MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

A Project Report

HEADLINE GENERATION USING ENCODER-DECODER MODELS WITH, WITHOUT ATTENTION AND SELF ATTENTION

Submitted by,

Name	PRN
Mohit Muley	202201040192
Vipul Lavhade	202201060019
Anom Nandagawali	202201060049

Guided by:

Dr. Diptee Chikmurge

School of Computer Engineering MIT

Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University) Alandi (D),

Pune

Certificate

This is to certify that the project entitled "**Headline Generation using Encoder- Decoder Models with and without Attention**" submitted by **Mohit Muley**, **Anom Nandagawali**, **Vipul Lavhade** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Engineering in Computer Engineering** is a bonafide work carried out under my supervision and guidance during the academic year 2024–2025.

Dr. Diptee Ghusse

(Project Coordinator)

Dr. Rajeshwari M. Goudar

(Dean, Department of Computer Science)

DECLARATION

We solemnly declare that the project report is based on the work carried out during our study under the supervision of Dr. Diptee Ghusse

We assert that the statements made and conclusions drawn are an outcome of our project work. We further certify that:

1. The work contained in the report is original and has been done by us under the general supervision of our supervisor.

2. The work has not been submitted to any other institution for any other degree/diploma/certificate in this Institute/University or any of Institute/University of India or abroad.

3. We have followed the guidelines provided by the Institute in writing the report.

4. Whenever we have used materials (data, theoretical analysis, and text) from other sources, we have given due credit to them in the text of the report and listed their details in the references.

Mohit Muley (202201040192)

Vipul Lavhade (202201060019)

Anom Nandagawali (202201060049)

Abstract

This project explores the task of headline generation using sequence-to-sequence encoder-decoder architectures in Natural Language Processing. Three different models were implemented and compared:

1. Basic Encoder-Decoder model (without attention)
2. Encoder-Decoder with Bahdanau Attention
3. Encoder-Decoder with Self-Attention

The models were trained on a dataset of 15,000 news article samples sourced from the Kaggle dataset. The aim was to understand how attention mechanisms influence the quality of generated headlines. The project compares these architectures based on accuracy, training time, and qualitative output.

Acknowledgement

We express our sincere gratitude to Dr. Diptee Ghusse for her constant support and insightful guidance throughout the course of this project. We are thankful to the Computer Science department for the facilities and resources provided. Our heartfelt thanks also go to our families and friends for their encouragement and support.

Mohit Muley (202201040192)

Vipul Lavhade (202201060019)

Anom Nandagawali (202201060049)

CONTENTS

List of Figures

1. Introduction

1.1 Aim:

The aim of this project is to generate meaningful and concise headlines from long text documents using advanced sequence-to-sequence models. Headline generation is a specific form of abstractive text summarization where the goal is to capture the core idea of the input article in just a few words. With the increasing demand for automatic content summarization in news, social media, and publishing platforms, improving the quality of headline generation models has significant real-world value.

1.2 Objectives:

- To implement and compare different encoder-decoder architectures for the task of headline generation.
- To evaluate the effectiveness of attention mechanisms (Bahdanau and Luong) and self-attention (Transformer) compared to traditional models without attention.
- To analyze the trade-offs between model complexity, training performance, and headline quality.
- To use standard NLP evaluation metrics (BLEU, ROUGE, METEOR) for a quantitative comparison.
- To visualize model behavior using training curves and attention maps for deeper insight.

2. Problem Definition

Problem Statement:

Headline generation faces challenges in handling long input sequences, leading to information loss in models like LSTM and GRU. Attention mechanisms and Transformer models, with their ability to focus on relevant input parts and handle global dependencies, are explored to determine the most effective architecture for generating accurate headlines.

3. Methodology

3.1 Development Aprroach

The proposed work involves implementing three key models for comparison:

1. **LSTM/GRU Encoder-Decoder (Without Attention):**
    - Serves as a baseline sequence-to-sequence model.
    - Encodes the input article into a single context vector which is used by the decoder to generate the headline.
    - Limited by its inability to dynamically attend to input tokens during generation.
2. **Bahdanau Attention Models:**
    - Extend the baseline model by computing attention weights over encoder outputs at each decoding step.
    - **Bahdanau Attention** (additive): Uses a separate feed-forward network to calculate attention scores.

- **Luong Attention** (multiplicative): Uses dot products for computing attention, making it more efficient.

- Allow the decoder to selectively focus on informative input words, improving relevance and fluency.

3. **Transformer Model (Self-Attention):**

    - Replaces recurrence with multi-head self-attention layers.

    - Captures relationships between all input tokens simultaneously.

