CS 6140: Machine Learning
Project Report

Fine-tuning Pretrained Language Model for Document Summarization

Sideeshwaran Lakkapuram Balasubramani
Sachin Palahalli Chandrakumar

Abstract

This project works on natural language processing (NLP), with a focus on enhancing document summarization by fine-tuning pretrained language models. The primary goal is to augment the capabilities of models like BERT or Pegausus, leveraging their extensive linguistic knowledge to generate concise and coherent summaries for lengthy documents. By undertaking a transfer learning approach, the project seeks to adapt these models to the specifics of document summarization, contributing to the advancement of language processing techniques.

The significance of this project extends to various domains, including information retrieval, content curation, news aggregation, and academic research. In a society where the rapid assimilation of vast amounts of information is essential, an advanced summarization model can be an important asset, saving time and resources. Furthermore, document summarization can be beneficial in scenarios where quick decision-making based on comprehensive yet concise information is necessary.

Overview

Document summarization is an important task in NLP, with applications in information retrieval, content summarization, and knowledge extraction. The project explores the fine-tuning of pretrained language models, facebook/bart and google/pegausus, to adapt them to the specifics of document summarization. The approach involves training the model on a specific summarization dataset, refining its understanding of document structures and content relevance. The goal of the project is to create a model that generates concise and informative document summaries. The main motivation behind the project is to provide a tool that can be useful in efficient information extraction and enables users to quickly grasp the key points in lengthy documents.

The significance of this project lies in its potential to streamline information consumption. Efficient document summarization can benefit a wide range of applications, including content curation, news aggregation, and academic research. In a society flooded with information, an advanced summarization model can be a valuable asset, saving time and resources. Additionally, the technology can be applied in scenarios where quick decision-making based on comprehensive yet concise information is essential.

The project proposes two approaches for fine tuning the language models available. The first approach is to fine tune pretrained generic language models for the purpose of document summarization, and the second approach involves fine tuning large language models pre-trained for document summarization for a specific use case or a dataset. The project fine tunes the the facebook/bart-base model, which is a generic language model on the cnn_dailymail dataset, and compares it with the facebook/bart-large-cnn model. The project fine tunes models facebook/bart-large-cnn on ccdv/govreport-summarization, google/pegasus-cnn_dailymail on on ccdv/govreport-summarization, google/pegasus-cnn_dailymail on samsum to achieve all the approaches specified above. The project uses evaluation metrics such as ROUGE scores to assess the models' performance in generating summaries that align with human-generated reference summaries.

The rationale behind fine-tuning pretrained language models is their ability to capture intricate language patterns and semantic structures for which they are already trained. The standard approaches to document summarization have been rule-based or extractive. Leveraging pretrained language models leads to a more abstract and context-aware summarization. The project analyzes how efficient the selected models are on each type of dataset chosen for summarization.

Datasets

CNN_Dailymail

The CNN/Daily Mail dataset is a widely used benchmark dataset in the field of natural language processing, particularly for text summarization tasks. The dataset consists of over 300k unique news articles written by journalists at CNN and the Daily Mail paired with human-generated summaries, providing a valuable resource for developing summarization algorithms. The articles cover a diverse range of topics, including current events, politics, sports, and entertainment.

For each data instance, there are attributes for the article, the highlight, and the id. The id is a string containing the heximal formated SHA1 hash of the url where the story was retrieved from, the article is a string containing the body of the news article and the highlights is a string containing the highlight of the article as written by the article author.

