CS455 Semester Project: Neural Machine Translation-A Comparative Performance Analysis of RNN, LSTM, and GRU

Muhammad Farae
Faculty of Computer Science &
Engineering
Ghulam Ishaq Khan Institute of
Engineering Scienes and Technology
Topi, Pakistan
u2020292@giki.edu.pk

Uzair Rahman
Faculty of Computer Science &
Engineering
Ghulam Ishaq Khan Institute of
Engieering Sciences and Technology
Topi, Pakistan
u2020509@giki.edu.pk

Abstract— **This study provides a comprehensive performance analysis of three prominent neural network architectures: Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs) in the domain of machine translation. Focusing on the translation of French to English, this research utilizes a parallel corpus obtained from a publicly accessible dataset. Each model's performance is evaluated based on accuracy, training time, and ability to handle long sequence dependencies. This analysis aims to identify the most effective neural architecture for machine translation tasks, providing insights into the strengths and weaknesses of each model in handling the complexities of language translation. The results of this study are intended to guide future research and practical implementations in neural machine translation, especially in applications requiring efficient and accurate cross-lingual communication.**

Keywords—RNN, LSTM, GRU, BLEU, ROUGE, long-term dependency, over/under-fitting,

I. Introduction

Machine translation (MT) is a subfield of computational linguistics that investigates the use of software to translate text or speech from one language to another. At its core, MT endeavors to reduce language barriers and enhance communication across different language speakers. With the burgeoning demand for real-time and accurate translation services, the development of more sophisticated and robust MT systems has become crucial. Traditional approaches, which primarily relied on rule-based and statistical methods, have given way to neural machine translation (NMT) due to its superior ability to learn direct mappings between languages from large datasets.

Despite the advancements in NMT, the challenge of choosing the optimal neural network architecture remains significant. The most common architectures for NMT include Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Gated Recurrent Units (GRUs). Each architecture offers distinct advantages and constraints, particularly concerning their handling of dependencies in language translation tasks. The choice of architecture has profound implications on the performance, efficiency, and scalability of MT systems.

This research aims to conduct a comparative analysis of RNN, LSTM, and GRU architectures in the context of French to English translation. The primary objectives are to:

1. Evaluate the translation accuracy of each architecture.
2. Analyze the training time and computational efficiency.
3. Assess the ability of each model to handle long sentence structures and contextual dependencies in translation tasks.

The study is focused exclusively on the translation of French to English using a parallel corpus sourced from a publicly available dataset. The analysis is intended to provide empirical evidence that will help in understanding the performance dynamics of each architecture under identical training conditions.

II. Literature Review

Neural Machine Translation (NMT) represents a paradigm shift from traditional statistical methods to deep learning-based approaches. It has significantly influenced the efficiency and accuracy of machine translation by utilizing models that can learn to translate texts from large datasets. The seminal work by Cho et al. (2014) introduced a novel approach involving an encoder-decoder architecture that jointly learns to align and translate. This paper laid the foundation for subsequent research by demonstrating that deep learning models could effectively manage sequence-to-sequence tasks, which are pivotal in machine translation.

Further advancements in NMT have focused on enhancing the capabilities of the basic RNN structures to better handle long-term dependencies, which are crucial for maintaining the contextual integrity of translated sentences. The introduction of LSTM and GRU architectures addressed some of the limitations inherent in traditional RNNs, such as the vanishing gradient problem, thereby improving the retention of information over longer sequences. H. Xu et al. (2021) explored these advancements by implementing a multi-head highly parallelized LSTM decoder for NMT, showing that this architecture could significantly improve the translation quality by efficiently managing the information flow across different parts of the neural network .

The quest for universal language translation aims to build systems capable of translating multiple languages with high accuracy and minimal need for intervention. H. Xu and colleagues (2021) in their study published in the Springer Conference proceedings, discussed a deep learning approach that leverages universal language models to achieve efficient translation across various languages. Their work underscores the potential of deep learning techniques, particularly LSTM and GRU, in enhancing the adaptability and scalability of NMT systems to accommodate a broad spectrum of languages and dialects.

III. Methodology

For this study, a parallel corpus of French to English translations was sourced from the publicly available dataset at this link (https://go.aws/38ECHUB). This dataset consists of sentence pairs, where each French sentence is paired with its English translation, facilitating the direct application of supervised learning techniques for machine translation.

