Summarisation of Indian Legal Text

Abhishek Prashant Chandurkar(BT22CSE104)

Supplementary Material

1 Fine-tuning BART Model (BART-fineTune-final.ipynb)

1.1 Model

- **BART (Bidirectional and Auto-Regressive Transformers)**: A pre-trained sequence-to-sequence model from Facebook, used for text summarization. The specific pre-trained model used is 'facebook/bart-large'.

1.2 Libraries and Dependencies

- **PyTorch Lightning**: Used for simplifying the training process and managing the model lifecycle.
- **Transformers**: Provides access to pre-trained models and tokenizers from Hugging Face.
- **NLTK**: Used for text preprocessing tasks such as tokenization.
- **Pandas**: For data manipulation and handling.
- **NumPy**: For numerical operations.
- **PyTorch**: The underlying deep learning framework.
- **Rouge Score**: For evaluating the quality of generated summaries.

1.3 Training Parameters

- **accelerator**: Specifies the hardware for training, set to `gpu` to enable GPU-based training.
- **devices**: Defines the number of devices (GPUs) to use, set to `1`.
- **max_epochs**: The maximum number of training epochs, set to `3`.
- **min_epochs**: The minimum number of training epochs, set to `2`.
- **callbacks**: Includes a callback to monitor training progress using `TQDMProgressBar` with a refresh rate of 5.
- **precision**: Defines the numerical precision for training, set to `16` for mixed precision training.

1.4 Training Results

The model ran for 3 epochs successfully. The training loop terminated as the `max_epochs` value reached 3.

2 Generating Summaries and ROUGE Metrics (generate-summaries-chunking-BART.ipynb)

2.1 Model

- **Fine-tuned BART Model**: The model trained in the first notebook is used for generating summaries.

2.2 Libraries and Dependencies

- **PyTorch Lightning**: For loading the trained model and generating summaries.
- **Transformers**: For handling the model and tokenizer.
- **NLTK**: For text processing and tokenization.
- **Pandas**: For data management.
- **Rouge Score**: For calculating ROUGE metrics and evaluating the summaries.

2.3 ROUGE Metrics

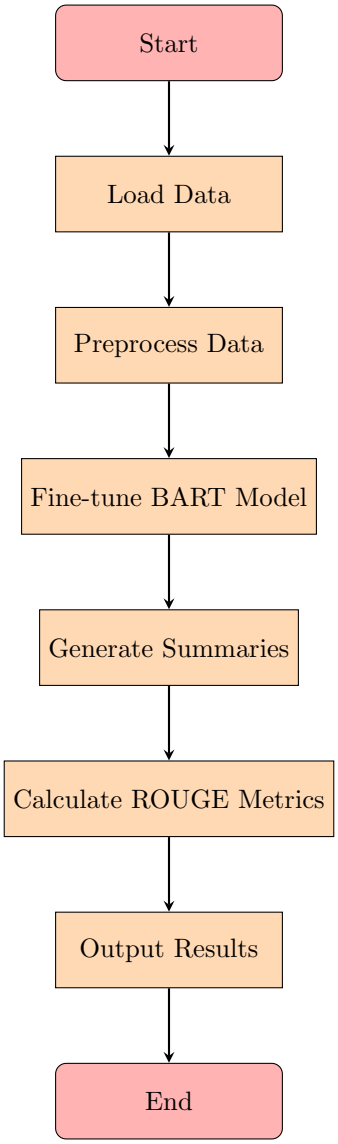
- **ROUGE-1**: 0.3652
- **ROUGE-2**: 0.1015
- **ROUGE-L**: 0.1846
- **ROUGE-Lsum**: 0.3425

2.4 Output

- The summary generated by the model can be found at: `/content/drive/MyDrive/IN-Abs/output/`.
- The reference summary can be compared with the generated summary at: `/content/drive/MyDrive/IN-Abs/test-data/summary`.
- The legal text that was summarized is available at: `/content/drive/MyDrive/IN-Abs/test-data/judgement`.

3 Flow Diagram

Flow Diagram for Text Summarization Process



4 References

Dataset: The dataset can be downloaded from the following link: `https://zenodo.org/record/7152317#.Yz6mJ9JByC0`.
Structure:

- `train-data` - Folder containing documents and summaries for training:
  - `judgement` - Contains legal judgments for training.
  - `summary` - Contains summaries for training.
  - `stats-IN-train.txt` - Text file with word and sentence count statistics for documents.
- `test-data` - Folder containing documents and summaries for testing:
  - `judgement` - Contains legal judgments for testing.
  - `summary` - Contains summaries for testing.
  - `stats-IN-test.txt` - Text file with word and sentence count statistics for test documents and summaries.

Code Reference: Part of the code was adapted from an existing implementation available at `https://github.com/Law-AI/summarization/tree/aacl/abstractive/BART_based_approaches` , which provided valuable insights for model training and evaluation.