TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

PROJECT PROPOSAL

ON

JHATTA SAMACHAR: LEVERAGING TRANSFORMER MODELS FOR PERSONALIZED NEWS DELIVERY

SUBMITTED BY:

AMRIT SHARMA (PUL077BCT009)

KRIPESH NIHURE (PUL077BCT037)

DARPAN KATTEL (PUL077BCT099)

SUBMITTED TO:

DEPARTMENT OF ELECTRONICS & COMPUTER ENGINEERING

SAMSUNG INNOVATION CAMPUS

AUGUST, 2024

Contents

1. Introduction

1.1 Background

Traditionally, news has been delivered through newspapers and TV. In recent times, platforms like Facebook and YouTube have modernized news sharing, but they often present lengthy news with extended references and background information. Jhatta Samachar will leverage its user with personalized and summarized news with the help of transformer model.

1.2 Problem statements

In today's fast-paced world, people seek more information in less time. We have observed that even when we listen to a 5-minute news segment on YouTube, only about 1 minute is actual news, while the rest is background information. This not only wastes time but also diminishes the news listening experience.

1.3 Objectives

This project aims to develop an app that helps users stay updated with the latest news and incidents according to their preferences as bulletin.

1.4 Scope

In Scope

- Development of a mobile app that summarizes news in English and reads it aloud in Nepali based on user preferences (sports, politics, etc.).
- Use of translation APIs to convert English summaries to Nepali for text-to-speech (TTS) functionality.

Out of Scope

- Direct handling of Nepali language for input and processing.
- The model will process news in English and then we will translate the summaries to Nepali, using services like, Google Translate, for TTS.

2. Literature Review

2.1 Related work

1. "Attention is All You Need" (2017) by Vaswani et al. The "Attention is All You Need" [1] paper by Vaswani et al. introduced the Transformer model, a novel architecture that eschews recurrence and instead relies entirely on self-attention mechanisms to draw global dependencies between input and output.

   Encoder-Decoder Architecture

   The Transformer model utilizes an encoder-decoder architecture, where the encoder processes the input sequence to a fixed-size representation, and the decoder generates the output sequence from this representation. Both the encoder and decoder are composed of multiple identical layers.

   Attention Mechanism

   The core innovation of the Transformer model is the self-attention mechanism, which enables the model to weigh the importance of different tokens in the input sequence dynamically. The self-attention mechanism computes a set of attention weights for each token, indicating its relevance to other tokens.

   Limitations of "Attention is All You Need"

   While the Transformer model has significantly advanced the field of natural language processing, it has several limitations:

   - Computational Complexity: The self-attention mechanism has a quadratic complexity with respect to the input sequence length, making it computationally expensive for long sequences.
   - Memory Usage: Transformers require a large amount of memory to store the attention weights, especially for long sequences.
   - Data Requirements: Training Transformers effectively requires large amounts of data, which can be a barrier for some applications.

2. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" (2019) by Colin Raffel et al. The paper "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" [2] by Colin Raffel et al. presents the T5 model, which unifies all text-based language problems into a text-to-text format.

Unified Text-to-Text Framework

The T5 model frames every NLP task as a text-to-text problem, meaning that both the input and output are always text strings. This approach simplifies the training and application of the model across diverse tasks, including translation, summarization, and question answering.

Model Architecture and Pre-training

The T5 model employs a similar encoder-decoder architecture as the Transformer, but with modifications to improve performance on text-to-text tasks:
1. Pre-training Objectives: The T5 model is pre-trained on a multi-task mixture of unsupervised and supervised tasks, using a denoising autoencoder objective. This involves corrupting the input text and training the model to reconstruct the original text.
2. Scaling and Training: The authors experimented with various model sizes, ranging from small to extremely large (up to 11 billion parameters), and demonstrated that larger models consistently perform better.

Key Findings and Contributions

(a) Transfer Learning: The T5 model leverages transfer learning effectively, achieving state-of-the-art results on a wide range of NLP benchmarks by fine-tuning the pre-trained model on specific tasks.

(b) Task-Agnostic Performance: The text-to-text framework allows the T5 model to be applied to any NLP task without task-specific modifications, making it highly versatile and adaptable.

(c) Efficiency and Scalability: Despite its large size, the T5 model is designed to be efficient in terms of both computation and memory usage, making it feasible to train and deploy at scale.

