CSC401 Homework Assignment #2
Analysis

Benson Li

1 Training Results

1.1 Training Loop Printout

1.1.1 Six Models

We have trained six models and we also present BLEU for each comparison.

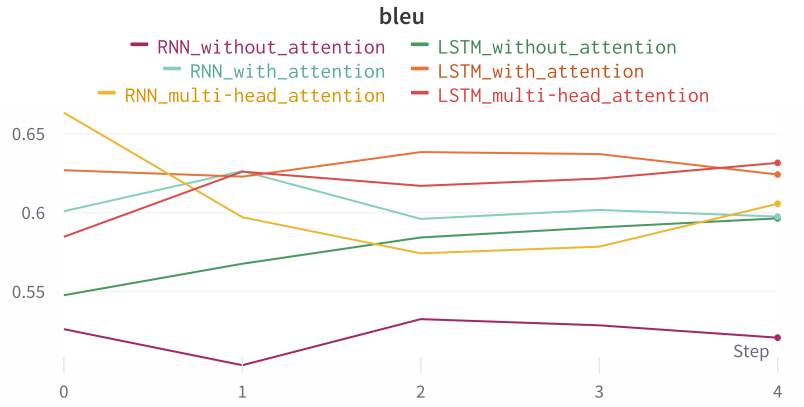

Figure 1: Wandb Training BLEU Score for six models

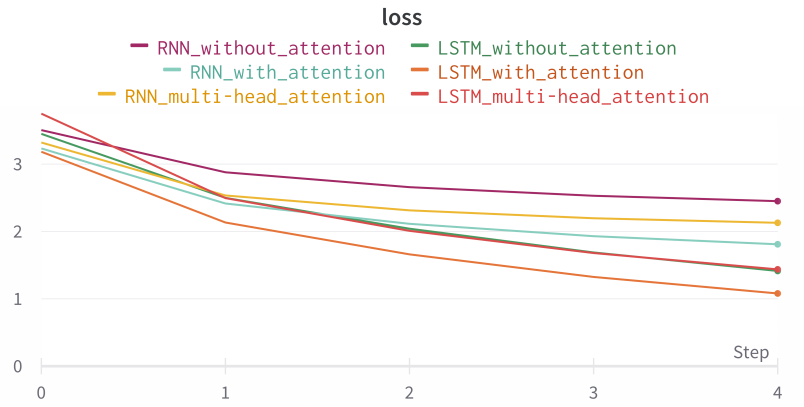

Figure 2: Wandb Training Loss for six models

1.1.2 Models without Attention

Model	Epoch1	Epoch2	Epoch3	Epoch4	Epoch5
RNN	0.5263	0.5035	0.5327	0.5288	0.5210
LSTM	0.5478	0.5678	0.5844	0.5908	0.5965