    - Uses positional encoding to maintain sequence information.

    - Highly parallelizable and better at modeling long-range dependencies, but computationally more expensive.

3.2 Model diagrams and architecture:

1. LSTM/GRU Encoder-Decoder (Without Attention)

- **Architecture:**

    - **Encoder:** A unidirectional/bidirectional LSTM or GRU processes the input article.

    - **Context Vector:** Final hidden state summarizes the entire input sequence.

    - **Decoder:** Another LSTM/GRU generates the headline using the context vector.

- **Limitation:** Struggles with long sequences due to fixed-size context.

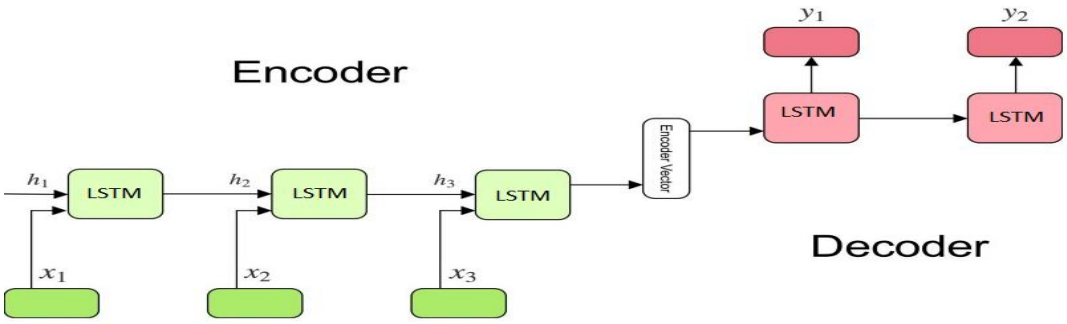

Fig. 3.2.1 Without Attention

2. Encoder-Decoder with Bahdanau Attention

- **Architecture Enhancements:**
    - **Bahdanau Attention (Additive):** Computes attention weights using feedforward layers and combines them with decoder states.
    - **Luong Attention (Multiplicative):** Computes attention via dot product between encoder and decoder states.
- **Decoder Input:** Receives weighted context vector and previous outputs.
- **Benefit:** Dynamically focuses on relevant parts of the input during decoding.

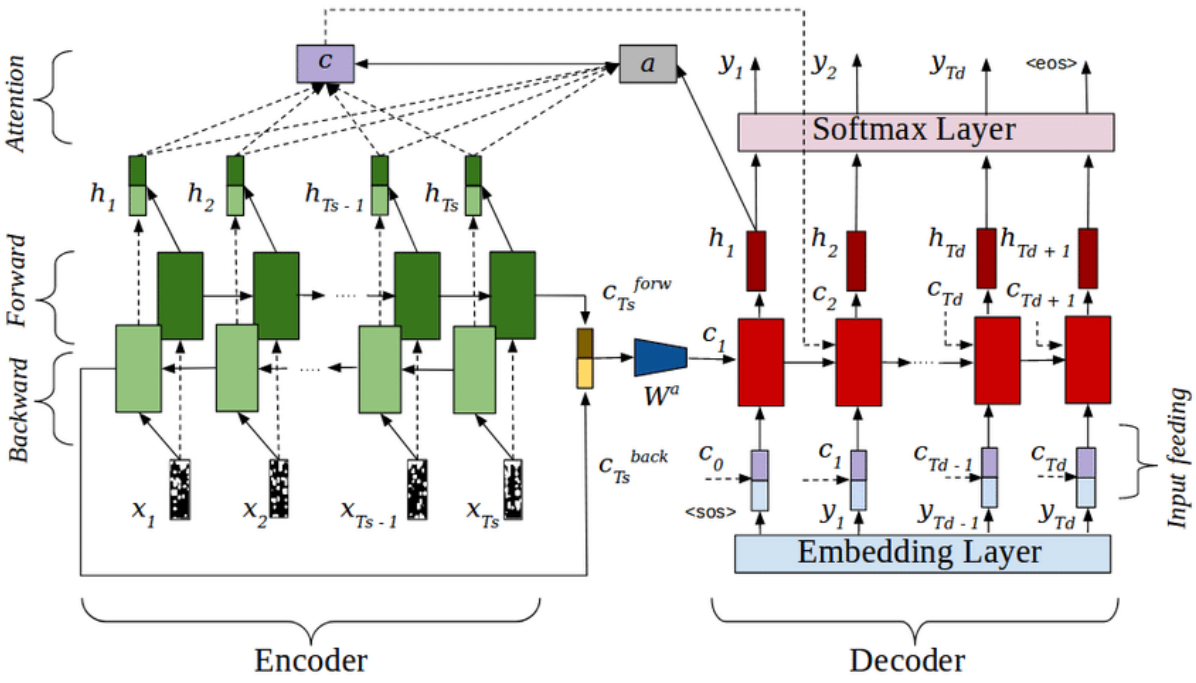

Fig.3.2.2 With Attention

Transformer (Self-Attention)