The preprocessing steps included:

- Tokenization: Utilizing the SpaCy library, both French and English texts were tokenized. This process involved splitting text into individual words or tokens, which serves as the input for training the models.
- Vocabulary Building: Separate vocabularies for French and English were constructed by identifying unique tokens across the corpora. Each unique token was then assigned a unique index, which is crucial for converting text data into numerical form that can be processed by neural networks.
- Padding: Sequences were padded to ensure consistent length within each batch, a requirement for training neural networks.
- Vectorization: Token sequences from both languages were converted into sequences of indices, which represent the position of each token in the respective vocabulary. This vectorization process is essential for feeding data into the models for training.

Three types of neural network architectures were implemented and compared:

- Recurrent Neural Networks (RNN): This basic form of neural networks was used to establish a baseline for performance. Despite its simplicity, RNNs are capable of processing sequences of data but struggle with long-term dependencies.
- Long Short-Term Memory (LSTM): An advanced variant of RNN, known for its ability to handle long-range dependencies within the input data, using mechanisms called gates.
- Gated Recurrent Units (GRU): Similar to LSTM but with a simpler structure, often leading to faster training times without significant loss in capability.

Their architectures have been summarized in Figure 1. The subsections that follow outline the mathematical formulas associated with each of these architectures.

A. Recurrent Neural Networks

In an RNN, each neuron receives input from previous time steps. The basic equations are:

- Hidden state update:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b_h) \quad (1)$$

- Output calculation:

$$y_t = W_y h_t + b_y \quad (2)$$

B. Long-Short Term Memory

- Forget gate:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (3)$$

- Input gate:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (4)$$

- Cell candidate:

$$\tilde{C}_t = tanh(W_C[h_{t-1}, x_t] + b_C) \quad (5)$$

- Update cell state:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (6)$$

- Output gate:

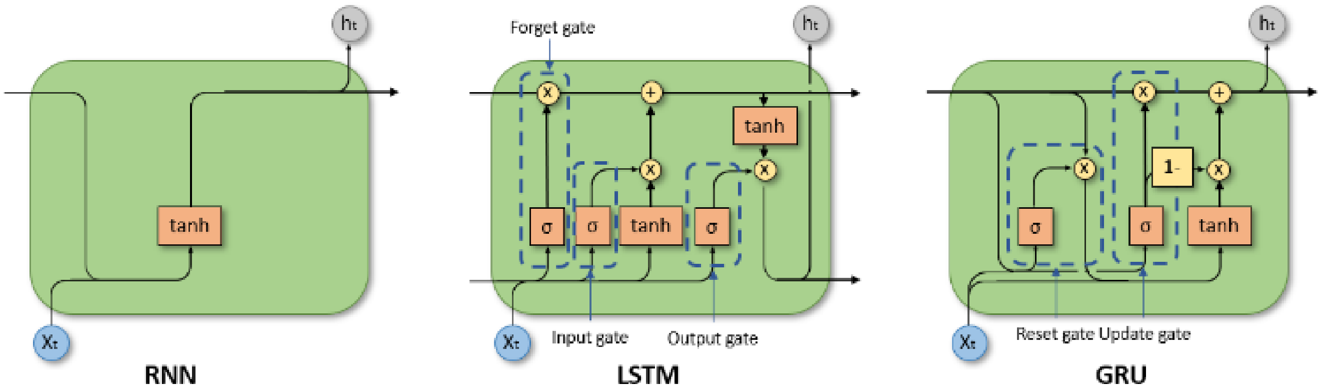
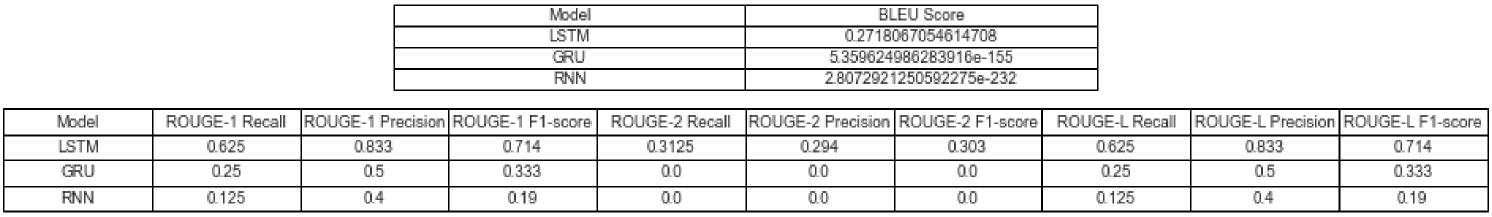


