Automatic Headline Generation using Deep Learning (LSTM + Attention)

By
Yash Nair (202201040075)
Kiran Shinde (202201040091)
Sanika Kundekar (202201040092)

Introduction

What is Headline Generation?

Headline Generation refers to the automatic creation of concise and informative headlines from longer news articles using Natural Language Processing (NLP) techniques.

Why is Headline Generation Important?

It enables faster content consumption, enhances news summarization, supports mobile-friendly reading, and improves content indexing in news aggregation platforms and search engines.

Challenges in Traditional Headline Generation:

- Reliance on manually written or rule-based headlines lacks scalability
- Difficulty in maintaining relevance, grammar, and context in generated summaries.
- Traditional extractive methods fail to produce novel, human-like headlines.
- Requires handling of diverse writing styles, lengths, and real-world ambiguities.

Summary of Paper

- The paper examines how large language models (LLMs) respond to irrelevant or distracting context within their inputs.
- It finds that LLMs are often misled by irrelevant information, which leads to a drop in performance across several NLP tasks.
- The study uses tasks like question answering and reading comprehension, inserting unrelated text to test the models' robustness.
- Even advanced models like GPT-3 and T5 show significant performance degradation when exposed to distractors.
- The impact of distractions varies by model size and task type, with larger models being more resistant but still affected.
- The results highlight a limitation in LLMs' ability to filter out irrelevant information and focus on the core task.
- The authors suggest future work should focus on improving models' relevance-awareness and context-filtering mechanisms.

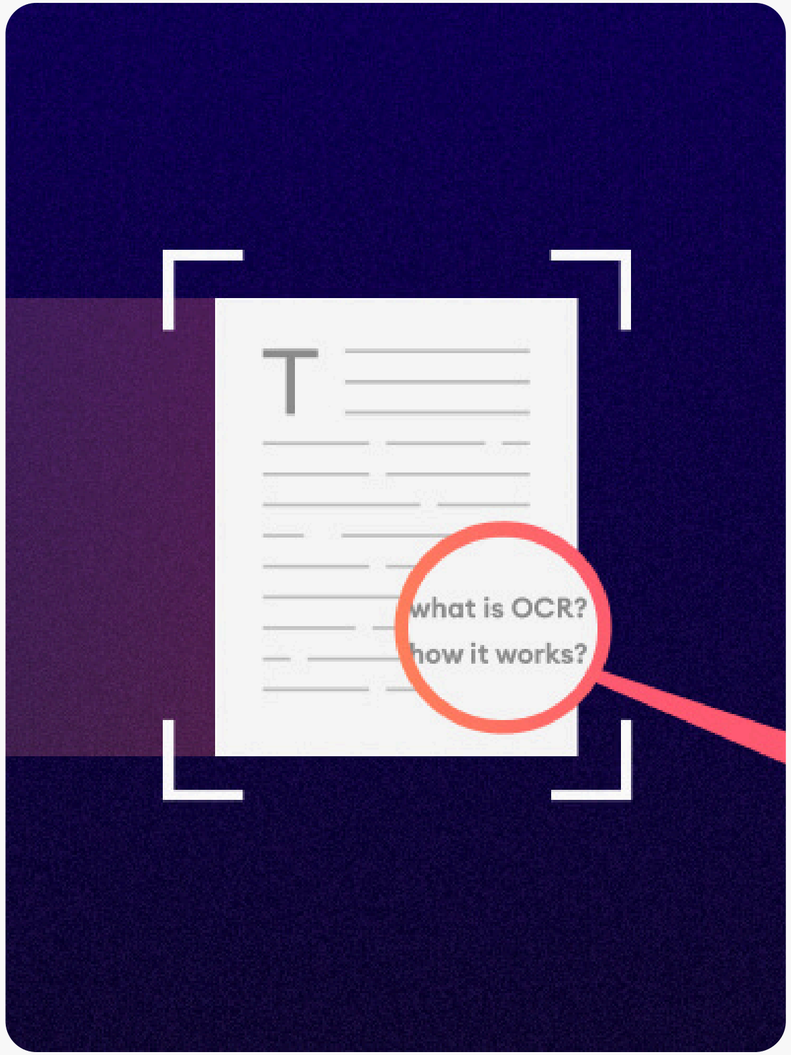
Problem Statement

Traditional headline generation systems face challenges

- Generating contextually accurate summaries from diverse news articles.
- Handling large-scale data efficiently for training deep learning models.
- Ensuring generalization across varied topics and writing styles in news content.

