CS617

Deep Learning Project

(Text Summarization)

Group 06

- Arpit Jain | 2015047
- Gautam Yadav | 2015093
- Narosenla Longkumer | 2015165

2018 International Conference on Robots & Intelligent System

Extractive Document Summarization Based on Hierarchical GRU

Yong Zhang, Jinzhi Liao, Jiuyang Tang, Weidong Xiao, and Yuheng Wang
National University of Defense Technology, Changsha, Hunan, 410072, China
495289761@qq.com

A Hierarchical Model for Text Autosummarization

Zhenpeng Zhou
Stanford University
Stanford, CA 94305
zhenpeng@stanford.edu

Abstract

Summarization is an important challange in natural language processing. Deep learning methods, however, have not been widely used in text summarization, although neural networks have been proved to be powerful in natural language processing. In this paper, an encoder-decoder neural network model is applied to text summarization, as an important step toward this task. Besides, a hierarchical model, which builds the sentence representations and then paragraph representations, enables the summarization for long documents.

https://machinelearningmastery.com/gentle-introduction-text-summarization/

Dataset

News Article Data

The data we will use here is the 'all-the-news'-dataset from Kaggle. It contains about 200000 news articles and the headlines of those articles. The headlines will serve as our summaries in this case. The articles are from several big news corporations.

https://www.kaggle.com/snapcrack/all-the-news

Content

- articles1.csv - 50,000 news articles (Articles 1-50,000)
- articles2.csv - 49,999 news articles (Articles 50,001-100,00)
- articles3.csv - Articles 100,001+

Pre-processing

- Lowercasing the entire text.
- Collapsing the text into a single line.
- Fixing the presence of back-slash escape characters.
- Substituting special characters and numeral with spaces.
- Reducing multiple spaces into a single space.
- Removing single letter words.
- Tokenization

Word Embeddings

- Pre-trained tensor flow hub embeddings.
- Token based text embedding trained on English Wikipedia corpus.

https://tfhub.dev/google/Wiki-words-250/1

Model

Sequence to Sequence Model

Model

- The basic RNN model we are going to use is seq2seq with LSTM celled layers.

(Basic Structure)

- Encoder-Decoder
- Attention Mechanism
- Cyclic Learning rate
- Beam Search (Using Beam size = 5)

Model Reference: thomasschmied (GitHub)

Parameters

Sequence to Sequence Model

Model hyperparameters

- num_layers_encoder = 4
- num_layers_decoder = 4
- rnn_size_encoder = 300
- rnn_size_decoder = 300

- batch_size = 32
- epochs = 100
- clip = 5 (Gradient Clip)
- keep_probability = 0.8

- learning_rate = 0.0005
- max_lr=0.005
- learning_rate_decay_steps = 100
- learning_rate_decay = 0.90

Results

(Our Model until Mid-Sem evaluation)

25 Epochs

Problems:

- Repetitive Terms.
- Large number of Unknowns <UNK>.
- Rouge Metric not implemented.
- Very poor summaries.

```
----------------------------------------------------------------------------------------------------
Actual Text:
in a major abortion ruling monday the supreme court <UNK> down <UNK> of a texas law that would have <UNK> dozens of <UNK> to
ose here are reactions from all <UNK> of the <UNK>

Actual Summary:
reactions to the supreme court ruling on texas abortion law

Created Summary:
reactions to the supreme court ruling ruling texas abortion abortion
```

```
----------------------------------------------------------------------------------------------------
Actual Text:
cnn president trump s travel ban will remain <UNK> a federal appeals court <UNK> thursday read the court s ruling below

Actual Summary:
full text <UNK> <UNK> rules against <UNK> travel ban

Created Summary:
full text <UNK> <UNK> rules against against travel
```

How we fixed these problems?

Changes

- Trained for increased number of Epochs (20,100)
- We were initializing our dictionary based on Minimum Occurrences of that particular word being 2 and labeling rest of the words as unknown in order to make our dictionary more generalized. We changed the Minimum occurrences to 1.
- Used Rouge Metric of Evaluation (Rouge-1, Rouge-2, Rouge-L)

Rouge Metric

To determine prediction accuracy

ROUGE stands for **Recall-Oriented Understudy for Gisting Evaluation.**

It works by comparing an automatically produced summary or translation against a set of reference summaries (typically human-produced). It works by matching overlap of n-grams of the generated and reference summary.

Rouge 1 : Refers to the overlap of 1 word (1-gram)
Rouge 2 : Refers to the overlap of 2 word (Bigram)
Rouge L : Longest Common Subsequence (LCS) based statistics. (Almost sentence structure level)

Results

100 Epochs

Early Stopping at around 85 Epochs

```
-----------------------------------------------------------------------------------------------------
Actual Text:
the senate voted to confirm former exxon ceo rex tillerson as president donald trump secretary of state wednesday afternoon <UNK> to 43 advertisement

Actual Summary:
senate confirms rex tillerson as secretary of state

Created Summary:
senate confirms rex tillerson as secretary state

Rogue-score: {'rouge-1': {'f': 0.9333333283555556, 'p': 1.0, 'r': 0.875}, 'rouge-2': {'f': 0.7692307642603551, 'p': 0.8333333333333334, 'r': 0.7142857142857143}, 'rouge-l': {'f': 0.9251461988300539, 'p': 1.0, 'r': 0.875}}
```

```
-------------------------------------------------------------------------------------------------------------
Actual Text:
new york ap many people have heard of twitter not enough of them are signing up to use it advertisement

Actual Summary:
celebrity megaphone fails to lure ordinary users to twitter

Created Summary:
new megaphone fails to lure ordinary users to twitter

Rogue-score: {'rouge-1': {'f': 0.874999995, 'p': 0.875, 'r': 0.875}, 'rouge-2': {'f': 0.874999995, 'p': 0.875, 'r': 0.875}, 'rouge-l': {'f': 0.8749999999995, 'p': 0.875, 'r': 0.875}}
```

Results

100 Epochs

Early Stopping at around 76 Epochs

```
Average Rouge (Recall-Oriented Understudy for Gisting Evaluation) Scores:

ROUGE-1:

Precision:  0.9821428571428572
Recall:  0.8746698671698673
F-Score:  0.9215101332729613

ROUGE-2:

Precision:  0.7872246272246272
Recall:  0.764080549080549
F-Score:  0.7747353012680038

ROUGE-L:

Precision:  0.9773809523809525
Recall:  0.8709661634661636
F-Score:  0.9050264758390452
```

Some Challenges

(Which still persist)

- Some terms are still repeating.

```
-----------------------------------------------------------------------------------------------------
Actual Text:
business advocates who want to import more foreign consumers and more foreign workers are developing plans to counter president donald trump popular call for immigration reform advertisement

Actual Summary:
gop senators hope to sneak amnesty into trump popular immigration reforms

Created Summary:
gop senators hope to sneak amnesty to popular popular reforms reforms
```