Name: Sean Goh Ann Ray
Matric No.: G2202190G

Task 2:

Using the provided code with the included seq2seq architecture, an experiment was performed on the fra-eng translation data. After some preprocessing of the French – English data sentences, there were 20639 pairs of sentences. This was then split 80/20 into train and test set. The seq2seq model provided includes both encoder and decoder, both of which uses a gated recurrent unit (GRU), with a hidden dimension of 512. During training, the negative log-likelihood (NLL) loss was used together with the stochastic gradient descent (SGD) optimizer with a fixed learning rate of 0.01 and other default parameters based on the pytorch library. The model was trained for 10 epochs and the Rouge scores were obtained on both the train and test sets after the training, as shown in Tables I and II respectively.

Table I. Rouge Scores of Train Set

Train Set	Rouge 1	Rouge 2
F-measure	0.86610216	0.8001527
Precision	0.799079	0.7251992
Recall	0.9489331	0.8978009

Table II. Rouge Scores of Test Set

Test Set	Rouge 1	Rouge 2
F-measure	0.67640704	0.5073378
Precision	0.62986386	0.46397305
Recall	0.7387624	0.56819886

Rouge 1 scores are defined as the overlap of unigrams (single words) between the ground truth text and the predicted translation while Rouge 2 scores are defined as the overlap of bigrams (2 words in a row) between the ground truth text and the predicted translation. As such, Rouge 2 scores will always be lower than their Rouge 1 counterparts. Precision is defined as the ratio of the number of n-grams appearing in both ground truth text and predicted translation to the number of n-grams in the predicted translation, while recall is the ratio of the number of n-grams appearing in both ground truth text and predicted translation to the number of n-grams in the ground truth text. F-measure is a combination of precision and recall by the formula below and will therefore be used for the comparisons in the results of all the experiments.

$$Fmeasure = 2 * \frac{precision * recall}{precision + recall}$$

As shown in Tables I and II, the trained model performs relatively well over the train set with Rouge 1 and Rouge 2 f-measure score of 0.866 and 0.800 respectively, but not as much over the test set with Rouge 1 and Rouge 2 f-measure score of 0.676 and 0.507 respectively. The Rouge scores in Tables I and II are the scores achieved by this baseline model and will be used to compare with those of the next few experiments.

Task 3:

The next experiment was performed with the GRU layer in both the encoder and decoder being replaced by a one-directional LSTM layer. Training parameters remain unchanged, and the Rouge scores achieved on the train and test set are shown in Tables III and IV respectively.

Table III. Rouge Scores on Train Set

Train Set	Rouge 1	Rouge 2
F-measure	0.8476615	0.77088743
Precision	0.78187186	0.6987244
Recall	0.92910296	0.8655181

Table IV. Rouge Scores on Test Set

Test Set	Rouge 1	Rouge 2
F-measure	0.66046494	0.48830846
Precision	0.6126053	0.4449118
Recall	0.7240006	0.5489743

Comparing Tables I, II, III and IV, the one-directional LSTM layer does not perform as well as compared to the original GRU layer. From Tables I and III, which are the results on the train set, there is an approximate 0.019 and 0.030 decrease in the Rouge 1 and Rouge 2 f-measure scores respectively. From Tables II and IV, which are the results on the test set, there is an approximate 0.016 and 0.019 decrease in the Rouge 1 and Rouge 2 f-measure scores respectively.

From this, it can also be concluded that the model with one-directional LSTM layer is able to generalize better as the decrease in performance on the test set is smaller than that on the train set, despite performing worse overall.

Task 4:

The next experiment was performed with the GRU layer in the encoder being replaced by a bi-directional LSTM layer, and the decoder remains unchanged, still having the GRU layer. Training parameters remain unchanged, and the Rouge scores achieved on the train and test set are shown in Tables V and VI respectively.

Table V. Rouge Scores on Train Set

Train Set	Rouge 1	Rouge 2
F-measure	0.8669647	0.80207396
Precision	0.7971948	0.7245477
Recall	0.9537548	0.9041542

Table VI. Rouge Scores on Test Set

Test Set	Rouge 1	Rouge 2
F-measure	0.679606	0.511453
Precision	0.62672544	0.46357793
Recall	0.75115275	0.57932574

Comparing Tables I, II, V and VI, the bi-directional LSTM layer in the encoder generalizes data better than the model with the original GRU layer. From Tables I and V, which are the results on the train set, there is an approximate 0.001 and 0.002 increase in the Rouge 1 and Rouge 2 f-measure scores respectively. From Tables II and VI, which are the results on the test set, there is an approximate 0.003 and 0.004 increase in the Rouge 1 and Rouge 2 f-measure scores respectively.

From this, it can be concluded that the bi-directional LSTM layer in the encoder helps to generalize unseen data better without overtraining the models.

Task 5:

The next experiment was performed with the attention mechanism (based on Lecture 8, as shown in Figure 1) added between the original encoder and decoder. Training parameters

remain unchanged, and the Rouge scores achieved on the train and test set are shown in Tables VII and VIII respectively.

