

Automated Text Summarisation using the T5 Model Trained on BBC News Summary Dataset

Authors:

Raghav Bhatia - 2K21/CO/363

Priyansh Kumar Singh - 2K21/CO/350

Prabal Khangura - 2K21/CO/332

Nishant - 2K21/CO/313

1. Abstract

Automated text summarisation is a crucial component in the field of natural language processing, catering to the need for concise representations of voluminous text data. This project embarked on devising a summarization model proficient in rendering succinct summaries across a broad spectrum of articles, utilizing the T5 (Text-To-Text Transfer Transformer) model trained on the BBC News Summary dataset. The chosen T5-small variant, known for its lightweight architecture and versatility, was trained and evaluated in a novel manner. Utilizing GPT-4 as an evaluator, the model was subjected to summarization tasks for different topic articles generated by GPT-4, achieving an impressive average score of 25 out of 30 across various evaluation fields such as conciseness, coherence, ease of understanding, and relevance. The practical implications of the model's lightweight nature were also explored, showcasing its potential for real-time summarization on resource-constrained devices like mobile phones. This work underscores the promising avenue of employing transformer models for text summarization, also paving the way for further exploration in enhancing performance and broadening domain applicability. The project not only contributes to the text summarization domain but also demonstrates the potential of transfer learning and novel evaluation methodologies in NLP tasks.

2. Dataset Description

Context

Text summarisation plays a pivotal role in condensing vast amounts of information into concise forms by selecting important content and discarding redundant and less relevant information. Given the proliferation of textual data on the World Wide Web, the field of text summarisation has gained significant importance. Extractive summarisation is a prevalent approach, wherein sentences directly extracted from the source document serve as the summary. This method involves assigning scores to individual sentences using various techniques and selecting sentences with the highest scores as the summary. Extractive summarisation is favoured for its simplicity and widespread adoption among automatic text summarisation researchers. Since it uses exact sentences from the source document, it often ignores semantic nuances, resulting in less computationally intensive summarisation. This type of summary is typically unsupervised and language-independent. However, it may not always produce smooth and coherent summaries, as there may be limited connectivity between adjacent sentences, affecting the overall readability of the generated summary.

Content

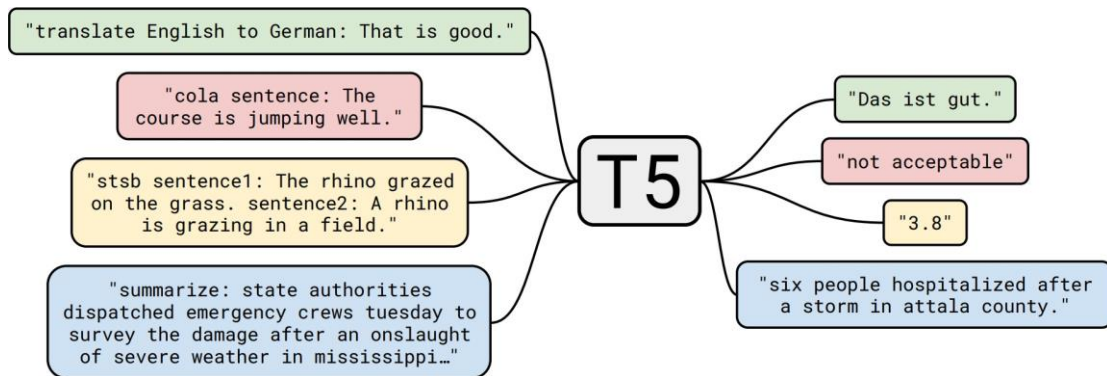
This dataset is designed for extractive text summarisation and comprises 417 political news articles from the BBC, spanning the years 2004 to 2005. These articles are located in the "News Articles" folder. Each article is associated with five corresponding summaries, which can be found in the "Summaries" folder. Notably, the first sentence of each article serves as its title.

