IT594: - DEEP NEURAL NLP

A Project on Research Paper Simplification

Sarvesh Bagwe (202211006)
Vedant Dave (202211042)
Kashyap Halavadia(202003040)
Hiren Thakkar (202211074)
Course Instructor:
Prof. Sourish Dasgupta

Problem Formulation

Objective: To reduce the complexity of vocabulary and sentence structures while preserving the original meaning to enhance readability and understanding.

Challenges:

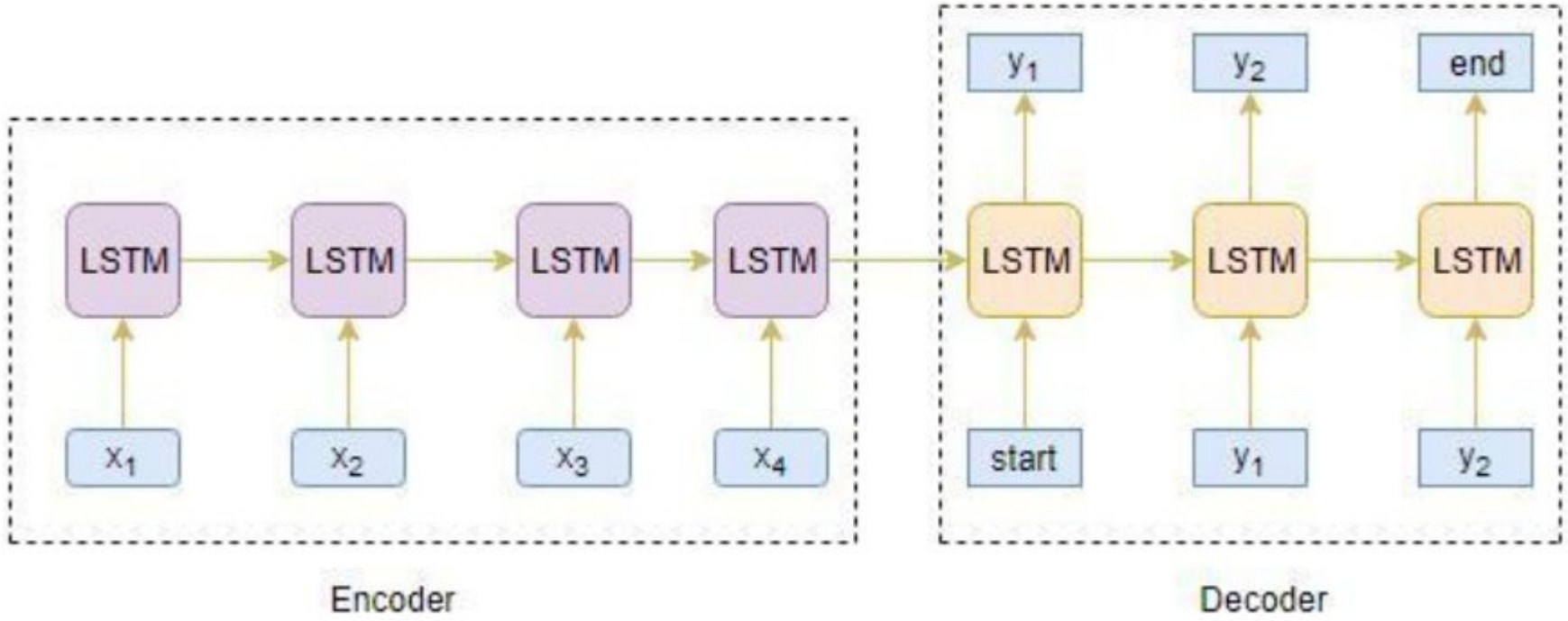
1. Preserving Meaning: Striking a balance between simplification and maintaining the core message.
2. Accuracy: To ensure precise simplification across diverse context and text types.

