CSC401 Homework Assignment #2
Analysis

Chen Shen
Student number: 1009724924
UTORid: shench40
pete.shen@mail.utoronto.ca

1 Training Results

1.1 Training Loop Printout

The followings are diagrams from WandB. I also put the training loop printout at the end of this report.

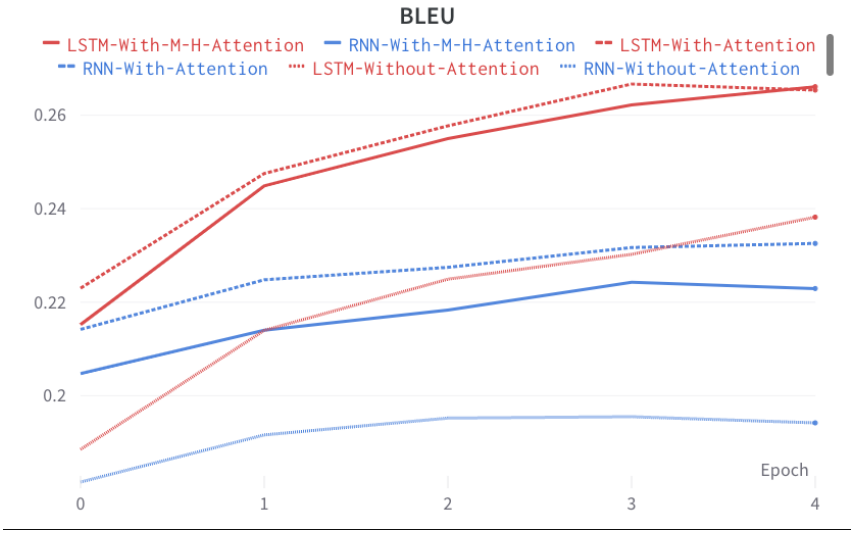

Figure 1: Wandlb Training BLEU Socre for Different Models

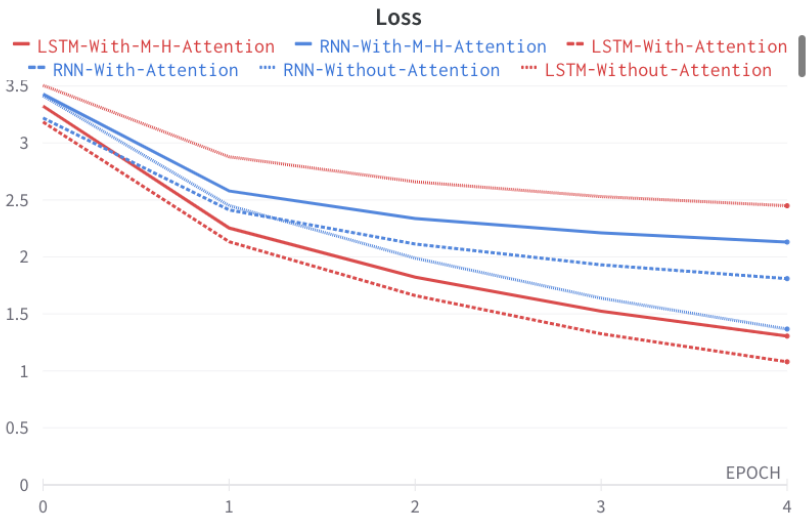

Figure 2: Wandlb Training Loss for Different Models

General Picture (1) Different typs of RNN models: blue curves represent RNN models, red curves represent LSTM models. (2) Different types of attentions: Dotted lines are without-attention models, dashed lines are with-single-attention models, solid lines are multi-head-attention models.