Table 1: BLEU comparision for Models without Attention

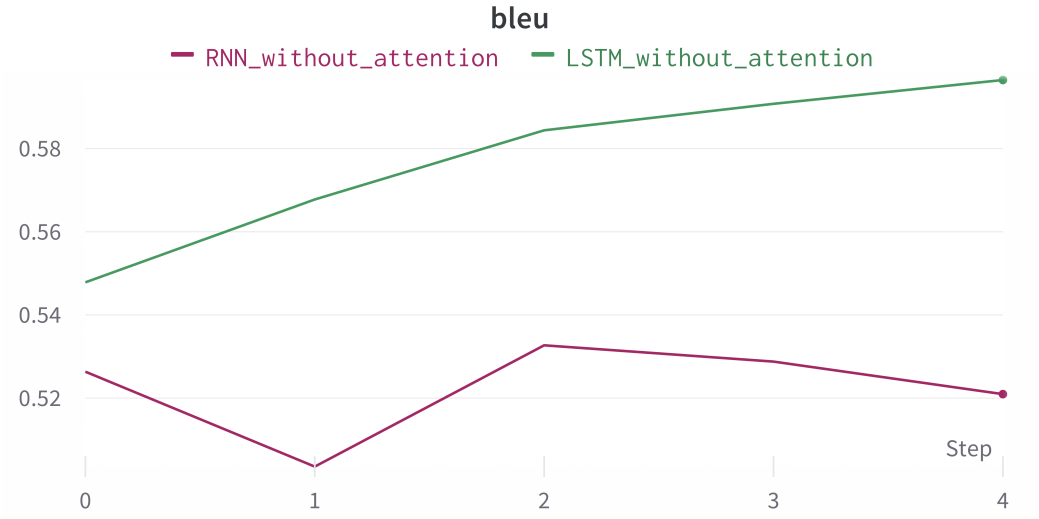

Figure 3: Wandb Training BLEU Score for Model without Attention

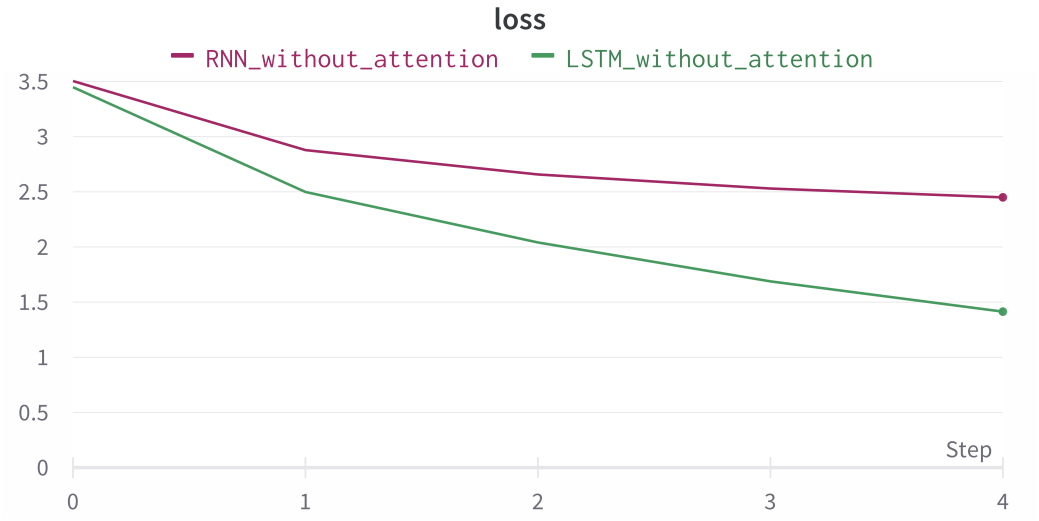

Figure 4: Wandb Training Loss for Model with Attention

1.1.3 Models with Attention

Model	Epoch1	Epoch2	Epoch3	Epoch4	Epoch5
RNN	0.6010	0.6265	0.5962	0.6019	0.5976
LSTM	0.6270	0.6229	0.6385	0.6372	0.6243

