Natural Language Processing Final Project: Transformer Summarizer

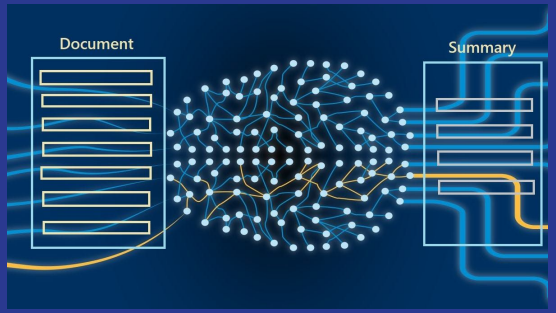
Team Members:

1.	Aditya Porwal	249645066
2.	Ankit Hiremath	462567778
3.	Manas Bhilare	881435179
4.	Samruddha Deshmukh	379991386
5.	Sahana Vasudevarao	976803547

Contents

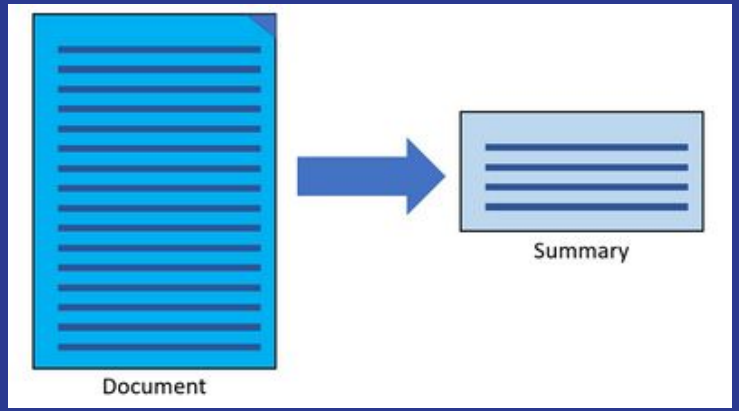
Introduction

- Text summarization involves creating a condensed version of a document while preserving its essential information, requiring natural language processing techniques to analyze semantics, context, and structure.
- Summarization algorithms identify and extract the core ideas within a text to generate a shortened summary capturing the essence of the original content, making it valuable for quickly digesting large volumes of text across various domains.
- Recent advances in deep learning, particularly transformer-based models like BERT and GPT, have significantly improved the performance of text summarization systems, achieving state-of-the-art results in abstractive summarization tasks.



Problem Statement

- 1. The exponential growth in daily published content has increased the demand for condensed article summaries to help people quickly grasp essential information amidst information overload.
- 2. Manual summarization is impractical due to the vast volume of content, necessitating the use of automatic methods and algorithms leveraging natural language processing and machine learning.
- 3. Key challenges in automated summarization include avoiding redundancy, maintaining coherence and fluency, and preserving the original tone and context of the source material, requiring sophisticated algorithms capable of nuanced comprehension and abstraction.



Significance

- This project aims to develop an abstractive text summarization model using multi-head attention transformers, which can generate concise summaries that capture the key information from long text documents.
- Abstractive summarization is a more advanced and challenging task compared to extractive summarization, as it requires understanding the semantic meaning of the text and generating new sentences.
- Transformer models with attention mechanisms have shown promising results for abstractive summarization by allowing the model to focus on the most relevant parts of the input text during summary generation.

Essential Processes in Text Summarization

1. Document parsing: The input document is obtained and converted into a suitable format for processing.

2. Preprocessing: Text is tokenized, stop words are removed, and stemming is applied to refine the data.

3. Lexical chain construction: Semantic relationships between words and phrases are identified to build lexical chains representing key concepts.

4. Strongest chain identification: The most salient and coherent lexical chain is selected as the basis for sentence extraction.

5. Sentence selection and assembly: Relevant sentences associated with the strongest chain are selected and assembled to form the summary.

6. Summary generation: The selected sentences are compiled into a concise narrative, effectively conveying the main ideas of the original document.

Note: We have proceeded with CNN/Daily Mail dataset as It is a popular benchmark dataset for abstractive text summarization.

