CS6001 DATA MINING

PROJECT REPORT

Abstractive Synopsis Generation

Reuel Samuel Sam	**2018103053**
Sayf Zakir Hussain	**2018103059**

GITHUB Link for Codes: https://github.com/ReuelSam/DataMiningProject

Base Paper:	Adelia, Rike, Suyanto Suyanto, and Untari Novia Wisesty. "Indonesian abstractive text summarization using bidirectional gated recurrent unit." *Procedia Computer Science* 157 (2019): 581-588.

I. Introduction

This project focuses on the generation of a synopsis for a given passage. To do so, this project attempts to adopt the important Natural Language Processing task of Text Summarization. Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning.

Text Summarization follows 2 methodologies:

- Extractive Summarization
  - The important phrases and sentences are extracted as is from the text
- Abstractive Summarization
  - New sentences are generated as the summary for this given text

This project attempts to provide abstractive text summarization for news headlines taken from the InShort dataset.

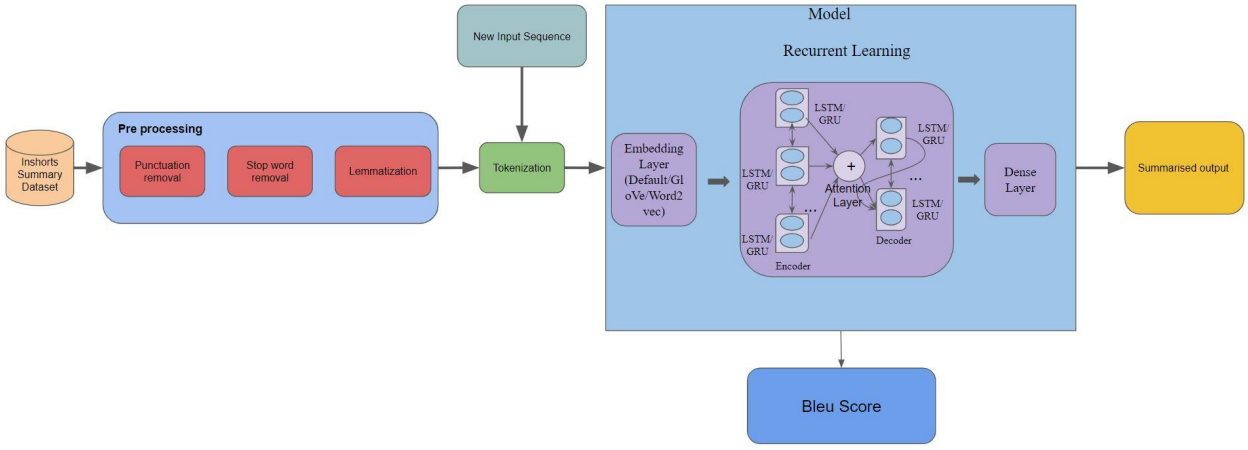
II. Objective

The main aim of this project is to provide Abstractive Text Summarization for news articles through the use of LSTM based Encoder and Decoder structure.

For training purposes, the InShorts News Article Dataset has been identified. Inshorts is a news service that provides short summaries of news from around the web. This dataset contains headlines and summary of news items along with its source. This dataset has 55104 different articles and their headlines that have been used for training purposes.

Dataset Link: https://www.kaggle.com/shashichander009/inshorts-news-data

III. Overall Block Diagram



GloVe:

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

Word2Vec:

Word2Vec is not a singular algorithm, rather, it is a family of model architectures and optimizations that can be used to learn word embeddings from large datasets. Embeddings learned through Word2Vec have proven to be successful on a variety of downstream natural language processing tasks.

LSTM:

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can process not only single

data points (such as images), but also entire sequences of data (such as speech or video). A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

GRU:

Gated recurrent units (GRUs) are a gating mechanism in recurrent neural networks. The GRU is like a long short-term memory (LSTM) with a forget gate, but has fewer parameters than LSTM, as it lacks an output gate. GRU's performance on certain tasks of polyphonic music modeling, speech signal modeling and natural language processing was found to be similar to that of LSTM.