Text Preprocessing

1. Removing extra white spaces in 'highlights', 'body', and 'text'.
2. Removing square brackets and numbers from the 'text'.
3. Extracting 'abstract' from 'text' using 'ABSTRACT' and 'INTRODUCTION' markers.
4. Filtering rows based on 'abstract' length (500 to 2000 characters).
5. Replacing URLs in 'text' with "<LINK>".
6. Stripping leading whitespaces in 'abstract' and 'highlights'.
7. Removing special characters (excluding alphanumeric, whitespace, dots, and hyphens) in 'abstract' and 'highlights'.
8. Replacing hyphens with spaces in 'abstract' and 'highlights'.

Text Summarization

Model Architecture (Encoder Decoder Architecture)



Rationale for using Encoder Decoder Architecture (LSTM-LSTM)

An architecture employing an Encoder-Decoder framework for text summarization.

Encoder:

- It utilizes LSTM to learn input text embeddings, capturing contextual information.

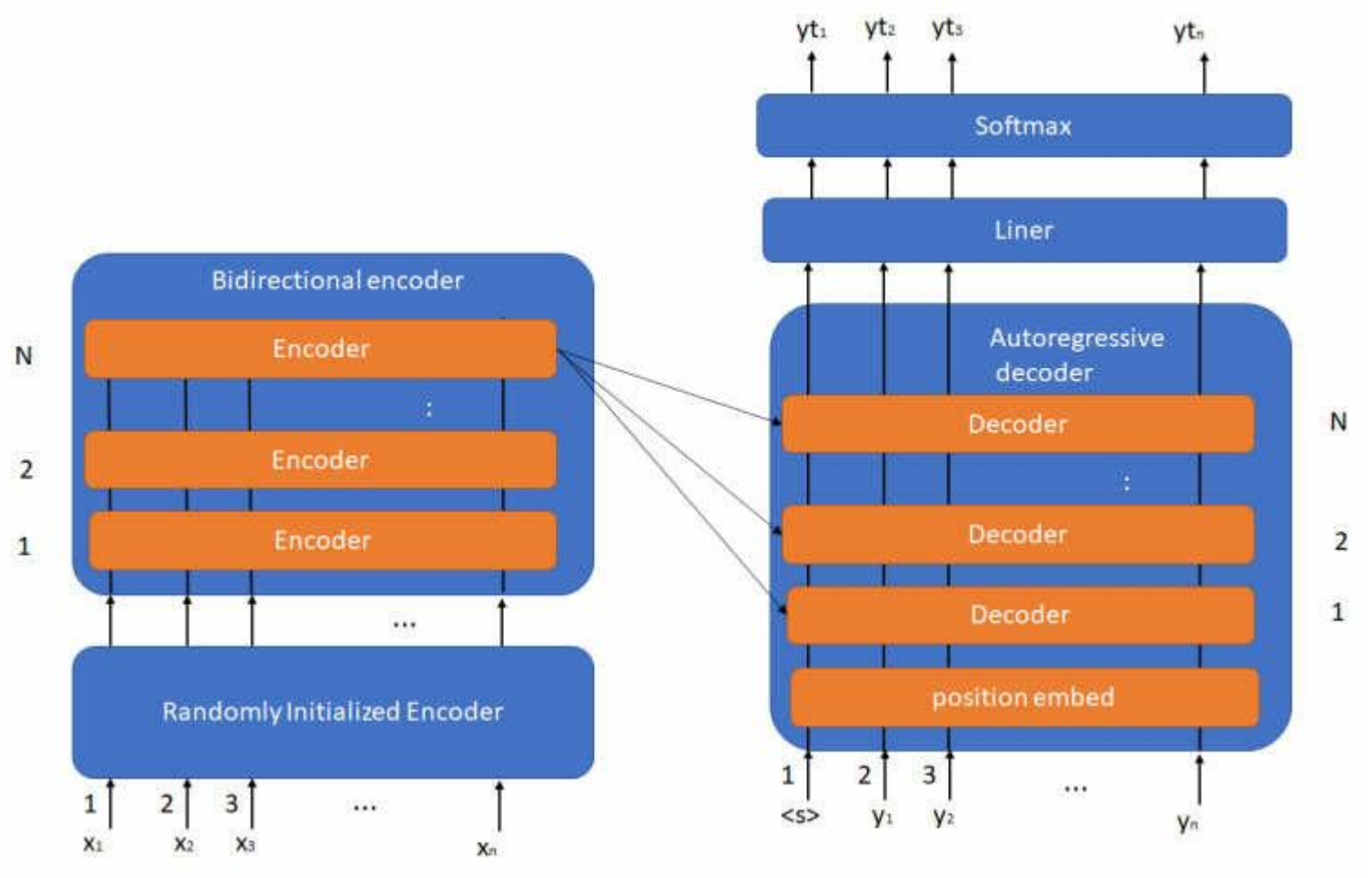
Decoder:

- Initially it focuses on Language Modelling to understand the structure and semantics.
- Takes text embeddings from encoder as condition (C) for generating summaries.
- Leverages the learned embeddings (while Language Modelling) and condition (C) for precise summary generation.

Specifics of the model architecture(LSTM-LSTM)

- Number of Total Parameters:- 8474316
- Hyperparameters :- Learning Rate (0.001), Batch Size (5-10), Embedding Dimension, Patience
- Loss function:- Sparse Categorical Cross Entropy
- Optimizer:- RMSProp

Model Architecture (BART)



Credits: Encoder-Decoder, a model works for sentence-to-sentence takes like translation and summarization, where an encoder encodes the input into a fixed-length vector and a decoder generates output from the encoded vector, such as T5 [13], BART [23], and MASS [24]. 3) Left-to-Right LM, a model is trained to predict the next word in a sentence given the previous words, such as GPT [25], GPT-2 [13], and GPT-3

Rationale for using Encoder Decoder Architecture (BART)

An architecture employing an Encoder-Decoder framework for text summarization.

Encoder:

- It utilizes LSTM to learn input text embeddings, capturing contextual information.

Decoder:

- Initially it focuses on Language Modelling to understand the structure and semantics.
- Takes text embeddings from encoder as condition (C) for generating summaries.
- Leverages the learned embeddings (while Language Modelling) and condition (C) for precise summary generation.

Rationale for using Encoder Decoder Architecture (BART)

- Attention Mechanism Advantage: Harnesses an attention mechanism to better capture intricate relationships between words.

Dual Pre-training Objective:

- Model Foundation: It is trained with a dual objective - to regenerate the correct version of a corrupted input, a feature that can prove to be beneficial for the summarization task.

Embracing Auto-Regressiveness:

- Dynamic Generation Process: it exhibits an Auto-Regressive nature (step-by-step generative process).

Specifics of the model architecture (BART)

- Hyperparameters :- Learning Rate (0.0001), Model Dimension, Optimizer
- Loss function:- Sparse Categorical Cross Entropy
- Optimizer:- AdamW

Rationale for using Decoder Only Architecture (GPT-2)

- It is Auto-Regressive in nature (step-by-step generative process).
- It can draw attention to potential redundancy in encoding text before generating summary.
- Human text summarization is linear and sequential process which can be leveraged by Auto-Regressive property.
- GPT-2 uses decoder only architecture which has been pre trained on large corpus of text.
- We can take advantage of pre trained embeddings to enhance our performance by fine-tuning it on our data.

Specifics of the model architecture(GPT-2)

- Hyperparameters :- Learning Rate (0.0002), Sequence Length (512), WarmUp Steps
- Loss function:- Sparse Categorical Cross Entropy
- Optimizer:- AdamW

Rationale for using Encoder Only Architecture (T5 Model)

- Encoder-Decoder framework for text summarization.
- Text-to-Text Framework instead of Sequence-to-Sequence Framework
- It is pre-trained on benchmark dataset of many NLP Tasks

Specifics of the model architecture(T5)

- Hyperparameters :- Learning Rate (0.0001), Weight Decay, Model Dimension
- Loss function:- Cross Entropy
- Optimizer:- AdamW

Results

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-3
BART	0.06343	0.3386	0.1278	0.3176
Sequence2Sequence(LSTM)	0.1022	0.3790	0.1596	0.3583
T5	0.1867	0.4439	0.2413	0.4274
GPT-2	0.2269	0.4821	0.2931	0.4669

Tabel 1: Evaluating our trained models on different metrics

Thank You