Fine-Tuning T5-Model for Summarization on News Data

Authors: Jake Carter, Ramy Qranfal

Institution: Wentworth Institute of Technology

Date: April 4, 2025

Introduction

AI-like summaries can sometimes be challenging to interpret and lack the writing style of human writers. We solve this problem of AI-like summaries by fine-tuning a model on a dataset that consists of human-written summaries, thus producing human-like summaries. The dataset is based on CNN and Daily Mail news articles, which contain the full articles and the authors' summaries. These summarized articles are helpful for people who want to keep up with the news but don't have the time to keep up with long-form news sources in their busy day-to-day lives.

Pre-Trained LLM Chosen

The pre-trained LLM chosen for this project is Google's T5-base model. The T5 models are pre-trained by taking in text as input and then outputting a modified version of the original text; therefore, it is useful for text summarization, as the output will be a condensed and modified version of the original input. In addition, a large portion of the training for the T5 model was done unsupervised, making it less susceptible to overfitting and more adjustable to specific tasks, which in our case will be summarization. We chose the T5-base version because it best balances computational cost and results. The T5-small model, while it would've cost fewer resources, likely would've led to a less intelligent model in the end, and the T5-large model would be more computationally expensive to fine-tune.

Methodology

Using the pre-trained T5-base model, we fine-tuned the model on the news dataset. We followed the proper steps to fine-tune the model listed below.

1. Load the dataset and tokenizer from the T5-base model.
2. Define sampling function and sampled 10,000 rows from the original dataset.
3. Preprocess the data by adding the prefix “summarize:” in front of each input text and then tokenizing both the modified input text and output (summarization by the authors).
4. Define the data collator and metric function to evaluate our fine-tuned models. In this project, we will be utilizing the ROUGE metric.
5. Load the pre-trained T5-base model, and define training arguments.
6. Pass in training arguments to the trainer class and begin training, which is fine-tuning in this case.

Results

Table:

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Rougelsum	Gen Len
1	0.770800	0.830198	0.234700	0.107700	0.163300	0.163600	53.630000
2	0.723600	0.823444	0.255100	0.118400	0.177100	0.177500	57.932000
3	0.727000	0.822413	0.261700	0.120800	0.181900	0.182400	60.128000
4	0.730100	0.821479	0.261400	0.120800	0.182100	0.182300	59.024000
5	0.730800	0.821524	0.263800	0.122800	0.183500	0.183800	59.633000
6	0.730400	0.821548	0.263900	0.122100	0.183500	0.183700	59.929000

Accuracy:

The Rouge1 metric shows that, on average, 26.38% of the unigrams (one-word matching) match those in the actual summaries within the datasets.

The Rouge2 metric means that around 12.28% of the bigrams (two-word sequences matching) match those in the actual summaries.

The Rougel metric measures the extent to which the longest matching sequences of words (appearing in order but not necessarily consecutively) align between the generated and reference sentences, with a match rate of 18.35%.

The Rougelsum is the same as Rougel, except it looks over the entire summary for the longest sequence instead of sentences. The 18.37% rate is matched to the actual summaries.

These results are reasonable considering that the base model is T5-base and not the large model, and we sampled only 10,000 rows from the data.

***Example*:**

```
GENERATED SUMMARY:
Kuku Kube was created by Canada-based Network365 and is available for free on
Facebook, Android, iOS and on desktop browsers. Players get a point for every
correct square identified, but if they click or tap the wrong square they lose a
point. Scores lower than 11 are poor, scores between 15 and 20 is 'lower than
average', 21 to 30 is considered normal or average, and a score higher than 31
means your eyesight is considered great.

REFERENCE SUMMARY:
Free app is available on Facebook, Android, iOS and on desktop browsers. It
starts with four squares and asks you to identify the different shade. Board
grows to up to 81 squares and differentiation is subtle each time. And a score
of 31 or above is a considered a sign of 'great eyesight'.
```

In this example, we can see that the generated summary is longer than the reference summary but also provides more details, such as the app name, the exact ranges of scores, and what they mean

about the game. Given that our rouge metric scores are not 100%, we can see that the generated summary didn't explicitly state how many squares you start within the game, showing some inaccuracy. However, from this example, we can see that the summary is both concise and informative.

Discussion

***Challenges*:**

- The initial challenge was attempting to train the entire dataset on three epochs, but we realized that the dataset was too large for what we had available.
- Faced low ROUGE metric scores in the initial fine-tuning phase.

Improvements:

- Sampled 10000 random rows from the entire dataset and split the dataset into 0.8, 0.1, and 0.1 splits for training, validation, and testing, respectively.
- Improved rouge metric score by increasing max length for the actual summaries tokenizer. Also, tweaking the training arguments, such as increasing epochs to 6, using warmup steps of 200, increasing weight decay to 0.03, and increasing the generation number of beams to 5.

Work Cited

Summarization, huggingface.co/docs/transformers/en/tasks/summarization. Accessed 3 Apr. 2025.