Fine-tune Bart to Summarize Medical Articles

Steve Abecassis, Chelomo Lubliner, Anna Assouline

The Hebrew University of Jerusalem

abecassis.steve@gmail.com, lubliner.chelomo@gmail.com, emram.anna@gmail.com

Abstract

We present BART, a denoising autoencoder built with a sequence-to-sequence model that can be fine-tuned on a very wide range of end tasks. There are some applications for BART and in this research, we will focus on Single Document Abstractive Summarization. We will fine-tune Bart for summarizing medical articles with a dataset of 552 articles from MEDIQA and achieve new results with the ROUGE metric.

1 BART Model

BART is a denoising autoencoder built with a sequence-to-sequence model that can be fine-tuned on a very wide range of end tasks. It is implemented with a bidirectional encoder over corrupted text and a left-to-right autoregressive decoder

1.1 Architecture

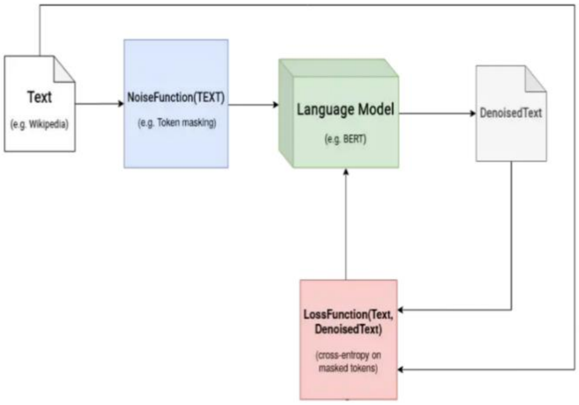
BART uses the standard sequence-to-sequence Transformer architecture, except, following GPT, that ReLU activation functions are modified to GeLUs and parameterinitializeded from N (0, 0.02). There is a base model and a large model that are available.

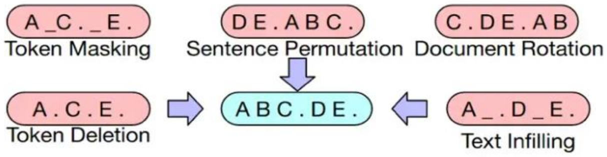
1.2 Training

It is trained by corrupting the text with a noising function, a language mode tries to reconstruct the text and by calculating the loss function (cross entropy over the original text) the language model is optimized (fig. 1).

The Noising functions that corrupt the text are:

- Token Masking: Random tokens are sampled unreplaced by [MASK] elements(following BERT).
- Token deletion: Random tokens are deleted. The model has to predict in which positions there is a missing token.
- Token infilling: A number of text spans are sampled, with span lengths drawn from a Poisson distribution ($\lambda = 3$). Each span is replaced with a single [MASK] token.
- **Sentence Permutation:** Random shuffling of the document's sentences.
- Document Rotation: a token is chosen randomly to be the start of the document, the section before the starting token is appended at the end.





1.3 Tasks

SQuAD: **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed to crowd workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

MNLI, the Multi-Genre Natural Language Inference (MultiNLI) corpus is a crowd-sourced collection of 433k sentence pairs annotated with textual entailment information. The corpus is modeled on the SNLI corpus, but differs in that covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation.

ELI5, a dataset for **long-form question answering**. It contains 270K complex, diverse questions that require explanatory multi-sentence answers. Web search results are used as evidence documents to answer each question.