Methodology

- We implemented abstractive text summarization technique, which generates entirely new text to summarize key concepts from the source text.
- Our proposed methodology integrates Transformer-based language models with attention mechanisms for enhanced abstraction.
- Key components include dot product attention, facilitating the model's focus on relevant input segments, and causal attention, restricting attention to past text for summarization tasks.
- The Transformer decoder block incorporates layers such as normalization, dense layers, and dropout layers to comprehend intricate relationships and generate accurate summaries.
- Training involves defining the cost function, optimizer, and implementing a training loop with Trax to tokenize and calculate probability distribution over vocabulary .
- This comprehensive approach leverages deep learning techniques to achieve state-of-the-art outcomes in abstractive text summarization.

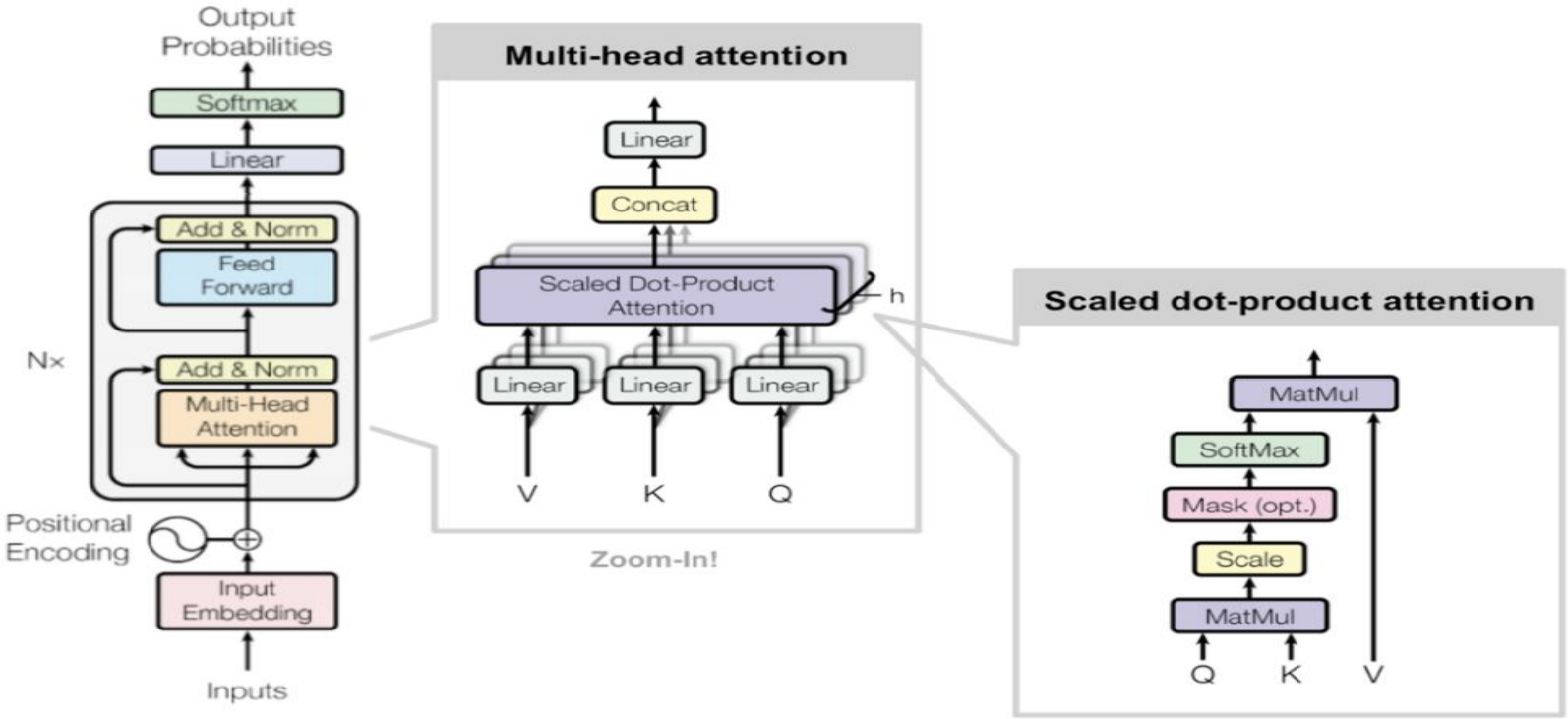
Attention Mechanism

- Attention mechanism in transformers helps the model to focus on certain text embeddings from the input text to better capture the context, and improving performance on text generation.
- Dot product attention in Transformers computes similarities between query and key vectors to focus on relevant parts of the input for generating summaries.
- Causal attention in Transformers focuses only on previous words in sequence tasks, using three specialized functions provided by Trax to compute and apply attention scores for text summarization.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)$$

d_k stands for the dimension of queries and keys.
Q stands for Query.
K stands for Key.
V stands for Value.
M stands for Mask.

