CSC401 Homework Assignment #2
Analysis

Fateme Sadat Haghpanah
Student number: 1007014561
UTORid: haghpan1
fateme.haghpanah@mail.utoronto.ca

1 Training Results

1.1 Training Loop Printout

Model without Attention - RNN

```
Epoch 1: loss=0.027417542339658634, BLEU=0.23113416391618719
Epoch 2: loss=0.022505874169482184, BLEU=0.2392896260902504
Epoch 3: loss=0.02079059126310163, BLEU=0.24366945718890734
Epoch 4: loss=0.0198007601263542, BLEU=0.24479469294731282
Epoch 5: loss=0.019174147500161047, BLEU=0.2458779134306805
```

Model without Attention - LSTM

```
Epoch 1: loss=0.0266884388195859, BLEU=0.23396378641042098
Epoch 2: loss=0.01931906481712649, BLEU=0.26218200912137485
Epoch 3: loss=0.015784949232469453, BLEU=0.27583366860653785
Epoch 4: loss=0.013131175401229666, BLEU=0.2859775718698979
Epoch 5: loss=0.011102035943770083, BLEU=0.2894192823229925
```

Model with Single-headed Attention - RNN

```
Epoch 1: loss=0.02527503864382393, BLEU=0.26386720028443195
Epoch 2: loss=0.018901744930327067, BLEU=0.27831559847695425
Epoch 3: loss=0.016560500331110367, BLEU=0.28251426547923925
Epoch 4: loss=0.015126939446399158, BLEU=0.28776641303668
Epoch 5: loss=0.014180067940269811, BLEU=0.28857703923011835
```

Model with Single-headed Attention - LSTM

```
Epoch 1: loss=0.025919217599905852, BLEU=0.26534758900122907
Epoch 2: loss=0.017644117801115616, BLEU=0.29471710223751296
Epoch 3: loss=0.014103900027669876, BLEU=0.311570235788219
Epoch 4: loss=0.011656312328864552, BLEU=0.3197946713365965
Epoch 5: loss=0.009881619183685901, BLEU=0.32075086693896676
```

Model with Multi-headed Attention - RNN

```
Epoch 1: loss=0.028214860720287734, BLEU=0.23790881972575445
Epoch 2: loss=0.021595614228512933, BLEU=0.2531353470564033
Epoch 3: loss=0.019614137970688877, BLEU=0.26013322752845247
Epoch 4: loss=0.018457143859437596, BLEU=0.2626344477377953
Epoch 5: loss=0.017714940358444117, BLEU=0.26694684889705056
```

Model with Multi-headed Attention - LSTM

```
Epoch 1: loss=0.025383769976655042, BLEU=0.26353697636652634
Epoch 2: loss=0.017693355877335934, BLEU=0.2947202832193179
Epoch 3: loss=0.014560080139734833, BLEU=0.30411221552042944
Epoch 4: loss=0.012388850307619083, BLEU=0.3108493882874471
Epoch 5: loss=0.010796328190850902, BLEU=0.31496691256370807
```

Wandb graphs

Figure 1 & 2 show the above numbers in two plots.

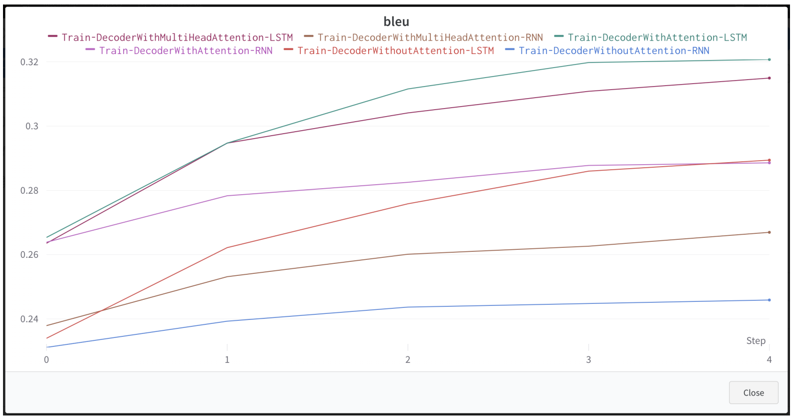

Figure 1: Wandb Training BLEU Score for Model without Attention

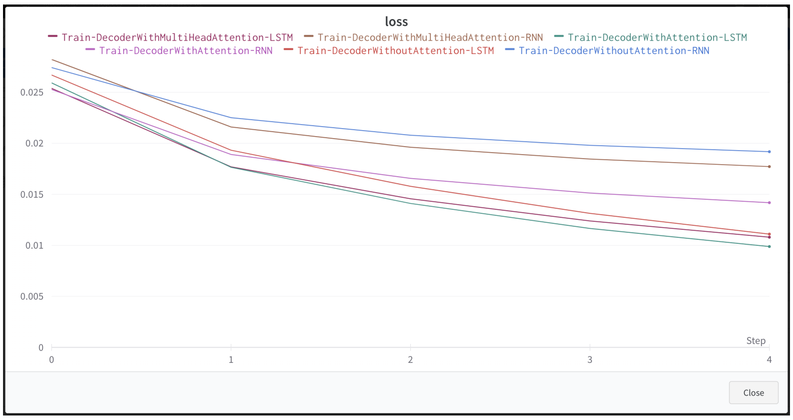

Figure 2: Wandb Training Loss for Model without Attention

1.2 Test Set BLEU Score

This section lists the test set BLEU score reported on the test set for each model in table 1.

Model	Test BLEU
Model without Attention - RNN	0.29519064029790
Model without Attention - LSTM	0.32843269208686
Model with Single-headed Attention - RNN	0.34238635601732
Model with Single-headed Attention - LSTM	0.36699024807924
Model with Multi-headed Attention - RNN	0.32403805053371
Model with Multi-headed Attention - LSTM	0.36585641167135

Table 1: 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?