Comparison with Other Models

The T5 model outperforms other models such as BERT and GPT in several key aspects:

- Flexibility: Unlike BERT, which is primarily designed for understanding tasks, T5 can handle both understanding and generation tasks.
- Unified Approach: T5's unified text-to-text framework simplifies the application of the model to different tasks, unlike GPT, which requires separate models or architectures for different tasks.

2.2 Related theory

This section provides a theoretical foundation for the project, focusing on key concepts and methodologies in deep learning and natural language processing.

1. **Deep Learning and Neural Networks**
   Deep learning [3], a subset of machine learning, involves training artificial neural networks on large datasets to perform complex tasks. Neural networks consist of layers of interconnected nodes (neurons) that process input data and learn patterns through backpropagation.

2. **Transformers**
   Transformers have revolutionized NLP by enabling models to process entire sequences simultaneously, rather than sequentially as in recurrent neural networks (RNNs). This parallelism allows for faster training and the ability to capture long-range dependencies.

3. **Encoder-Decoder Structure**
   Transformers use an encoder-decoder structure, where the encoder processes the input sequence and the decoder generates the output sequence. Each layer in the encoder and decoder contains self-attention and feed-forward sub-layers, enabling the model to learn complex relationships between tokens.

   Attention Mechanism
   The attention mechanism is a key innovation in Transformers, allowing the model to focus on relevant parts of the input sequence when generating each token in the output sequence. This mechanism computes attention weights that dynamically adjust the importance of different tokens based on their relevance to the current task.

   Self-Attention
   Self-attention enables the model to weigh the importance of each token in the input

sequence relative to all other tokens. This is achieved by computing dot products between query, key, and value vectors derived from the input embeddings.

Multi-Head Attention

Multi-head attention extends self-attention by allowing the model to attend to different parts of the input sequence simultaneously. Each attention head captures distinct aspects of the input, and their outputs are combined to form a comprehensive representation.

4. **T5 Model** The T5 model builds on the Transformer architecture, framing all NLP tasks as text-to-text problems. Features of T5 model:

   Text-to-Text Framework

   By unifying all tasks into a text-to-text format, the T5 model simplifies training and application, allowing for consistent performance across diverse tasks.

   Pre-training and Fine-tuning

   The T5 model is pre-trained on a large corpus using a denoising objective and fine-tuned on specific tasks. This transfer learning approach enables the model to achieve state-of-the-art results with relatively few task-specific adjustments.

   Comparison with GPT While both T5 and GPT are powerful language models, T5's text-to-text framework offers greater flexibility and ease of use across different tasks. GPT, on the other hand, excels in generating coherent and contextually rich text, making it suitable for tasks requiring high-quality text generation.

3. Proposed Methodology

3.1 Feasibility Study

We find the project feasible because:

- We can collect data from websites like Ekantipur, Annapurna Post and other e-news platforms in Nepal, that has news summaries in meta description tags and categorized news.
- Google Colab provides sufficient computational resources for testing.

3.2 Proposed Architectural Design

We propose a client-server architecture with a Flutter mobile app as the client and Django as the server. PostgreSQL will be used for database management.

