English-Vietnamese Machine Translation: Applying Gated Recurrent Unit and Long Short-Term Memory Models

Nguyen Tien An
20021080
20021080@vnu.edu.vn

Ho Tu Minh
22022674
22022674@vnu.edu.vn

Nguyen Phan Hien
22022534
22022534@vnu.edu.vn

Abstract

This study focuses on the application and exploitation of the strengths of two popular deep learning models, Transformer and Long Short-Term Memory (LSTM), in the task of automatic translation from English to Vietnamese. By building, training, and evaluating these models on bilingual data, the research aims to determine how these architectures can be optimized to enhance translation quality, while also analyzing the factors affecting the performance of each model.

1 Introduction

Machine translation is a core field in natural language processing, playing a crucial role in various industries such as information technology, education, and media. Its applications go beyond breaking language barriers between nations, supporting businesses in market expansion, enhancing access to information, and improving global communication efficiency. With advancements in artificial intelligence, machine translation models have been continuously refined to achieve higher accuracy and efficiency. In the field of machine translation, two commonly used models are Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM).

Gated Recurrent Unit (GRU), with its simple yet effective design, has become a popular choice in natural language processing. GRU employs gating mechanisms to control the flow of information, enabling the model to retain essential data in sequential inputs without requiring an excessive number of parameters. This design reduces computational complexity, accelerates training and inference processes, and remains highly effective when handling sequential data, especially longer sequences.

Meanwhile, Long Short-Term Memory (LSTM) is a pioneering recurrent neural network known for its ability to process long-term sequential data. LSTM utilizes more complex gating mechanisms to manage the retention and forgetting of information, offering superior performance in capturing sequential relationships between words in a sentence. Despite being introduced long ago, LSTM continues to play a vital role in natural language processing, particularly in tasks that demand the modeling of intricate contextual relationships.

This project aims to evaluate and compare the performance of two models on the English-Vietnamese translation dataset, a language pair characterized by significant differences in grammar structure and vocabulary. By building and training both models on a standard bilingual dataset, the research will analyze their translation effectiveness, focusing on aspects such as semantic accuracy and the ability to handle long and complex sentences. The comparison results will highlight the strengths and weaknesses of each model and provide practical recommendations for improving machine translation quality in the future.

Preprint. Under review.

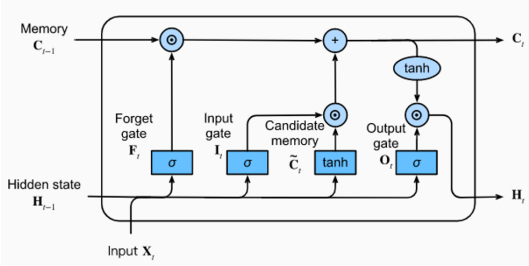

Figure 1: Long Short-Term Memory.

2 Models

2.1 Long Short-Term Memory

Long Short-Term Memory (LSTM) is an improved version of recurrent neural network. A traditional RNN has a single hidden state that is passed through time, which can make it difficult for the network to learn long-term dependencies. LSTMs model address this problem by introducing a memory cell, which is a container that can hold information for an extended period. LSTM architectures are capable of learning long-term dependencies in sequential data, which makes them well-suited for tasks such as language translation, speech recognition, and time series forecasting.

The Long Short-Term Memory architectures involves the memory cell which is controlled by three gates: the input gate, the forget gate, and the output gate. These gates decide what information to add to, remove from, and output from the memory cell.

- The input gate controls what information is added to the memory cell.
- The forget gate controls what information is removed from the memory cell.
- The output gate controls what information is output from the memory cell.

This allows LSTM networks to selectively retain or discard information as it flows through the network, which allows them to learn long-term dependencies. The LSTM maintains a hidden state, which acts as the short-term memory of the network. The hidden state is updated based on the input, the previous hidden state, and the memory cell's current state. The model is illustrated in Figure 1.