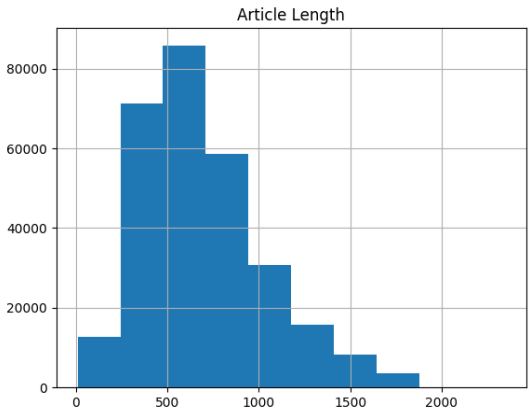
Dataset Split	Number of Instances
Train	287,113
Test	13,368
Validation	11,490

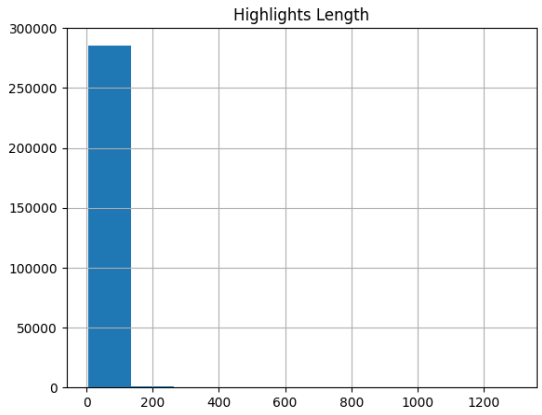
The dataset is relatively large, containing three hundred thousands of articles with their corresponding summaries. It is split into training, validation, and test sets for training and evaluating models.

The summaries provided in the data set are fairly small when compared to the other datasets such as ccdv/govreport - summarization which provide long summaries and also have very long articles to work with.

Feature	Mean Token Count
article	287,113
highlight	13,368

The distribution of size of articles and the summaries in the dataset is given in the graph below.





ccdv/govreport - summarization

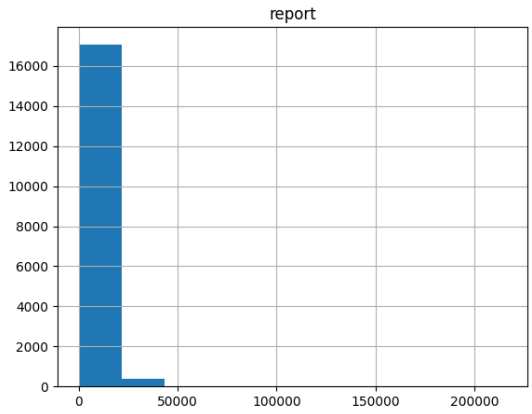
The GOVREPORT dataset is a new large-scale dataset consisting of about 19.5k U.S. government reports with abstractive summaries written by experts. The words in the dataset are significantly longer, and the dataset contains long summaries. Therefore, this dataset would serve as a perfect benchmark to compare with the model trained with ccn_dailymail and evaluate how the base model fairs with the two types of datasets.

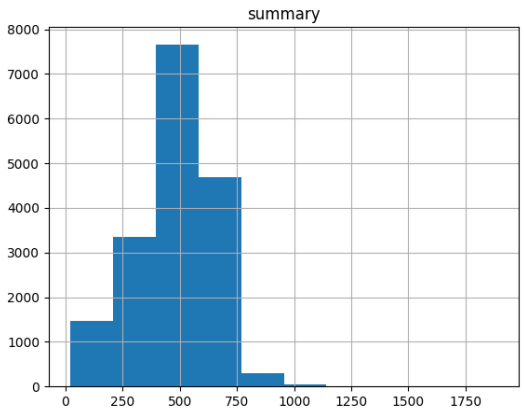
For each data instance, there are attributes for the report, the summary, and the id. The reports attribute consists of the long US government reports, and the summary consists of a string that contains expert written summaries. Each summary, on average, contains 500 words, and the reports contain about 9000 words.

Dataset Split	Number of Instances
Train	17,517
Test	973
Validation	973

The dataset contains about seventeen thousand reports with their corresponding summaries. It is split into training, validation, and test sets for training and evaluating models.

The distribution of the size of reports and the summaries in the dataset are given in the graph below.





