Task 2

Working of the code:

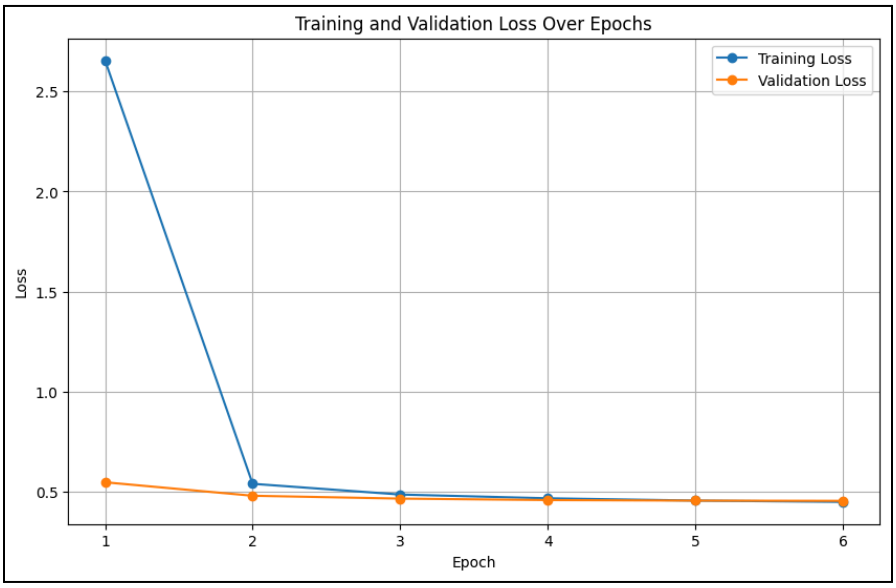
- **Preprocessing:** Expands contractions, removes URLs/special characters, and lowercases text. Maintained an exhaustive list of contractions.
- **Dataset Preparation:** Splits the data into train, validation, and test sets; converts to HuggingFace `DatasetDict`.
- **Tokenization:** Inputs and targets are tokenized with padding and truncation.
- **Model Training:** Fine-tunes `T5-base/Facebook-bart-large` using `Seq2SeqTrainer` over 6 epochs.
- **Evaluation:** Computes training and validation loss, plots them, and evaluates the model using ROUGE-L, BERTScore, and BLEU-4 on the test set.
- **Saving/Loading:** Saves the fine-tuned model and tokenizer; reloads them for inference.
- **Inference:** Generates predictions for the test set and saves them along with evaluation metrics to a CSV file.

Hyperparameters:

Parameter	Value
Learning Rate	2e-5
Epochs	6
Batch Size (Train/Eval)	8 / 8
Max Input Length	512 tokens
Max Target Length	128 tokens
Weight Decay	0.01
Loss Function	CrossEntropy (Seq2Seq)
Optimizer	AdamW (implicitly used by Trainer)

Results:

1. **T5-Base:**



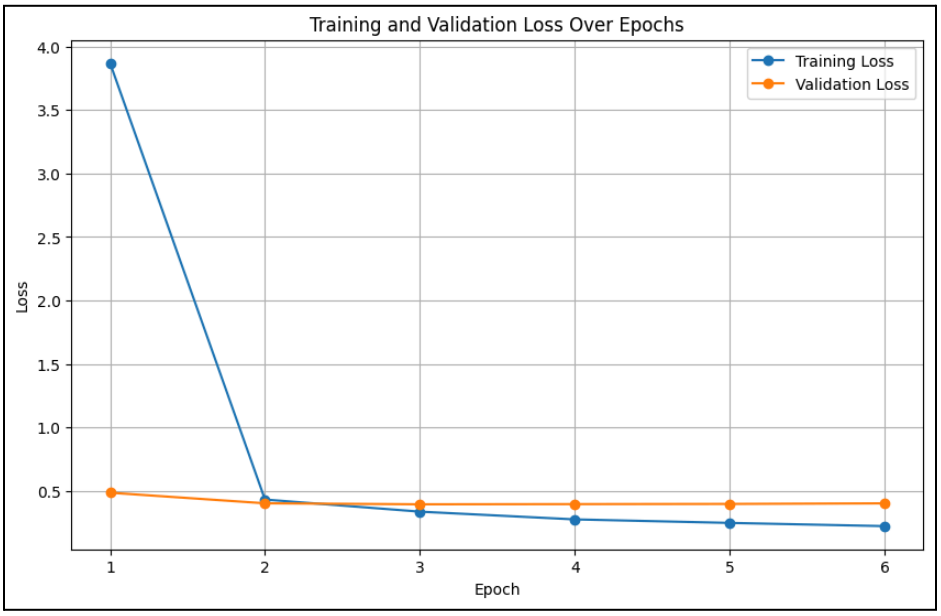

```
ROUGE-L Precision: 0.3405
ROUGE-L Recall: 0.3628
ROUGE-L F1: 0.3245
BERTScore F1: 0.8649
BLEU-4: 0.2100
```

Testing Results

T5 Model - Summary of Results:

The training and validation loss curves show a quick convergence, with both losses stabilizing after just a few epochs, indicating effective learning without significant overfitting. These scores reflect moderate text generation quality, with high BERTScore suggesting good semantic similarity, though ROUGE and BLEU indicate room for improvement in lexical overlap and phrasing. This will serve as a useful baseline for comparing with other models.

2. **Bart-large:**




```
ROUGE-L Precision: 0.4176
ROUGE-L Recall: 0.3757
ROUGE-L F1: 0.3715
BERTScore F1: 0.8811
BLEU-4: 0.2452
```

on testing set

The BART-large model shows smooth convergence with training and validation loss stabilizing after the second epoch. Evaluation metrics like ROUGE-L and BLEU-4 are slightly better than the previous model, and BERTScore F1 is reasonably high at 0.8811. The results suggest that BART-large is able to capture more semantic relevance, but further comparison will help solidify this.

Comparison:

- **BART-large** outperforms **T5** on all evaluation metrics.
- Gains are especially notable in **ROUGE-L Precision** and **BLEU-4**, indicating better lexical overlap and fluency.
- **BERTScore F1** is also higher, suggesting stronger semantic similarity to reference outputs.

While both models perform well, **BART-large** demonstrates consistently better results across all key metrics, along with slightly more stable training dynamics. However, further analysis on different datasets or tasks would help generalize this observation.

Resource Constraints and Model Selection

Due to limitations in computational resources, particularly GPU memory and session time on Kaggle, the selection of models was influenced significantly. **T5-base** was chosen for its balance between performance and efficiency—it requires less memory, trains faster, and is suitable for experimentation within limited runtime environments. On the other hand, **BART-large**, while more resource-intensive, was included for its strong performance on generation tasks. However, training it required careful tuning (e.g., smaller batch sizes, fewer epochs) and longer execution time. Storage constraints also meant saving only essential outputs like the final model. These factors shaped the overall training strategy and comparisons.