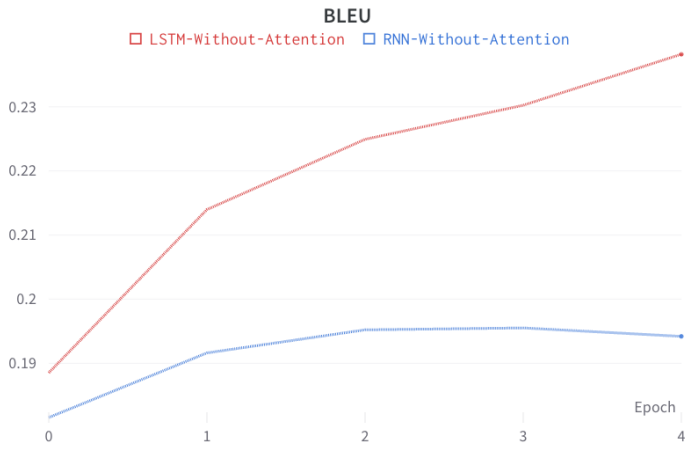

Figure 3: Wandlb Training BLEU Socre for **Models without Attention**

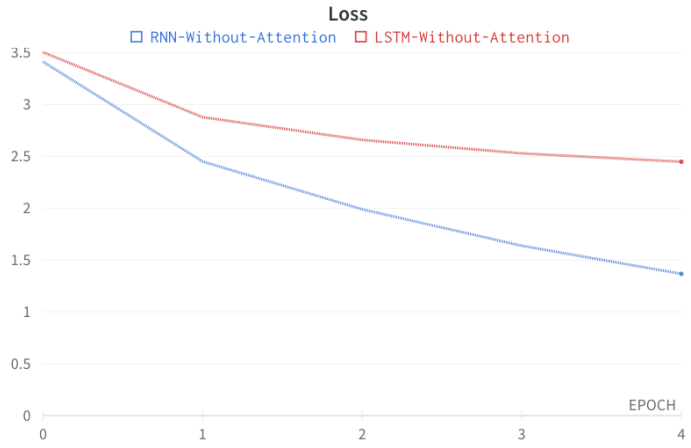

Figure 4: Wandlb Training Loss for **Models without Attention**

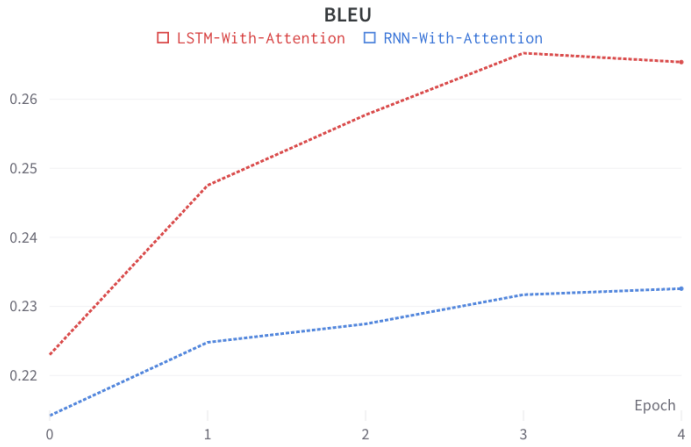

Figure 5: Wandlb Training BLEU Socre for **Models with Single-head Attention**

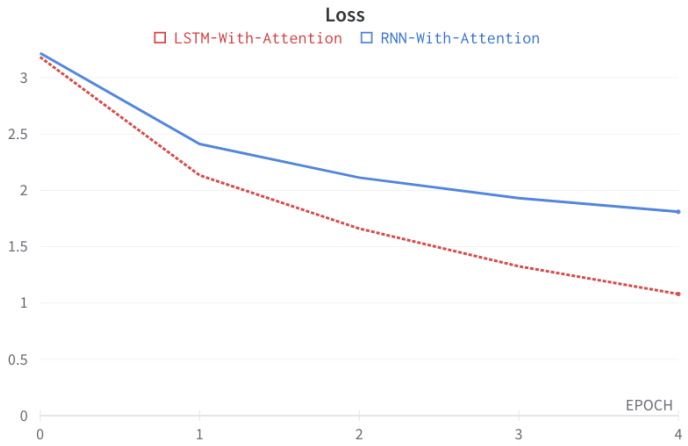

Figure 6: Wandlb Training Loss for **Models with Single-head Attention**

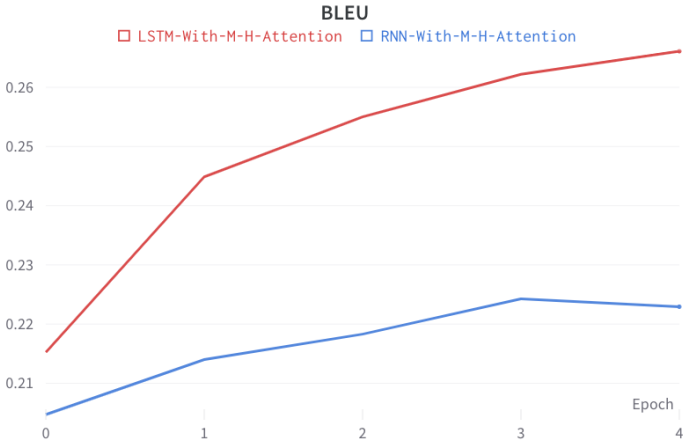

Figure 7: Wandlb Training BLEU Socre for **Models with Multi-head Attention**

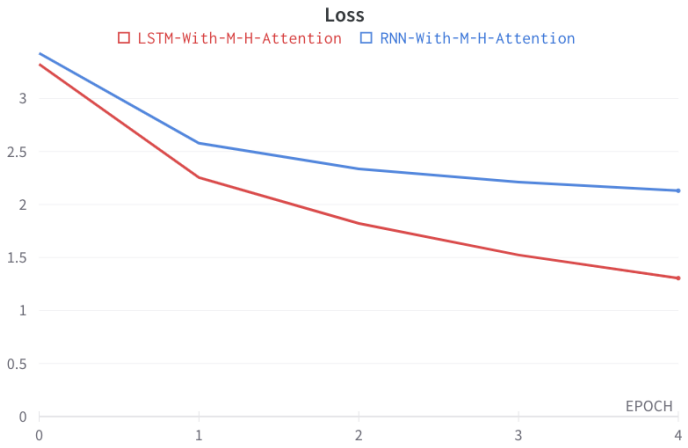

Figure 8: Wandlb Training Loss for **Models with Multi-head Attention**

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)	0.2538
Model without Attention (LSTM)	0.2772
Model with Single-headed Attention (RNN)	0.2954
Model with Single-headed Attention (LSTM)	0.3185
Model with Multi-headed Attention (RNN)	0.2871
Model with Multi-headed Attention (LSTM)	0.3209

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?

There is a discrepancy in between training and testing results: for all the models, the average BLEU score in the test set is higher than those in the training set. Explanations: There are dropout layers in the encoder during the training. Hence, the ability of encoder is reduced to guarantee its generalization ability. While they are removed during the testing, that is why all of those model perform better in testing set.

LSTM performs constantly better than RNN in all 3 attention modes: (1) [training set] BLEU is over 0.3 higher, loss is approximately 1 smaller. (2) [testing set] BLEU is 0.2-0.4 higher. Explanations: As a variant of RNN, LSTM is designed with 3 extra gates to control information flow better than simple RNN. Apart from that, it is also helpful for solving gradient vanishing so a model with LSTM converges faster.

Models with attention are always better than the ones without attention. However, it seems **multi-head attention not always beat single-head attention**: (1) LSTM with multi-head performs equal good as LSTM with single-head attention. (2) RNN with multi-head performs less good as RNN with single-head attention. Explanations: The assumption of using multi-head is 'different heads are able to focus on different information of data'. However, some questions about it would be how many heads are good enough? Why wouldn't different heads focus on the same thing as time goes up? Does initialization affect them in certain cases? This project only locked in a certain head number so it is inequitable to draw further conclusion. Furthermore, some previous research found 16 heads are better in practice.