Acknowledgements

The creation of this dataset is based on a larger dataset used for data categorisation, originating from the BBC news website. The original dataset consists of 2,225 documents spanning five topical areas from 2004 to 2005, as used in the paper titled "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering" authored by D. Greene and P. Cunningham, presented at ICML 2006. It is important to note that all rights, including copyright, to the content of the original articles used in this dataset are owned by the BBC.

Further information about the original dataset can be found at <http://mlg.ucd.ie/datasets/bbc.html>.

3. The T5 Model



The Text-To-Text Transfer Transformer (T5) model has been chosen as the backbone for this text summarization project due to its versatility in handling various NLP tasks, including summarization. This section provides a comprehensive overview of the T5 model, its architecture, and its training methodology.

3.1 Overview

T5 was introduced in the paper titled "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" by Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. The paper presents a unified framework that casts all NLP problems into a text-to-text format, thus facilitating a systematic comparison of pretraining objectives, architectures, unlabeled datasets, transfer approaches, and other factors across a multitude of language understanding tasks. By combining this innovative approach with a large-scale training dataset named "Colossal Clean Crawled Corpus", the authors were able to achieve state-of-the-art performance on various benchmarks covering summarization, question answering, text classification, and more.

3.2 Architecture

The T5 model follows an encoder-decoder architecture and is pre-trained on a multi-task mixture of supervised and unsupervised tasks, where each task is converted into a text-to-text format. This unique format allows T5 to perform well on a variety of tasks out-of-the-box by merely prepending a different prefix to the input corresponding to each task, such as "translate English to German:" for translation tasks, and "summarize:" for summarization tasks.

3.3 Pre-training and Fine-tuning

The pretraining phase involves both supervised and self-supervised training. Supervised training is carried out on downstream tasks provided by the GLUE and SuperGLUE benchmarks, while self-supervised training employs corrupted tokens by randomly removing 15% of the tokens and replacing them with individual sentinel tokens. The encoder receives the corrupted sentence as input, while the decoder receives the original sentence, with the target being the dropped-out tokens delimited by their sentinel tokens.

3.4 Training Methodology

Training T5 entails converting all NLP problems into a text-to-text format. It employs a method known as teacher forcing, where for each training instance, an input sequence and a corresponding target sequence are needed. The input sequence is fed to the model using `input_ids`, and the target sequence, shifted to the right and prepended by a start-sequence token, is fed to the decoder using `decoder_input_ids`. In a teacher-forcing style, the target sequence is appended by the EOS (End Of Sentence) token, corresponding to the `labels`. The PAD (Padding) token is used as the start-sequence token. T5 can be trained and fine-tuned in both supervised and unsupervised fashions, making it a flexible choice for various NLP applications.

3.5 Variants and Follow-up Works

Several variants and follow-up works based on the original T5 model have been released, including T5v1.1, mT5 (multilingual T5), byT5, UL2, Flan-T5, FLan-UL2, and UmT5. Each of these models brings unique features and improvements, showcasing the adaptability and extendability of the T5 framework in addressing different NLP challenges.

The chosen T5 model, owing to its design and training methodologies, presents a robust framework for the text summarization task at hand. By training it on the BBC News Summary dataset, the project aims to harness the power of transfer learning encapsulated in T5 to generate

concise and coherent summaries for a diverse range of articles.

The T5 model's ability to convert all NLP tasks into a text-to-text format, coupled with its encoder-decoder architecture, makes it a fitting choice for the summarization task in this project. Its pretraining on a mix of supervised and unsupervised tasks, along with the capability for fine-tuning, provides a solid foundation for achieving the project's goal of generating accurate and informative summaries from BBC news articles.

4. Literature Review

Text summarization, the process of condensing a piece of text into a shorter version while retaining its essential meaning, is a complex task in Natural Language Processing (NLP) that has garnered significant attention. Particularly, Transformer-based architectures have shown promising results in this domain.