Table 2: BLEU comparision for Models without Attention

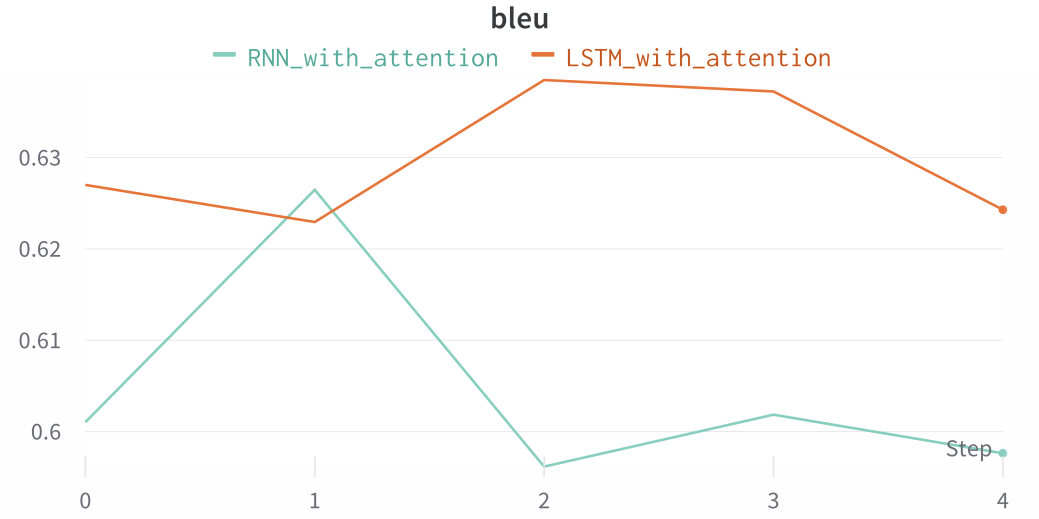

Figure 5: Wandb Training BLEU Score for Model with Attention

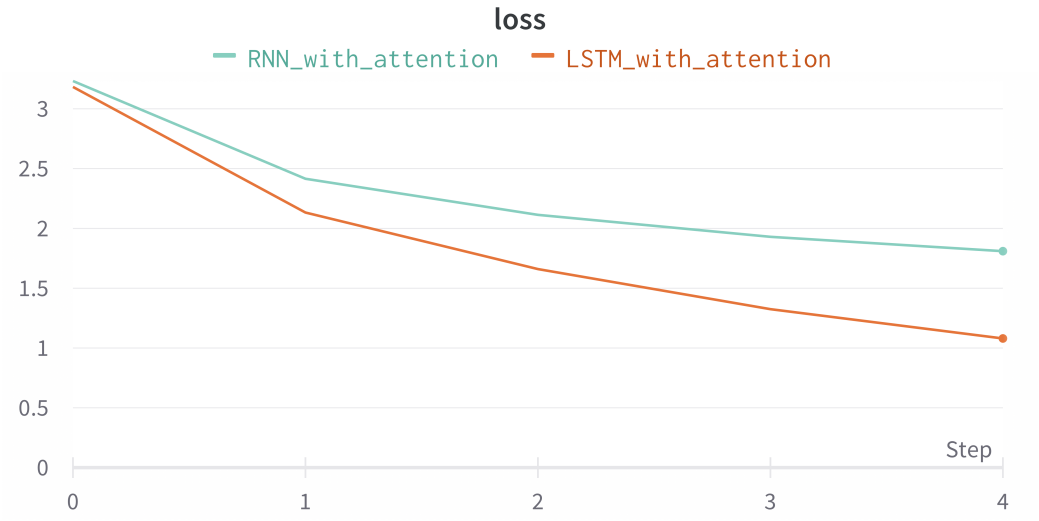

Figure 6: Wandb Training Loss for Model with Attention

1.1.4 Models with Multi-head Attention

Model	Epoch1	Epoch2	Epoch3	Epoch4	Epoch5
RNN	0.6635	0.5972	0.5744	0.5786	0.6058
LSTM	0.5848	0.6261	0.6171	0.6217	0.6317