Transformer-based language model



Source: 'Attention is all you need' Ashish Vaswani (2017)

Results

- Given a starting prompt, generate text using greedy decoding method.
- Greedy decoding works by selecting the most likely next word or token at each step in the text generation process.
- Iteratively, appends next word in the sequence until some stopping criterion is met.
- Our model was tested out on various articles, using greedy decoding to generate the text output.

```
[ ] # Test it out with a whole article!
    article = "It's the posing craze sweeping the U.S. after being brought to fame by skier Lindsey Vonn, soccer star Omar Cummings, baseball player Albert Pujols - and even Republ
    print(wrapper.fill(article), '\n')
    print(greedy_decode(test_sentence, model, 50, show_progress=True))
```

```
It's the posing craze sweeping the U.S. after being brought to fame by
skier Lindsey Vonn, soccer star Omar Cummings, baseball player Albert
Pujols - and even Republican politician Rick Perry. But now four
students at Riverhead High School on Long Island, New York, have been
suspended for dropping to a knee and taking up a prayer pose to mimic
Denver Broncos quarterback Tim Tebow. Jordan Fulcoly, Wayne Drexel,
Tyler Carroll and Connor Carroll were all suspended for one day
because the 'Tebowing' craze was blocking the hallway and presenting a
safety hazard to students. Scroll down for video. Banned: Jordan
Fulcoly, Wayne Drexel, Tyler Carroll and Connor Carroll (all pictured
left) were all suspended for one day by Riverhead High School on Long
Island, New York, for their tribute to Broncos quarterback Tim Tebow.
Issue: Four of the pupils were suspended for one day because they
allegedly did not heed to warnings that the 'Tebowing' craze at the
school was blocking the hallway and presenting a safety hazard to
students.

GENERATING SUMMARY:
 The U.S. The New York year-year-yaer high-school was found player to-be warning-warning year-year-year-year-....49
```

Conclusion

- The exponential growth of digital data necessitates automated text summarization to efficiently manage overwhelming information volumes, as manual methods are impractical.
- Advancements in natural language processing and machine learning have enabled sophisticated algorithms to generate accurate and coherent summaries across diverse content types, marking significant progress in text summarization.
- Ongoing research collaboration between researchers, industry, and policymakers is crucial to address remaining challenges, enhance effectiveness, capture nuances, ensure accuracy, and uphold ethical considerations in text summarization, empowering informed decision-making in the big data era.

Future Work

- Utilize more powerful hardware (e.g., GPUs) to increase training speed and enable longer training, allowing the model to capture more language nuances and potentially improve performance.
- Explore techniques like hyperparameter tuning, different architectures, larger datasets, or additional data sources to enhance the model's accuracy and capabilities.
- Employ evaluation metrics such as ROUGE scores, which measure the similarity between generated and reference texts, providing a comprehensive assessment of the model's summarization performance.

References

Paperswithcode: https://paperswithcode.com/sota/abstractive-text-summarization-on-cnn-daily

Papers:

- Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond https://arxiv.org/pdf/1602.06023v5.pdf
- Attention is all you need: https://arxiv.org/abs/1706.03762
- Transformer Based Implementation for Automatic Book Summarization: https://ijisae.org/index.php/IJISAE/article/view/2421

Articles:

- Abstractive Text Summarization Using Transformers: https://medium.com/swlh/abstractive-text-summarization-using-transformers-3e774cc42453
- Get started with Google Trax for NLP: https://towardsdatascience.com/get-started-with-google-trax-for-nlp-ff8dcd3119cf

Online Courses:

- Coursera Deep Learning Specialization by Andrew Ng https://www.coursera.org/specializations/deep-learning