How can deep learning models, particularly sequence-to-sequence with attention mechanisms, improve the accuracy and relevance of automatically generated headlines?

Objectives



Develop an advanced headline generation model that:

- Leverages sequence-to-sequence models with attention mechanisms to improve summary accuracy.
- Enhances text processing capabilities for diverse topics and writing styles.
- Utilizes deep learning for effective handling of large-scale datasets.
- Ensures the relevance and contextual accuracy of generated headlines across different news domains.

Methodology

OCR System Workflow Using Deep Learning

1️⃣ Data Preprocessing

- Text normalization and cleaning: Lowercasing, removing special characters, and correcting spelling errors.
- Tokenization: Converting articles and summaries into sequences of words for better model training.
- Padding: Ensuring uniform input lengths for the model through padding sequences.

2️⃣ Model Architecture

- Encoder-Decoder Model: Utilizes an LSTM-based architecture to encode input articles and decode them into summaries.
- Attention Mechanism: Implements Bahdanau attention to focus on relevant parts of the article for more accurate summary generation.

Methodology

3️⃣ Training Process

- Tokenization and padding of input data to prepare for model training.
- Use of categorical cross-entropy loss function and optimization through RMSprop.
- Training the model over multiple epochs with validation to monitor performance and avoid overfitting.

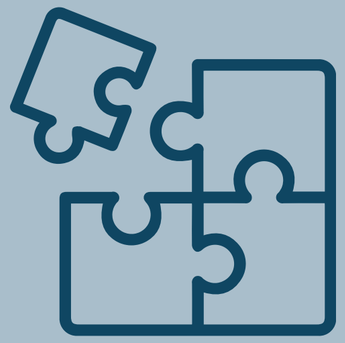
4️⃣ Post-Processing

- Decoding of predicted sequences into human-readable summaries.
- Evaluation using metrics like BLEU score and ROUGE score to assess summary quality.

Project Objectives

Without Attention Mechanism:

- Generate summaries from context
- Process sequential input data
- Evaluate model accuracy and loss



With Attention Mechanism (Bahdanau Attention):

- Focus on relevant input regions
- Learn long-range dependencies
- Improve headline generation quality

With Self-Attention Mechanism:

- Capture inter-word relationships
- Improve global context understanding
- Handle long input texts

With Attention Mechanism (Luong Attention):

Objective:

- To enhance text generation quality by integrating an attention mechanism into the encoder-decoder model.

Architecture Overview:

- **Input → Embedding → Attention Layer → Output (Next Token)**
- Custom attention (Luong-style): focuses on key parts of the input during generation.

How Attention Works:

- Calculates importance using:
  $$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
- This helps the model "attend" to relevant tokens like *"Blockchain"* and *"Supply Chain"* when generating the next word.

With Self-Attention Mechanism:

Improve Sequence Understanding:

Self-attention enhances model understanding by considering all parts of input text.

Focus on Key Relationships:

It allows the model to focus on important relationships between distant tokens.

Capture Global Context:

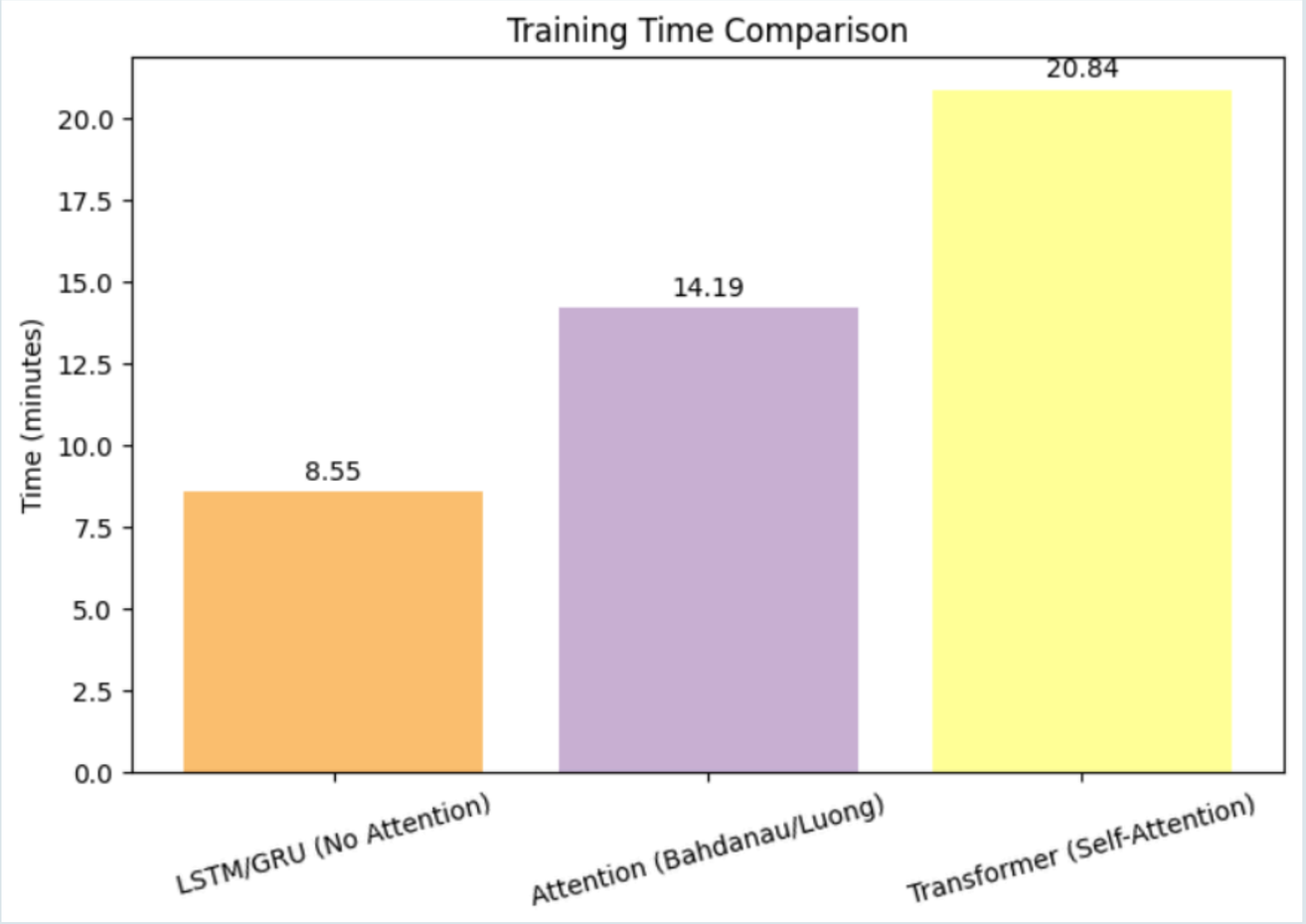
Self-attention enables the model to understand the entire context of the sequence.

Enhance Parallelization Efficiency:

It allows faster training by processing all tokens simultaneously, unlike sequential methods.