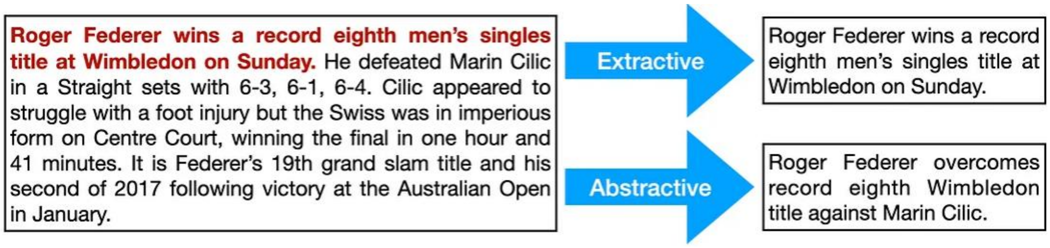
XSum, The Extreme Summarization dataset is a dataset for the evaluation of abstractive single-document summarization systems. The goal is to create a short, one-sentence new summary answering the question "What is the article about?". The dataset consists of 226,711 news articles accompanied by a one-sentence summary. The articles are collected from BBC articles (2010 to 2017) and cover a wide variety of domains (e.g., News, Politics, Sports, Weather, Business, Technology, Science, Health, Family, Education, Entertainment, and Arts). **ConvAI2**: a dialogue response generation task, conditioned on context and a persona.

CNN/Daily Mail. a dataset for text summarization. Human-generated abstractive summary bullets were generated from news stories on CNN and Daily Mail websites as questions (with one of the entities hidden), and stories as the corresponding passages from which the system is expected to answer the fill-in-the-blank question. The authors released the scripts that crawl, extract and generate pairs of passages and questions from these websites.

1.4 Applications

Bart has a lot of applications, such as Text Generation, Question Answering, Machine Translation, and Text Summarization. We will focus on the application of BART for Text Summarization. Summarization is the task of producing a shorter version of a document while preserving its important observation. Some models can extract text from the original input, while other models can generate entirely new text.

There are different types of Text Summarization, Abstractive and Extractive Summarization, and Single document or Multi-document summarization on another side.
Extractive Summarization selects a subset of sentences or words from the text to form a summary.
Abstractive Summarization reorganizes the language in the text and adds novel words/phrases into the summary.
A single Document is an approach in that only one document is summarized.
Multi-Document summarizes information from multiple input documents. The main problem in Multi-Document MDS occurred due to the collection of multiple resources from where the data is extracted, because it may contain the risk of higher redundant information than is generally found in a single document.



Roger Federer wins a record eighth men's singles title at Wimbledon on Sunday. He defeated Marin Cilic in a Straight sets with 6-3, 6-1, 6-4. Cilic appeared to struggle with a foot injury but the Swiss was in imperious form on Centre Court, winning the final in one hour and 41 minutes. It is Federer's 19th grand slam title and his second of 2017 following victory at the Australian Open in January.

Roger Federer wins a record eighth men's singles title at Wimbledon on Sunday.

Roger Federer overcomes record eighth Wimbledon title against Marin Cilic.

This research will focus only on a single document, abstractive summarization. Of course that the next step can be to train our data for extractive and multi-document summarization.

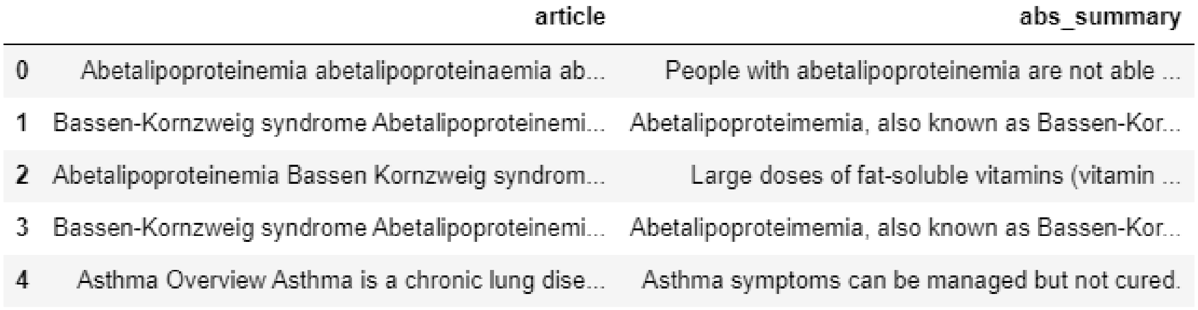
2 Evaluation Metrics

ROUGE or Recall-Oriented Understudy for Gusting Evaluation,[1] is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. (*Wikipedia*)

