AI6127 Deep Neural Networks for Natural Language Processing
Assignment 2
Seq2Seq Model for Machine Translation

Seow Yee Ying
G2304259C

1 Introduction

This report presents a comparative analysis of five sequence-to-sequence models for machine translation. The models include the GRU Encoder-Decoder, LSTM Encoder-Decoder, Bi-LSTM Encoder - GRU Decoder, GRU Encoder-Decoder with Attention Mechanism, and Transformer Encoder - GRU Decoder. We describe the architecture of each model, showcase their translation outputs, and evaluate their performance using ROUGE-1 and ROUGE-2 metrics.

2 Description of Model Architecture

2.1 GRU Encoder and Decoder

The GRU Encoder-Decoder model is a sequence-to-sequence framework widely used in machine translation task. This model consists of two main components: the GRU encoder and the GRU decoder. The encoder processes the input sequence and converts it into context vector. The context vector, which is also known as the hidden state, encapsulates the information for all the input data for the decoder to make accurate predictions. GRU is applied to prevent the vanishing gradient issue. It has two internal gates: the reset gate and the update gate. These gates manage the information flow, deciding the past information to discard or integrate with the current input. The GRU in decoder receives the context vector from the encoder and generates the output sequence one at a time. The GRU architecture is simple and easy to train. It allows the model to capture dependencies in the output data but may struggle with long-term dependencies.

2.2 LSTM Encoder and Decoder

This architecture employs LSTM units in both the encoder and decoder. LSTM include three gates: input, output, and forget gates. This structure allows LSTMs to maintain or discard information dynamically. The encoder generates hidden states based on input data and the decoder uses it to generate the outputs. Due to the comprehensive gate mechanisms, this model is more robust for handling long input sequences and complex data. The rouge scores for LSTM Encoder-Decoder architecture are shown in the table below.

2.3 Bi-LSTM Encoder and GRU Decoder

The Bi-LSTM encoder processes the input sequence in both forward and backward directions. This allows the model to capture the context from both directions, having a better understanding about the relationship between the sequences. The resulting hidden states from both directions are concatenated before passing to the decoder. The GRU decoder then uses the hidden states to generate the output sequence. This hybrid approach leverages the Bi-LSTM's ability to capture the context and the GRU's efficiency in the output generation.

2.4 GRU Encoder and Decoder with Attention Mechanism

The GRU Encoder-Decoder with Attention Mechanism is built upon the basic GRU seq2seq model. The attention mechanism allows the decoder to focus on specific part of the input for translation tasks. The encoder processes the input sequence into a series of hidden states, while the attention mechanism calculates and normalizes attention weights across these states. These weights form an attention distribution that helps create a context vector for decoder. The decoder can focus adaptively on the specific input sequence for translation process based on the context vector to generate the output sequence.

2.5 Transformer Encoder and GRU Decoder

The transformer encoder processes the entire input sequence simultaneously using the self-attention mechanisms. This process weights the relevance of all parts of the input relative to each other. In addition, the positional encoding in transformer helps to keep track of the word positions in the input sequence, enriching the contextual representations. These representations are then fed into the GRU decoder to generate the output sequentially.

3 Model Setting

The dataset is split into training set and test set with ratio of 90:10. The models are trained for 20 epochs, using SGD optimizer with learning rate of 0.01. Cosine annealing learning rate scheduler with a maximum iteration of 20 is applied on GRU Encoder-Decoder with Attention model and Transformer Encoder-GRU Decoder model. These models use negative log likelihood as their loss function and are evaluated using the ROUGE scores, specifically ROUGE-1 and ROUGE-2 for fmeasure, precision and recall.

4 Result Analysis

Model	Rouge 1			Rouge 2		
	F Measure	Precision	Recall	F Measure	Precision	Recall
GRU Encoder-Decoder	0.6749	0.6308	0.7335	0.5102	0.4686	0.5679
LSTM Encoder-Decoder	0.6864	0.6374	0.7504	0.5294	0.4837	0.5916
Bi-LSTM Encoder - GRU Decoder	0.6964	0.6481	0.7598	0.5385	0.4926	0.6015
GRU Encoder-Decoder with Attention Mechanism	**0.7006**	0.6546	0.7613	0.5388	0.4950	0.5995
Transformer Encoder - GRU Decoder	0.7003	0.6522	0.7636	**0.5458**	0.4992	0.6096

The model with the best performance in ROUGE-1 score is the GRU Encoder-Decoder with Attention Mechanism. This model achieves an F Measure of 0.7006, which is slightly higher than the Transformer Encoder-GRU Decoder model with the score of 0.7003. The higher ROUGE-1 score for the GRU Encoder-Decoder with Attention Mechanism model suggests that it is slightly more effective at capturing correct individual words in the translation outputs. The attention mechanism in the GRU Encoder-Decoder allows the model to align its output closely with individual words from the reference translation.

In contrast, the Transformer Encoder is highly capable of capturing the complex relationships in the encoding phase. However, the subsequent GRU decoder may not effectively decode this complex representation into an equally nuanced output. The simplicity of the GRU causes the model may not fully utilize the detailed context provided by the encoder. Consequently, this can lead to the Transformer Encoder - GRU Decoder model underperforming in ROUGE-1.

In terms of ROUGE-2 scores, the Transformer Encoder-GRU Decoder demonstrates superior performance, achieving the highest F Measure of 0.5458. This is attributed to the model's self-attention mechanism, which excels at understanding the context and nuances of language. This mechanism enables the model to maintain the sentence structure and coherence effectively, which is important for constructing coherent bigrams. The ability to process the entirety of the input sequence effectively ensures that the model not only captures but also preserves the structural integrity and contextual relevance of phrases throughout the translation process. Therefore, the Transformer Encoder-GRU Decoder model is adept at generating translations that are both linguistically and contextually aligned with the target language.

