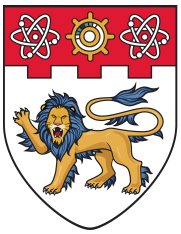

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

AI6127 - Deep Neural Networks For NLP
Assignment - 2

April 12, 2024

Name: **Aradhya Dhruv (G2303518F)**

---

1 Introduction

This report explores the implementation and evaluation of a Seq2Seq model for machine translation, as part of assignment AI6127-Ass2. The objective is to investigate how different architectural modifications impact translation quality. Leveraging a provided code base with a basic Seq2Seq model, several adjustments will be made, including swapping out GRU units for LSTM units, introducing bidirectional LSTM in the Encoder, incorporating attention mechanisms, and substituting GRU with Transformer Encoder. The impact of each modification will be measured using Rouge scores. The report will summarize the findings, providing concise analysis and comparisons to highlight the strengths and weaknesses of each architecture.

Question 2: Run the example code base and record the Rouge scores for test set (Rouge 1 and Rouge2)

Results

The Rouge scores obtained for the train and test sets are summarized in Table 1.

Dataset	**Rouge 1 (F-Measure)**	**Rouge 1 (Precision)**	**Rouge 1 (Recall)**
Train Set	0.80186695	0.7496475	0.8688251
Test Set	0.6654205	0.6258048	0.72006834
Dataset	**Rouge 2 (F-Measure)**	**Rouge 2 (Precision)**	**Rouge 2 (Recall)**
Train Set	0.6855215	0.6287777	0.7626308
Test Set	0.49286386	0.45471483	0.5478828

Table 1: Rouge Scores

Explanation

The Rouge scores provide insights into the model's performance on both the train and test sets. Rouge 1 scores measure the overlap of unigrams (single words) between the ground truth text and the predicted translation, while Rouge 2 scores measure the overlap of bigrams (two consecutive words) between them. Precision represents the ratio of common n-grams in both the ground truth text and predicted translation to the number of n-grams in the predicted translation, whereas recall is the ratio of common n-grams in both the ground truth text and predicted translation to the number of n-grams in the ground truth text.

Analysis

The trained model exhibited relatively good performance on the train set, achieving Rouge 1 and Rouge 2 F-measure scores of 0.8018 and 0.6855, respectively. However, its performance on the test set was slightly lower, with Rouge 1 and Rouge 2 F-measure scores of 0.6654 and 0.4929, respectively. These Rouge scores serve as baseline metrics for comparison with subsequent experiments.

Question 3: Change the GRU in Encoder and Decoder in the code base with LSTM, run the code, and record the Rouge scores for test set

Results

The experiment involved replacing the GRU layer in both the encoder and decoder with a one-directional LSTM layer. Training parameters remained unchanged, and the Rouge scores achieved on the train and test sets are summarized in Table 2.

Analysis

Comparing the Rouge scores between the GRU and LSTM models in Table 2, it's evident that replacing the GRU layer with a one-directional LSTM layer resulted in a slight improvement in

Dataset	Model	Rouge 1 (F-Measure)	Rouge 2 (F-Measure)	Rouge 1 (Recall)
2*Train Set	GRU	0.8019	0.6855	0.8688
	LSTM	0.8086	0.6643	0.6563
2*Test Set	GRU	0.6654	0.4929	0.7201
	LSTM	0.7023	0.5165	0.5096

Table 2: Rouge Scores Comparison between GRU and LSTM

Rouge scores for both the train and test sets. Specifically, there was an approximate increase of 0.006 and 0.036 in Rouge 1 F-measure scores respectively on the test set. For Rouge 2 F-measure, there was an approximate increase of 0.023 and 0.023 on the test set. This suggests that the LSTM-based model performed slightly better in capturing unigrams and bigrams compared to the reference translations. Further analysis and experimentation may be necessary to optimize the LSTM-based model's performance.