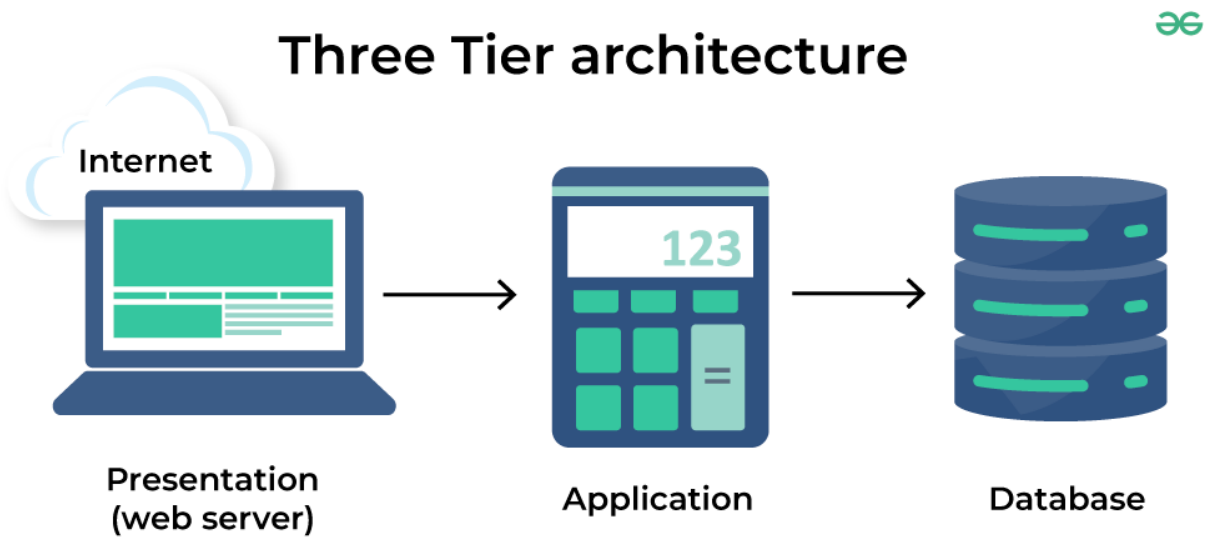

Figure 3.1: Client-Server Architecture

3.3 Testing and Implementation

Training will be conducted on Google Colab with GPU support, using subscription packages if necessary. Validation will be performed using metrics like ROUGE score, and testing will involve friends, family, and a beta program.

4. Proposed Experimental Setup

This section describes the experimental setup, including data collection, model training, and the technology stack.

4.1 Tech Stack

- Python, Django (DRF), Flutter, TensorFlow, and Keras.
- Google Colab for GPU training, potentially using the pro version.

4.2 Data Collection and Preparation

Data will be collected through web scraping from sites like Ekantipur and Annapurna Post. The data will be cleaned and prepared for NLP tasks, including tokenization, lowercasing, and special character removal, before being formatted as JSON.

We will scrape data from those popular Nepali websites, focusing on English news. These sites have categorized news and summaries in meta description tags. Additional datasets may be sourced from the internet.

4.3 Model Training

We will use the pretrained T5 model, fine-tuning it to our specific task and training it on our dataset.

5. Proposed System Design

The proposed system can be described with the help of a system flow chart as shown in the figure 5.1, as well as the proposed transformer model's architecture in as per figure 5.2.