2.2 Gate Recurrent Unit

Gated Recurrent Unit (GRU) is a type of recurrent neural network (RNN) that was introduced by Cho et al. in 2014 as a simpler alternative to Long Short-Term Memory (LSTM) networks. Like LSTM, GRU can process sequential data such as text, speech, and time-series data.

The basic idea behind GRU is to use gating mechanisms to selectively update the hidden state of the network at each time step. The gating mechanisms are used to control the flow of information in and out of the network. The GRU has two gating mechanisms, called the reset gate and the update gate.

The reset gate determines how much of the previous hidden state should be forgotten, while the update gate determines how much of the new input should be used to update the hidden state. The output of the GRU is calculated based on the updated hidden state. The Gated Recurrent Unit (GRU) model is illustrated in Figure 2.

3 Experiments

3.1 Data

The dataset used in the experiment is **IWSLT'15 English-Vietnamese** from the Stanford NLP Group, a widely used and high-quality bilingual translation dataset. It includes 133,000 bilingual sentence pairs in the training set (*train.en* and *train.vi*) and two test sets: *tst2012* (1,544 bilingual sentences) and *tst2013* (1,269 bilingual sentences).

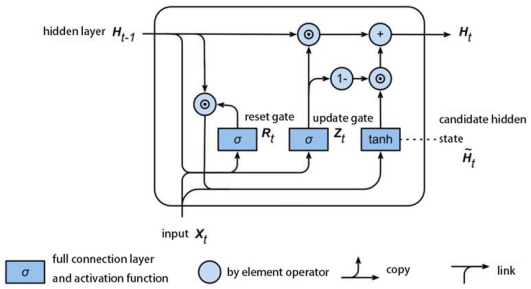

Figure 2: Gated Recurrent Unit.

The English vocabulary (*vocab.en*) contains 17,262 words, while the Vietnamese vocabulary (*vocab.vi*) contains 7,709 words. Additionally, the dataset provides a contextual dictionary (*dict.en-vi*) consisting of English phrases, their Vietnamese translations, and associated alignment probabilities or scores.

Regarding data split ratios, the training set accounts for 95% of the total data (133,000 sentences), the validation set uses 1,544 sentences from *tst2012* (2%), and the test set uses 1,269 sentences from *tst2013* (3%). This dataset is a rich resource that effectively supports research and applications in machine translation and natural language processing between English and Vietnamese.

3.2 Settings

We used Long Short-Term Memory[1][2]
We used Gated Recurrent Unit[3]

3.3 Results

3.3.1 Long Short-Term Memory

The Seq2Seq model with LSTM encoder and decoder was trained on the IWSLT'15 English-Vietnamese dataset for 20 epochs. The training process tracked loss after each epoch to monitor the convergence and performance of the model. BLEU score of 18 is a reasonable baseline for machine translation, specifically using the Vietnamese - English dataset. BLEU 18 indicates that the model produces translations with decent n-gram overlap compared to reference translations. The model demonstrates stable training behavior, showing no abrupt spikes or plateaus in loss, which is a common issue in complex models like Seq2Seq.

3.3.2 Gated Recurrent Unit

The model was trained on the IWSLT'15 English-Vietnamese dataset for 20 epochs. Both the training and test data were filtered to remove any sentences with more than 60 words. This ensured that the GRU model avoided the vanishing gradient problem and allowed for a fair comparison with the LSTM model. Training loss and validation loss were monitored after each epoch to assess the model's convergence and performance. The key results during training is illustrate in Figure3.

From the Figure4, it can be observed that the training loss steadily decreases throughout the training process, while the validation loss gradually declines and stabilizes at a low level after approximately 10 epochs. Although there are minor fluctuations in the later epochs, the validation loss generally remains low, indicating that the model does not exhibit significant overfitting.

