Logo, company name

Description automatically generated

**CSE523 Machine Learning**

**Prof. Mehul Raval**

**Weekly report**

**Group number: 17**

**Group name: The Mandelbrot set**

| **Name** | **Enrolment Number** |
| --- | --- |
| Aastha Gaudani | AU2040032 |
| Khushi Patel | AU2040068 |
| Devyash Shah | AU2040152 |
| Simran Khoja | AU1910606 |

**Report**

Now in our project with text classification we are planning to do text summarization. For that we have taken reference of a new paper. This week we studied a research paper related to our topic to get more information on different and clearer ways of approaches for our project. Below is the summary of what we understand.

**Link of the paper** - <https://arxiv.org/pdf/1909.08089.pdf>

**Summary of the paper:**

**Introduction:**

The difficulties of summarizing lengthy papers are discussed in the paper since they are a growing problem in numerous professions, including law, medicine, and finance. They point out that conventional methods for extractive summarization sometimes rely on regional variables like sentence position and frequency, which may not be adequate for lengthy documents. The authors offer a novel strategy that combines global and local context information to find key lines in lengthy publications as a solution to this problem.

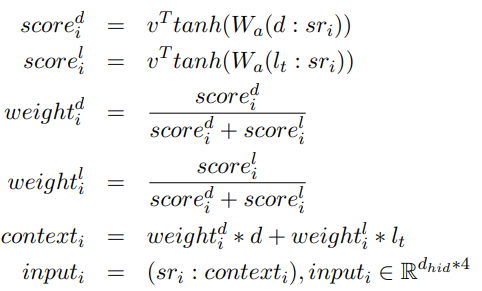
**Approach:**

The authors explain the extraction summary method they have developed. By devising a technique that takes both national and local context information into account, the authors hope to alleviate the difficulty of summarizing lengthy papers. In order to capture both the hierarchical structure of the sentences and the content similarity, they utilize a graph-based representation of the document. Then, they provide a cutting-edge scoring algorithm that identifies crucial sentences in the manuscript by fusing global and local context data. The efficiency of the suggested strategy for condensing lengthy texts is demonstrated through evaluation on two benchmark datasets and comparison to SOTA baselines.

According to the authors, the document can be visualized as a graph, with each node denoting a sentence and edges between nodes signifying similarity between sentences. Based on how connected each sentence is to other sentences in the graph, they utilize the TextRank algorithm to determine how important each statement is. This is comparable to the PageRank algorithm that Google uses to rank websites according to their significance.

The authors then suggest a novel scoring system that identifies crucial sentences in the manuscript by fusing local and global context data. The scoring mechanism takes into account the sentence's placement within the text, the number of named things it refers to, and how similar it is to other phrases in the text.

And the final summary is based on the importance of the sentences based on their scores.



**Sentence Encoder:**

The primary goal of the sentence encoder is to convert word embedding sequences into a fixed-length vector. Several methods, including RNN, CNN, and average word embedding are utilised for sentence embedding. After experimenting with all three methods, Kedzie et al. (2018) came to the conclusion that Word Embedding Averaging performs as well as or better than RNNs or CNNs for embedding sentences across various domains and summarizer architectures.

Since the sentence embedding is the mean of its word embeddings and se belongs to R demb, we utilize Average Word Embedding as our sentence encoder.

**Document Encoder:**

The document encoder is a technique for extractively summarizing lengthy documents that is discussed in the paper. The local and global context of the page is encoded using a hierarchical method. For each sentence in the document, the local context is encoded using a different bidirectional GRU network, while the global context is encoded using a bidirectional GRU network. The last hidden state of the global encoder and the last hidden state of the local encoder for each sentence are combined to create the final document embedding. In order to maximize the ROUGE score between the summary and the ground truth summary, the model is trained using reinforcement learning. The tests demonstrate that on benchmark datasets for summarization, the proposed document encoder beats a number of previous approaches.

**Decoder:**

To determine if a sentence belongs in the summary, the decoder stage combines the sentence representation, topic segment, and global context. Concatenation and attentive context are the two techniques under consideration for this combination. The vectors of these three factors are merely concatenated in concatenation. The attentive context approach determines the weight of each context vector by means of an attention mechanism. A multi-layer perceptron (MLP) with a sigmoid activation function is then employed as the input along with the sentence representation for each phrase's weighted context vector. The confidence score for each sentence's selection is shown in the MLP output. The weighted negative log-likelihood is minimized during training to address the issue of excessively unbalanced data. The proposed strategy is assessed using automatic metrics like ROUGE and METEOR scores and contrasted with earlier works.



In the upcoming week we are planning to implement the text summarization.