Faculty of Computers and
Information Sciences -Computer
Science Dept.

Auto-summarization

By:

Lama Tarek Abbas Moustafa
Emad Magdy Telmeez
Kareem Mohammed Shaaban
Abdelrahman Mohamed Misbah
Faten Elsaid Mohamed Elbialy

2022/2023

1. Project idea

Text Summarization refers to the technique of shortening long pieces of text while capturing its essence. This is useful in capturing the bottom line of a large piece of text, thus reducing the required reading time. In this context, rather than relying on manual summarization, we can leverage a deep learning model built using an Encoder-Decoder Sequence-to-Sequence Model to construct a text summarizer.

In this model, an encoder accepts the actual text and summary, trains the model to create an encoded representation, and sends it to a decoder which decodes the encoded representation into a reliable summary. As the training progresses, the trained model can be used to perform inference on new texts, generating reliable summaries from them. This can help users save time when processing large amounts of information, such as news or research articles. Additionally, it can be used to automatically generate summaries for social media posts or other short-form content, which can be useful for quickly conveying key information to readers. However, the task of text summarization is challenging due to the complexity and variability of natural language, as well as the need to capture the key ideas and concepts expressed in the original text. Therefore, developing effective approaches to this problem is an active area of research.

2. Software

We implemented a neural network model for extractive text summarization. The idea behind the design of this model is to enable it to process input where we do not constrain the length. One RNN will be used as an encoder, and another as a decoder. The output vector generated by the encoder and the input vector given to the decoder will possess a fixed size. However, they need not be equal. The output generated by the encoder can either be given as a whole chunk or can be connected to the hidden units of the decoder unit at every time step.

Our model architecture consists of Encoder, the input length that the encoder accepts is equal to the maximum text length which you've estimated, In the decoder, an embedding layer is defined followed by an LSTM network. The initial state of the LSTM network is the last hidden and cell states taken from the encoder. The output of the LSTM is given to a Dense layer wrapped in a TimeDistributed layer with an attached SoftMax activation function.

3. Dataset

- **News Summary (Arabic dataset)**

4. Related Papers

1."Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond" by R. Nallapati, F. Zhai, and B. Zhou (2016):

In this paper, the author proposes a sequence-to-sequence framework for abstractive text summarization that utilizes a novel intra-attention mechanism to align the input and output sequences. the model incorporates both word-level and sentence-level attention mechanisms to better capture the semantic relationships between words and sentences in the input document. he also introduces a hybrid pointer-generator network that can copy words from the input sequence, enabling the model to generate rare or unseen words. His experimental results demonstrate that the model outperforms state-of-the-art abstractive and extractive summarization methods on the Gigaword and DUC datasets. He also present human evaluations that show that his system-generated summaries are more informative, readable, and enjoyable to read than those produced by other systems.

2.”SciBERTSUM: Extractive Summarization for Scientific Documents” by Athar sefid, C Lee Giles:

They created an extractive summarization framework, SciBERTSUM, based on BERTSUM for long documents with multiple sections (e.g. scientific papers). They generate sentence vectors based on their sections. The section information is important for the summarization task since sentences in the abstract or method sections are more important compared to the acknowledgement parts. To build a computationally efficient model that scales linearly with the number of sentences in the document, we employed the sparse attention mechanism of LongFormer to embed the inter sentence relations. All sentences attend to a limited number of sentences before and after the current sentence and only a small number of random sentences attend globally to all other sentences. The model is computationally efficient and improves the ROUGE scores on the dataset of paper-slide pairs.

3. "A Deep Reinforced Model for Abstractive Summarization" by Romain Paulus, Caiming Xiong, Richard Socher:

It presented a new model and training procedure that obtains state-of-the-art results in text summarization for the CNN/Daily Mail, improves the readability of the generated summaries and is better suited to long output sequences. They also run the abstractive model on the NYT dataset for the first time. They saw that despite their common use for evaluation, ROUGE scores have their shortcomings and should not be the only metric to optimize on summarization model for long sequences. their intra-attention decoder and combined training objective could be applied to other sequence-to-sequence tasks with long inputs and outputs, which is an interesting direction for further research.

2. Teammates

1. Lama Tarek Abbas Moustafa - Data collection and preprocessing.

2. Emad Magdy Telmeez - Creating the Model architecture.

3. Kareem Mohammed Shaaban - Training the Model and Generating

Predictions.

4. Faten Elsaid Mohamed Elbialy - Writing the report.

5. Abdelrahman Mohamed Misbah - Writing the proposal.