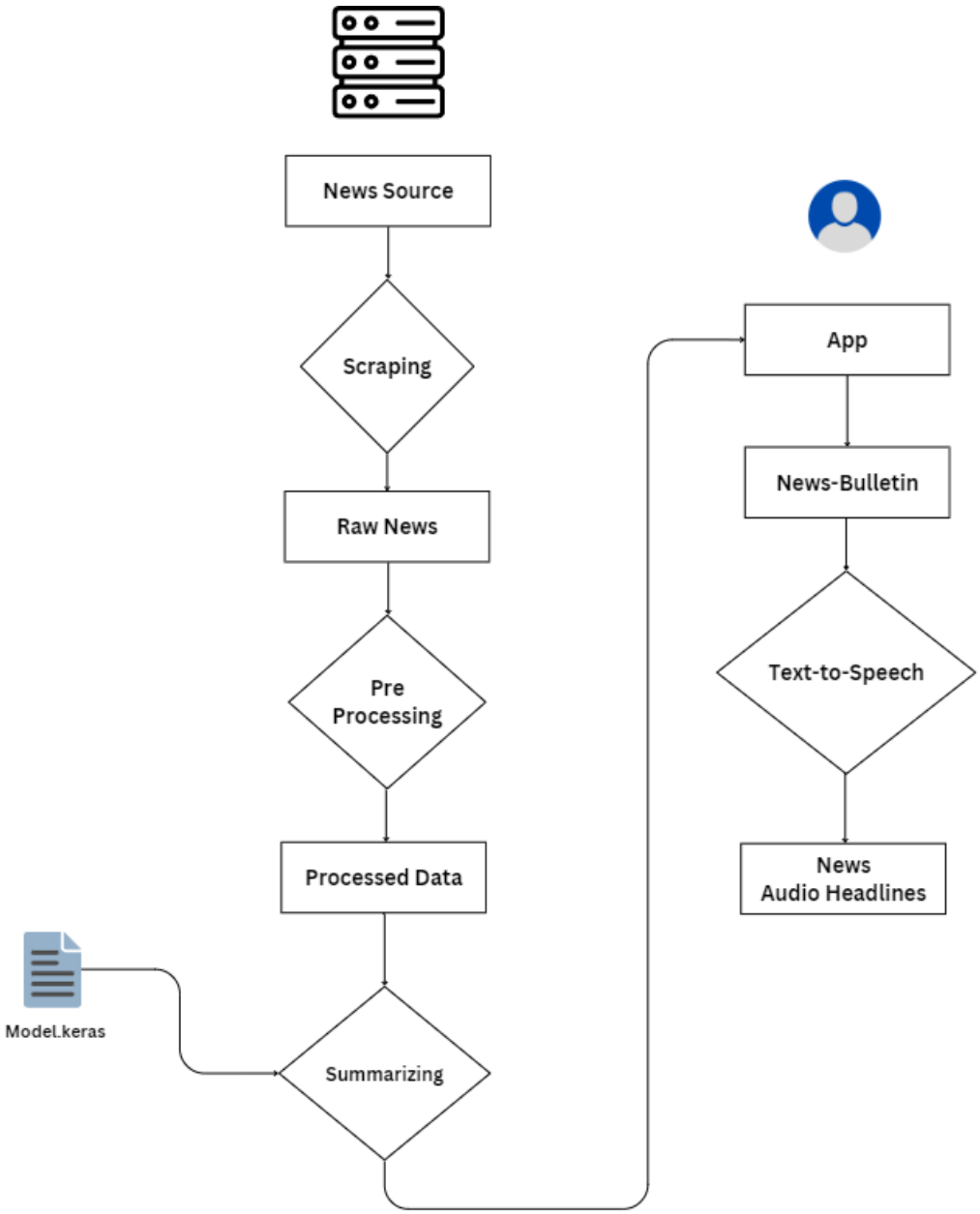

Figure 5.1: System Flowchart Diagram

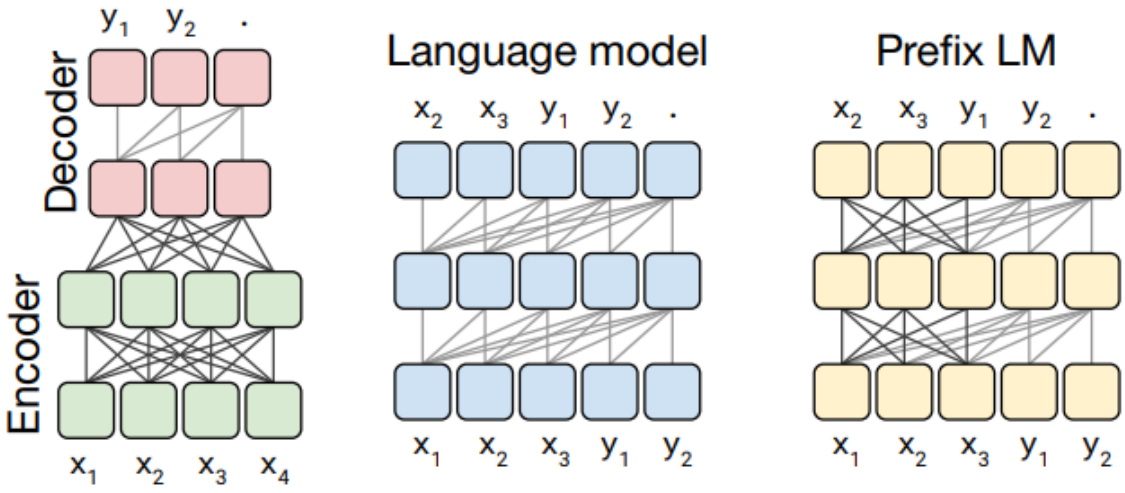

Figure 5.2: Model Architecture

6. Expected Output

The expected outcome of this project is to develop a mobile application that offers personalized and concise news summaries tailored to the user's preferences. The key features and expected outputs include:

User Preferences

Users will have the ability to select their preferred news categories such as sports, politics, international news, cultural events, and more. This customization ensures that users receive news updates relevant to their interests, making the app highly personalized and user-centric.

News Display

The home page of the app will display a curated list of news articles. Each article will be presented with the following elements:

- **Thumbnail Image:** A small image representing the news article.
- **Title:** A brief and engaging headline of the news article.
- **Description:** A short summary or teaser of the news article.
- **Link:** A clickable link directing users to the full news article for detailed reading.

This layout ensures that users can quickly scan through the news articles and decide which ones they want to explore further.

Summarized News Playback

One of the standout features of the app will be the "Listen to Bulletin" button. This feature will enable users to listen to a summarized version of the news, providing a hands-free and time-efficient way to stay informed. The summarized news playback will offer:

- **English Summaries:** Concise and clear news summaries in English, generated using the trained T5 model.
- **Translated Nepali Summaries:** For users who prefer Nepali, the English summaries will be translated into Nepali using translation APIs, and then read aloud in Nepali.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.