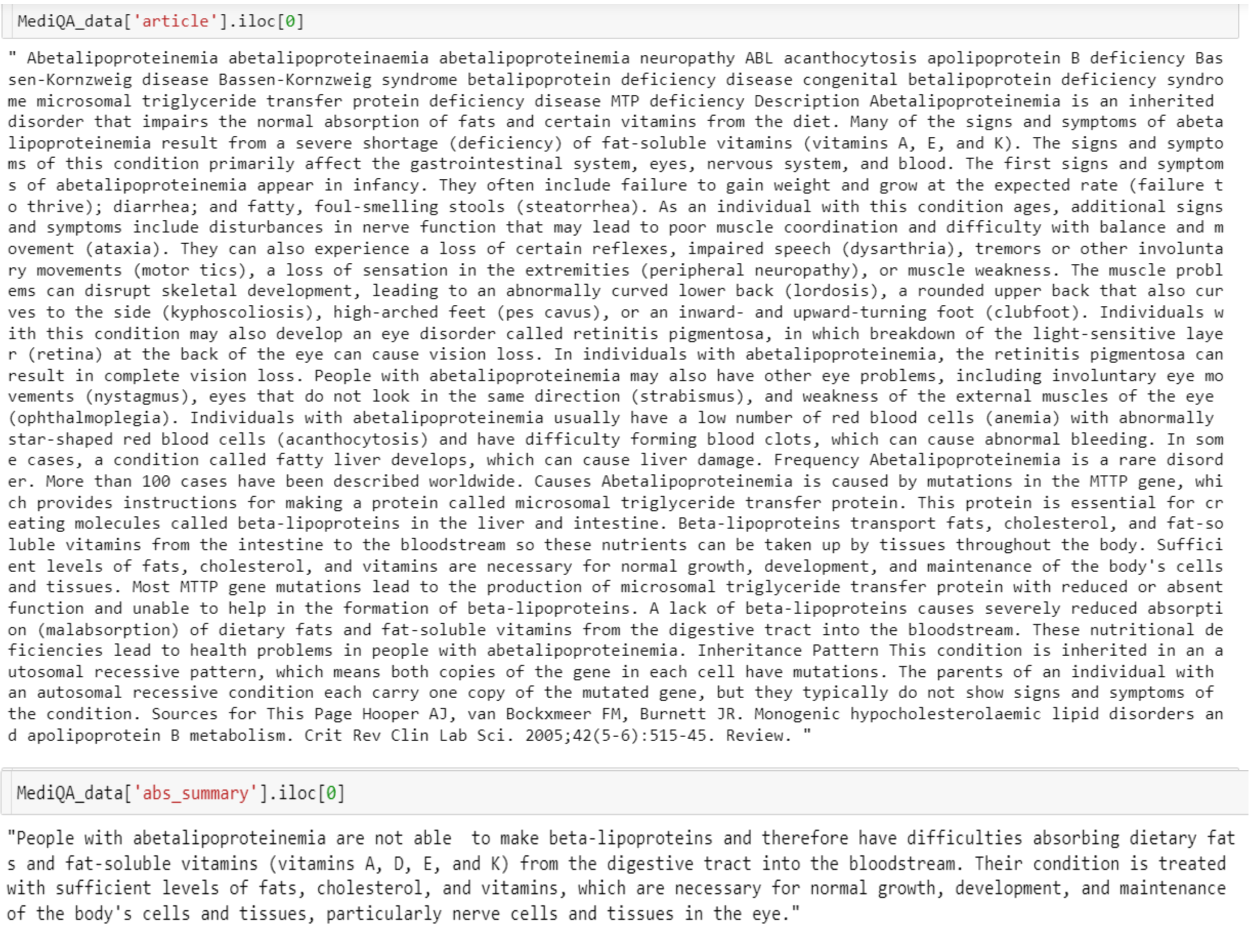
ROUGE measures the number of contiguous words (n-grams in NLP) occurring in a candidate summary compared to a reference summary. For example, ROUGE-2 measures the number of contiguous two-word (bigram) sequences that occur in both the candidate and reference summary, penalizing the candidate for missing bigrams. This means that ROUGE is oriented toward recall We report ROUGE-1, ROUGE-2, and ROUGE-L. Note that ROUGE-L is computed slightly differently than the other ROUGE variations, measuring the longest common subsequence between a candidate and reference.