Table 3: BLEU comparision for Models with multi-head Attention

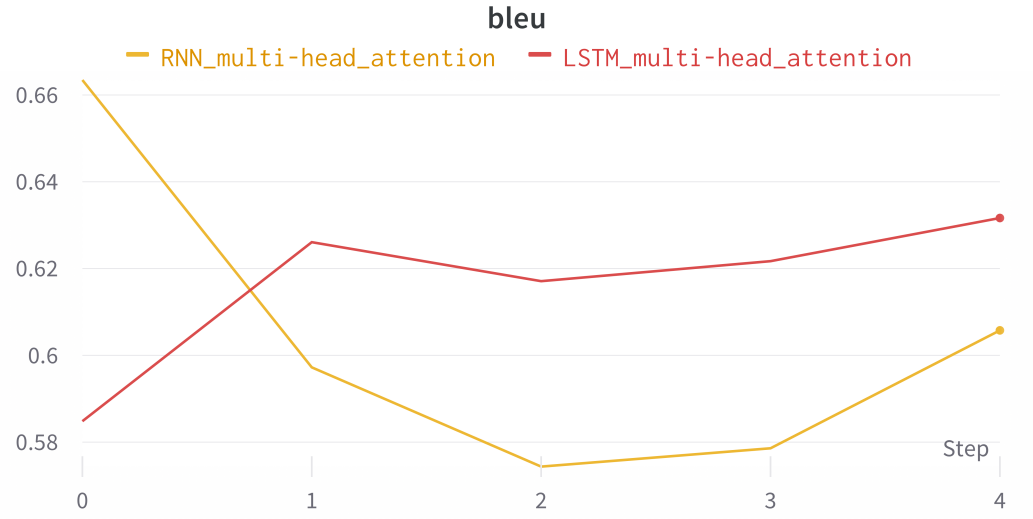

Figure 7: Wandb Training BLEU Score for Model with Multi-head Attention

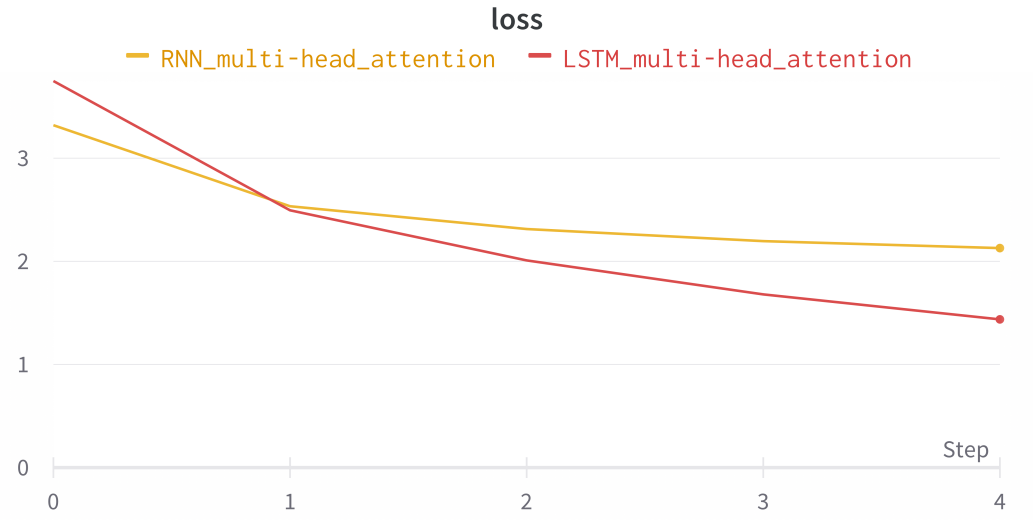

Figure 8: Wandb Training Loss for Model with Multi-head Attention

1.2 Test Set BLEU Score

According to the training output above, LSTM models beat RNN models for all different conditions (without_attention, with_attention, multi-head_attention).

Therefore, we select LSTM models for testing only.

This is the test ouput on each model:

```
LSTM_without_attention: The average BLEU score over the test set was 0.6131839463172473
LSTM_with_attention: The average BLEU score over the test set was 0.6470358437412032
LSTM_multi-head_attention: The average BLEU score over the test set was 0.6480662785308159
```

This section lists the test set BLEU score reported on the test set for each model in table 4.

Model	Test BLEU
LSTM without Attention	0.6131839463172473
LSTM with Single-headed Attention	0.6470358437412032
LSTM with Multi-headed Attention	0.6480662785308159

Table 4: The BLEU score reported on the test set for each model.

1.3 Discussion

In this section, write a brief discussion on your findings. Was there a discrepancy in between training and testing results? Why do you think that is? If one model did better than the others, why do you think that is?

<i> According to out result, there is not a discrepancy in between training and testing results. For each model, the difference between the final training BLEU score is very similar to the testing BLEU score.