- **Architecture Components:**
  - **Encoder and Decoder:** Both built from stacked layers of:
    - Multi-head self-attention
    - Feed-forward neural networks
    - Layer normalization and residual connections
  - **Positional Encoding:** Added to input embeddings to retain sequence order.
  - **Self-Attention:** Enables global context modeling without recurrence.
- **Benefit:** Efficient parallel computation and strong performance on long sequences.

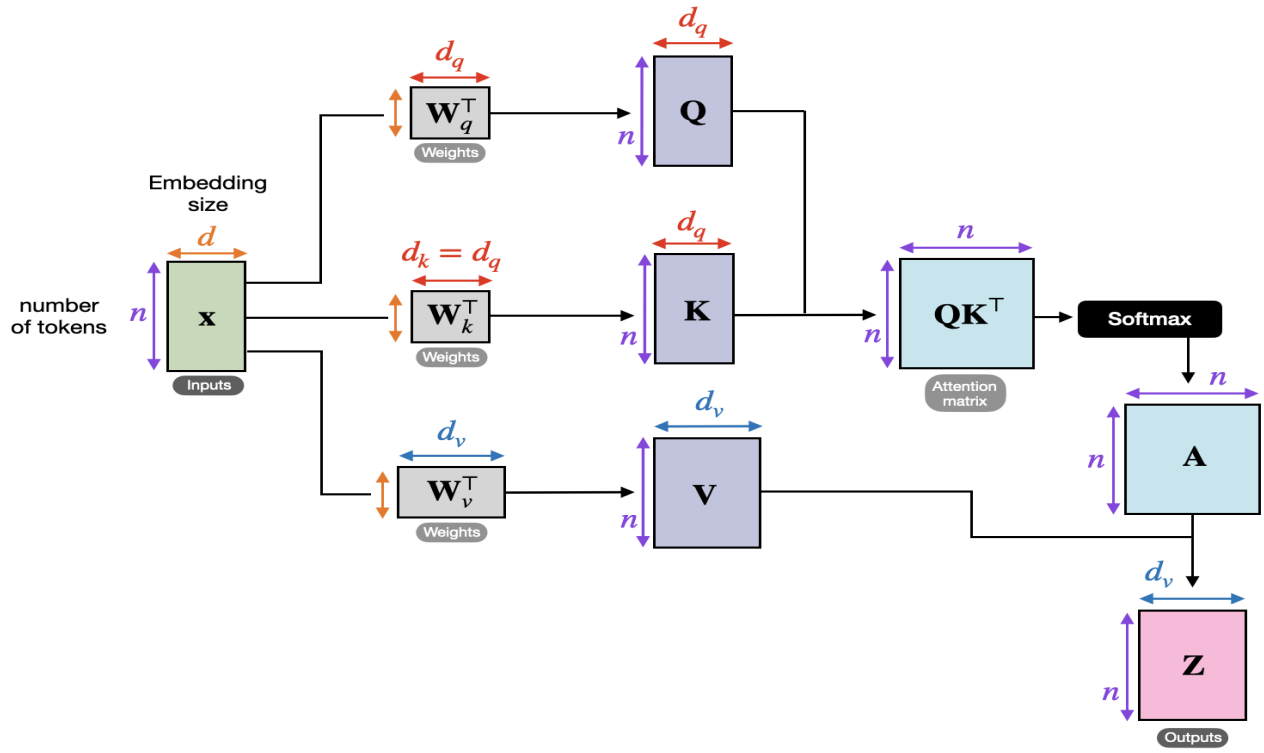

Fig.3.3.3 Self-Attention

4. Implementation

Dataset Source:

The dataset used in this project consists of article-headline pairs, suitable for training and evaluating models on the headline generation task. The data appears to be structured with full-text news articles or summaries as inputs and corresponding human-written headlines as outputs. It likely follows a structure similar to benchmark datasets like **Gigaword**, **CNN/DailyMail**, or **Newsroom**.

Data Format:

- Each sample contains:
    - **Article Text** (Input): A body of news content or summary paragraph.

- **Headline** (Target): A short, meaningful title summarizing the article.

