Comparative Analysis of LSTM and GRU Networks for Text Summarization with Attention Mechanisms

Abstract

This project investigates the effectiveness of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, both enhanced with attention mechanisms, for the task of text summarization. Utilizing the extensive and diverse WikiHow dataset, which comprises over 230,000 article-summary pairs, this project aims to enhance the adaptability of sequence-to-sequence (Seq2Seq) models to the varied lengths and styles of texts typical in everyday applications. By focusing on quantitative metrics such as ROUGE scores, the project highlights the strengths and limitations of each model in managing the complex structure of language and producing concise, relevant summaries from longer texts. Preliminary results suggest that while GRU models train more rapidly, the LSTM models demonstrate superior performance in handling long-range dependencies, a crucial aspect of generating coherent and comprehensive summaries.

1 Introduction

The exponential growth of digital text from sources such as academic papers, blogs, and news articles necessitates effective text summarization techniques in natural language processing (NLP). Text summarization systems aim to transform lengthy texts into concise summaries that capture essential details and the main idea. Recurrent neural networks (RNNs) such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks are particularly effective in these tasks due to their ability to manage long-range dependencies in text. This project explores the enhancement of LSTM and GRU models with attention mechanisms to improve summarization performance on the diverse WikiHow dataset.

2 Problem Statement

Despite significant advancements in RNN technology for text summarization, challenges remain in handling longer texts, where maintaining contextual integrity and relevance becomes computationally demanding and prone to errors. While traditional LSTM and GRU models excel at modeling long dependencies, they often struggle with attention distribution across extended sequences. This can lead to summarizations that either omit crucial information or misrepresent the original text's intent and factual content, diminishing the effectiveness of the summaries.

3 Research Objectives

1. Evaluate the effectiveness of LSTM and GRU networks in generating accurate and contextually relevant text summaries.

2. Implement and test the impact of attention mechanisms integrated with LSTM and GRU models to enhance the focus on significant segments of text, thus potentially improving the quality of the summarizations.

3. Compare the performance of LSTM and GRU models in handling diverse and complex datasets, specifically the WikiHow dataset, to determine which model architecture offers better efficiency and accuracy in real-world applications.

4 Literature Review

4.1 Key Research Findings

Long Short-Term Memory (LSTM) networks are widely recognized in the literature for their ability to handle long sequences in text summarization tasks, largely due to their internal memory mechanism. This capability makes them particularly adept at maintaining context over extended textual inputs, which is crucial for producing coherent summaries. On the other hand, Gated Recurrent Unit (GRU) networks are noted for their efficiency, possessing a simpler architecture with fewer parame-

ters, which can lead to reduced computational overhead while maintaining comparable performance. Comparative studies, such as those reviewed in this project, frequently assess these models' capabilities to balance computational efficiency with high-quality summarization outcomes, often highlighting the specific conditions under which one model might outperform the other in handling diverse and complex datasets like WikiHow.

5 Related Work

5.1 Common Challenges and Limitations

- **Diversity in Styles**: Many datasets are primarily composed of news articles, which tend to follow a specific writing style. This lack of diversity can limit the generalizability of models trained on such data to other domains.
- **Size of Datasets**: Datasets like DUC are not large enough to effectively train sequence-to-sequence models, which require substantial amounts of data to perform well.
- **Level of Abstraction**: The abstraction level in datasets like CNN/Daily Mail might be limited, reducing the capability of models to perform abstractedness which is essential for human-like summarization.
- **Compression and Attention**: Large datasets with varied sentence lengths and styles necessitate more complex models that can handle variable compression ratios and pay attention to the most relevant parts of the text for summarization.

5.2 Model Performance and Evaluation Metrics

Performance metrics are employed to evaluate the performance of the projected system by using metrics like ROUGE and loss.

ROUGE

The ROUGE metric is used to assess the quality of text summaries by measuring the overlap between the generated summaries and reference summaries. It is an important indicator of the employability of the summarization model.

$$ROUGE = \frac{\sum_{S \in \text{reference\_summaries}} \sum_{gram_n \in S} \text{Count}_{\text{match}}(gram_n)}{\sum_{S \in \text{reference\_summaries}} \sum_{gram_n \in S} \text{Count}(gram_n)} \quad (1)$$

Loss

The loss function is a critical component of the training process as it provides a measure of the model's predictions accuracy. In the context of this research, it is used to estimate the performance during both training and testing phases.