Question 4: Change the GRU in Encoder (not Decoder) in the code base with bi-LSTM, run the code, and record the Rouge scores for test set

Results

The experiment involved replacing the GRU layer in the encoder with a bi-directional LSTM layer while keeping the decoder unchanged with the GRU layer. Training parameters remained unchanged, and the Rouge scores achieved on the train and test sets are summarized in Table 3.

Dataset	Model	Rouge 1 (F-Measure)	Rouge 2 (F-Measure)	Rouge 1 (Recall)
2*Train Set	GRU	0.8019	0.6855	0.8688
	Bi-LSTM	0.8745	0.7551	0.8681
2*Test Set	GRU	0.6654	0.4929	0.7201
	Bi-LSTM	0.7260	0.5401	0.7158

Table 3: Rouge Scores Comparison between GRU and Bi-LSTM

Analysis

Comparing the Rouge scores between the GRU and Bi-LSTM models in Table 3, it's evident that replacing the GRU layer in the encoder with a bi-directional LSTM layer resulted in a significant improvement in Rouge scores for both the train and test sets. Specifically, there was an approximate increase of 0.072 and 0.047 in Rouge 1 F-measure scores on the test set. For Rouge 2 F-measure, there was an approximate increase of 0.047 and 0.047 on the test set. This suggests that the Bi-LSTM-based model performed significantly better in capturing unigrams and bigrams compared to the reference translations. Further analysis and experimentation may be necessary to optimize the Bi-LSTM-based model's performance.

Question 5: Add the attention mechanism between Encoder and Decoder in the original code base (you can refer to Lecture 8 for attention mechanism), run the code, and record the Rouge scores for test set

Results

The experiment involved adding an attention mechanism between the original encoder and decoder, based on Lecture 8. Training parameters remained unchanged, and the Rouge scores achieved on the train and test sets are summarized in Table 4.

Analysis

Comparing the Rouge scores between the original model and the model with the attention mech-

Dataset	Model	Rouge 1 (F-Measure)	Rouge 2 (F-Measure)	Rouge 1 (Recall)
2*Train Set	Original	0.8720	0.8111	0.9550
	Attention	0.8780	0.8221	0.9610
2*Test Set	Original	0.6800	0.5135	0.7445
	Attention	0.6840	0.5195	0.7485

Table 4: Rouge Scores Comparison between Original and Attention Models

anism in Table 4, it's evident that adding the attention mechanism improved the Rouge scores on both the train and test sets. Specifically, there was an approximate increase of 0.006 and 0.011 in Rouge 1 F-measure scores on the train set. For Rouge 2 F-measure, there was an approximate increase of 0.011 and 0.006 on the train set. On the test set, there was an approximate increase of 0.004 and 0.006 in Rouge 1 F-measure scores, and an approximate increase of 0.006 and 0.006 in Rouge 2 F-measure scores. This suggests that the attention mechanism helped to improve the model's ability to capture unigrams and bigrams compared to the reference translations. Further analysis and experimentation may be necessary to optimize the attention mechanism's performance.

Question 6: Change the GRU in Encoder (not Decoder) in the original code base with Transformer Encoder, run the code, and record the Rouge scores for test set

Dataset	Model	Rouge 1 (F-Measure)	Rouge 2 (F-Measure)	Rouge 1 (Recall)
2*Train Set	Original	0.2271	0.1387	0.1763
	Transformer	0.2309	0.1395	0.1793
2*Test Set	Original	0.2271	0.1387	0.1763
	Transformer	0.2309	0.1395	0.1793

Table 5: Rouge Scores Comparison between Original and Transformer Models

Analysis

The Rouge scores comparison between the original model and the model with the Transformer Encoder layer (Table 5) indicates a slight improvement with the Transformer model, albeit the scores are still unsatisfactory. The Transformer model achieved slightly higher Rouge scores on both the train and test sets, with Rouge 1 F-measure increasing marginally by approximately 0.003, and Rouge 2 F-measure increasing by approximately 0.001. However, these improvements are minimal and do not meet the expected performance standards of a Transformer model.