SAMSum

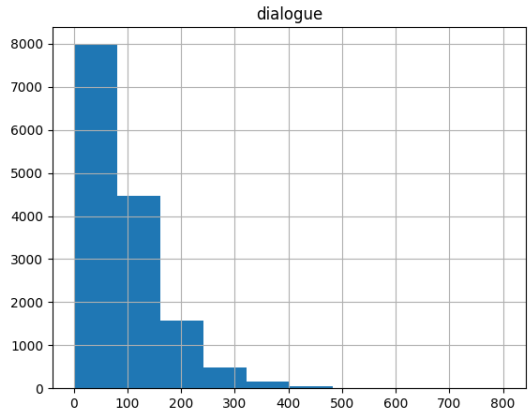
The SAMSum dataset contains about 16k messenger-like conversations with summaries. The conversations were written by linguists fluent in English. Linguists were asked to create conversations similar to those they write on a daily basis, reflecting the proportion of topics in their real-life messenger conversations. The style and register are diversified - conversations could be informal, semi-formal or formal; they may contain slang words, emoticons, and typos.

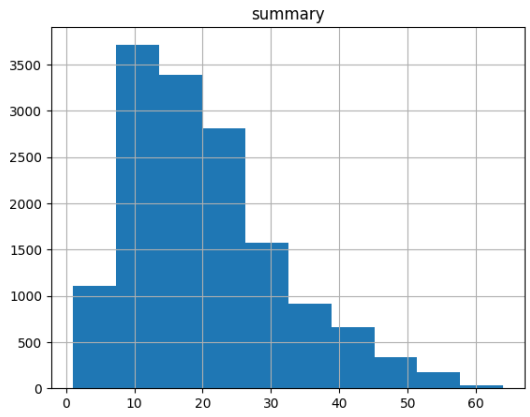
The dataset is made up of 16369 conversations distributed uniformly into 4 groups based on the number of dialogues in conversations: 3-6, 7-12, 13-18 and 19-30. Each dialouge contains the name of the speaker. Most conversations consist of dialogues between two participants (about 75% of all conversations); the rest are between three or more people.

Dataset Split	Number of Instances
Train	14,732
Test	818
Validation	819

The dataset contains about fourteen thousand reports with their corresponding summaries. It is split into training, validation, and test sets for training and evaluating models.

The distribution of the size of the dialouges and the summaries in the dataset are given in the graph below.





Large Language Models (LLMs) chosen

facebook/bart-base

The "facebook/bart-base" refers to a pretrained language model known as BART (Bidirectional and Auto-Regressive Transformers) developed by Facebook AI. BART is designed for various natural language processing (NLP) tasks.

When fine-tuned for specific tasks, the "facebook/bart-base" model has demonstrated remarkable effectiveness in text generation. Fine-tuning allows the model to adapt its pre-trained knowledge to particular domains or tasks, making it highly versatile for applications such as content summarization, document understanding, and other text-related tasks.

facebook/bart-large-cnn

The facebook/bart-large-cnn is a variant of the BART (Bidirectional and Auto-Regressive Transformers) model developed by Facebook AI. Similar to the base BART model, "facebook/bart-large-cnn" is pretrained using a denoising autoencoder objective. It learns to reconstruct the original text from corrupted or masked versions, enhancing its understanding of language structures.

The model is well-suited for fine-tuning on specific tasks, particularly in text generation applications. It can be adapted for tasks such as abstractive summarization, and it excels at producing coherent and relevant summaries of input documents. This particular varian has been fine-tuned on the CNN Daily Mail dataset.

Due to its increased parameter size, facebook/bart-large-cnn offers improved performance on tasks compared to the smaller variants. It's particularly beneficial for tasks that demand a more extensive language understanding.

google/pegasus-cnn_dailymail

The google/pegasus-cnn_dailymail is a transformer-based model developed by Google, specifically designed for abstractive text summarization. The model is pretrained on the CNN/Daily Mail dataset, a widely used benchmark for summarization tasks. The model learns to predict masked or missing tokens in a sequence, which helps it capture contextual relationships and semantic understanding.

