PUNE INSTITUTE OF COMPUTER TECHNOLOGY

Department of Computer Engineering

(2024-2025)

LP-VI

Batch:- R3

NLP Mini Project

Finetune a pre trained transformer for the task of summarization using a relevant dataset

Submitted By -

41352 - Advait Naik

41353 - Kaustubh Netke

41363 - Rhea Shah

Guided By-Prof. Hema Kumbhar

TABLE OF CONTENTS

1. ABSTRACT

This project explores the fine-tuning of a pretrained transformer model for the task of abstractive text summarization. Using the powerful BART model (facebook/bart-base) and the XSum dataset, the primary objective is to enable the model to generate concise, informative summaries that accurately reflect the content of longer articles. Abstractive summarization, unlike extractive methods, requires the generation of new text rather than the selection of sentences directly from the original input. This makes the task significantly more challenging and computationally intensive.

The use of transfer learning in this project allows the application of a model that has already learned general language representations, reducing the amount of labeled data and training time required. The HuggingFace Transformers library and the Datasets API are used for model access and data loading, while tokenization, padding, and truncation are applied for effective input formatting. The model is trained for a single epoch on a subset of 1,000 training and 200 validation examples to demonstrate proof of concept.

2. INTRODUCTION

Natural Language Processing (NLP) has witnessed a revolutionary change with the advent of transformer-based architectures. These models have proven highly effective for a wide range of NLP tasks, including machine translation, question answering, and summarization. In particular, summarization remains a critical yet complex task in NLP, as it requires both syntactic and semantic understanding of lengthy documents to generate coherent, concise representations. This project aims to fine-tune a pretrained transformer—specifically the BART model—for the task of abstractive summarization using the XSum dataset.

BART (Bidirectional and Auto-Regressive Transformers) is designed as a sequence-to-sequence model that combines the strengths of both encoder-decoder and language generation mechanisms. The model, pretrained on massive corpora, is adept at understanding contextual relationships within and across sentences, making it an ideal choice for tasks requiring text generation. By fine-tuning BART on a specific summarization dataset, it becomes possible to adapt its general linguistic knowledge to a specific application domain.

The XSum dataset, composed of BBC news articles and their corresponding one-sentence summaries, is well-suited for this task. Each summary is independently generated to capture the core meaning of the article, demanding a high level of abstraction from the model. This project demonstrates how transformer-based models can be effectively fine-tuned for real-world summarization tasks using modern machine learning tools and frameworks.

3. PROBLEM STATEMENT

The task of text summarization is essential for managing the overwhelming amount of textual content generated daily across various domains. Traditional summarization techniques either rely on heuristic-based methods or extractive approaches that copy sentences directly from the source document. However, abstractive summarization—which requires generating novel sentences that capture the essence of the original text—poses a greater challenge.

The problem addressed in this project is to fine-tune a pretrained transformer model, specifically BART (facebook/bart-base), for generating abstractive summaries using the XSum dataset.
The dataset consists of news articles paired with one-sentence summaries that are not simple extractions, but rather new formulations written to reflect the key points of the articles. This demands a model that can comprehend long contexts and generate fluent, contextually appropriate outputs.

4. MOTIVATION

The ever-increasing availability of online textual content, from news articles to academic papers, has led to a growing demand for automatic summarization tools. Users often do not have the time or capacity to read through long documents, and therefore need concise representations that preserve the essential information. This has motivated extensive research in the field of text summarization, especially through the lens of deep learning and transformers.

Transformer models like BART, T5, and Pegasus have redefined the state-of-the-art in sequence-to-sequence tasks, offering robust solutions for language generation problems. The ability of these models to learn contextual representations and generate fluent text has made them a prime candidate for summarization tasks. However, training such models from scratch is often impractical due to the massive computational resources and large-scale datasets required.

This project is motivated by the success of transfer learning, where a powerful pretrained model like BART can be fine-tuned on a smaller dataset such as XSum to produce impressive results. This makes advanced NLP accessible even to those with limited resources. Furthermore, the real-world utility of such a summarization system—in news summarization, executive briefings, or content recommendation engines—adds practical value and relevance to the project.

5. LITERATURE SURVEY

1. Transformers in NLP: The transformer architecture, introduced by Vaswani et al. (2017), has become the foundation of state-of-the-art NLP systems. Models such as BERT, GPT, T5, and BART are built on transformers and have demonstrated superior performance in various tasks including summarization, translation, and question answering.

2. BART Model: According to Lewis et al. (2020), BART (Bidirectional and Auto-Regressive Transformers) combines the strengths of both BERT (for understanding) and GPT (for generation), making it well-suited for sequence-to-sequence tasks. The model has shown exceptional results in abstractive summarization benchmarks, especially when fine-tuned on datasets like CNN/DailyMail or XSum.