3 Dataset

MEDIQA-Answer-Summarization is a data set of health questions asked by consumers and the summary, With this dataset, we will fine-tune Bart for health questions.
We were very interested to take a dataset from a newspaper, but for example, the data of the New York Times isn't free, and since we are very interested in AI in Healthcare we have chosen MEDIQA. The dataset for Single document abstractive summarization contains 552 articles and their summaries. We will plot the distribution of the count of words in the article and the summary and set 256 for the maximum target length of the words summary because the upper fence is 168 and the closest exponent of 2 is 256. About the maximum input length, BART is limited to 1024. The data is composed of two columns, the article, and the summarization 'abs_summary' We can see an example of an article and its summary. The distribution of the number of words is plotted both for the Article and the Summary.



	article	abs_summary
0	Abetalipoproteinemia abetalipoproteinaemia ab...	People with abetalipoproteinemia are not able ...
1	Bassen-Kornzweig syndrome Abetalipoproteinemi...	Abetalipoproteimemia, also known as Bassen-Kor...
2	Abetalipoproteinemia Bassen Kornzweig syndrom...	Large doses of fat-soluble vitamins (vitamin ...
3	Bassen-Kornzweig syndrome Abetalipoproteinemi...	Abetalipoproteimemia, also known as Bassen-Kor...
4	Asthma Overview Asthma is a chronic lung dise...	Asthma symptoms can be managed but not cured.