Table 1: Percentage Split of Data

Set	Percentage
Training Set	60% (approximately 138,000 reviews)
Testing Set	20% (approximately 46,000 reviews)
Validation Set	20% (approximately 46,000 reviews)

The data split in Table 1 provides a substantial basis for training the model while also allowing for a significant portion to be used for testing and validation. This distribution can be adjusted based on specific needs and requirements of the research.

6 Data

6.1 WikiHow Dataset

With more than 230,000 article-summary pairings extracted from a vast online knowledge base, the dataset offers a wide range of starting points for model training. The articles cover a broad spectrum of subjects and are written in a variety of styles by several writers, which increases the dataset's suitability for intricate summarizing tasks. The primary article material is compiled from the following in-depth descriptions, whilst summaries are produced by selecting and synthesizing the most important lines from each paragraph.This dataset requires more sophisticated summarizing skills than basic extraction due to its high degrees of abstraction and non-standard article formats.The dataset's resilience for testing sophisticated summarizing methods is demonstrated by benchmarking using measures like ROUGE and METEOR.The WikiHow dataset is an important benchmarking resource for text summarizing research since it is a publicly accessible resource.

6.2 Comparative Advantages and Selection Rationale

- **Existing Datasets Limitations**: Current large-scale summarization datasets predominantly comprise news articles, characterized by their specific and consistent writing style. This homogeneity limits the applicability of trained models to real-world scenarios where text styles and structures vary widely.

- **Reason for Choosing WikiHow**: Unlike these datasets, WikiHow offers an exceptional mix of topics and authorial styles, making it superior for training summarization models that require high levels of abstraction and adaptability. The rich diversity in content also tests and enhances the generalizability of summarization algorithms.

Dataset Size	230,843
Average Article Length	579.8
Average Summary Length	62.1
Vocabulary Size	556,461

Table 2: The WikiHow dataset statistics.

6.3 WikiHow Properties

The distinctiveness of the WikiHow dataset in text summarization research lies in its variety of topics and authorial voices. It poses unique challenges that push the boundaries of abstractive summarization, encouraging the development of models that can generate novel content and generalize across diverse text forms.

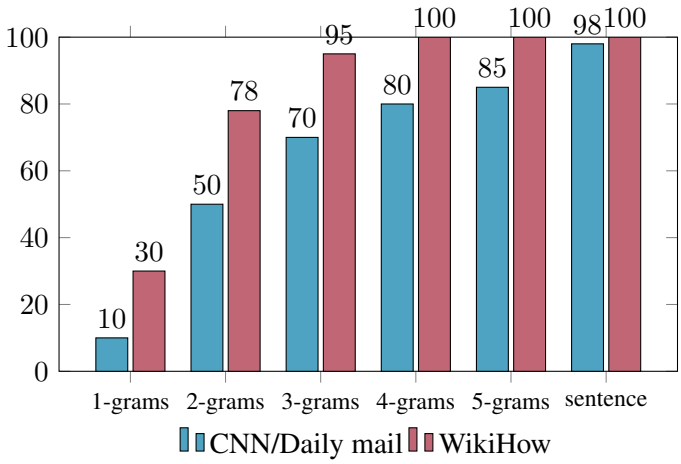

Figure 1: Uniqueness of n-grams in CNN/Daily mail and WikiHow datasets.

6.4 Compression Ratio

To describe the summary, we define the compression ratio. First, we figure out how long sentences typically are in both the summaries and the articles. The ratio between the average length of sentences and the average length of summaries is then determined to be the compression ratio. The summarizing effort becomes increasingly challenging as the compression ratio increases because greater layers of abstraction and semantics must be captured. The outcomes for CNN/Daily Mail and WikiHow are displayed in Table 3. greater layers of abstraction are required, as seen by WikiHow's greater compression ratio.

Table 3: The ROUGE-F1 scores of different methods on the non-anonymized version of the WikiHow dataset. The ROUGE scores are given by the 95% confidence interval of at most ± 0.25 in the official ROUGE script.