Attention Layer:

An Attention Layer allows a model to look over all the information the original sentence holds, then generate the proper word according to the current word it works on and the context. It can even allow the model to zoom in or out (focus on local or global features).

IV. Models Used

1. Vanilla LSTM

A LSTM based auto encoder - decoder structure has been designed which takes the news article as inputs to generate its corresponding summary. Before feeding the data into the model, the data has been cleaned by performing different preprocessing strategies such as stop word removal and lemmatization. Word embeddings are learned while the model is trained.

Model Description:

- Embedding: Default Keras Embedding
- Number of LSTMs in Encoder: 3
- Number of LSTMs in Decoder: 1
- Optimizer: RMSProp
- Loss: Sparse Categorical Cross Entropy
- Learning Rate: 0.001 (Default)
- Epochs: 45
- Batch Size: 512

2. LSTM with Glove embeddings

A LSTM based auto encoder - decoder structure has been designed which takes the news article as inputs to generate its corresponding summary. A pre-trained 300 dimensional Glove model is used to generate the embedding matrix which is then used while training both the encoder and decoder structure. Before feeding the data into the model, the data has been cleaned by performing different preprocessing strategies such as stop word removal and lemmatization.

Model Description:

- Embedding: Glove 300 Dimensions trained on Common Crawler 42B
- Number of LSTMs in Encoder: 3
- Number of LSTMs in Decoder: 1
- Optimizer: RMSProp
- Loss: Sparse Categorical Cross Entropy
- Learning Rate: 0.001 (Default)
- Epochs: 30
- Batch Size: 512

3. LSTM with attention and Glove embeddings

A LSTM based auto encoder - decoder structure has been designed which takes the news article as inputs to generate its corresponding summary. A pre-trained 300 dimensional Glove model is used to generate the embedding matrix which is then used while training both the encoder and decoder structure. Additionally, an attention layer is added between the encoder and decoder outputs so that the model knows which parts of the encoder outputs it should pay more attention to. Before feeding the data into the model, the data has been cleaned by performing different preprocessing strategies such as stop word removal and lemmatization.

Model Description:

- Embedding: Glove 300 Dimensions trained on Common Crawler 42B
- Number of LSTMs in Encoder: 3
- Number of LSTMs in Decoder: 1
- Optimizer: RMSProp
- Loss: Sparse Categorical Cross Entropy
- Learning Rate: 0.001 (Default)
- Epochs: 100
- Batch Size: 512

4. LSTM with attention and word2vec embeddings

A LSTM based auto encoder - decoder structure has been designed which takes the news article as inputs to generate its corresponding summary. Pretrained Word2vec embeddings from the gensim module are used to generate the embedding matrix which is then used while training both the encoder and decoder structure. Additionally, an attention layer is added between the encoder and decoder outputs so that the model knows which parts of the encoder outputs it should pay more attention to. Before feeding the data into the model, the data has been cleaned by performing different preprocessing strategies such as stop word removal and lemmatization.

Model Description:

- Embedding: Word2vec embeddings from gensim
- Number of LSTMs in Encoder: 3
- Number of LSTMs in Decoder: 1
- Optimizer: RMSProp
- Loss: Sparse Categorical Cross Entropy
- Learning Rate: 0.001 (Default)
- Epochs: 150
- Batch Size: 512

5. Vanilla GRU

A GRU based auto encoder - decoder structure has been designed which takes the news article as inputs to generate its corresponding summary. Before feeding the data into the model, the data has been cleaned by performing different preprocessing strategies such as stop word removal and lemmatization. Word embeddings are learned while the model is trained.

Model Description:

- Embedding: Default Keras Embedding
- Number of LSTMs in Encoder: 3
- Number of LSTMs in Decoder: 1
- Optimizer: RMSProp
- Loss: Sparse Categorical Cross Entropy
- Learning Rate: 0.001 (Default)
- Epochs: 30
- Batch Size: 512

6. GRU with Glove embeddings

A GRU based auto encoder - decoder structure has been designed which takes the news article as inputs to generate its corresponding summary. A pre-trained 300 dimensional Glove model is used to generate the embedding matrix which is then used while training both the encoder and decoder structure. Before feeding the data into the model, the data has been cleaned by performing different preprocessing strategies such as stop word removal and lemmatization.