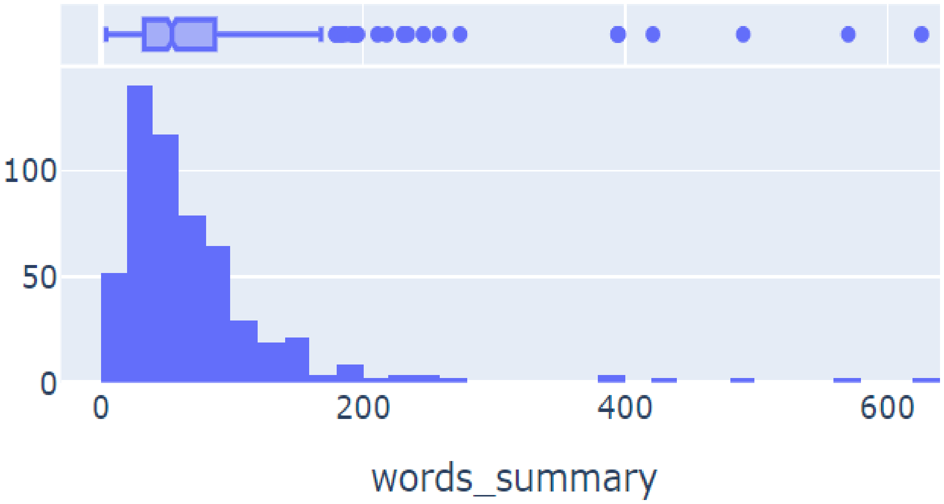


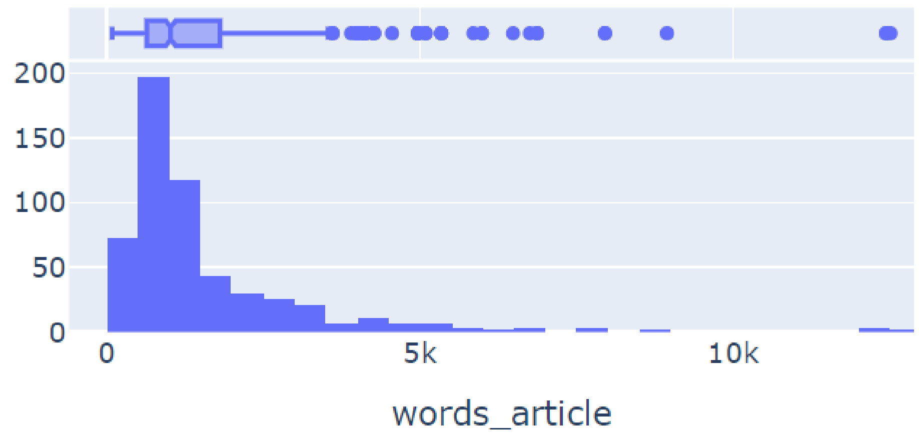
```
MediQA_data['article'].iloc[0]
```

```
" Abetalipoproteinemia abetalipoproteinaemia abetalipoproteinemia neuropathy ABL acanthocytosis apolipoprotein B deficiency Bas
sen-Kornzweig disease Bassen-Kornzweig syndrome betalipoprotein deficiency disease congenital betalipoprotein deficiency syndro
me microsomal triglyceride transfer protein deficiency disease MTP deficiency Description Abetalipoproteinemia is an inherited
disorder that impairs the normal absorption of fats and certain vitamins from the diet. Many of the signs and symptoms of abeta
lipoproteinemia result from a severe shortage (deficiency) of fat-soluble vitamins (vitamins A, E, and K). The signs and sympto
ms of this condition primarily affect the gastrointestinal system, eyes, nervous system, and blood. The first signs and symptom
s of abetalipoproteinemia appear in infancy. They often include failure to gain weight and grow at the expected rate (failure t
o thrive); diarrhea; and fatty, foul-smelling stools (steatorrhea). As an individual with this condition ages, additional signs
and symptoms include disturbances in nerve function that may lead to poor muscle coordination and difficulty with balance and m
ovement (ataxia). They can also experience a loss of certain reflexes, impaired speech (dysarthria), tremors or other involunta
ry movements (motor tics), a loss of sensation in the extremities (peripheral neuropathy), or muscle weakness. The muscle probl
ems can disrupt skeletal development, leading to an abnormally curved lower back (lordosis), a rounded upper back that also cur
ves to the side (kyphoscoliosis), high-arched feet (pes cavus), or an inward- and upward-turning foot (clubfoot). Individuals w
ith this condition may also develop an eye disorder called retinitis pigmentosa, in which breakdown of the light-sensitive laye
r (retina) at the back of the eye can cause vision loss. In individuals with abetalipoproteinemia, the retinitis pigmentosa can
result in complete vision loss. People with abetalipoproteinemia may also have other eye problems, including involuntary eye mo
vements (nystagmus), eyes that do not look in the same direction (strabismus), and weakness of the external muscles of the eye
(ophthalmoplegia). Individuals with abetalipoproteinemia usually have a low number of red blood cells (anemia) with abnormally
star-shaped red blood cells (acanthocytosis) and have difficulty forming blood clots, which can cause abnormal bleeding. In som
e cases, a condition called fatty liver develops, which can cause liver damage. Frequency Abetalipoproteinemia is a rare disord
er. More than 100 cases have been described worldwide. Causes Abetalipoproteinemia is caused by mutations in the MTTP gene, whi
ch provides instructions for making a protein called microsomal triglyceride transfer protein. This protein is essential for cr
eating molecules called beta-lipoproteins in the liver and intestine. Beta-lipoproteins transport fats, cholesterol, and fat-so
luble vitamins from the intestine to the bloodstream so these nutrients can be taken up by tissues throughout the body. Suffici
ent levels of fats, cholesterol, and vitamins are necessary for normal growth, development, and maintenance of the body's cells
and tissues. Most MTTP gene mutations lead to the production of microsomal triglyceride transfer protein with reduced or absent
function and unable to help in the formation of beta-lipoproteins. A lack of beta-lipoproteins causes severely reduced absorpti
on (malabsorption) of dietary fats and fat-soluble vitamins from the digestive tract into the bloodstream. These nutritional de
ficiencies lead to health problems in people with abetalipoproteinemia. Inheritance Pattern This condition is inherited in an a
utosomal recessive pattern, which means both copies of the gene in each cell have mutations. The parents of an individual with
an autosomal recessive condition each carry one copy of the mutated gene, but they typically do not show signs and symptoms of
the condition. Sources for This Page Hooper AJ, van Bockxmeer FM, Burnett JR. Monogenic hypocholesterolaemic lipid disorders an
d apolipoprotein B metabolism. Crit Rev Clin Lab Sci. 2005;42(5-6):515-45. Review. "
```

```
MediQA_data['abs_summary'].iloc[0]
```

```
"People with abetalipoproteinemia are not able  to make beta-lipoproteins and therefore have difficulties absorbing dietary fat
s and fat-soluble vitamins (vitamins A, D, E, and K) from the digestive tract into the bloodstream. Their condition is treated
with sufficient levels of fats, cholesterol, and vitamins, which are necessary for normal growth, development, and maintenance
of the body's cells and tissues, particularly nerve cells and tissues in the eye."
```