Model	ROUGE 1	ROUGE 2	ROUGE L	M exact
TextRank	27.53	7.40	20.00	**12.92**
Seq-to-seq	22.04	6.27	20.87	10.06
Pointer-gen	27.30	9.10	25.65	9.70
Ptr-gen + cov	**28.53**	**9.23**	**26.54**	10.56
Lead-3 base	26.00	7.24	24.25	12.85

7 Methods

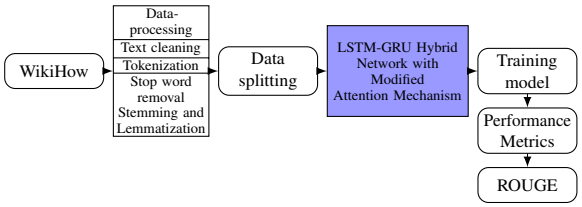

Figure 2: Overall Method of the Proposed Work.

7.1 Data Preprocessing

The raw dataset is cleaned to remove any non-relevant content, HTML tags, and special characters. Texts and summaries are then tokenized converted into a sequence of integers where each integer represents a unique word in the corpus. Stop words are removed, and contraction mappings are applied to ensure text normalization.

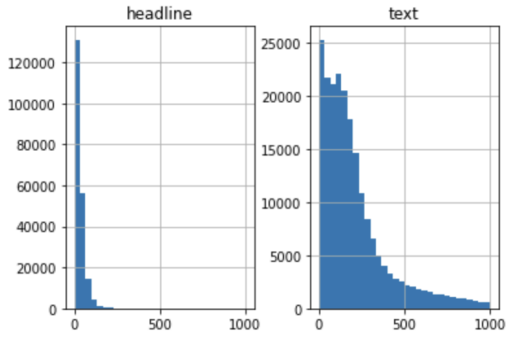

Figure 3: Word Frequency in Headlines and Text

7.2 Model Architecture

Both LSTM and GRU models are structured in a sequence-to-sequence (Seq2Seq) framework which consists of an encoder-decoder architecture. The encoder reads the input sequence and compresses the information into context vectors, which are the final hidden states of the LSTM/GRU layer. The decoder is initialized with the context vectors and is trained to predict the next word in the sequence given the previous words (auto-regressive).

7.3 Seq2Seq LSTM Modelling for Text Summarization

The LSTM model architecture used in this study implements a sequence-to-sequence (Seq2Seq) framework tailored for text summarization tasks. This methodology leverages the LSTM's ability to handle long-term dependencies, making it well-suited for processing and generating sequences such as sentences in natural language.

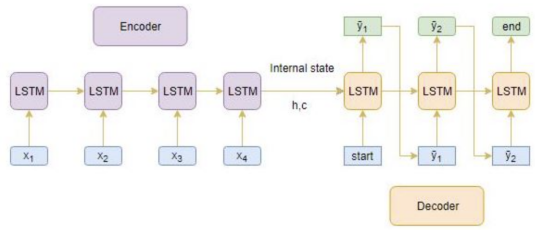

Figure 4: Seq2Seq LSTM Modelling

7.4 Attention Mechanism

Alongside the standard Seq2Seq architecture, an attention mechanism is integrated to allow the decoder to focus on specific parts of the input sequence during the generation of the summary. This attention-enhanced approach facilitates the model in learning alignments between the input text and the summary, significantly improving the performance over the baseline LSTM model.

7.5 Training Procedure

During training, the model utilizes a teacher-forcing strategy, which speeds up convergence by using the actual target outputs as inputs for the next time step, instead of the model's predictions. The model is optimized using the RMSprop optimizer, with loss computed via the sparse categorical cross-entropy function, enabling it to effectively handle the multi-class classification nature of the problem.

8 Results and Discussion

8.1 Presentation of Findings

The findings from the study on LSTM and GRU models for text summarization include:

- **Performance Metrics:** The use of ROUGE metrics, such as ROUGE-1 and ROUGE-L, to measure the overlap of unigrams and the longest common subsequence between the generated and reference summaries.
- **Quality of Summaries:** Illustration of the generated summaries and comparison with the actual summaries to evaluate fluency and readability. The models' performance could be showcased through tables or graphs indicating the ROUGE scores for a clear comparison.
- **Model Efficiency:** Highlighting the efficiency in training the GRU model with teacher forcing and the LSTM model's effectiveness in managing longer text dependencies. Mention the LSTM model's capability to understand detailed textual information, which is crucial for creating summaries.