After completing the training, the model was evaluated on the test set using the BLEU score—a widely used metric in machine translation tasks. The model achieved a BLEU score of 48 on the test set, indicating strong translation performance with outputs that are semantically and grammatically accurate compared to the reference data.

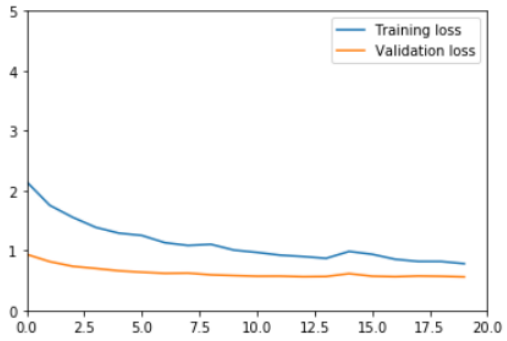

Figure 3: The convergence and performance of the GRU.

Epoch	Training Loss	Validation Loss
0	2.1455	0.9352
1	1.7541	0.8140
2	1.5531	0.7353
3	1.3857	0.6995
4	1.2893	0.6601
5	1.2504	0.6371
6	1.1299	0.6185
7	1.0839	0.6227
8	1.1024	0.5931
9	1.0062	0.5811
10	0.9680	0.5706
11	0.9219	0.5717
12	0.8985	0.5618
13	0.8686	0.5656
14	0.9842	0.6130
15	0.9366	0.5706
16	0.8531	0.5624
17	0.8169	0.5738
18	0.8163	0.5690
19	0.7796	0.5589

Figure 4: The convergence and performance of the GRU.

3.4 Analyses

3.4.1 Long Short-Term Memory

After obtaining the results, we tried fine-tuning a few parameters in the model and received the results shown in the tables below. The parameters we fine-tuned include the number of layers, batch size, and learning rate. The results obtained are shown in the Figure 5, 6, and 7

3.4.2 Gated Recurrent Unit

We provide an analysis of the results obtained from applying the GRU model to the machine translation task:

- **Efficient Training:** The model shows a consistent decrease in loss with each epoch, demonstrating that the optimization process is functioning well. Additionally, the hyperparameters, such as learning rate and batch size, were appropriately chosen to contribute to stable and effective training.
- **Good Convergence:** The validation loss decreases steadily and converges to a low value (0.5589 at epoch 19), confirming that the model generalizes well to unseen data and is not overfitting.
- **BLEU Score Performance:** With a BLEU score of 48, our model performs competitively compared to modern machine translation systems on the IWSLT'15 English-Vietnamese dataset. However, if the learning rate is increased, training becomes unstable. While both training loss and validation loss continue to decrease, they exhibit significant oscillations, particularly near a local minimum.

Learning rate	Convergence Speed	Result
0.0001	Slow convergence, requiring many epochs to reduce loss.	It can achieve high accuracy given enough training time, but it is resource-intensive and prone to getting stuck in local minima.
0.01	Fast convergence, but unstable.	Unstable, with loss oscillating significantly or failing to decrease. It may not converge, leading to poor translation results.
0.001	Fast and stable convergence.	Achieves better results within a limited number of epochs. This is often an optimal choice for Seq2Seq models.

Figure 5: The results of fine-tuning the learning rate.

Batch size	Pros	Cons
16	Frequent gradient updates, requiring RAM.	Slow training, with significant loss oscillations, making convergence difficult.
64	Stable loss, effectively utilizing the GPU if large memory is available	Fewer gradient updates, longer convergence time, and requiring more hardware resources.
32	A balance between computational efficiency and stability.	It is often chosen because it is suitable for the available resources and provides a sufficiently fast convergence speed.

Figure 6: The results of fine-tuning batch size.

Layer Number	Pros	Cons
1	Simple, low resource usage, fast convergence.	Not powerful enough to learn complex contexts, especially for long sentences.
2	Increases the ability to learn longer contexts.	More resource-intensive, but often a suitable level to balance learning efficiency and resource usage.

