AI6127: Deep Neural Networks For Natural Language Processing

Names: Joseph UWINEZA
Matric No: G2303477F
Program: MSAI

NTU, April 2024

Assignment 2

Abstract

This report explores the development and evaluation of a seq2seq model for machine translation, focusing on various architectural modifications and their impact on translation accuracy. Initially, the model utilizes GRUs[3] in both the encoder and decoder, serving as our baseline. Subsequent modifications include replacing GRUs with LSTMs[2] and bi-LSTMs[5] in the encoder, and incorporating an attention mechanism[1] and a Transformer [4]encoder. Each configuration is assessed using Rouge scores for the test set to evaluate performance. The results reveal distinct performance variations across the different architectural setups, providing insights into the effectiveness of each model configuration in handling the complexities of language translation. The analysis underscores the potential improvements in translation accuracy through strategic model enhancements and highlights the critical role of model architecture in machine translation tasks.

1 Introduction

Machine translation (MT) stands as a pivotal area within the field of natural language processing (NLP), continually pushing the boundaries of how machines understand and translate human languages. The sequence-to-sequence (seq2seq) model, characterized by its encoder-decoder architecture, has been a cornerstone in advancing MT due to its ability to handle sequences of variable lengths with remarkable effectiveness.

In recent advancements, variations in the basic seq2seq model architecture have shown potential for improving translation accuracy and model robustness. This report delves into an experimental investigation where the traditional seq2seq model, initially equipped with Gated Recurrent Units (GRUs), is systematically modified by introducing Long Short-Term Memory (LSTM) units, bidirectional LSTMs (bi-LSTMs), and an attention mechanism. Furthermore, we explore the integration of a Transformer encoder as an alternative to the conventional RNN-based approach.

These modifications are aimed at enhancing the model's capacity to capture deeper linguistic structures and contextual nuances critical for accurate translation. Each architectural variant is evaluated using Rouge metrics, specifically Rouge-1 and Rouge-2 scores, which serve as indicators of the quality of the generated translations compared to human-translated texts.

No	Hyperparameter	Values
1	Optimizer	SGD
2	Learning rate	0.01
3	Patience	5
4	Hidden layers	512
5	Dropout	0.1
6	Number of epochs	5

Table 1: shows the parameter settings that have been used all along the experiment.

2 *Task 2: Run the example code base and record the Rouge scores for test set.*

ROUGE		Train	Test
Rouge 1	F-measure	0.80065733	0.66147625
	Precision	0.74582875	0.61951584
	Recall	0.8708215	0.7182766
Rouge 2	F-measure	0.68419397	0.4864638
	Precision	0.62558454	0.44780356
	Recall	0.76382875	0.5414094

Table 2: Shows the Rouge scores obtained for both the training and testing datasets under fixed training parameters. we use base code where the GRU layer is in both the encoder and decoder.

Rouge scores help us understand how well the model performs on the training and testing data. Rouge 1 scores assess how many single words (unigrams) in the predicted translation match with the original text, and Rouge 2 scores look at how many pairs of consecutive words (bigrams) match. *Precision* calculates the percentage of matching n-grams (groups of words) in the predicted translation compared to all the n-grams in that translation. *Recall*, on the other hand, calculates the percentage of matching n-grams compared to all the n-grams in the original text.

Analysis

The seq2seq model with GRU units performs well on the training set but shows a noticeable decline in performance on the test set, as evidenced by both Rouge-1 and Rouge-2 scores. This drop suggests that while the model effectively captures training data, it struggles with generalization to unseen test data, particularly in maintaining bigram relationships. The results underscore the need for enhancing the model's generalization capabilities to handle new and diverse examples more effectively.

3 *Task 3: Change the GRU in Encoder and Decoder in the code base with LSTM.*

ROUGE		Train	Test
Rouge 1	F-measure	0.7325168	0.63673013
	Precision	0.6877228	0.60081947
	Recall	0.79072773	0.68588054
Rouge 2	F-measure	0.588044	0.4600741
	Precision	0.5427725	0.4270779
	Recall	0.65027845	0.50747013

Table 3: Shows the Rouge scores obtained for both the training and testing datasets under fixed training parameters. GRU layer in both the encoder and decoder was swapped out for a one-directional LSTM layer.

After replacing GRU units with LSTM units in both the encoder and decoder, the seq2seq model exhibits a decrease in performance on the training set. On the test set, there is only a minor decrease in performance metrics for both Rouge-1 and Rouge-2 compared to those from table 2. The precision and recall for both metrics are lower than those seen with the GRU-based model. This suggests that the transition to LSTM units did not significantly enhance the model's ability to generalize to unseen data. The findings imply that the different gating mechanisms of LSTM did not provide a notable advantage over GRU in this context.