Translation Output

Model	Translation Output
GRU Encoder-Decoder	> vous n etes pas chanteuse . = you re no singer . < you re not singer . <EOS> > je suis pour ainsi dire heureux . = i m kind of happy . < i m kind of happy . <EOS> > il est probablement toujours vivant . = he is probably still alive . < he is probably still alive . <EOS> > elle etudie l anglais tous les jours . = she studies english every day . < she studies english every day . <EOS> > je creve d envie d une boisson fraiche . = i m dying for a cold drink . < i am dying for a cold drink . <EOS>
LSTM Encoder-Decoder	> il est marie et a deux enfants . = he is married with two children . < he is married with two children . <EOS> > il joue avec mon chat . = he s playing with my cat . < he is playing with my cat . <EOS> > ce que tu penses ne m interesse pas . = i m not interested in what you think . < i m not interested in what you think . <EOS> > je suis toujours heureux . = i m always happy . < i m still happy . <EOS> > tu es plus grande que moi . = you are taller than me . < you re taller than me . <EOS> > vous etes le seul a pouvoir m aider . = you re the only one who can help me . < you re the only one who can help me . <EOS>
Bi-LSTM Encoder - GRU Decoder	> sortir avec elle et manger a notre restaurant prefere m a manque . = i missed going out with her and eating at our favorite restaurant . < i missed going out with her and eating at our favorite restaurant . <EOS> > je ne dis pas que tu as tort . = i m not saying you re wrong . < i m not saying that you re wrong . <EOS> > il est confiant dans ses capacites . = he is confident of his ability . < he is confident of his ability . <EOS> > je me specialise en sociologie . = i m majoring in sociology . < i m majoring in sociology . <EOS> > je n ai pas tres faim a l instant . = i m not very hungry right now . < i m not very right right now . <EOS>
GRU Encoder-Decoder with Attention Mechanism	> je suis depressif . = i m depressed . < i m bad . <EOS> > vous etes fort effrontes . = you re very forward . < you re very forward . <EOS> > j ai tres soif . = i m really thirsty . < i m very thirsty . <EOS> > nous ne sommes pas ouvertes . = we re not open . < we re not open . <EOS> > j essaye d imaginer ca . = i m trying to imagine that . < i m trying to figure out . <EOS>
Transformer Encoder - GRU Decoder	> elles ont de la chance d etre vivantes . = they re lucky to be alive . < they re lucky to be alive . <EOS> > nous cherchons un endroit ou dormir . = we re looking for a place to sleep . < we re looking for a place to sleep . <EOS> > nous ne sommes pas seules . = we re not alone . < we aren t alone . <EOS> > tu es un peu bizarre . = you re a little weird . < you re a little weird . <EOS> > c est un gentleman il ne peut avoir dit une telle chose . = he is a gentleman . he cannot have said such a thing . < he is a gentleman . he cannot have said such a thing . <EOS>

Five random translation pairs of sentences are selected from each model. The analysis for the translation output is shown in the table above.

The GRU Encoder-Decoder shows a basic level of semantic correctness but struggles with the contractions and grammatical consistency. For example, the model can correctly translate 'I m' to 'I am' in the first and last example but couldn't identify it in the second example. However, it can accurately translate the expressions and sentences reflecting daily life activities. This is shown by the translation of ‘kind of happy’ and ‘she studies english every day’. The model can capture the text context, showing the overall meaning with some room of improvement for the linguistic details.

The LSTM Encoder-Decoder model maintains the overall meaning of the sentence during translation. However, it shows the similar challenges with contractions as seen with the GRU Encoder-Decoder model. However, there is an incorrect translation for the adverb of frequency in the fourth example. The original translation is ‘I m always happy’, but the model’s translation shows ‘I m still happy’. This might lead to inaccurate semantic understanding, showing the area of improvement for this model.

Bi-LSTM Encoder-GRU Decoder model shows an accurate translation for a complex sentence, shown in the first example. This indicates its ability to handle complex sentence structures. However, it consistently struggles with the use of contractions. There are some missing nuanced translations in the second, but it doesn’t affect the understanding of the whole sentence. Nevertheless, there is a clear error in the fifth example, showing the duplication of the word ‘right’. This could be due to the model generation process by the GRU Decoder.

GRU Encoder-Decoder with Attention Mechanism shows a reasonable overall context understanding and is capable to handle the complex sentence meaning. It demonstrates good translation skills, by providing output with different word or phrases, representing the same meaning as the original text. This can be shown by the first, third and fifth examples. Although there is some consistent error in the contraction translation, this model still exhibits the good semantic understanding to the context and translates the message accurately.

Transformer Encoder - GRU Decoder model shows a high-level accuracy in translating each word. This indicates the model’s superior performance in context understanding and semantic translation. However, the model has minor error with the contraction translation, without affecting the understanding of the overall sentence context.

Conclusion

In conclusion, our analysis shows the distinct performance across five sequence-to-sequence machine translation models. The GRU Encoder-Decoder with Attention Mechanism model shows the best performance in terms of individual word accuracy. At the same time, the Transformer Encoder-GRU Decoder demonstrates its superior performance in preserving the sentence structure and capturing the overall context understanding. Throughout the qualitative analysis on the translation output, we found that there is a consistent challenge across all model about the handling of English contractions, pointing to a potential area for future research and improvement.