3. XSum Dataset: Narayan et al. (2018) introduced the XSum dataset as a challenging corpus for single-sentence abstractive summarization. Each summary is human-written and designed to be highly abstractive, making it ideal for testing the generation capabilities of transformer models.

4. Transfer Learning: As highlighted in Howard and Ruder (2018), transfer learning significantly reduces the need for large task-specific datasets. Pretrained language models, when fine-tuned, can achieve strong performance with relatively limited training data and compute.

5. Evaluation Metrics: The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric is commonly used to evaluate summarization models by comparing generated summaries to human-written references. BLEU and METEOR are also used in broader NLP tasks to evaluate fluency and semantic similarity.

6. Tools and Frameworks: The HuggingFace Transformers library provides easy access to pretrained models and training utilities. It has been widely adopted for fine-tuning transformer models on specific NLP tasks, including summarization.

6. SUMMARY OF PROPOSED APPROACH

The proposed approach in this project follows a structured and modular pipeline, enabling the effective fine-tuning of the BART transformer model for summarization. The pipeline begins with data preparation, using the HuggingFace Datasets library to load the XSum dataset, followed by selecting a manageable subset of 1,000 training and 200 validation examples. This allows for faster experimentation while maintaining the integrity of the task. Next, the pretrained BART model (**facebook/bart-base**) and its tokenizer are loaded. Text data is tokenized using appropriate truncation and padding to ensure that both input and output sequences fit within the model's

constraints—512 tokens for input and 128 for output.The HuggingFace Trainer API is utilized to manage the training process. Key components such as optimizer selection (AdamW), learning rate (2e-5), and batch size (4) are chosen to balance memory usage and training efficiency.

After training, model performance is evaluated using ROUGE metrics, which compare generated summaries with the reference summaries. The project also demonstrates the model's capabilities by generating a summary for a randomly selected article.This approach showcases how to integrate pretrained models with real datasets, process data efficiently, and evaluate performance—all in a reproducible, modular manner using standard NLP tools.

7. METHODOLOGY

The methodology adopted in this project consists of a clearly defined series of steps to fine-tune a transformer model using modern NLP libraries. The project leverages HuggingFace's Datasets and Transformers libraries along with PyTorch for model training and evaluation.

1. Dataset Loading and Subsetting
   The XSum dataset is loaded using datasets.load_dataset(). A subset (1,000 training and 200 validation examples) is selected for quick experimentation.

2. Model and Tokenizer Initialization
   The pretrained facebook/bart-base model and its corresponding tokenizer are loaded. The tokenizer processes both input articles and target summaries, applying truncation and padding.

3. Preprocessing
   Each article-summary pair is tokenized into input and target sequences. Token lengths are set to 512 (input) and 128 (summary), ensuring that most samples fit within the BART model's capacity.

4. Fine-Tuning
   HuggingFace's Trainer API is configured with:

   - Model: BART (facebook/bart-base)
   - Batch size: 4
   - Epochs: 1 (proof of concept)
   - Optimizer: AdamW with learning rate 2e-5
   - Weight decay: 0.01

5. Evaluation
   ROUGE-1, ROUGE-2, and ROUGE-L metrics are calculated on the validation set to measure summary overlap with reference summaries.

6. Sample Generation
   The model is tested on unseen validation articles. Predicted summaries are compared to the reference summaries and the original articles for qualitative assessment.

8. CONCLUSION

This project successfully demonstrates the use of a pretrained transformer model—BART—for the task of abstractive text summarization using the XSum dataset. The primary goal was to implement and evaluate a complete pipeline from data loading and preprocessing to model fine-tuning and performance evaluation.

Despite being trained for only one epoch on a small subset of data, the model was capable of generating semantically meaningful summaries. This highlights the strength of transfer learning and the ability of pretrained models to generalize even with limited fine-tuning. Evaluation using ROUGE metrics indicated that the model was able to capture key ideas from the original articles effectively, confirming its potential for deployment in real-world applications.

9. REFERENCES

[1] Vaswani, A., et al. (2017). *Attention is All You Need*. NeurIPS.

[2] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. arXiv:1910.13461.

[3] Narayan, S., Cohen, S. B., & Lapata, M. (2018). *Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization*. EMNLP.

[4] Lin, C. Y. (2004). *ROUGE: A Package for Automatic Evaluation of Summaries*. ACL Workshop.

[5] Wolf, T., et al. (2020). *Transformers: State-of-the-Art Natural Language Processing*. EMNLP Demos.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805.