4 *Task 4: Change the GRU in Encoder (not Decoder) in the code base with bi-LSTM.*

ROUGE		Train	Test
Rouge 1	F-measure	0.77224183	0.6508719
	Precision	0.72080034	0.61267936
	Recall	0.83815247	0.70283407
Rouge 2	F-measure	0.63934374	0.47044268
	Precision	0.5860248	0.43505514
	Recall	0.71162575	0.52098143

Table 4: Shows the Rouge scores obtained for both the training and testing datasets under fixed training parameters. The GRU layer in the encoder (not Decoder) was swapped out for a bi-directional LSTM layer(bi-LSTM).

Implementing a bi-LSTM in the encoder while retaining the GRU in the decoder led to a moderate improvement in Rouge-1 F-measure on the test set, but a decline in Rouge-2 F-measure. The training set maintained high performance in Rouge-1, indicating that the bi-directional context provided by the bi-LSTM aids in better processing of the input sequence, improving unigram matching on unseen data. However, the smaller gains in Rouge-2 scores on the test set suggest that the impact on bigram matching is less significant, highlighting an area where the model's ability to generalize bigram patterns could still be enhanced.

5 *Task 5: Add the attention mechanism between Encoder and Decoder in the original code base.*

ROUGE		Train	Test
Rouge 1	F-measure	0.75284285	0.632525643
	Precision	0.71169413	0.60161106
	Recall	0.80852262	0.6774530928
Rouge 2	F-measure	0.6086932	0.447649697
	Precision	0.56474171	0.417563018
	Recall	0.67172156	0.49304019

Table 5: Shows the Rouge scores obtained for both the training and testing datasets under fixed training parameters. The GRU layer in the encoder and Decoder was swapped out for attention mechanism.

Integrating an attention mechanism into the original GRU-based seq2seq model significantly enhances its performance on the test set, with notable increases in both Rouge-1 and Rouge-2 F-measures. This improvement suggests that the attention mechanism provides a more refined contextual focus during the decoding process, leading to better generalization, particularly in unigram matching as indicated by the Rouge-1 score. The training set also shows high performance, underscoring the effectiveness of the attention mechanism in enabling the model to focus more on relevant segments of the input sequence, thereby improving the quality of translations.

6 *Task 6: Change the GRU in Encoder (not Decoder) in the original code base with Transformer Encoder.*

ROUGE		Train	Test
Rouge 1	F-measure	0.20368144	0.20699301
	Precision	0.25756693	0.26047578
	Recall	0.17630842	0.17930742
Rouge 2	F-measure	0.12028131	0.12094837
	Precision	0.16141184	0.16150153
	Recall	0.10252892	0.10299744

Table 6: Shows the Rouge scores obtained for both the training and testing datasets under fixed training parameters. The GRU layer in the encoder (not the Decoder) was swapped out for the Transformer Encorder

Replacing the GRU with a Transformer Encoder in the seq2seq architecture resulted in a decrease in Rouge scores across both training and test sets. The Rouge-1 and Rouge-2 scores are notably low, with marginal improvements in precision and recall. These results fall short of the expectations typically held for Transformer models. A closer analysis suggests that the Transformer architecture may not have been fully leveraged, possibly due to limitations such as restricted sequence lengths of up to 15 tokens and fewer training epochs, which could have hindered the model's ability to showcase its capabilities fully.

7 General Analysis

GRU Baseline: The baseline GRU model delivered respectable performance metrics, establishing a robust standard for subsequent comparisons.

LSTM Transition: Transitioning to LSTM units showed no notable performance enhancements; instead, there was a slight dip. This could suggest that the added complexity of LSTM cells does not inherently translate into superior generalization for this specific task.

bi-LSTM Upgrade: The introduction of bi-LSTM led to a modest rise in the Rouge-1 F-measure on the test set, hinting at the value of bi-directional context in marginally enhancing unigram detection.

Attention Mechanism Addition: The integration of an attention mechanism presented a small increase in both Rouge-1 and Rouge-2 F-measures. This indicates its potential in directing the model's focus towards pertinent parts of the input, thus improving translation quality.

Transformer Encoder Shift: Implementing a Transformer Encoder, contrary to expectations, resulted in a significant decline in performance. Typically robust in NLP tasks, the low scores observed here point to possible issues such as incompatibility with the model's architecture, suboptimal parameter settings, or limitations in the training approach.

References

[1] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need. 2023. Accessed: 2024-04-18.

[2] Erling Stray Bugge Christian Bakke Vennerød, Adrian Kjærran. Long short-term memory rnn. 2021. Accessed: 2024-04-14.

[3] Michael W. Mahoney N. Benjamin Erichson, Soon Hoe Lim. Gated recurrent neural networks with weighted time-delay feedback, 2022. Accessed: 2024-04-13.

[4] Richard E. Turner. An introduction to transformers. 2025. Accessed: 2024-04-19.

[5] Ziyuan Pu Yinhai Wang Zhiyong Cui, Ruimin Ke. Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. 2019. Accessed: 2024-04-17.

-end-