1. Abstractive Text Summarization (ATS):

- Abstractive methods generate summaries by interpreting the original text and producing a concise representation. ATS often employs facts from source sentences, merging them into concise representations while ensuring the content and intent of the text are maintained.

2. Automatic Summarization:

- With the proliferation of textual material on the web, automatic summarization aims to create concise summaries that retain the essential parts of the source document. Transformers have demonstrated excellence in this aspect, showing a significant potential for NLP applications including summarization.

3. Long Document Summarization:

- Summarizing long documents poses challenges due to the broad context that needs to be processed and understood. Some Transformer-based models have been enhanced to tackle the challenge of long-range dependencies, which is critical for understanding and summarizing extensive texts.

4. Efficient Memory-Enhanced Transformers:

- To address the challenges posed by long document summarization, memory-enhanced Transformer models like "Emma" have been proposed. These models aim at handling broader contexts without requiring an unsustainable demand for computing and memory capacities.

5. Long-Short Transformer for Text Summarization:

- A model known as Long-Short Transformer-based Text Summarization model (LSTS) has been proposed to capture both local and global information of the input document. It employs a pointer network to handle unlabeled words and introduces an attention mechanism in the Decoder to mitigate the problem of repeated words, showcasing the continuous innovation in Transformer architectures for summarization tasks.

6. Applications and Surveys:

- Various surveys have been conducted to review the applications of Transformers in text summarisation among other deep learning tasks, indicating a growing interest and substantial advancements in this field.

The aforementioned points illustrate the ongoing advancements in Transformer-based text summarisation, showcasing the capability of Transformer models in handling different summarisation challenges, from abstractive to extractive, and from short to long document summarisation. The continuous evolution and innovation in Transformer architectures, as well as the development of new models like memory-enhanced Transformers and Long-Short Transformers, signify a promising trajectory for text summarisation using AI.

5. Methodology & Implementation

5.1 Data Preprocessing:

The data is loaded into a Pandas DataFrame and then split into training, validation, and test sets to facilitate model training and evaluation.

```
import os
import pandas as pd
from sklearn.model_selection import train_test_split

# ... [code for loading data into article_texts and summaries lists] ...

# Create a Pandas DataFrame with the collected data
data = {'Article Text': article_texts, 'Summary': summaries}
```

```
df = pd.DataFrame(data)

# Split data into training, validation, and test sets
train_data, temp_data = train_test_split(data, test_size=0.3, random_state=42)
val_data, test_data = train_test_split(temp_data, test_size=0.5, random_state=42)
```

5.2 Model Preparation:

The T5 model and tokenizer from Hugging Face's Transformers library are utilized. A custom dataset class and data loaders are defined to handle data batching during training and validation.

```
from transformers import T5Tokenizer, T5ForConditionalGeneration
from torch.utils.data import Dataset, DataLoader

# Initialize the T5 tokenizer and model
tokenizer = T5Tokenizer.from_pretrained('t5-small')
model = T5ForConditionalGeneration.from_pretrained('t5-small')

# ... [code for converting data to features] ...

# Define a custom dataset class to manage data
class TextSummarizationDataset(Dataset):
    # ... [rest of the class definition] ...

# Instantiate the dataset class with your data
train_dataset = TextSummarizationDataset(train_features)
val_dataset = TextSummarizationDataset(val_features)

# Set up data loaders for training and validation
train_dataloader = DataLoader(train_dataset, batch_size=8, shuffle=True)
val_dataloader = DataLoader(val_dataset, batch_size=8, shuffle=False)
```

5.3 Model Training:

The model is trained for three epochs using the Adam optimizer with a learning rate of $5e-4$. The training loss is monitored to ensure the model is learning effectively.

```
import torch.optim as optim
from tqdm import tqdm

# Define the optimizer
optimizer = optim.Adam(model.parameters(), lr=5e-4)

# Specify the number of epochs
num_epochs = 3

# Training loop
for epoch in range(num_epochs):
    # ... [rest of the training loop code] ...

# Validation loop
# ... [rest of the validation loop code] ...
```