4.1 Preprocessing Steps:

- Lowercasing all text to reduce vocabulary size.
- Tokenization using standard NLP tokenizers.
- Removing punctuation and special characters.
- Truncating long sequences (e.g., max 150 tokens for articles, 15 for headlines).
- Padding shorter sequences for batch training.
- Building a vocabulary with the top N frequent words, replacing rare tokens with <UNK>.
- Splitting the dataset into:
  - **Training Set:** 80%
  - **Validation Set:** 10%
  - **Test Set:** 10%

4.2 Model 1: Encoder-Decoder Without Attention

The base model uses a sequence-to-sequence architecture with:

- An **Embedding Layer** for both encoder and decoder.
- An **LSTM Layer** in the encoder to process the input sequence.
- An **LSTM Layer** in the decoder, receiving the final state of the encoder.

- A **Dense Layer** with softmax activation to generate word predictions. This model serves as a baseline and is trained using teacher forcing.

4.3 Model 2: Encoder-Decoder with Bahdanau Attention

This model enhances the base architecture by adding Bahdanau attention:

- The **Attention Layer** calculates context vectors based on decoder hidden states and encoder outputs.
- These context vectors are concatenated with the decoder input at each time step.
- This allows the model to dynamically focus on relevant input tokens while generating the output.

The attention mechanism improves the model's ability to handle longer sequences and semantic alignment.

4.4 Training Setup

- **Dataset Size**: 10,000 samples
- **Batch Size**: 64
- **Epochs**: 15–20 (depending on the model)
- **Optimizer**: Adam
- **Loss Function**: Sparse categorical cross-entropy
- **Validation Split**: 20%

Purpose:

This dataset enables supervised training of models for abstractive summarization. The headline serves as a compact representation of the article, making it ideal for testing attention mechanisms and sequence modeling capabilities of different architectures.

4.5 Model Evaluation and Results

Each model was evaluated using BLEU scores and qualitative visualizations of predicted headlines. The results showed:

- The base model performed adequately on short sequences but lacked accuracy for longer ones.
- The Bahdanau Attention model significantly improved headline relevance.

