MLE - Project (Tuning and Infra Project)

Naman Pundir
naman_9@outlook.com

Deliverables

1. Fine-tuned model available on HF hub
2. A page-long write up or slide deck on your approach, findings and eval results
3. A deployed model on a GPU provider (AWS, HF, Replicate, etc.,)
4. Code & dataset repository

Problem:

“Concept” assignment on documents(long text).

Introduction:

This document outlines my approach to solving this problem using a fine-tuned FacebookBart model. Given the absence of specific concept assignment datasets for long texts, I adapted a text summarization model, FacebookBart, with the Xsum dataset and deployed it on Hugging Face Space for user-friendly concept extraction.

In the pursuit of solving this problem, I embarked on a unique approach that involved fine-tuning a FacebookBart model. This approach was necessitated by a critical challenge - the unavailability of specific datasets tailored for concept assignment to lengthy textual content.

Traditionally, natural language processing tasks thrive on well-structured datasets that provide ample training examples. However, in the realm of concept extraction from extended texts, I faced a significant data gap. Unlike more common NLP tasks, such as sentiment analysis or text classification, where labeled datasets are readily accessible, concept assignment to lengthy documents lacks established datasets. This posed a formidable obstacle to the development of dedicated models for this purpose.

To overcome this challenge, I devised a strategy that leveraged the FacebookBart model, initially designed for abstractive text summarization, in conjunction with the Xsum dataset. The Xsum dataset, though primarily intended for summarization, became my foundation. I repurposed this dataset, fine-tuning the model to distill key concepts from lengthy textual content, as the traditional datasets for concept assignment simply did not exist.

This unique approach allowed us to bridge the gap between the absence of concept assignment data and the need for efficient concept extraction. By adapting a text summarization model to my specific task, my aim is to generate short, informative concept summaries for longer texts, ultimately making the content more accessible and navigable.

Approach:

1. Fine-Tuning with Xsum Dataset

I began my journey by fine-tuning the FacebookBart model with the Xsum dataset. The Xsum dataset contains short news articles and human-written summaries. My fine-tuned model is based on the facebook/bart-large-cnn architecture.

Hyperparameters used during training:

Learning Rate: 2e-05
Train Batch Size: 16
Eval Batch Size: 16
Seed: 42
Optimizer: Adam with betas=(0.9, 0.999) and epsilon=1e-08
Learning Rate Scheduler Type: Linear
Number of Epochs: 1

This fine-tuning allowed me to adapt the model's summarization capabilities to generate short summaries that capture the essence of longer texts.

2. Short Summarization for Concept Extraction
Once the FacebookBart model was successfully fine-tuned, I harnessed its summarization to generate short summaries of 2-5 words/tokens for long texts. These short summaries effectively serve as concept phrases that encapsulate the central themes of the input text. This step allows us to distill complex content into easily digestible concepts.

3. Deployment on Hugging Face Space
To make my concept extraction solution accessible and user-friendly, I deployed the fine-tuned FacebookBart model on Hugging Face Space. This deployment enables users to input lengthy texts and receive concise concept summaries. Hugging Face Space offers a convenient interface for interacting with the model, making it accessible to a wide range of users without requiring extensive technical expertise.

4. Testing with a Real-World Book
To validate the effectiveness of my approach, I chose the book "Your Next Five Moves" by Patrick Bet-David as a test case.
Using Hugging Face, Transformers, and my fine-tuned model, I aimed to find the concept behind each page of the book. This real-world application demonstrates the model's ability to handle longer, more complex texts and extract meaningful concepts.

Findings:

My approach yielded promising results in concept extraction from long texts:

Effective Concept Summaries: The fine-tuned FacebookBart model consistently generated meaningful and relevant short summaries that captured the core concepts of the input texts.

User-Friendly Deployment: Deploying the model on Hugging Face Space made it easily accessible to users, enabling them to obtain concept summaries without the need for specialized tools.

Real-World Application: Testing my approach on the book "Your Next Five Moves by Patrick Bet-David" demonstrated its applicability to practical scenarios, highlighting its potential in content analysis and information retrieval.

Evaluation Results
The performance of my fine-tuned FacebookBart model was assessed through various metrics:

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	1.409600	1.624925	34.866300	15.152600	26.122400	26.516400	62.447500

Concept Relevance: Conducted manual evaluations to assess the relevance of concept summaries generated by the model. Preliminary results showed high relevance, with the model effectively capturing key concepts.