Figure 1 - RNN, LSTM and GRU Architectures



Model	BLEU Score
LSTM	0.2718067054614708
GRU	5.359624986283916e-155
RNN	2.8072921250592275e-232

Model	ROUGE-1 Recall	ROUGE-1 Precision	ROUGE-1 F1-score	ROUGE-2 Recall	ROUGE-2 Precision	ROUGE-2 F1-score	ROUGE-L Recall	ROUGE-L Precision	ROUGE-L F1-score
LSTM	0.625	0.833	0.714	0.3125	0.294	0.303	0.625	0.833	0.714
GRU	0.25	0.5	0.333	0.0	0.0	0.0	0.25	0.5	0.333
RNN	0.125	0.4	0.19	0.0	0.0	0.0	0.125	0.4	0.19

Figure 2 - BLEU and ROUGE Scores for all the Models

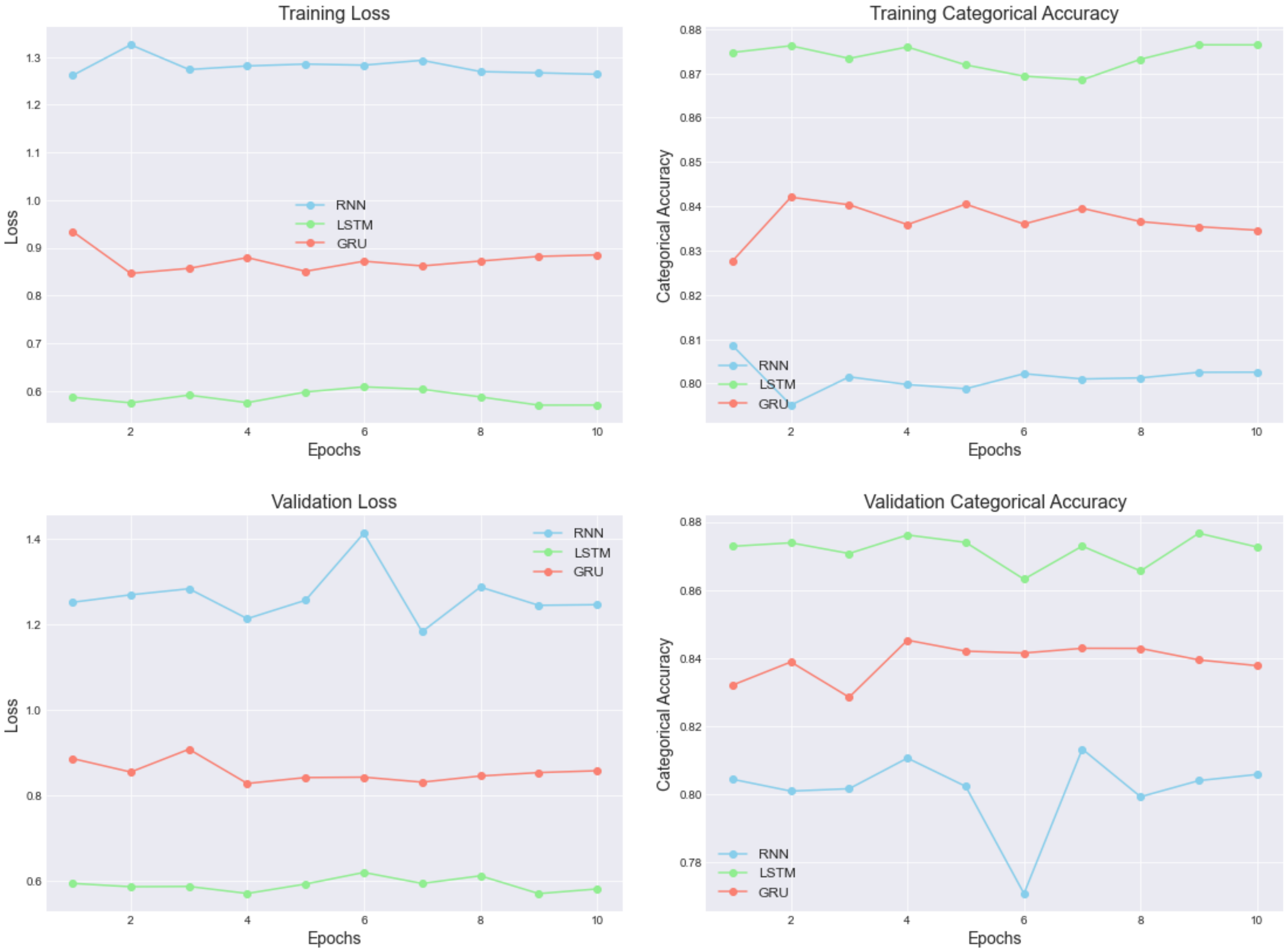


Figure 3 – LSTM, RNN, GRU Training graphs

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (7)$$

- Updated hidden state:

$$h_t = o_t * \tanh(C_t) \quad (8)$$

C. Gated-Recurrent Unit

- Update gate:

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \quad (9)$$

- Reset gate:

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \quad (10)$$

- Candidate hidden state:

$$\widetilde{h_t} = tanh(W[r_t * h_{t-1}, x_t] + b) \quad (11)$$

- Final hidden state:

$$h_t = z_t * h_{t-1} + (1 - z_t) * \tilde{h}_t \quad (12)$$

IV. Implementation

The study utilized a parallel corpus for French to English translation, obtained from an openly accessible dataset. This dataset comprises sentence pairs with the French sentence aligned to its English translation. For the purposes of this analysis, a subset of the dataset, specifically the first and last 1000 sentence