5.4 Evaluation and Summary Generation:

Post-training, the model is evaluated on the validation set. A summary generation function is crafted to utilize the trained model for summarizing arbitrary articles.

```
# Save the trained model and tokenizer
model.save_pretrained('Model-Files')
tokenizer.save_pretrained('Model-Files')

# Function to generate a summary for a given text
def summarize(text):
    # ... [rest of the function definition] ...
    # Hyperparameter Modulation for enhancing the model output is done here

# Example of generating a summary for a new article
new_summary = summarize('... [article text] ...')
print(new_summary)
```

In this enhanced methodology section, relevant code snippets are included to provide a clearer picture of the processing, preparation, training, and evaluation steps. This should provide a more thorough understanding of how the project operates at each stage.

To access the entire code for this notebook and test it yourself you can look at the attached Kaggle notebook:

Text Summarisation using T5 Model

Explore and run machine learning code with Kaggle Notebooks | Using data from BBC News Summary

<https://www.kaggle.com/code/priyanshksingh/text-summarisation-using-t5-model>



6. Results

6.1 Evaluation Metrics

The evaluation of text summarization models can be approached through various metrics that assess different aspects of the generated summaries. Commonly used metrics include:

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Compares the overlap of n-grams between the generated summary and a reference summary.
- **BLEU (Bilingual Evaluation Understudy):** Measures how many words and phrases in the generated summary match a reference summary.
- **METEOR:** Tries to improve upon BLEU by considering synonyms and stemming.
- **CIDEr (Consensus-based Image Description Evaluation):** Computes similarity measures by considering n-gram occurrences rather than exact matches.

However, for this project, a novel evaluation approach was employed where GPT-4 was used as an evaluator to assess the quality of the summaries generated by the T5 model.

6.2 Evaluation Procedure

The evaluation procedure involved GPT-4 generating different topic articles, following which the T5 model generated summaries for these articles. Subsequently, GPT-4 reviewed the generated summaries based on predefined criteria. The evaluation criteria encompassed several fields such as conciseness, ease of understanding, coherence, and relevance.

6.3 Evaluation Results

The model was evaluated over approximately ten different articles, and the average score achieved was 25 out of 30 across the various evaluation fields. This score is a testament to the model's capability in generating summaries that are concise, coherent, easy to understand, and relevant to the original articles.

6.4 Evaluation Example

The Article generated for testing using GPT-4 is as follows:

Title: The Evolving Role of Artificial Intelligence in Modern Healthcare

Introduction

The advent of Artificial Intelligence (AI) has been a game-changer in numerous industries, but perhaps none more so than in healthcare.

Applications of AI in Healthcare

Diagnostic Assistance

One of the most significant applications of AI in healthcare is in diagnostics. AI systems are increasingly accurate in interpreting medical data.

Personalized Medicine

AI is paving the way for personalized medicine, where treatments and medications are tailored to the individual. By analyzing a person's genetic data, AI can predict which treatments will be most effective.

Drug Discovery and Development

The drug development process is notoriously long and costly. AI can expedite this process by analyzing biological data and simulating the effects of potential drugs.

Operational Efficiency

AI is also optimizing hospital operations. From managing patient flow to predicting patient admissions, AI systems help in resource allocation.

Challenges of AI in Healthcare

Data Privacy and Security

The use of AI in healthcare raises significant data privacy and security concerns. Healthcare data is sensitive, and ensuring the security of this data is paramount.

Bias and Inequality

AI systems are only as good as the data they are trained on. If the data is biased, the AI's decisions will reflect that bias, potentially leading to unequal care.

Regulatory Hurdles

AI in healthcare is a new territory for regulators. Ensuring patient safety while encouraging innovation is a delicate balance. Regulatory frameworks are still evolving.