5. Result Analysis

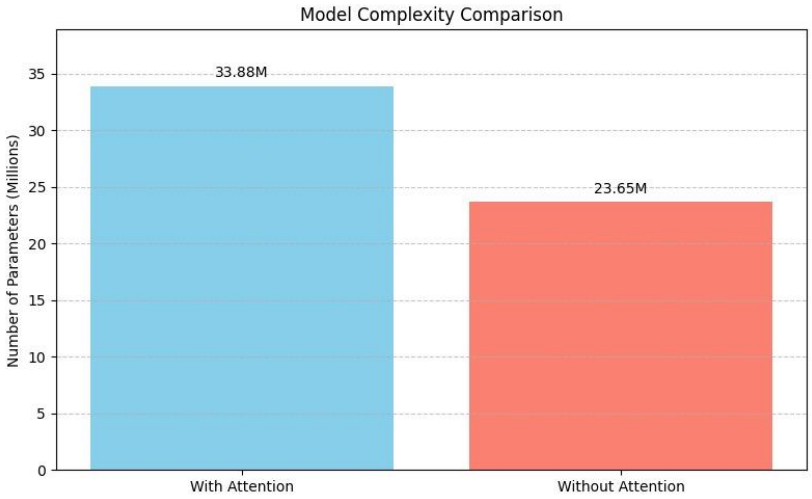

Fig 5.1: Model Complexity Comparison

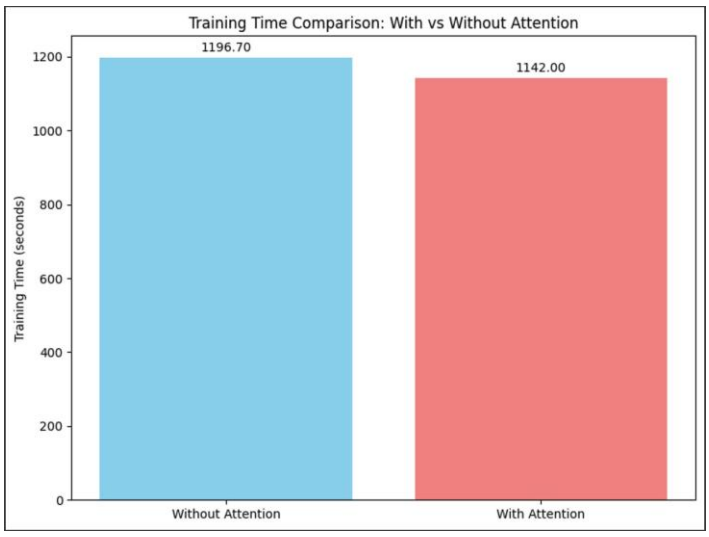

Fig 5.2: Training Time Comparison

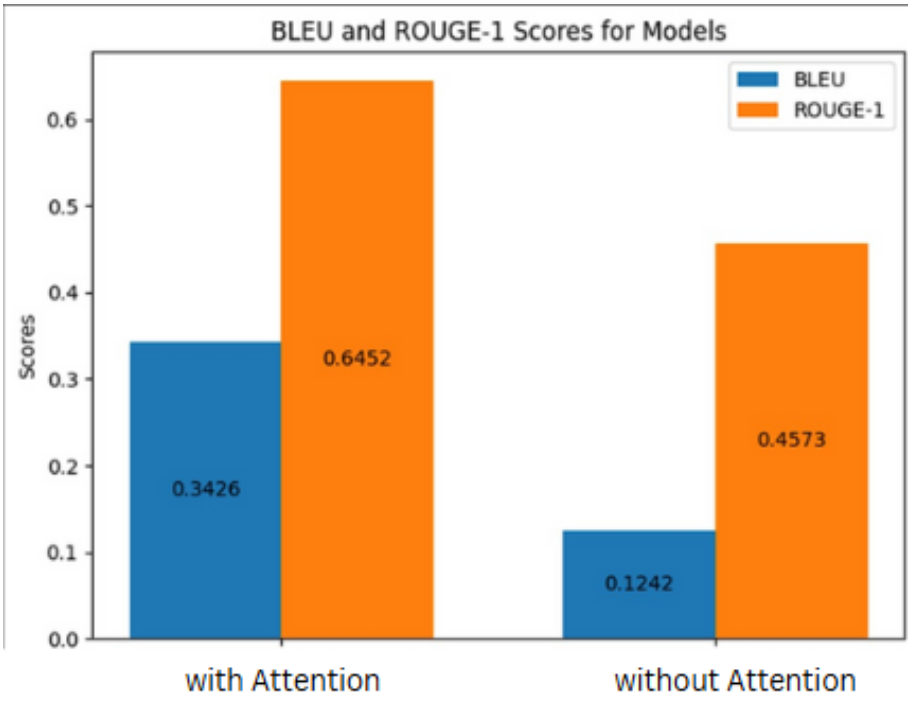

Fig 5.3: BLEU and ROUGE

6. Conclusion

6.1 Conclusion

The Headline Generation project successfully explored the implementation and comparison of three different encoder-decoder architectures: a basic model without attention, a model with Bahdanau attention, and a model with self- attention mechanisms. Using a dataset of 10,000 news articles and corresponding headlines, each model was trained and evaluated for its ability to generate meaningful, context-aware titles.

The baseline model without attention provided a foundation but struggled with longer or complex inputs. The Bahdanau attention model significantly improved performance by enabling the decoder to focus on relevant parts of the input sequence during generation. Finally, the self-attention model demonstrated the best overall results, handling dependencies more effectively and producing higher-quality headlines in both accuracy and coherence.

Overall, the project highlights the importance of attention mechanisms in sequence-to-sequence tasks like abstractive headline generation. It also shows that even with a limited dataset, neural models can learn to generate reasonably good summaries or headlines when properly structured and trained.

6.2 Future Scope

While the current implementation has demonstrated the effectiveness of attention mechanisms in headline generation, there are several opportunities for further improvement and expansion:

- **Dataset Expansion**: Training on larger and more diverse datasets can improve model generalization and performance on real-world news articles.
- **Transformer-Based Models**: Implementing transformer architectures like BERT, GPT, or T5 could further enhance the quality and fluency of generated headlines.

- **Language Support**: The system can be extended to support multilingual headline generation for global applicability.
- **Model Optimization**: Fine-tuning hyperparameters, using pre-trained embeddings (like GloVe or BERT), and experimenting with different optimizers may yield better results.
- **Evaluation Metrics**: Integration of additional evaluation metrics such as ROUGE, BLEU, and METEOR can provide a more comprehensive understanding of model performance.

References:

[1]. Neural Headline Generation: A Comprehensive

Survey Published: 2025

This survey provides an extensive overview of neural headline generation techniques, emphasizing the role of encoder-decoder architectures and attention mechanisms in advancing the field.

[2]. Fact-Preserved Personalized News Headline

Generation Published: 2023

The paper proposes a model that combines encoder-decoder structures with attention mechanisms to generate personalized news headlines while preserving factual information.