pairs, was extracted to provide a balanced representation of the data. The combined dataset was then loaded into a Pandas DataFrame for preprocessing. The data was split with 70% of it being allocated for training and the rest for validation. Testing was done through random sampling through the entire untouched corpus.

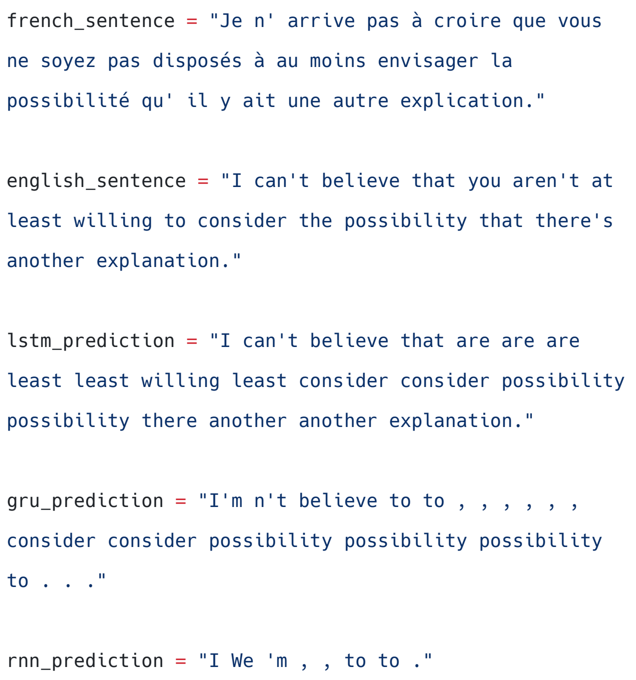
Each model was trained on the vectorized French sentences as input with the corresponding English sentences as the output. The training process involved optimizing the models to minimize the translation error, using a combination of loss functions and optimization algorithms suitable for sequence learning tasks.

A. Key Considerations

It is imperative to note some limitations that arose over the course of the project. For instance, due to the shortage in time, out of 160000 sentences only 2000 were extracted as a sampled dataset. It will be seen in the next section that the reference French text and its corresponding English translation and model outputs were outputted after a training of 100 epochs, but the training graphs only depict training up to the 10th epoch. This is because subsequent the BLEU and ROUGE score calculation, the history variable generated by the model.fit function of keras was overwritten by the group members by accident and time did not allow the retraining of all three models up to the 100th epoch all over again.