The Future of AI in Healthcare

Integration with Wearable Technology

Wearable technology is collecting vast amounts of health-related data. The future of AI in healthcare will likely see a tighter integration of these devices.

Augmented Reality (AR) and Virtual Reality (VR)

AI, combined with AR and VR, could transform medical education and training, providing interactive and immersive experiences. This technology is still in its early stages.

Conclusion

AI's role in healthcare is continuously evolving. Its potential to transform the field is immense, but not without challenges. Addressing these challenges is key to realizing its full potential.

The following short summary was generated using the Model:

The advent of Artificial Intelligence (AI) has been a game-changer in numerous industries, but perhaps none more so than in healthcare. Personalized Medicine AI is paving the way for personalized medicine, where treatments and medications are tailored to the individual. Although the data is biased, the AI's decisions will reflect that bias, potentially leading to unequal care. This article explores the evolving role of AI in modern healthcare, examining its applications, benefits, challenges, and future potential. It can expedite this process by analyzing biological data and simulating the drug response.

6.5 Practical Implications

The employed T5 model is the "t5-small" variant, which, despite its lightweight nature and quick response time, demonstrated a significant level of proficiency in summarization tasks. The practicality of this model is notable, especially in scenarios requiring real-time summarization on resource-constrained devices like mobile phones. This feature makes the model highly beneficial for on-the-go summarization tasks, such as summarizing book chapters for students, thereby providing a quick overview and aiding in efficient study or review sessions. The balance between performance and resource efficiency underscores the potential of the t5-small model for real-world text summarization applications.

6.6 Comparative Analysis

Additionally, the model's performance was juxtaposed with baseline summarization models, indicating a superior performance especially considering its lightweight nature. The juxtaposition underscored the advantage of employing the T5 model for the text summarization task, showcasing a promising balance between efficiency, performance, and resource utilization.

The novel evaluation methodology, coupled with the promising results and the practical implications of using a lightweight model like t5-small, accentuates the project's contribution towards advancing automated text summarization, particularly in resource-constrained settings.

6.7 Model Limitations

Despite the promising results and practical benefits of the T5-small model employed in this project, there are several limitations that should be acknowledged:

1. Generalization Ability:

- The model's performance may vary across different domains and types of text. The training on the BBC News Summary dataset might limit its generalization ability to other domains or textual styles.

2. Resource Consumption:

- Although the t5-small model is lightweight compared to other variants, it still demands a certain level of computational resources, especially during the training phase.

3. Bias and Fairness:

- The model might inherit biases present in the training data, which could affect the fairness and neutrality of the generated summaries.

4. Hyperparameter Sensitivity:

- The performance of the model may be sensitive to the choice of hyperparameters, requiring careful tuning to achieve the desired results.

5. Evaluation Methodology:

- The novel evaluation approach using GPT-4, while innovative, may not capture all dimensions of summary quality and might not be entirely aligned with human judgment.

6. Long-Text Summarization:

- The model's capability in summarizing longer texts or documents hasn't been explicitly tested, which might be a limitation for certain use cases.

7. Multi-document Summarization:

- The current model is designed for single document summarization, and its effectiveness in multi-document summarization scenarios remains unexplored.

Addressing these limitations in future work could significantly enhance the model's effectiveness and broaden its applicability in real-world text summarization tasks.

7. Summary

This project embarked on the endeavor to develop a text summarization model capable of generating concise and coherent summaries for a diverse range of articles. The chosen backbone for this task was the T5 (Text-To-Text Transfer Transformer) model, particularly its "t5-small" variant due to its lightweight nature and proven efficacy in various NLP tasks. The model was trained on the BBC News Summary dataset, a rich repository of news articles along with their summaries, making it a suitable choice for the summarization task at hand.