Model Description:

- Embedding: Default Keras Embedding
- Number of LSTMs in Encoder: 3
- Number of LSTMs in Decoder: 1
- Optimizer: RMSProp
- Loss: Sparse Categorical Cross Entropy
- Learning Rate: 0.001 (Default)
- Epochs: 45
- Batch Size: 512

7. GRU with word2vec embeddings

A GRU based auto encoder - decoder structure has been designed which takes the news article as inputs to generate its corresponding summary. Pretrained Word2vec embeddings from the gensim module are used to generate the embedding matrix which is then used while training both the encoder and decoder structure. Before feeding the data into the model, the data has been cleaned by performing different preprocessing strategies such as stop word removal and lemmatization.

Model Description:

- Embedding: Word2vec embeddings from gensim
- Number of LSTMs in Encoder: 4
- Number of LSTMs in Decoder: 3
- Optimizer: RMSProp
- Loss: Sparse Categorical Cross Entropy
- Learning Rate: 0.001 (Default)
- Epochs: 100
- Batch Size: 512

V. Performance Metric

For this project, we have adopted the BLEU Score performance metric for evaluating the performance. BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate translation of text to one or more reference translations. Although developed for translation, it can be used to evaluate text generated for a suite of natural language processing tasks. Ideally, the Perfect BLEU Score is given by 1.

$$Bleu = \rho . e^{\sum_{n=1}^{n} (\frac{1}{N} * log \rho n)}$$

where,

ρ = Brevity Penalty given by:

$$\rho = 1 \; if \; c > r \; else \; e^{(1-\frac{r}{c})}$$

VI. Results

The Bleu score for each of the models have been recorded and tabulated below:

Model	**Bleu Score**
Model 1: LSTM Vanilla	0.5504
Model 2: LSTM with GloVe	0.6016
Model 3: LSTM with GloVe and Attention	0.7254
Model 4: LSTM with Word2Vec and Attention	0.7481
Model 5: GRU Vanilla	0.5320
Model 6: GRU with GloVe	0.6259
Model 7: GRU with Word2Vec	0.6208

As seen in the above table, Model 4 that makes use of an LSTM based Encoder-Decoder Structure along with an Attention Layer performs the best with a Bleu score of 0.7481.

For this model, some example summaries are provided:

```
print("Review:",seq2text(padded_xtest[13]))
print("Original summary:",seq2summary(padded_ytest[13]))
print("Predicted summary:",decode_sequence(padded_xtest[13]))
```

```
Review: rafael benitez appoint new manager relegation threaten premier league outfit newcastle united magpies sack previous manager steve mclaren early
today benitez sack real madrid january seven month charge sign three year deal newcastle previously manage liverpool inter milan napoli valencia
Original summary: rafael benitez hired as newcastle united manager
Predicted summary:  rafael benitez hired as newcastle united manager
```

```
print("Review:",seq2text(padded_xtest[25]))
print("Original summary:",seq2summary(padded_ytest[25]))
print("Predicted summary:",decode_sequence(padded_xtest[25]))
```

```
Review: mall china 39 s shanxi create rooster statue resemble us president elect donald trump small replicas statue available sale price start 3 800 11
lakh 32 foot tall statue notably new year china begin january 28 year rooster accord chinese zodiac
Original summary: china mall installs 39 donald trump like 39 rooster statue
Predicted summary:  china mall installs 39 donald trump statue 39 china
```

```
print("Review:",seq2text(padded_xtest[48]))
print("Original summary:",seq2summary(padded_ytest[48]))
print("Predicted summary:",decode_sequence(padded_xtest[48]))
```

```
Review: usa become first nation register gold medal rio olympics 2016 19 year old virginia thrasher gold 10 metre air rifle day one event hold saturday
thrasher finish ahead china 39 s du li yi sile silver bronze medal open china 39 s medal tally event
Original summary: usa wins first gold medal of rio olympics 2016
Predicted summary:  usa wins first gold medal at rio olympics
```

NOTE: The code and results for all the models can be found at the following Github Link: https://github.com/ReuelSam/DataMiningProject