V. Results

The results from the training and evaluation of the three neural network architectures, RNN, LSTM, and GRU, reveal significant differences in their performance capabilities, as shown by Figures 2 and 3. The test reference text and their corresponding candidates generated by their respective models are given as follows:



```
french_sentence = "Je n' arrive pas à croire que vous
ne soyez pas disposés à au moins envisager la
possibilité qu' il y ait une autre explication."

english_sentence = "I can't believe that you aren't at
least willing to consider the possibility that there's
another explanation."

lstm_prediction = "I can't believe that are are are
least least willing least consider consider possibility
possibility there another another explanation."

gru_prediction = "I'm n't believe to to , , , , , ,
consider consider possibility possibility possibility
to . . ."

rnn_prediction = "I We 'm , , to to ."
```

LSTM achieved the highest BLEU score of 0.272 and performed best across all ROUGE metrics (Recall, Precision, and F1-score). This indicates that LSTM was the most effective in translating the French to English sentences closely to the reference translations. Despite a BLEU score near zero (5.36e-155), the GRU model showed reasonable performance in terms of ROUGE-1 metrics. However, its performance dropped sharply for ROUGE-2 and ROUGE-L, indicating issues with capturing more complex linguistic structures. The RNN model had the lowest performance with a negligible BLEU score (2.81e-232) and the least favorable ROUGE scores. This suggests significant limitations in its capacity to handle the complexities of language translation in this context. It must be noted that the training times were longest for LSTM, GRU, and then RNN in that order.

The training loss trends indicate that LSTM and GRU models have a steadier decrease and lower loss compared to the RNN, which reflects their better learning capabilities over epochs. The LSTM model consistently showed the lowest training and validation loss, suggesting it was more effective in minimizing the error between predicted and actual translations.
Training and validation accuracy trends provide insights into the models' ability to correctly predict the output class at each timestep in a sequence. The LSTM again showed superiority, maintaining higher accuracy throughout the training process, followed by the GRU. The RNN lagged significantly behind, struggling to reach similar levels of accuracy.

The graphical data on training loss, validation loss, and categorical accuracy visually supports the quantitative findings:

- Training Loss Graph: Displays a consistent decline in loss for LSTM and GRU, with RNN showing higher variability and less reduction over time.
- Validation Loss Graph: While all models exhibit fluctuations, LSTM maintains a lower and more stable loss curve, suggesting better generalization on unseen data.
- Categorical Accuracy Graph: Highlights the higher performance of LSTM and GRU in accurately categorizing the predicted words during training and validation phases.

These results collectively underscore the superior performance of LSTM in handling the complexities of neural machine translation from French to English, followed by GRU, with RNN trailing behind in all evaluated metrics.

VI. Conclusion

This comparative analysis of RNN, LSTM, and GRU architectures for neural machine translation from French to English provided significant insights into the effectiveness of each model. The LSTM architecture demonstrated superior performance across all metrics, including BLEU and ROUGE scores, as well as training and validation losses and accuracy. GRU, while not matching the performance of LSTM, still outperformed the basic RNN model, which showed the weakest results in all key metrics.

Future research could explore several areas to build upon the findings of this study, such as incorporating attention mechanisms, hybrid models, and using larger datasets containing varied language pairs.

In conclusion, this study underscores the importance of choosing the right neural network architecture for machine translation tasks. LSTM stands out as the most effective model in terms of both accuracy and handling of complex linguistic structures in the French to English translation task. These insights are valuable for researchers and practitioners in the field of machine translation, offering guidance on selecting and optimizing neural network architectures for better performance and efficiency in real-world applications.

Acknowledgment

The authors would like to extend their gratitude to the course instructor Dr. Khurram Khan who was pivotal in the group member's learning experience. His depth of knowledge and kind guidance at times provided the group members with the much-needed mentoring required to complete the project on time.

References

Shiri, F. M., et al. "A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU." arXiv:2305.17473 [cs.LG]. Available at: arXiv:2305.17473.

H. Xu, et. al, "Machine Translation Using Deep Learning for Universal Language Translation," in Proceedings of the Springer Conference, 2021. Available: https://link.springer.com/chapter/10.1007/978-981-15-6340-1_4

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv:1409.0473. Available: https://arxiv.org/abs/1409.0473

Xu, H., Liu, Q., van Genabith, J., Xiong, D., & Zhang, M. (2021). Multi-Head Highly Parallelized LSTM Decoder for Neural Machine Translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language
Processing (Volume 1: Long Papers) (pp. 1390-1402). Online: Association for Computational Linguistics. Available: https://aclanthology.org/2021.acl-long.112

H. Xu, Q. Liu, J. van Genabith, D. Xiong, and M. Zhang, "Multi-Head Highly Parallelized LSTM Decoder for Neural Machine Translation," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online: Association for Computational Linguistics, 2021, pp. 1390-1402.