The results show that the performances of all variants of models are not significantly different, with a difference of only 0.07 between the best and worst models on the test set. This may be due to several factors, such as a small training data set, training for only 5 epochs, and hyperparameters not being fine-tuned. However, the differences in the results are still meaningful and aligned with my expectations.

For all settings, models with LSTM cells perform better than those with RNN cells, as expected. This is because RNN cells may have difficulty learning long-term dependencies and are more prone to vanishing gradients, while LSTM cells address this issue by incorporating a memory mechanism and gating mechanism.

According to training BLEU scores, the model performances from worst to best are as follows:
1) RNN(without att)
2) RNN (multi-head)
3) RNN (with att)
4) LSTM (without att)
5) LSTM (multi-head)
6) LSTM (with att)

Based on the results, it is evident that the performance of the LSTM model on the training set is better than that of the RNN model, regardless of the attention mechanism used. This can be attributed to the flexibility of LSTM, which can handle the varying sentence lengths in the dataset. Another noteworthy finding is that both the RNN and LSTM models with attention mechanism perform better than the models without attention, while the performance of the models with multi-head attention is not as good as that of the models with simple attention. It can be result of two different scenario. First, this is likely because multi-head attention is computationally expensive and can cause overfitting, particularly when the dataset is small. The better performance of the models with simple attention suggests that the models with multi-head attention may be overfitting, although this cannot be confirmed without additional metrics. Second interpretation of the results can be that the models with multi-head attention require longer training time to reach their full potential to may eventually exceed the performance of the models with simple attention. Currently, the models with multi-head attention may be underfitting on this dataset. However, since the BLEU score of training is not significantly higher than the test (or even higher) in this study, it is more probable that it is underfitting.

The order of performances of BLEU scores on the test set is the same as on the training set, except that on the test set, "model without attention - LSTM" performs better than "model with attention - RNN" with a difference of about 0.014, which is not significant. This suggests that the flexibility of LSTM and its long and short-memory capabilities can be comparable to attention mechanisms. There is no absolute answer as to which is better, as it depends on the task and dataset.

2 Translation Analysis

2.1 Translations

2.1.1 Toronto est une ville du Canada.

’Toronto is a city in Canada.’: Google Translate
’miss deborah grey edmonton north’: Model without Attention - RNN
’canada s multicultural community’: Model without Attention - LSTM
’city is a community in canada’: Model with Attention - RNN
’city is a city of canada’: Model with Attention - LSTM

’a toronto star is a rural riding’: Model with Multi-head Attention - RNN
’toronto is a part of canada’: Model with Multi-head Attention - LSTM

2.1.2 Les professeurs devraient bien traiter les assistants d’enseignement.

’Professors should treat teaching assistants well.’: Google Translate
’the separatists will be matched’: Model without Attention - RNN
’the <unk> should be coordinated too’: Model without Attention - LSTM
’internship <unk> flap <unk>’: Model with Attention - RNN
’the uncertainties should be grounded’: Model with Attention - LSTM
’expanding opportunities are expanding the shots’: Model with Multi-head Attention - RNN
’they should be able to adjust the <unk>’: Model with Multi-head Attention - LSTM

2.1.3 Les etudiants de l’Universite de Toronto sont excellents.

’The students at the University of Toronto are excellent.’: Google Translate
’world class aerospace projects’: Model without Attention - RNN
’the toronto star made approximately million tonnes’: Model without Attention - LSTM
’the students of ontario hydro students are good’: Model with Attention - RNN
’the students of halifax are very high’: Model with Attention - LSTM
’the elevators are owned by bricks’: Model with Multi-head Attention - RNN
’the students of toronto are full of course’: Model with Multi-head Attention - LSTM

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?

Based on the translations produced by the models, it is evident that the quality of the translations varies significantly depending on the type of attention used (or not used) and the type of model architecture. The translations produced by the models without attention (both RNN and LSTM) are generally nonsensical and ungrammatical. On the other hand, the models with attention (both RNN and LSTM) produced better translations than the models without attention, but there are still some issues with grammar and word choice. The models with attention mechanisms (both RNN and LSTM, and both simple and multi-head) produced translations that are difficult to compare semantically. The performances of these models are very close, making it hard to determine the best model based on only these three provided examples.

There does appear to be a correlation between the model’s BLEU score and the quality of its translations. The model with the highest BLEU score generally produced the most accurate translations, while the model with the lowest BLEU score produced the least accurate translations. The BLEU scores of the LSTM-based models are consistently greater than those of the RNN-based models, indicating that the translations produced by LSTM models are better than those produced by RNN models. However, it is challenging to compare the performance of LSTM models with different attention mechanisms and determine which one is better.

Based on the translated sentences, the quality of the translation of second and third examples are not good enough and way worse than the first sentence. It can be because the models in this experiment were trained on a particular dataset, the Canadian Hansards, a specific and not general corpus, and we are evaluating their performance on general sentences. The first sentence is relatively short and more likely to be present in the Hansard corpus, whereas the other two sentences are longer and semantically less similar to the corpus and their embeddings are far from it. As a result, the translations produced for the second and third sentences were more random and less relevant compared to the first sentence and these can directly affect the quality of the translations.

Overall, these results suggest that attention mechanisms can be helpful for improving the accuracy and quality of machine translation systems and that LSTM cells improve translators compared to RNN networks. However, further examples from both the corpus and more general sequences, as well as experimentation with fine-tuning all the hyperparameters, are necessary to fully understand the strengths and limitations of different attention mechanisms and model architectures.