8.2 Comparative Analysis of GRU and LSTM Training Loss

Training Stability

GRU: Exhibits significant volatility with sharp, regular spikes in loss.

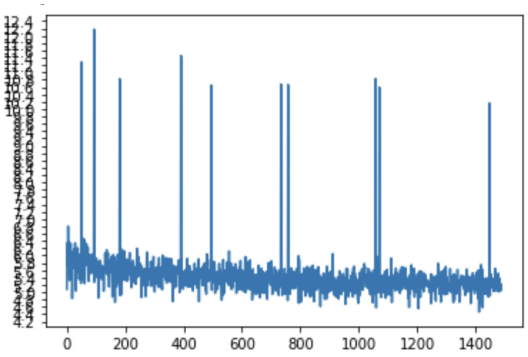

Figure 5: GRU Training Loss Graph

LSTM: While still volatile, the spikes are less pronounced and less frequent.

Implications: The LSTM model appears to be more stable during training compared to the GRU model. This suggests that LSTM may be better at handling anomalies in the data or that its architecture inherently provides more stable gradients during backpropagation.

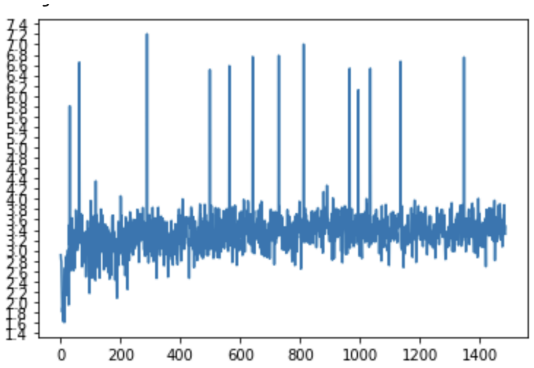

Figure 6: LSTM Training Loss Graph

Convergence Rate

GRU: Loss drops quickly initially and then descends at a slower, more variable rate.
LSTM: Displays a smoother but similarly rapid initial drop in loss, with a gradual reduction over time.

Implications: Both models learn quickly at the start, but the LSTM's smoother decline in loss suggests it may have a more consistent learning rate across different training iterations.

8.3 Experimental Result

Our most recent work on text summarizing with the WikiHow review dataset shows that the LSTM model can effectively condense long reviews into brief summaries. In one sad crime report scenario, for example, the model managed to distill a long and complex narrative into a concise summary that conveyed the main points of the incident, however it did so by somewhat changing the original summary's main points. This indicates that the model can preserve important information components, but it also implies that more improvement would be needed to improve the accuracy of the context representation.

In the comparative analysis of text summarization on the WikiHow dataset, the LSTM model has outperformed the GRU model across all ROUGE metrics, as detailed in Table 5. The results reflect the LSTM's advanced capability to handle the structured data of the WikiHow dataset, leading to summaries that are more aligned with reference standards. This affirms the LSTM's suitability for complex summarization tasks where understanding the nuanced context is crucial.

Table 4: LSTM Text Summarization Results

Review
An Indian origin couple was killed by their daughter's ex-boyfriend in an apparent revenge crime in the US on Friday. The 24-year-old suspect, Mirza Tatlic, fatally shot Naren Prabhu, a Silicon Valley tech executive, and his wife in San Jose. The police called the SWAT team after a standoff with Tatlic, who was later killed after a bullet hit him.
Original summary
Start Indian origin couple killed in revenge crime in US end
Predicted summary
Start Indian origin woman shot dead in US end

Table 5: ROUGE Score Comparison for LSTM and GRU Models on WikiHow Dataset

Neural Networks	**Evaluation Metrics**	
	ROUGE-1	**ROUGE-2**
LSTM	**49.15**	**26.83**
GRU	39.3	18.0

8.4 Challenges and Limitations

This study identified several challenges and limitations in the deployment of LSTM and GRU networks for text summarization tasks. Key among these were trends in training loss, which sometimes indicated difficulties in model convergence, particularly when dealing with texts exceeding a certain length. Such trends highlight the limitations of both models in processing very long sequences without losing contextual accuracy or experiencing increased training times.