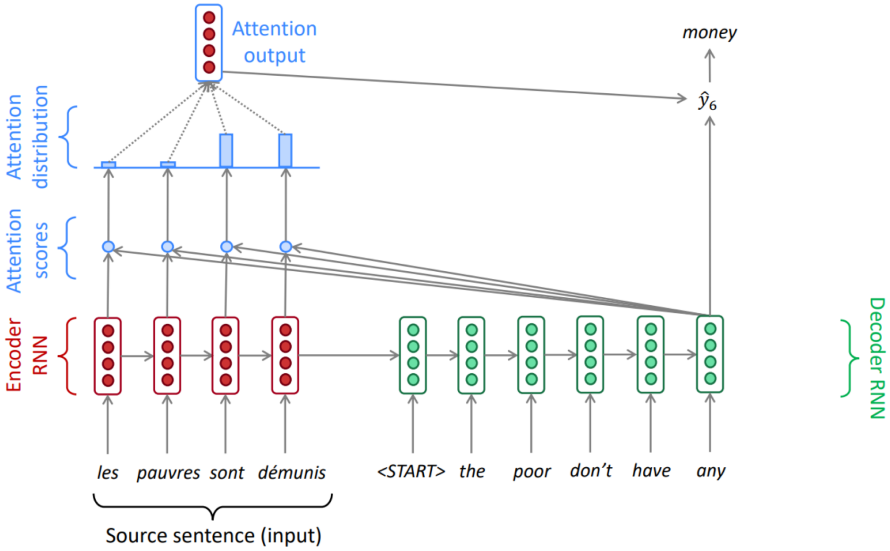

Figure 1. Attention Mechanism between Encoder and Decoder

Table VII. Rouge Scores on Train Set

Train Set	Rouge 1	Rouge 2
F-measure	0.8720434	0.81110376
Precision	0.80465776	0.7356296
Recall	0.9550148	0.9094266

Table VIII. Rouge Scores on Test Set

Test Set	Rouge 1	Rouge 2
F-measure	0.6800175	0.513534
Precision	0.6320626	0.46883753
Recall	0.74454427	0.57655424

Comparing Tables I, II, VII and VIII, the attention mechanism between the encoder and decoder helps to improve the scores on the train set while not compromising those on the test set. From Tables I and VII, which are the results on the train set, there is an approximate 0.006 and 0.011 increase in the Rouge 1 and Rouge 2 f-measure scores respectively. From Tables I and VIII, which are the results on the test set, there is an approximate 0.004 and 0.006 increase in the Rouge 1 and Rouge 2 f-measure scores respectively.

From this, it can be concluded that the attention mechanism helps to generalize unseen data better while also increasing the accuracy on the train set.

Task 6:

The last experiment was performed with the GRU layer in the encoder being replaced by a transformer encoder layer, and the decoder remains unchanged, still having the GRU layer. Training parameters remain unchanged, and the Rouge scores achieved on the train and test set are shown in Tables IX and X respectively. It is obvious that the scores are very poor in comparison to what was expected of a transformer model. A better conclusion can be made by plotting the training curves of all the experiments, which are shown in Figure 2.

Table IX. Rouge Scores on Train Set

Train Set	Rouge 1	Rouge 2
F-measure	0.1993742	0.11692218
Precision	0.2517856	0.15676597
Recall	0.17285438	0.09986526

Table X. Rouge Scores on Test Set

Test Set	Rouge 1	Rouge 2
F-measure	0.19942978	0.11680654
Precision	0.2520901	0.15629354
Recall	0.1723949	0.09946037

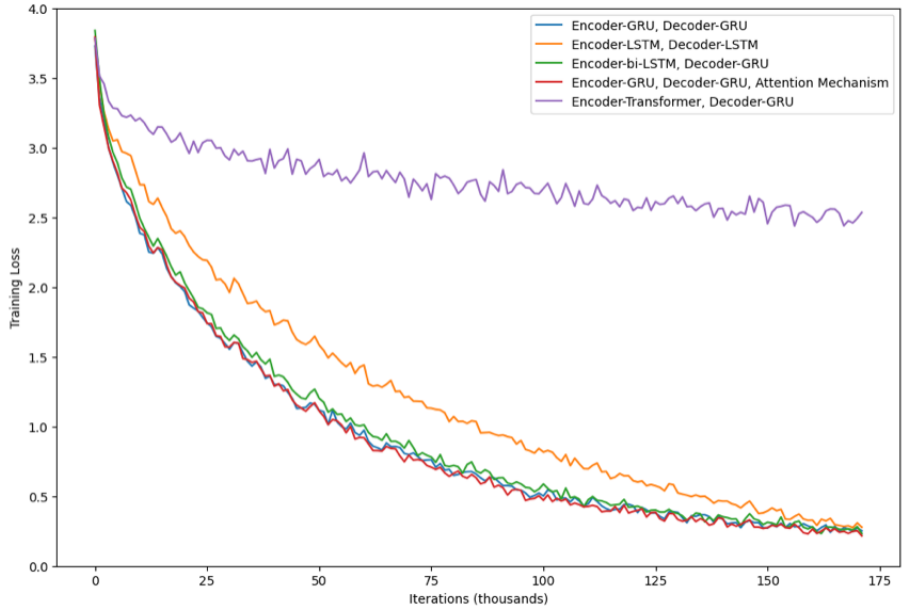

Figure 2. Training Loss of All Experiments

From Figure 2, it can be observed that all models perform relatively similar by the end of the training, with the exception on the transformer model, where the training loss decreases at a much slower rate. Upon further research, it was discovered that a transformer model would normally require much more training iterations and performs much better than LSTMs when the sequences are long, with more than 200 tokens. However, in the experiments, the longest sequence is set to 15 tokens long. As such, the benefits of the transformer model is not utilized. Additionally, with only 10 epochs, which amounts to approximately 172,000 iterations, the transformer model will perform much worse than the other models.

An improvement in the transformer model may be achieved by training over a longer number of epochs and iterations, but this was not possible due to computational limitations.