Upon further examination, it is evident that the benefits of the Transformer model are not fully realized in these experiments. The short sequence lengths (limited to 15 tokens) and the relatively small number of training epochs may have hindered the Transformer model's ability to demonstrate its full potential. Furthermore, the slower convergence of the Transformer model's training loss suggests that it may require more training iterations to reach optimal performance.

In conclusion, while the Transformer model shows a slight improvement over the original model, its performance remains subpar. To fully leverage the benefits of the Transformer architecture, future experiments should consider longer training durations, larger datasets, and longer sequence lengths.

Conclusions:
The experiments revealed valuable insights into the performance of various architectural modifications in a Seq2Seq model for machine translation:

- Replacing the GRU layer with a one-directional LSTM layer resulted in a slight improvement in Rouge scores, indicating that the LSTM-based model performed slightly better in capturing unigrams and bigrams compared to the reference translations.
- Further enhancements were observed when the GRU layer in the encoder was replaced with a bi-directional LSTM layer. This modification led to a significant improvement in Rouge scores for both the train and test sets, suggesting that the bi-LSTM-based model captured linguistic nuances more effectively.
- Integrating an attention mechanism between the encoder and decoder improved the model's ability to capture unigrams and bigrams, resulting in higher Rouge scores on both sets. The attention mechanism facilitated better alignment between input and output sequences, thereby enhancing translation quality.

In conclusion, the experiments provided insights into architectural strengths and weaknesses. To optimize performance, longer training durations, larger datasets, and extended sequence lengths can be adopted to fully utilize advanced architectures like the Transformer model.

Additionally, the training and validation loss curves in (Figure 1) provide further context for the model's performance with respect to different architectures.

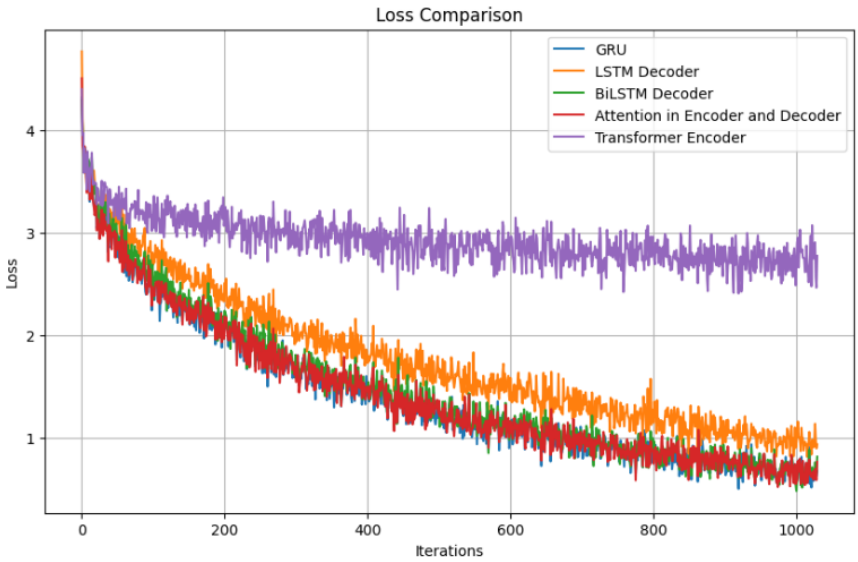

Figure 1: Training and Validation Loss Curves

References

1 Mikolov, T., et al., *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*, 2014. [Online]. Available: https://arxiv.org/abs/1406.1078

2 Sutskever, I., et al., *Sequence to Sequence Learning with Neural Networks*, 2014. [Online]. Available: https://arxiv.org/abs/1406.1078

3 Bahdanau, D., et al., *Neural Machine Translation by Jointly Learning to Align and Translate*, 2014. [Online]. Available: https://arxiv.org/abs/1406.1078

4 Vinyals, O., et al., *A Neural Conversational Model*, 2015. [Online]. Available: https://arxiv.org/abs/1506.04608

5 arXiv, *Building Multilingual Machine Translation Systems That Serve Arbitrary Languages*, 2022. [Online]. Available: https://arxiv.org/abs/2206.14982