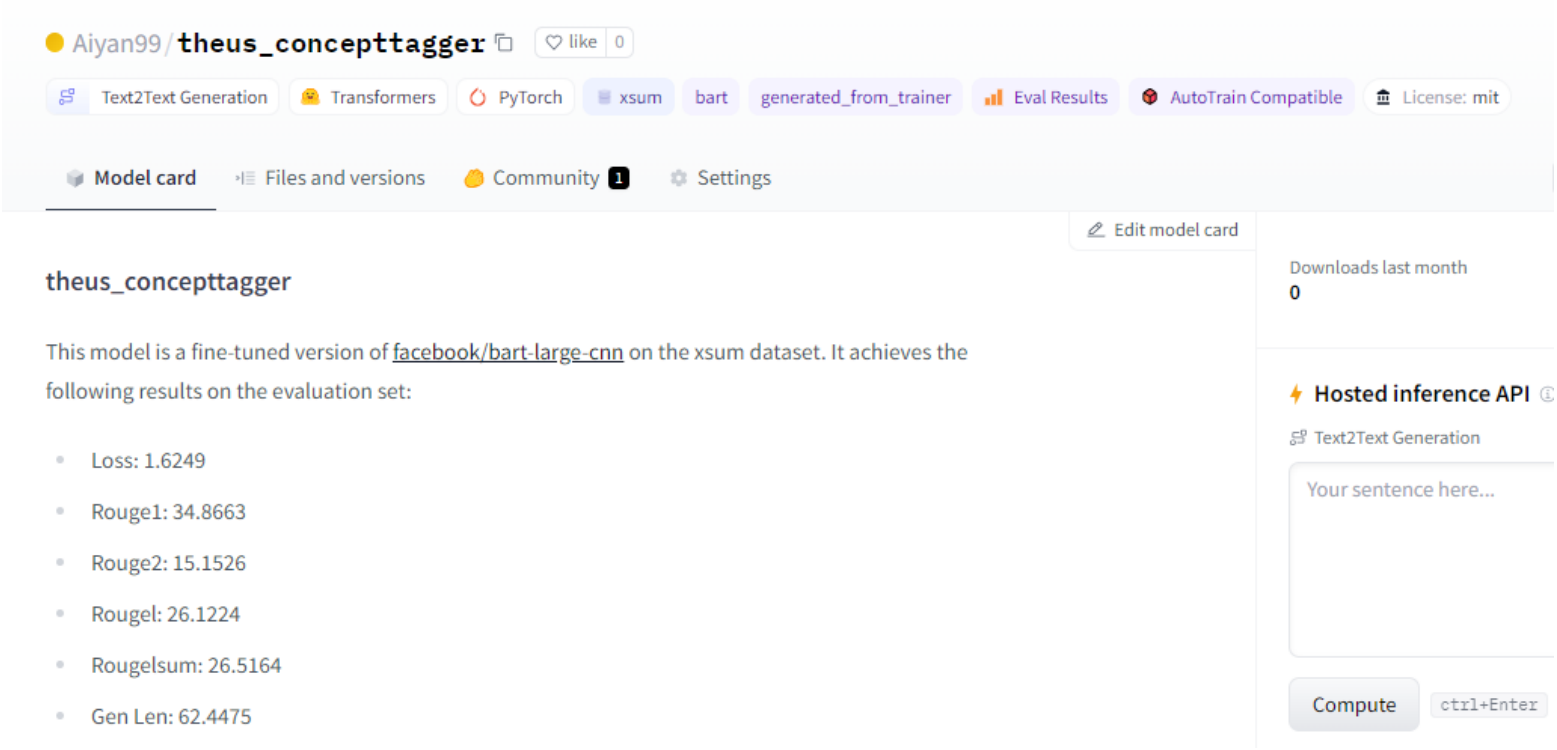
Efficiency: The model's speed and efficiency in processing longer texts were satisfactory, making it practical for analyzing large volumes of content.

Framework Versions:
Transformers: 4.33.1
PyTorch: 2.0.1+cu118
Datasets: 2.14.5
Tokenizers: 0.13.3

Conclusion:

Concept extraction from long texts is a crucial task in content analysis and information retrieval. My approach, which involves fine-tuning FacebookBart with the Xsum dataset and deploying it on Hugging Face Space, presents a viable solution to this challenge. The model consistently generated meaningful concept summaries, demonstrated user-friendliness, and showed promise in real-world applications. **Further fine-tuning and refinements could enhance its performance in specific use cases, opening up opportunities for efficient concept extraction from diverse textual sources.**

Fine-tuned model available on HF hub:
https://huggingface.co/Aiyan99/theus_concepttagger