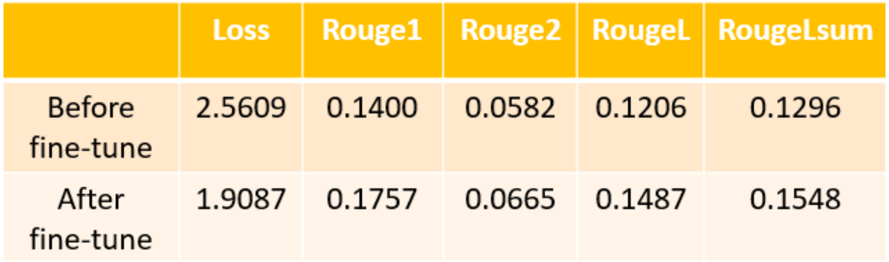




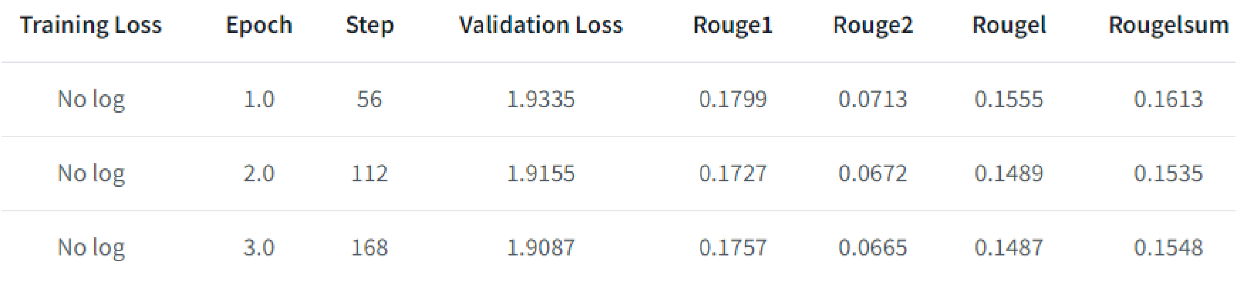
4 Model

One interesting part of this research was to load the Bart model, check that it's working well, and doing a success fine-tune. We will first load the dataset, split to train test check metrics before training the model and also after. The dataset has 552, and we chose to split only to train/test (80/20) without a validation step because the data is very small.

We have trained that data on 3 epochs, and we can see that the Validation Loss decreases while the Rouge metrics don't change so much between the epochs. Before and after fine-tuning we can see a clear difference in the Validation Loss and the Rouge metrics. Because the data is relatively small, maybe it was not mandatory to train the data on multiple epochs and maybe one would be enough.

