Abstractive Text Summarization

Prakash Nath Jha
2018201013

Nitish Srivastava
2018201012

Introduction:

Text summarization is the task of reducing a text, such as a sentence, article, to a shorter version that retains the core ideas of the original text. Before 2014, the majority of work regarding text summarization made use of the extractive approach with relative success.

After the development of sequence to sequence models, abstractive summarization (based on seq2seq models) became widespread, competing with extractive models for state-of-the-art results. In this project we propose to build a RNN based Seq2seq model using attention mechanism for the task of abstractive text summarization.

Literature Survey:

For the task of text summarization there are broadly two main approches:

- **Extractive methods** which build summaries from selecting and compressing input text sequences. Without actually understanding the sentences, it acts as a highlighter for the document.
- **Abstractive methods** which build summaries more like a human would, by creating new text passages containing words selected from a wider vocabulary.

Most of the abstractive techniques for text summarization differ in one of these three categories:

- Network/Model structure
- Parameter inference
- Decoding/generation

Following are the state-of-the-art models proposed for the task of neural based text summarization:

- RNN based Seq2seq model with attention
- CNN based Seq2seq model with attention like Quasi-Recurrent Neural Network (QRNN)
- Pointing/Copying Mechanism based Seq2seq model
- Reinforcement Learning Approaches based model

Different types of attention mechanism has been proposed which is used for text summarization:

- Hierarchical Attention
- Discourse-Aware Attention
- Coarse-to-Fine Attention
- Graph-based Attention

Baseline Model:

We propose to implement a baseline RNN based Sequence-to-Sequence deep learning model with attention for the task of text summarization. The model consists of two parts, an encoder, that understands the input, and represent it in an internal representation, and feed it to another part of the network which is the decoder.

Here we intend to use Long Short Term Memory (LSTM) as our building block for constructing RNN based aggregator model for Encoder as well as Decoder component in our architecture.

In the Encoder we mainly use a multi-layer bidirectional LSTM, while in the decoder we use attention mechanism.

Proposed Methodology:

This project is divided into two main components:

- Building text summarization model:

  We propose to implement abstractive text summarization model and for this purpose we have choosen to implement the model presented in the following reasearch paper

  Bidirectional Attentional Encoder-Decoder Model and Bidirectional Beam Search for Abstractive Summarization
  Link: https://arxiv.org/pdf/1809.06662v1.pdf

- Building a web based interface to interact with the model

  We will provide an interface using which user can provide the document text to summarize and will receive a summary as an output.

Dataset:

We are propose to use CNN/Daily Mail dataset in form of news and their headers, the news body is used as the feature for our model, while the header would be used as the summary target output.

There are two possible ways:

1. using the raw data itself, and manually applying processing on them
2. using a preprocessed version for the data

We intend to use preprocessed data available at https://github.com/abisee/cnn-dailymail for building our model.