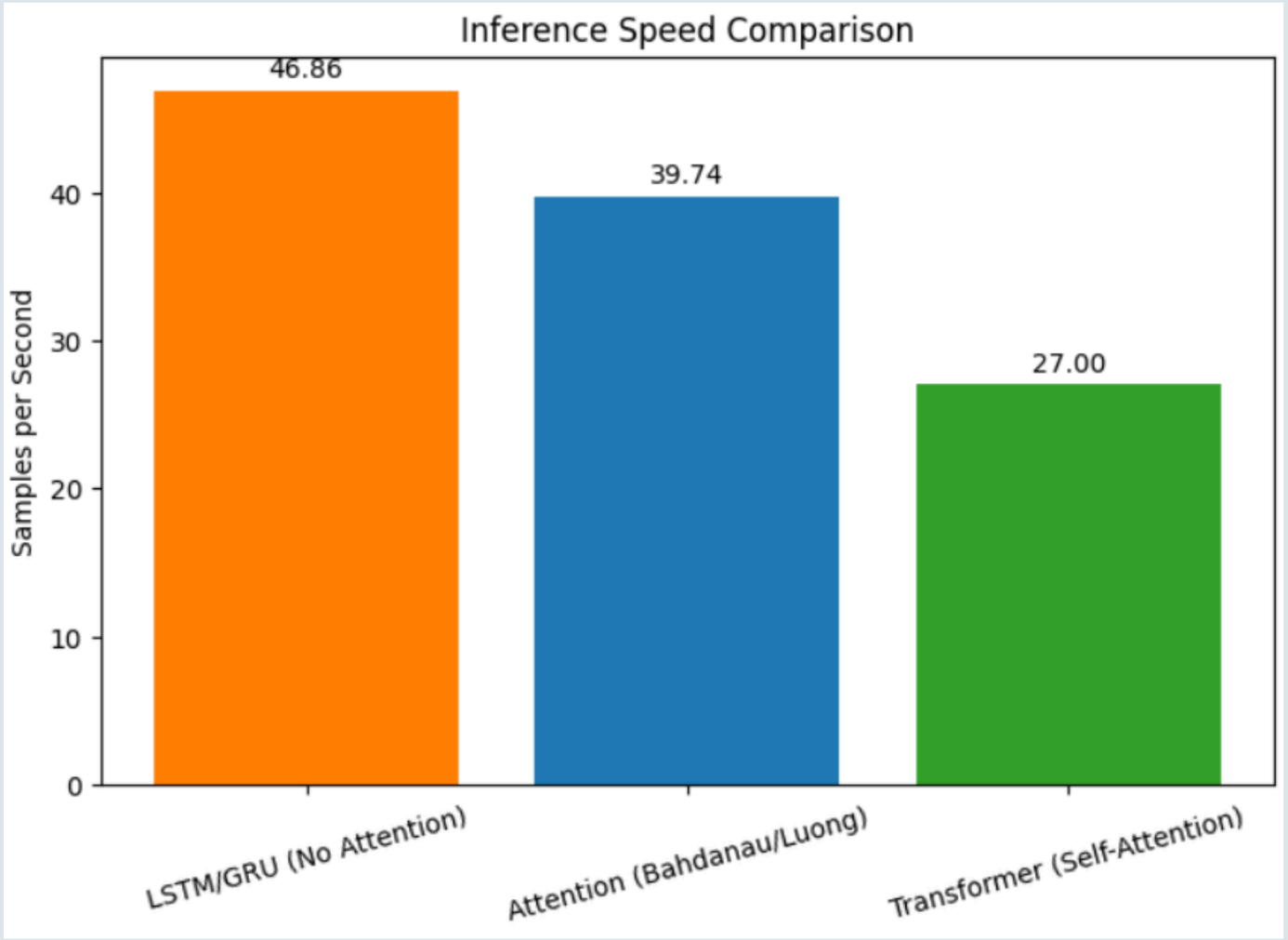
Model Comparison

Training Time Comparison



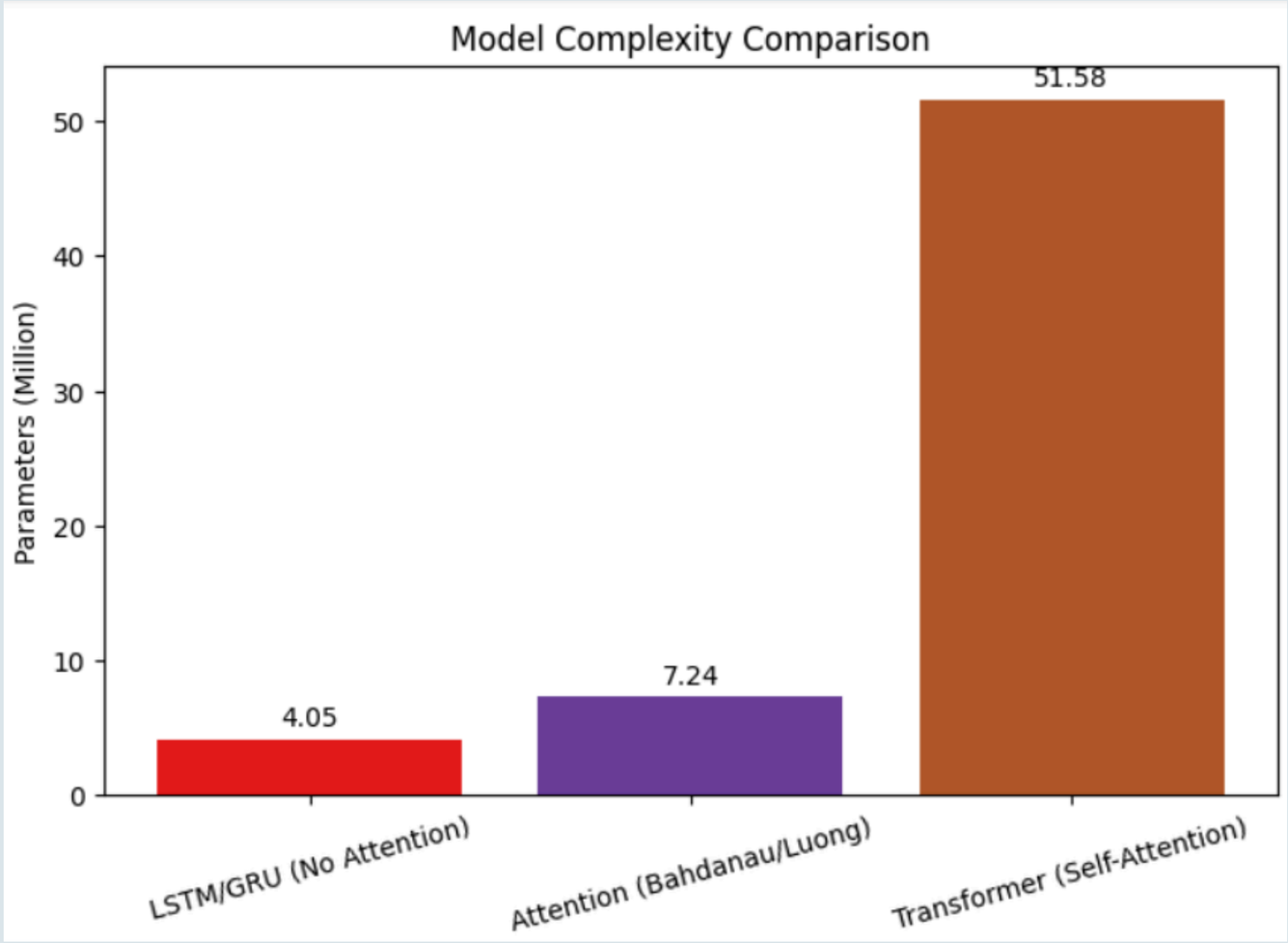
Model Comparison

Inference Speed Comparison



Model Comparison

Model Complexity Comparison



Expected Outcomes

Without Attention Mechanisms:

- Faster processing of high-quality images with basic text recognition accuracy.

With Attention Mechanisms (Badnau):

- Improved contextual accuracy by focusing on relevant parts of the text.

With Self-Attention (Transformer-based):

- Better long-range dependency handling for complex text recognition tasks.

Conclusion

- Attention mechanisms enhance model accuracy by focusing on important input parts.
- Without attention, models process inputs sequentially, limiting context and performance.
- With attention, models capture long-range dependencies for better task performance.
- Self-attention enables efficient context utilization, improving handling of complex sequences.
- Future improvements lie in refining attention mechanisms for broader applications in AI.

Thank you