2 Translation Analysis

2.1 Translations

List all of the translations in this section. (**bold part** means perfect alignment)

Correct translations by Google
»> model.translate("Toronto est une ville du Canada.")
'<s> Toronto is a city in Canada </s>'
»> model.translate("Les professeurs devraient bien traiter les assistants d'enseignement.")
'<s> Professors should treat teaching assistants well </s>'
»> model.translate("Les etudiants de l'Universite de Toronto sont excellents.")
'<s> University of Toronto students are excellent </s>'

RNN - WITHOUT ATTENTION
'<s> royal **canadian** mounted police </s>'
'<s> the hon member for saint bruno saint hubert </s>'
'<s> judges salaries for healing </s>'

LSTM - WITHOUT ATTENTION
'<s> correctional service **canada** is in </s>'
'<s> they should be grounded </s>'
'<s> the residents of guelph wellington has done a great deal </s>'

RNN – WITH SINGLE-HEAD ATTENTION
'<s> **toronto is a canadian city** </s>'
'<s> iraqis should be <unk> </s>'
'<s> larry many of the **students are excellent** </s> </s></s>'

LSTM – WITH SINGE-HEAD ATTENTION
'<s> ottawa is a **city of canada** </s>'
'<s> the serbs should remember the rail lines </s>'
'<s> the **university of toronto** made some very high tech experience </s>'

RNN – WITH MULTI-HEAD ATTENTION
'<s> it is a town **of canada** </s>'
'<s> they should be pressured </s>'
'<s> the **students** are tired of the great fortune </s>'

LSTM – WITH MULTI-HEAD ATTENTION
'<s> **toronto** is a town of **canada** </s>'
'<s> it should be gavel to gavel filming </s>'
'<s> the **students of toronto** are very few </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?

The most intuitive finding is most translations are super bad compared to the mature models that are used in the commercial. I bolded the parts that are related to the meaning of the original sentences. It can be seen that with a higher BLEU score (which are those models with attention mechanism), the quality of the translation is higher. In contrast, the models without attention merely translate nothing relevant.

3 Appendix: training loop digits

3.1 RNN - without attention

Epoch 1: loss = 3.50543, BLEU = 0.18153
Epoch 2: loss = 2.87853, BLEU = 0.19161
Epoch 3: loss = 2.65925, BLEU = 0.19521
Epoch 4: loss = 2.52932, BLEU = 0.19551
Epoch 5: loss = 2.44883, BLEU = 0.19419
The average BLEU score over the test set was 0.25378

3.2 LSTM - without attention

Epoch 1: loss = 3.41309, BLEU = 0.18851
Epoch 2: loss = 2.45068, BLEU = 0.21397
Epoch 3: loss = 1.98928, BLEU = 0.22493
Epoch 4: loss = 1.63796, BLEU = 0.23024
Epoch 5: loss = 1.36694, BLEU = 0.23820
The average BLEU score over the test set was 0.27716

3.3 RNN - with single attention

Epoch 1: loss = 3.21887, BLEU = 0.21418
Epoch 2: loss = 2.41207, BLEU = 0.22479
Epoch 3: loss = 2.11314, BLEU = 0.22746
Epoch 4: loss = 1.93065, BLEU = 0.23169
Epoch 5: loss = 1.80937, BLEU = 0.23257
The average BLEU score over the test set was 0.29547

3.4 LSTM - with single attention

Epoch 1: loss = 3.18375, BLEU = 0.22300
Epoch 2: loss = 2.13335, BLEU = 0.24752
Epoch 3: loss = 1.66069, BLEU = 0.25772
Epoch 4: loss = 1.32566, BLEU = 0.26668
Epoch 5: loss = 1.07976, BLEU = 0.26537
The average BLEU score over the test set was 0.31850

3.5 RNN - with multi-head attention

Epoch 1: loss = 3.42743, BLEU = 0.20471
Epoch 2: loss = 2.57844, BLEU = 0.21400
Epoch 3: loss = 2.33682, BLEU = 0.21830
Epoch 4: loss = 2.21083, BLEU = 0.22426
Epoch 5: loss = 2.13049, BLEU = 0.22293
The average BLEU score over the test set was 0.28717

3.6 LSTM - with multi-head attention

Epoch 1: loss = 3.32441, BLEU = 0.21522
Epoch 2: loss = 2.25362, BLEU = 0.24488
Epoch 3: loss = 1.82188, BLEU = 0.25501
Epoch 4: loss = 1.52379, BLEU = 0.26221
Epoch 5: loss = 1.30520, BLEU = 0.26610
The average BLEU score over the test set was 0.32099