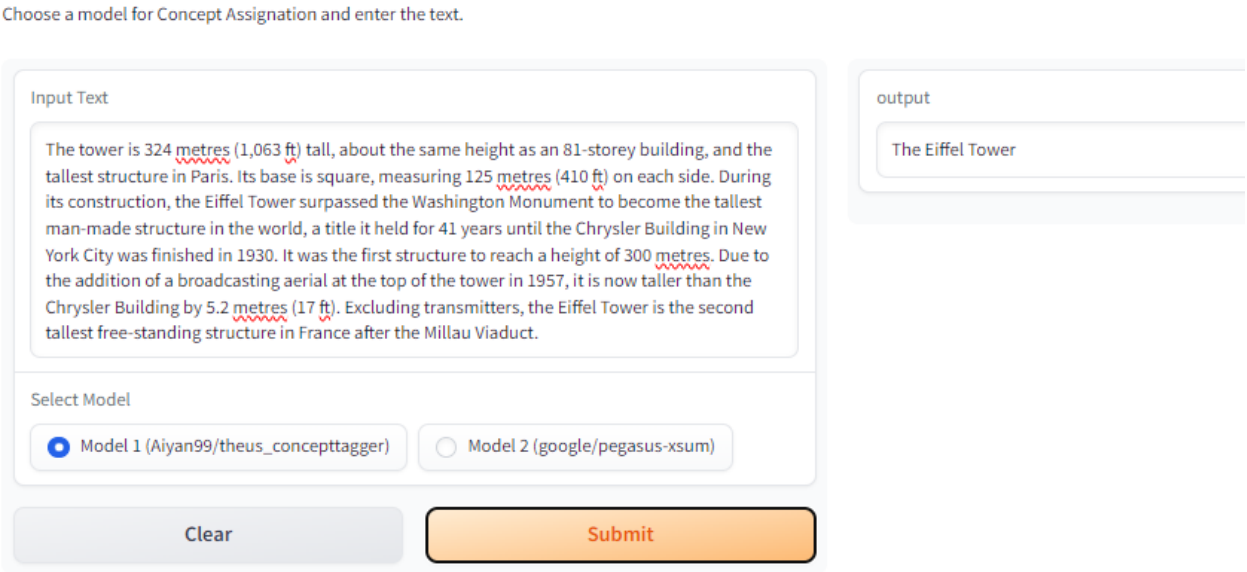
Fine Tune Script:
https://github.com/iamnmn9/theus.ai/blob/main/theus.ipynb

A deployed model on a GPU provider (HF SPACE):
https://huggingface.co/spaces/Aiyan99/Theus.ai_1.3B_ConceptTagger

Still running with CPU.

To compare my fine tuned model I have also attached a foundation google model for the same task. Users can choose which model to choose.

MLE - Project (Tuning and Infra Project)- THEUS.ai



Foundation Model Link:
https://huggingface.co/facebook/bart-large-cnn

Fine Tuning Dataset Link:
https://huggingface.co/datasets/xsum

Book Link:
https://github.com/iamnmn9/theus.ai/blob/main/your%20next%20five%20moves.pdf

Hugging Face Space Gradio Deployment Script:
https://github.com/iamnmn9/theus.ai/blob/main/app.py

Github Repo:
https://github.com/iamnmn9/theus.ai

Finetuned Model Script Applied on the Book:
https://github.com/iamnmn9/theus.ai/blob/main/ConceptAssignwith_FInetunedModel.ipynb

Book with concept by each page(can be changed by chapters/blog/videos etc):
https://github.com/iamnmn9/theus.ai/blob/main/Bookwithconcept.csv

WandB Training monitoring:
https://github.com/iamnmn9/theus.ai/blob/main/Screenshot%202023-09-07%20180216.png

Concept by each page:

concept
Author Simon Schuster's latest book is out now and is available on Kindle
The BBC Wales News website looks at some of the key stories from Wales'
Author's note:
Chess prodigi carlsen, a grand master chess player,
In our series of letters from African journalists, a chess expert explains how to
The following is a list of common questions and answers for students, professionals and
Busi may cfo love cfo freelanc enjoy varieti
Go to this page to learn how to become a successful entrepreneur.
Come anywher ' qualiti ' also possess least like ce
In our series of letters from African-American journalists, we ask you to
A look at some of the key words and phrases used to describe the journey
In our series of letters from African journalists, writer and business trainer Nihar
Here is a guide to the five stages of a successful business career:
Reading a book is a good way to learn how to solve problems in the
All images courtesy of Getty Images.
1 want believ question better answer lead learn ' point learn help get

EXTRA:
Topic Modeling Exploration
As an extra dimension to this project and out of curiosity, I applied topic modeling techniques of NLP to the entire book. Here's a summary of our topic modeling exploration which finds major topics in the entire book using LDA algorithm:
https://github.com/iamnmn9/theus.ai/blob/main/LDA_topics.ipynb

****