SCTR's Pune Institute of Computer Technology, Pune

A PROJECT REPORT ON

Fine Tuning a Pretrained Transformer for Summarization

SUBMITTED BY

Class : BE 4 (R4)

Name : Om G. Panchwate

Roll No. : 41454

Under the guidance of

Prof. N. Y. Kapadnis

DEPARTMENT OF COMPUTER ENGINEERING

Title:

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

Problem Statement:

Finetune a pre trained transformer model for the text summarization task using the XSum dataset. The aim is to build a system that can generate concise and coherent summaries from given articles.

Learning Objectives:-

- Understand and apply transfer learning with transformers, including the concepts of encoder-decoder architecture, attention mechanisms, and the benefits of fine-tuning versus training from scratch.
- Explore preprocessing and tokenization strategies for text data using tools like HuggingFace Tokenizers, and examine the impact of padding, truncation, and maximum sequence lengths on model performance.
- Evaluate the model's summarization capability using ROUGE metrics, while also considering additional evaluation methods such as BLEU or METEOR for comprehensive performance assessment.

Learning Outcomes :-

CO4: Analyze performance of an algorithm.
CO5: Implement an algorithm that follows one of the following algorithm design strategies: divide and conquer, greedy, dynamic programming, backtracking, branch and bound.

Theory:

Transformer models such as BART (Bidirectional and Auto-Regressive Transformers) are powerful architectures for natural language understanding and generation tasks, especially in scenarios requiring context-rich comprehension and fluent generation of text. These models utilize self-attention mechanisms to process and relate different parts of the input sequence efficiently, making them particularly effective for sequence-to-sequence applications like summarization, translation, and question answering.

Pretrained models like "facebook/bart-base" are initially trained on large-scale corpora to learn a broad understanding of language patterns, syntax, and semantics. These models are then fine-tuned on task-specific datasets, which helps them specialize in generating outputs that are closely aligned with the requirements of a particular task. This transfer learning approach greatly reduces the need for massive labeled datasets and computational resources for each new application.

In this project, the BART model is fine-tuned using the XSum dataset, which consists of BBC news articles and their corresponding single-sentence summaries written to capture the essence of each article. The dataset is well-suited for abstractive summarization because the summaries are not merely extracted sentences but are generated independently, demanding a high level of abstraction and language generation ability from the model. By training the model on this dataset, we aim to enable it to generate concise, informative, and contextually accurate summaries from longer articles.

Methodology

1. Load the XSum dataset using the HuggingFace Datasets library.
2. Select a subset of 1000 training and 200 validation examples for quicker training.
3. Load the pretrained BART model and tokenizer.
4. Preprocess the data using tokenization with truncation to fit maximum input and target lengths.
5. Use the HuggingFace Trainer API to fine-tune the model.
6. Define and compute ROUGE metrics to evaluate summarization performance.
7. Generate a summary for a randomly selected validation article to demonstrate the model's output.

Experimental Setup:

- Environment: The experimental setup included a Python-based programming environment using the PyTorch framework for deep learning and the Transformers library by HuggingFace for accessing pretrained models and tools. The XSum dataset was employed for training and evaluation, providing a rich source of real-world summarization examples.

- Model: The model used was "facebook/bart-base," a transformer-based sequence-to-sequence model that combines the capabilities of both encoder and decoder for high-quality text generation tasks.

- Epochs: The training was conducted for 1 epoch to demonstrate proof of concept and obtain preliminary results. Longer training can lead to improved generalization.

- Batch Size: A mini-batch size of 4 was used during both training and evaluation to accommodate memory constraints while maintaining learning efficiency.

- Max Input Length: 512 tokens, which is the standard maximum for BART and accommodates the majority of article lengths without truncation.
- Max Target Length: 128 tokens, sufficient for producing concise and informative summaries.

- Optimizer Settings: The AdamW optimizer was used with a learning rate of 2e-5 and a weight decay of 0.01 to promote stable learning and avoid overfitting.

Results:

After completing the training, the fine-tuned BART model was capable of generating abstractive summaries that effectively captured the essential meaning and key ideas from the input documents. Despite being trained for a short duration on a limited dataset subset, the model demonstrated reasonable summarization quality. Evaluation using ROUGE metrics provided a quantitative assessment of performance and confirmed the model's potential for further improvements with extended training and larger datasets.

Sample Output: Original Article:

The hyperbaric chamber, which treats divers with "the bends", was operated by St John's Ambulance on a donation basis until it broke in April 2014. The health department replaced it in 2015, but says it needs to "balance the books". Diving instructor Steve Bougourd said he was "gobsmacked". "I'm just worried that this kind of cost will put people off of actually going to the [hospital] and notifying them if they suspect a problem," he said. "We may find it's going to be very expensive to get out divers insured." In the UK hyperbaric oxygen treatment is covered by the NHS, but Guernsey has its own health care system. Source: NHS Assistant director at Guernsey's health and social care department (HSC) Ed Freestone said renting the chamber was costing the government £60,000 a year. He said the department would not make a profit from the new charges, which were based on "the average usage that we could identify over the previous few years". In addition to paying for the training of staff and the maintenance of a 24 hour service, the department had to fund plans to buy its own chamber for about £250,000, Mr Freestone said. Commercial divers already pay a £150 notification fee to dive which raises about £10,000 a year, according to HSC.It is a legal requirement to provide a hyperbaric chamber facility for commercial diving activity to take place within Guernsey's 12-mile limit.

Reference Summary:

Divers in Guernsey will be hit with a £30,000 charge if they require treatment for decompression sickness, the government has confirmed.

Predicted Summary:

The health department of Guernsey is planning to buy its own hyperbaric chamber to pay £60,000 a year.

Conclusion:

This project clearly illustrates the practical utility and efficiency of using pretrained transformer-based architectures for the task of text summarization. With only minimal training and computational resources, the BART model was able to generate fluent, accurate summaries that were contextually appropriate and semantically meaningful. The experiment underscores the advantages of transfer learning, where a powerful base model can be quickly adapted to a specific task with impressive results. The fine-tuning process proved effective and scalable, making it an ideal solution for real-world summarization applications, especially in domains where concise content is critical.

Future Scope:

- Fine-tune with a larger dataset size, higher number of training epochs, and hyperparameter tuning to achieve improved accuracy and generalization.

- Explore and compare performance with alternative architectures such as T5, Pegasus, and LongT5 to evaluate strengths and weaknesses across various tasks.

- Integrate the trained summarization model into a user-friendly web-based or desktop application capable of providing on-demand summaries for input articles, documents, or web content.

- Investigate domain-specific fine-tuning by using specialized datasets, such as medical, legal, or scientific documents, to improve summarization quality for niche applications.