Figure 7: The results of change number of layers.

3.5 Limitations

3.5.1 Long Short-Term Memory

Limitations of Long Short-Term Memory in Machine Translation:

- **Slow Training and High Computational Cost:** LSTM's complex structure with multiple gates leads to slower training times and higher computational resource requirements.
- **Difficulty with Long-Term Semantic Dependencies:** While LSTM can store information over time, it struggles with maintaining long-term context in very long sentences.
- **Overfitting with Limited Data:** LSTM is prone to overfitting, especially when training data is limited, reducing its ability to generalize.
- **Inability to Capture Global Context:** LSTM processes data sequentially, which limits its ability to capture the global context of the entire sentence or paragraph.
- **Challenges with Language Pairs Having Different Structures:** LSTM can struggle when translating between languages with significantly different grammatical and syntactic structures.

3.5.2 Gated Recurrent Unit

Limitations of Gated Recurrent Unit in Machine Translation:

- **Limited Long-Term Context Handling:** GRU, while better than traditional RNNs, still struggles with capturing long-range dependencies in complex sentences, especially compared to LSTM.
- **Difficulty with Complex Syntax:** GRU models may struggle to handle complex syntactic structures and nuanced grammatical rules, which are common in machine translation tasks.
- **Challenges with Imbalanced or Insufficient Data:** GRU performance may degrade when training data is imbalanced or lacks sufficient context, especially for rare words or unusual grammatical constructions.
- **Optimization Challenges:** Despite its simplicity, GRU can be difficult to optimize in complex machine translation tasks, particularly when dealing with large-scale datasets or nuanced language pairs.
- **Inability to Capture Global Context:** GRU processes information sequentially, limiting its ability to capture global sentence-level context, which can affect translation quality in longer or more complex sentences.

3.6 Future Work

3.6.1 Long Short-Term Memory

Based on the results achieved, here are some suggested future work and improvements for the LSTM model:

- **Extended Training:** Increasing the number of epochs beyond 20 could lead to further reduction in loss and improved model performance.
- **Fine-tuning Hyperparameters:** Experimenting with additional hyperparameters such as the number of layers, hidden size, and dropout rate to further optimize the model.

3.6.2 Gated Recurrent Unit

Also based on the results achieved, here are some suggested future work and improvements for the GRU model:

- **Increase the Number of Epochs:** Training with more epochs could further improve the model's performance.
- **Hyperparameter Tuning:** Experimenting with different learning rates, the number of GRU layers, and dropout rates to optimize performance.

- **Comparison with Other Architectures:** Comparing the GRU model with LSTM and Transformer models to further evaluate their benefits and limitations.

4 Conclusion

In this report, our team primarily applied two models, LSTM and GRU, to address the machine translation task, specifically for English-Vietnamese translation. Based on the BLEU score, the GRU model outperformed the LSTM model in terms of translation quality. However, the dataset used in this study is relatively limited, so the results should be considered as an initial step in exploring the application of these two models to machine translation tasks. Further research with a more diverse and extensive dataset is needed to fully assess the potential of both models and refine their performance in real-world translation scenarios.

The experiments presented in this paper are considered preliminary. To gain a deeper understanding of how gated units, such as those in LSTM and GRU, contribute to the learning process and to better isolate the impact of each component (e.g., the gating mechanisms in LSTM or GRU), more comprehensive experiments are needed. Additionally, further optimization of the machine translation task, including fine-tuning of hyperparameters and utilizing more diverse datasets, will be crucial for improving the performance and assessing the true potential of these models in real-world applications.

References

[1] Ralf C. Staudemeyer, E. R. M. Understanding lstm – a tutorial into long short-term memory recurrent neural networks. 2019.

[2] Minh-Thang Luong, C. D. M. Stanford neural machine translation systems for spoken language domains. 2015.

[3] Junyoung Chung, K. C. Y. B., Caglar Gulcehre. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014.