Process

1. In sentiment Analysis we were able to replicate and validate the accuracy.
2. Impact of less data has been observed in the obtained accuracy of the classification task of amazon dataset

We Ensured for NLP task can use LSTM recurrent network, along with it the team were able to find and validate sequence auto-encoder can stabilize the LSTM learning process. LSTM Outperformed all other evaluation categories.

The End