Additionally, computational constraints emerged as a significant challenge. The practical aspects of training these models specifically computational costs and time—require substantial resources, which may limit their applicability in resource-constrained environments. This is particularly relevant for real-time applications where quick processing is essential. Moreover, the scalability of these models is hampered by their intensive memory and processing requirements, which can become prohibitive as the dataset size increases or as the complexity of the text to be summarized grows.

These findings suggest that while LSTM and GRU networks are powerful tools for text summarization, their deployment in real-world scenarios must be carefully managed to balance performance with computational efficiency.

9 Conclusion

The sequence-to-sequence (Seq2Seq) models employing LSTM and GRU networks, enhanced with attention mechanisms, have shown promising performance in the task of text summarization. This project's key findings include:

- **Performance:** Both LSTM and GRU models demonstrated the ability to generate coherent and concise summaries, as evaluated by ROUGE metrics. Notably, LSTM models excelled in capturing detailed textual information, benefiting from their superior long-term memory capabilities.
- **Efficiency:** GRU models were observed to train faster than LSTM models, indicating a more efficient use of computational resources, which could be particularly advantageous in scenarios requiring quick model training.
- **Attention Mechanism:** The integration of attention mechanisms significantly enhanced the models' ability to focus on relevant parts of the input text, thereby improving the quality of the generated summaries.

These findings suggest that while both LSTM and GRU networks are viable for high-quality text summarization, the choice between them may depend on specific requirements regarding detail and processing time. Future research could explore the scalability of these models in larger, more diverse datasets and the integration of more sophisticated attention mechanisms to further refine summary quality. Additionally, these models could be adapted for real-time summarization tasks in dynamic environments, such as news aggregation or content curation platforms, where both speed and accuracy are crucial.

Future Research Directions

The results of this project provide several pathways for future research in enhancing text summarization technologies. Key areas include:

- **Refinement of Attention Mechanisms:** Further development of attention mechanisms could focus on optimizing their application to longer text sequences. This refinement aims to improve the models' ability to maintain context and relevance across extensive inputs.
- **Hybrid Models:** Exploring hybrid approaches that combine the strengths of LSTM and GRU models could yield a new class of models that leverage both detailed memory retention and computational efficiency.
- **Cross-Domain Validation:** Testing the models on datasets from varied domains would help assess their generalizability and robustness. This could facilitate broader applications in different text summarization scenarios, such as legal documents, scientific articles, or informal blog posts.
- **Computational Efficiency:** Developing methods to reduce the computational demands of Seq2Seq models without compromising the quality of the generated summaries is critical. This could involve algorithmic improvements or more efficient training techniques.

These directions not only aim to enhance the functionality and applicability of Seq2Seq models but also address the pressing need for models that can operate efficiently in resource-constrained environments.

References

[1] Cho, K., Courville, A. & Bengio, Y. (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

[2] Rush, A.M., Chopra, S. & Weston, J. (2015). "A Neural Attention Model for Abstractive Sentence Summarization." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 379–389.

[3] Chopra, S., Auli, M. & Rush, A.M. (2016). "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks." In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT)*, pages 93–98.

[4] Bahdanau, D., Cho, K. & Bengio, Y. (2014). "Neural Machine Translation by Jointly Learning to Align

and Translate." In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[5] Pennington, J., Socher, R., & Manning, C.D. (2014). "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[6] Lin, C.-Y. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries." In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, volume 8.

[7] Banerjee, S. & Lavie, A. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

[8] Barrios, F., López, F., Argerich, L., & Wachenchauzer, R. (2016). "Variations of the Similarity Function of TextRank for Automated Summarization." arXiv preprint arXiv:1602.03606.

[9] Durrett, G., Berg-Kirkpatrick, T., & Klein, D. (2016). "Learning-Based Single-Document Summarization with Compression and Anaphoricity Constraints." arXiv preprint arXiv:1603.08887.

[10] Grusky, M., Naaman, M., & Artzi, Y. (2018). "Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.

[11] Harman, D. & Over, P. (2004). "The Effects of Human Variation in DUC Summarization Evaluation." In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*.

[12] Hong, K. & Nenkova, A. (2014). "Improving the Estimation of Word Importance for News Multi-Document Summarization." In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721.