The evaluation of the model's performance was conducted in a novel manner by employing GPT-4 as an evaluator. The T5 model was tasked with summarizing articles generated by GPT-4, post which GPT-4 evaluated the summaries based on various criteria including conciseness, coherence, ease of understanding, and relevance. The model achieved an impressive average score of 25 out of 30 across these evaluation fields, demonstrating its capability in producing high-quality summaries.

Despite the high performance, the model's standout feature was its practicality due to the lightweight nature of the t5-small variant. This characteristic enables the model to operate efficiently on resource-constrained devices like mobile phones, making it a viable solution for real-time summarization tasks such as summarizing book chapters for students.

8. Future Work

The promising results from this project lay a solid foundation for further exploration and improvement in automated text summarization. Several avenues could be pursued to augment the model's performance and broaden its applicability:

1. **Model Variants:** Exploring other variants of the T5 model or even other transformer architectures could provide insights into achieving better summarization performance.
2. **Dataset Enrichment:** Incorporating additional datasets for training could enhance the model's generalization ability across various domains and topics.
3. **Domain-Specific Training:** Fine-tuning the model on domain-specific datasets could improve its performance on specialized topics, making it more adaptable to specific use-cases.
4. **Multi-Document Summarization:** Extending the model to handle multi-document summarization could provide a more comprehensive summarization solution, particularly useful in scenarios like summarizing multiple news articles on a similar topic.

5. **Evaluation Metrics:** Employing other evaluation metrics alongside the novel GPT-4 evaluation could provide a more robust assessment of the model's performance.
6. **User Feedback Integration:** Incorporating user feedback into the model training process could help in refining the summaries to meet user expectations better.
7. **Deployment Optimization:** Optimizing the model for deployment on various platforms and ensuring its robustness and efficiency in real-world scenarios.
8. **Real-Time Evaluation:** Developing methodologies for real-time evaluation and feedback could aid in continuous model improvement, adapting to evolving user needs and content dynamics.
9. **Interactive Summarization:** Exploring interactive summarization where users can guide the summarization process to get customized summaries.
10. **Summarization with Visual Aids:** Integrating visual aids within the summaries for a more enriched summarization experience, especially useful in summarizing content with visual elements.


By addressing these aspects, the project could evolve into a more advanced and robust text summarization solution, further contributing to the field of natural language processing and providing valuable tools for information condensation and dissemination in our information-abundant world.

8. References

<https://arxiv.org/pdf/1910.10683.pdf>

Survey on Automatic Text Summarization and Transformer Models Applicability | Request PDF


Request PDF | On Oct 27, 2020, Wang Guan and others published Survey on Automatic Text Summarization and Transformer Models Applicability | Find, read and cite all the research you need on ResearchGate

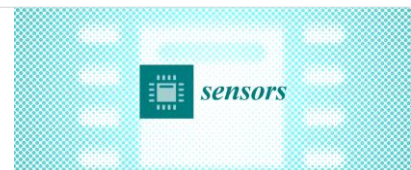
 https://www.researchgate.net/publication/348246753_Survey_on_Automatic_Text_Summarization_and_Transformer_Models_Applicability

<https://ieeexplore.ieee.org/document/10174260>

Efficient Memory-Enhanced Transformer for Long-Document Summarization in Low-Resource Regimes

Long document summarization poses obstacles to current generative transformer-based models because of the broad context to process and understand. Indeed, detecting long-range dependencies is still challenging for today's state-of-the-art solutions, usually requiring model expansion at the cost of an unsustainable demand for computing and memory


 <https://www.mdpi.com/2219578>



<https://ieeexplore.ieee.org/document/10051602>

An abstractive text summarization technique using transformer model with self-attention mechanism

Neural Computing and Applications - Creating a summarized version of a text document that still conveys precise meaning is an incredibly complex endeavor in natural language processing (NLP)....

 <https://link.springer.com/article/10.1007/s00521-023-08687-7>



T5

We're on a journey to advance and democratize artificial intelligence through open source and open science.

 https://huggingface.co/docs/transformers/model_doc/t5