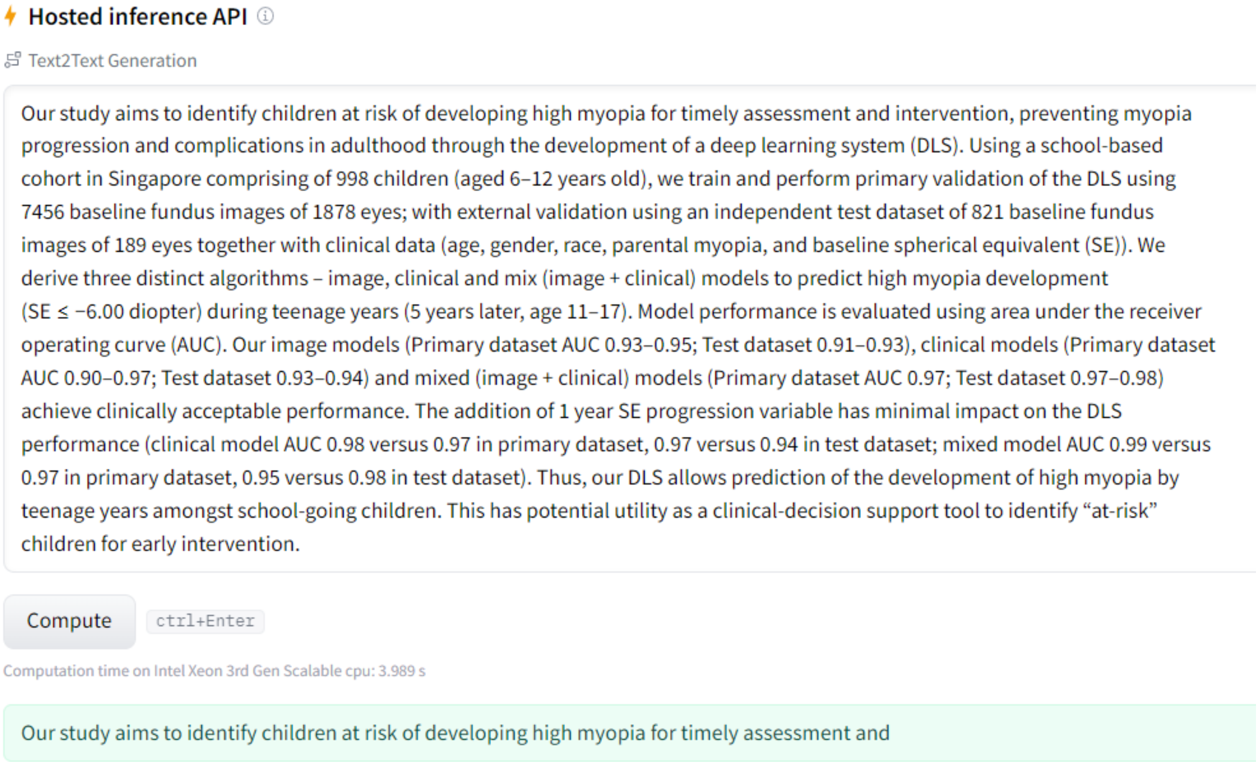


	Loss	Rouge1	Rouge2	RougeL	RougeLsum
Before fine-tune	2.5609	0.1400	0.0582	0.1206	0.1296
After fine-tune	1.9087	0.1757	0.0665	0.1487	0.1548



Training Loss	Epoch	Step	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum
No log	1.0	56	1.9335	0.1799	0.0713	0.1555	0.1613
No log	2.0	112	1.9155	0.1727	0.0672	0.1489	0.1535
No log	3.0	168	1.9087	0.1757	0.0665	0.1487	0.1548

We can see an example of a summarization, available at https://huggingface.co/Chelomo/summarizer_MediQA