Model	Final Train BLEU	Test BLEU
LSTM without Attention	0.597	0.613
LSTM with Single-headed Attention	0.624	0.647
LSTM with Multi-headed Attention	0.632	0.648

Table 5: Comparison for Training and Testing BLUEs for each model

<ii> The above results indicate that the model is not overfitting or underfitting the training data, and that it has learned to generalize well to new data. This may due to the following reasons:

1. Sufficient Training Data: The model is trained on a sufficiently large and diverse dataset (2hours for each), which allows it to learn general patterns in the data and avoid memorizing specific examples.
2. Appropriate Model Complexity: The model architecture may be appropriate for the complexity of the task at hand. If the model is too simple, it may underfit the trainning data, while if it is too complex, it may overfit the training data.
3. Consistency in Data: The training and testing data are similar in their distribution, so the model is not faced with potential bias during testing that it has not seen during training.

<iii>

We notice LSTM is better than RNN. This is because LSTM can selectively remember or forget information over time which has ability to handle long-term dependencies and explode the gradient vanish

problem which traditional RNN suffer from.

We also notice that the model with multi-attention is better than models without attention. This is because it is able to attend to multiple parts of the source sentence simultaneously, allowing it to capture more complex dependencies between the source and target languages comparing with others.

2 Translation Analysis

2.1 Translations

We present our translation results for the following three sentences (translated by Google):

- Toronto est une ville du Canada.
  (Toronto is a city in Canada.)
- Les professeurs devraient bien traiter les assistants d'enseignement.
  (Professors should treat teaching assistants well.)
- Les etudiants de l'Universite de Toronto sont excellents.
  (The students at the University of Toronto are excellent.)

2.1.1 Translation result for the model: LSTM without Attention.

```
model.translate("Toronto est une ville du Canada.")
>>> 'louis j robichaud in canada </s>'
model.translate("Les professeurs devraient bien traiter les assistants d'enseignement. ")
>>> 'the serbs should be gavel to gavel filming </s>'
model.translate("Les etudiants de l'Universite de Toronto sont excellents. ")
>>> 'the students of guelph wellington are a lot of people </s>'
```

2.1.2 Translation result for the model: LSTM with Single-headed Attention.

```
model.translate("Toronto est une ville du Canada.")
>>> 'toronto is a city of canada </s>'
model.translate("Les professeurs devraient bien traiter les assistants d'enseignement. ")
>>> 'the <unk> should be grounded </s>'
model.translate("Les etudiants de l'Universite de Toronto sont excellents. ")
>>> 'the students of toronto in ottawa have been very good </s>'
```

2.1.3 Translation result for the model: LSTM with Multi-headed Attention.

```
model.translate("Toronto est une ville du Canada.")
>>> 'toronto is a city of canada </s>'
model.translate("Les professeurs devraient bien traiter les assistants d'enseignement. ")
>>> 'the teens should be given the <unk> of <unk> <unk> </s>'
model.translate("Les etudiants de l'Universite de Toronto sont excellents. ")
>>> 'the students of toronto maple syrup is very high </s>'
```

2.2 Discussion

In this section, write a brief discussion on your findings. Describe the quality of those sentences. Can you observe any correlation with the model's BLEU score?

<i> Among these three models, the LSTM with single-headed attention model performs better than the model without attention, but still generates some nonsensical words or phrases (sentence 2&3). The LSTM with multi-headed attention model performs the best among the three, generating more

coherent and accurate translations, although it still makes mistakes (sentence 3).

<ii>

There is a general trend of higher BLEU scores correlating with better translation quality. Specifically, the LSTM without attention model has the lowest BLEU score and the worst translation quality, while the LSTM with multi-headed attention model has the highest BLEU score and the best translation quality among the three.

However, there is also some limitations if we use BLEU score as the measure for model's translation quality. According to the test result, the model with single-headed attention is only 0.001 BLEU score from the model with multi-headed attention (0.648 & 0.647 according to table 4), but their translation results has many differences (sentence 2&3).

Overall, BLEU score can be a rough indicator of the model's translation quality but not perfect. It would be better to use in conjunction with some other evaluation metrics (e.g. ROUGE, METEOR) to get a more comprehensive understanding of the model's performance. (please refer to **bonues.pdf**)