XSum Summarization

Data Loading/Preprocessing

For basic preprocessing,

- Removed the id column
- Removed leading and trailing whitespaces
- Since a few rows had only a "." for the summary and only 6-7 word documents i chose rows who have string length more than 5 for the summaries and string length more than 100 for training documents, whereas test and val documents with string length greater than 40.
- Since I am using a T5 model for generation, I added a prefix "summarize: " to all documents.

Model

I took the following steps to prepare my model,

- I chose t5-small out of all sizes due to the memory constraints when running the model.
- I used the encode_plus method to tokenize the sentences in the dataset and add special tokens to the sentences.
- Since the average length of words in the document was 431, i chose a max length of 512 which is a little above that and chose a max length of 100 for the summary length since when i set the length lesser than that, i.e. closer to the average sentence length (23), the summary predicated was short in length and incomplete.
- I used the AdamW optimizer with a learning rate 3e-4. I previously used learning rates 3e-5 and 2e-5 however the model train loss was not decreasing even though the valid loss was decreasing significantly. Since the model takes a long time to run (at least 1.5 hrs for a single epoch) I could not try out different learning rates.
- I used autocast for the forward call to possibly reduce the neural net training time for the model.
- The number of epochs=2 and batch size=16 were mainly chosen due to time and memory constraints as the training time for every epoch was long and batch size greater than 16 gave a CUDA out of memory error.
- First epoch *training loss = 0.79 validation loss = 0.69*
- Second epoch *training loss = 0.73 validation loss = 0.67*
- I set the following configuration options for my T5 Model
    - do_sample = True and top_k = 40 to enable the model to choose the 40 most likely next words and distribute the probability mass only amongst those words.
    - Additionally, I set top_p = 0.9 so that the model keeps a small set of words when the next word is relatively more predictable and a larger set of words for when the next word is less predictable.
    - max_length = 200, since keeping the max_length a lesser number (the number i originally kept was 50 and 100) was leading to extremely short summaries which were mostly incomplete sentences.

- num_beams = 4, to explore all 4 hypotheses which have a probability and I set early_stopping = True to stop the generation as soon as all 4 hypotheses reach the end of the string.
- no_repeat_ngram_size=3, to avoid repetition in the generated summary, no three consecutive tokens should be repeated again.
- temperature=0.7, to make the distribution sharper and to avoid low probability words in the generated summary.

Evaluation metric

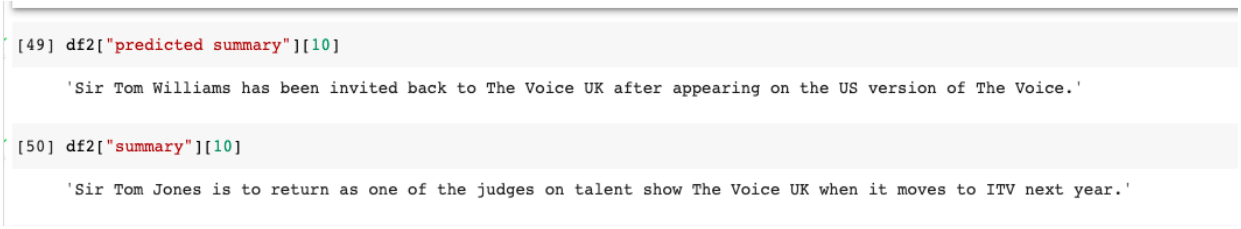
```
{'rouge-1': {'f': 0.2826572409038569,
  'p': 0.32043419604278683,
  'r': 0.26188185145327736},
 'rouge-2': {'f': 0.08640420556106883,
  'p': 0.09795060114085956,
  'r': 0.08043462810621867},
 'rouge-l': {'f': 0.23019530243034508,
  'p': 0.2609580362292668,
  'r': 0.2133451160549097}}
```

Improving the model

Firstly, training the model for more epochs will help the model reduce the loss. The training loss reduced substantially after one epoch. Therefore if trained over more epochs the generated results could be better as the model will learn more.

Fine-tuning the parameters for the models' generation can also make significant changes to the generated summaries.

As visible in the screenshots, even though the model does not have an overlap of bi-grams with the reference summary, it still captures the essence of the summary in some cases.



As visible in the screenshot below, the model captures a short summary (headline of a sort) from the document which is coherent as well, however it is not the same as the reference summary.

```
[51] df2["predicted summary"][111]

'At least 10,000 people have been evacuated from their homes in Chennai due to heavy rains.'

[52] df2["summary"][111]

'More than 70 people have been killed as incessant rains continue to batter the southern Indian city of Chennai, media reports say.'
```

However in some cases, the model generates incorrect information. With fine tuning of the parameters of generation issues like these can be solved. The number of beams can be increased and the model can be tested.

```
[57] df2["predicted summary"][866]

'A man has been charged with murder after the bodies of two men were found in Dundee.'

df2["summary"][866]

'A 37-year-old man has been charged in connection with the deaths of a man and woman at a Dundee flat.'
```