Our study aims to identify children at risk of developing high myopia for timely assessment and intervention, preventing myopia progression and complications in adulthood through the development of a deep learning system (DLS). Using a school-based cohort in Singapore comprising of 998 children (aged 6–12 years old), we train and perform primary validation of the DLS using 7456 baseline fundus images of 1878 eyes; with external validation using an independent test dataset of 821 baseline fundus images of 189 eyes together with clinical data (age, gender, race, parental myopia, and baseline spherical equivalent (SE)). We derive three distinct algorithms – image, clinical and mix (image + clinical) models to predict high myopia development (SE ≤ −6.00 diopter) during teenage years (5 years later, age 11–17). Model performance is evaluated using area under the receiver operating curve (AUC). Our image models (Primary dataset AUC 0.93–0.95; Test dataset 0.91–0.93), clinical models (Primary dataset AUC 0.90–0.97; Test dataset 0.93–0.94) and mixed (image + clinical) models (Primary dataset AUC 0.97; Test dataset 0.97–0.98) achieve clinically acceptable performance. The addition of 1 year SE progression variable has minimal impact on the DLS performance (clinical model AUC 0.98 versus 0.97 in primary dataset, 0.97 versus 0.94 in test dataset; mixed model AUC 0.99 versus 0.97 in primary dataset, 0.95 versus 0.98 in test dataset). Thus, our DLS allows prediction of the development of high myopia by teenage years amongst school-going children. This has potential utility as a clinical-decision support tool to identify "at-risk" children for early intervention.

Computation time on Intel Xeon 3rd Gen Scalable cpu: 3.989 s

Our study aims to identify children at risk of developing high myopia for timely assessment and

We can see that unfortunately, the results aren't good at all. I think that there are some reasons that impact to the performance of this model. The first one is that it's a little model, bart-base (and not bart-large). The dataset is little with only 552 articles and bart-base have a limitation of 1024 chars for input length.

5 Conclusion and Discussion

This project was very interesting and constructive for us, we learn a lot, starting with fine-tuning a big Deep Learning model, using a known platform such as Hugging Face. A summarization task is very interesting in NLP and there are some sub-domains such as Multi-document, and Abstractive/Extractive summarization. The dataset was great because we are very interested in medical research and it's the domain of our final project and thesis. It was a great challenge to deal with a little dataset, but unfortunately, we realized too late that our model don't give a good summary. We don't have enough experience to say exactly how it would be better with a bigger dataset, a larger model, or other hyperparameters but we found an article (7) that deals with this topic. I think that the first option it offers is more likely and an interesting follow to this project is to do a Multi-document summarization for each article. Thank you for this interesting project.

Data Availability

The dataset for single-document, abstractive summarization is

Hugging Face Model

The link to the Hugging Face project can be found at https://huggingface.co/Chelomo/summarizer_MediQA "question_driven_answer_summarization_primary_dataset", available at https://osf.io/fyg46/

Code Availability

The code of this study can be found at https://github.com/ChelomoLubliner/fine-tune-bart

References

1. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, https://ai.facebook.com/research/publications/bart-denoising-sequence-to-sequence-pre-training-for-natural-language-generation-translation-and-comprehension/.
2. Question-driven summarization of answers to consumer health questions, https://www.nature.com/articles/s41597-020-00667-z.
3. Summarization with Hugging Face, https://huggingface.co/course/chapter7/5?fw=tf.
4. Bart base model with Hugging Face, https://huggingface.co/facebook/bart-base.
5. Medium, Revealing BART: A denoising objective for pretraining, https://medium.com/analytics-vidhya/revealing-bart-a-denoising-objective-for-pretraining-c6e8f8009564.
6. Medium, BART: Are all pretraining techniques created equal? https://medium.com/dair-ai/bart-are-all-pretraining-techniques-created-equal-e869a490042e,
7. Towards Improving Faithfulness in Abstractive Summarization, https://arxiv.org/abs/2210.01877.