In abstractive summarization, the model doesn't simply extract sentences from the input but also generates summaries that may include novel phrases and rephrased content.

Results:

Model = Bart Base, Data Set = CNN Daily Mail

In this experiment, we took the base model Bart Base which is a sequence to sequence model which is pre-trained for learning a model to reconstruct the original text. In this, we are fine tuning the model to do text summarization for CNN Daily Mail data Set. The reason why we took this is, there is already a fine tuned Bart model, which is called the Bart-CNN model. So, we did this to compare the results of how it fares when compared to the already existing fine tuned model.

Hyper Parameters used:

Batch size: 8

Warm-up steps: 500

Weight-decay: 0.01

Gradient_accumulation_steps: 16

For 5 epochs, training_loss = 0.102

Below is the results for a test data:

Sample Output:

```
Article:
LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a reported £20 million ($41.1 million) fortune as he turn
s 18 on Monday, but he insists the money won't cast a spell on him. Daniel Radcliffe as Harry Potter in "Harry Potter and the Order of th
e Phoenix" To the disappointment of gossip columnists around the world, the young actor says he has no plans to fritter his cash away on
fast cars, drink and celebrity parties. "I don't plan to be one of those people who, as soon as they turn 18, suddenly buy themselves a m
assive sports car collection or something similar," he told an Australian interviewer earlier this month. "I don't think I'll be particul
arly extravagant. "The things I like buying are things that cost about 10 pounds -- books and CDs and DVDs." At 18, Radcliffe will be abl
e to gamble in a casino, buy a drink in a pub or see the horror film "Hostel: Part II," currently six places below his number one movie o
n the UK box office chart. Details of how he'll mark his landmark birthday are under wraps. His agent and publicist had no comment on his
plans. "I'll definitely have some sort of party," he said in an interview. "Hopefully none of you will be reading about it." Radcliffe's
earnings from the first five Potter films have been held in a trust fund which he has not been able to touch. Despite his growing fame an
d riches, the actor says he is keeping his feet firmly on the ground. "People are always looking to say 'kid star goes off the rails,'" h
e told reporters last month. "But I try very hard not to go that way because it would be too easy for them." His latest outing as the boy
wizard in "Harry Potter and the Order of the Phoenix" is breaking records on both sides of the Atlantic and he will reprise the role in t
he last two films.  Watch I-Reporter give her review of Potter's latest » . There is life beyond Potter, however. The Londoner has filmed
a TV movie called "My Boy Jack," about author Rudyard Kipling and his son, due for release later this year. He will also appear in "Decem
ber Boys," an Australian film about four boys who escape an orphanage. Earlier this year, he made his stage debut playing a tortured teen
ager in Peter Shaffer's "Equus." Meanwhile, he is braced for even closer media scrutiny now that he's legally an adult: "I just think I'm
going to be more sort of fair game," he told Reuters. E-mail to a friend . Copyright 2007 Reuters. All rights reserved.This material may
not be published, broadcast, rewritten, or redistributed.

Highlights:
Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday .
Young actor says he has no plans to fritter his cash away .
Radcliffe's earnings from first five Potter films have been held in trust fund .

Model Highlight:
Harry Potter star Daniel Radcliffe gains access to a reported £20 million fortune .<n>The young actor says he has no plans to fritter his
cash away .<n>Radcliffe's earnings from the first five Potter films have been held in a trust fund .
```

Rouge scores are as follows:

Rouge 1: 0.53

Rouge 2: 0.3

RougeL: 0.26

Rouge-Lsum: 0.25

We can see that the fine tuned model does well w.r.t the summarization. So, we can say that the model does well when fine tuned w.r.t CNN Daily Base model.

Model = Bart CNN, Data Set = Samsum

In this experiment, we took the fine tuned Bart CNN model which is finetuned on CNN daily mail which is pre-trained for learning a model to reconstruct the original text. In this, we are fine tuning the model to do text summarization for Samsum data Set. The reason why we took this is, there it is already fine tuned for Text Summarization. However, the input is a dialogue instead of a text and the output is a summary of it. So, we did this to compare the results of how it fares when trained on a text summarization fine tuned model.

Batch size: 8

Warm-up steps: 500

Weight-decay: 0.01

Gradient_accumulation_steps: 16

For 3 epochs, training_loss = 1.8

Below is the results for a test data:

```
Dialogue:
Tim: Hi, what's up?
Kim: Bad mood tbh, I was going to do lots of stuff but ended up procrastinating
Tim: What did you plan on doing?
Kim: Oh you know, uni stuff and unfucking my room
Kim: Maybe tomorrow I'll move my ass and do everything
Kim: We were going to defrost a fridge so instead of shopping I'll eat some defrosted veggies
Tim: For doing stuff I recommend Pomodoro technique where u use breaks for doing chores
Tim: It really helps
Kim: thanks, maybe I'll do that
Tim: I also like using post-its in kaban style

Summary:
Kim may try the pomodoro technique recommended by Tim to get more stuff done.

Model Summary:
Kim was going to do lots of stuff but ended up procrastinating. She'll move her ass and do everything tomorrow. Tim recommends Pomodoro t
echnique for doing chores.
```

Rouge scores are as follows:

Rouge 1: 0.12

Rouge 2: 0.001

RougeL: 0.1

Rouge-Lsum: 0.09

We can see that the fine tuned model does not do well on the new training dataset as the new data set is little different than the actual articles data. Also, the data set is of very less size and hence, probably not well trained for very less epochs.

So, increasing the epochs to 10, it gives training_loss of around 0.9 which is better than how it fared when it only ran for 3 epochs.

Model = Pegasus-cnn_dailymail, Data Set = Government Data

In this experiment, we have taken a fine tuned Pegasus-cnn_dailymail model which is designed for sequence to sequence tasks and is pre-trained on denoising auto encoder objectives. In this, we are fine tuning the model to do text summarization for the Government data Set in which all the articles are very huge. We want to analyse the model which is fine tuned on Daily Mail CNN for text Summarization on how it fares for summarization of a data relatively different than daily Mail CNN

Hyper Parameters used:

Batch size: 8

Warm-up steps: 500

Weight-decay: 0.01

Gradient_accumulation_steps: 16

For 2 epochs, training_loss = 2.6

Below is the results for a test data:

```
Report Summary:
As the Department of Defense (DOD) has expanded its involvement in overseas military operations, it has grown increasingly reliant on its
federal civilian workforce to support contingency operations. The Senate Armed Services Committee required GAO to examine DOD's policies
concerning the health care for DOD civilians who deploy in support of contingency operations in Afghanistan and Iraq. GAO analyzed over
3,400 deployment-related records for deployed federal civilians and interviewed department officials to determine the extent to which DOD
has established and the military services and defense agencies (hereafter referred to as DOD components) have implemented (1) force healt
h protection and surveillance policies and (2) medical treatment policies and procedures for its deployed federal civilians. GAO also exa
mined the differences in special pays and benefits provided to DOD's deployed federal civilians and military personnel. DOD has establish
ed force health protection and surveillance policies to assess and reduce or prevent health risks for its deployed federal civilian perso
nnel, but it lacks procedures to ensure implementation. Our review of over 3,400 deployment records at eight component locations found th
at components lacked documentation that some federal civilian personnel who deployed to Afghanistan and Iraq had received, among other th
ings, required pre- and post-deployment health assessments and immunizations. These deficiencies were most prevalent at Air Force and Nav
y locations, and one Army location. As a larger issue, DOD lacked complete and centralized data to readily identify its deployed federal
civilians and their movement in theater, further hindering its efforts to assess the overall effectiveness of its force health protection
and surveillance capabilities. In August 2006, DOD issued a revised policy which outlined procedures that are intended to address these s
hortcomings. However, these procedures are not comprehensive enough to ensure that DOD will know the extent to which its components are c
omplying with existing health protection requirements. In particular, the procedures do not establish an oversight and quality assurance
mechanism for assessing the implementation of its force health protection and surveillance requirements. Until DOD establishes a mechanis
m to strengthen its force health protection and surveillance oversight, it will not be effectively positioned to ensure compliance with i
ts policies, or the health care and protection of deployed federal civilians. DOD has also established medical treatment policies for its
deployed federal civilians which provide those who require treatment for injuries or diseases sustained during overseas hostilities with
care that is equivalent in scope to that provided to active duty military personnel under the DOD military health system. GAO reviewed a
sample of seven workers' compensation claims (out of a universe of 83) filed under the Federal Employees' Compensation Act by DOD federal
civilians who deployed to Iraq. GAO found in three cases where care was initiated in theater, that the affected civilians had received tr
eatment in accordance with DOD's policies. In all seven cases, DOD federal civilians who requested care after returning to the United Sta
tes had, in accordance with DOD's policies, received medical examinations and/or treatment for their deployment-related injuries or disea
ses through either military or civilian treatment facilities. DOD provides certain special pays and benefits to its deployed federal civi
lians, which generally differ in type and/or amount from those provided to deployed military personnel. For example, both civilian and mi
litary personnel are eligible to receive disability benefits for deployment-related injuries; however, the type and amount of these benef
its vary, and some are unique to each group. Further, while the survivors of deceased federal civilian and military personnel generally r
eceive similar types of cash survivor benefits, the comparative amounts of these benefits differ.

Model Summary:
The Department of Defense (DOD) relies on the federal civilian personnel it deploys to support essential missions .<n>DOD has established
force health protection and surveillance policies aimed at assessing and reducing or preventing health risks for its deployed federal civ
ilian personnel .<n>DOD’s policies did not require the centralized collection of data on the identity of its deployed civilians, their mo
vements in theater, or their health status .<n>DOD’s force health protection and surveillance policies stipulate that all DOD deploying f
ederal civilians receive theater-specific immunizations to address possible health threats in deployment locations .
```

Rouge scores are as follows:

Rouge 1: 0.001

Rouge 2: 0.00003

RougeL: 0.01

Rouge-Lsum: 0.01

We can clearly see that the model does very badly when trained on this dataset. It is because we have shortened the output summary more than the expected summary and also, the input single

data is very large. The other reason is, it's only trained for only 2 epochs and is not trained well. For 10 epochs, we can see that the training loss is around 1.8. Still not very good, because the output summary is shorter than the actual summary here and hence the bad results. However, if we read through the summary output, it does a fair job in returning approximately good summaries.

Here, for each model and training data set, we have run for 10 epochs. Training each model takes a lot of time and training for 10 epochs will take hours with the computational powers we used. However, with increasing epochs, we might be able to get better results. We did try on multiple batch sizes to increase the time of training but faced memory issues even if the ram memory used is very high. The ideal batch size w.r.t time and space we found was 8 without causing out of memory issues for the datasets we have taken.

Conclusion:

Text Summarization is a very helpful tool in understanding and working on large reports, articles when you don't have much time to go through the details. Evn, multiple applications like news articles can use these models to give gist of the news instead of the whole details which will help the users who just want to get the brief of the articles. There are many such uses of text summarization and we can clearly see that we can fine tune the base model for our required data sets and use it to our use case. Hence, we have tried multiple models and datasets to see how data sets fare across multiple models and which kind of model fits best for which datasets. For our analysis, we can see that the bart base model is fine tuned really very well for the CNN data set. However, the other two datasets and models did not perform as expected. With fine tuning the hyper parameters in a better way and training for more epochs, we can get better results.

Github Code repo link:

https://github.com/Sachin-PC/FineTuningLargeLanguageModels