Homework 4

Kevin Lee
2995359247

For all the tasks, I have added in <s> and </s> in order to mark the start and end of the sentence. Not only that, but I have added in <p> to ensure that each batch size was correct. I've also utilized multiple computing systems to expedite the tuning process.

I have also separated out the tasks to three different files (batch, batch2, batch3) so that I can keep track of the models and Run All.

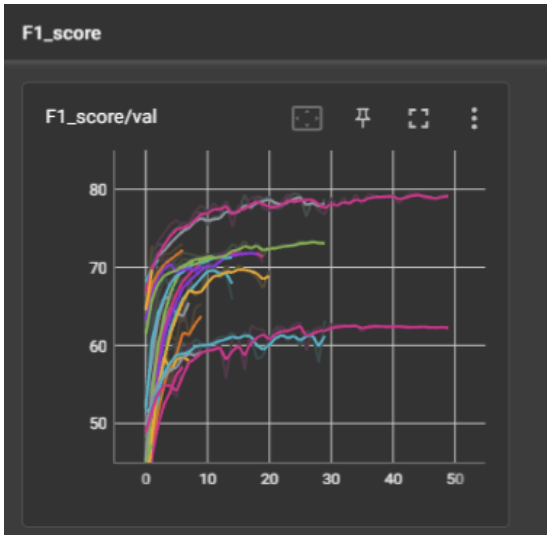
Task 1

In task 1, some of the ways I tried hyperparameters used was first trying to see if lowercasing all the words had an effect. This didn't have too big of an impact, but in the end, I've used the uppercase as that was useful in Task 2. Some of the hyperparameters that I've used to tune the model to get to the accuracy, precision, recall and F1 score posted at the bottom was by messing around with the batch size of the data, SGD's learning rate, weight decay, momentum, early stopping, CosineAnnealing/StepLR/ReduceLROnPlateau, and more. The scheduler also required hyperparamter tuning to, as I had to play with the patience to achieve the best score.

```
C:\Users\jaehw\OneDrive\Desktop\CSCI544\hw4>perl conll03eval.txt < pred.txt
processed 51577 tokens with 5942 phrases; found: 5413 phrases; correct: 4557.
accuracy:  95.99%; precision:  84.19%; recall:  76.69%; FB1:  80.26
              LOC: precision:  87.14%; recall:  87.04%; FB1:  87.09  1835
             MISC: precision:  84.67%; recall:  74.30%; FB1:  79.15  809
              ORG: precision:  79.75%; recall:  70.17%; FB1:  74.65  1180
              PER: precision:  83.83%; recall:  72.31%; FB1:  77.65  1589
```

To get this, run perl conll03eval.txt < dev1.out

I've also utilized TensorBoard to see how the parameters had affected my dev's F1 Score.



Task 2

Word2Idx, unlike Task 1, was all lowercase. But this meant that the uppercase or lowercase feature wasn't accounted for, decreasing the mode's performance. To ensure that we capture this feature, I added a feature that was 1 if the first letter of the word was uppercase, and 0 otherwise. This improved the performance, allowing me to achieve the precision, recall and F1 score shown at the bottom, and I played around with the hyperparameters similar to the ones in Task 1.

```
C:\Users\jaehw\OneDrive\Desktop\CSCI544\hw4>perl conll03eval.txt < pred.txt
processed 51577 tokens with 5942 phrases; found: 6002 phrases; correct: 5262.
accuracy:  97.96%; precision:  87.67%; recall:  88.56%; FB1:  88.11
              LOC: precision:  93.22%; recall:  92.00%; FB1:  92.60  1813
             MISC: precision:  90.76%; recall:  79.93%; FB1:  85.01  812
              ORG: precision:  76.32%; recall:  84.86%; FB1:  80.37  1491
              PER: precision:  89.98%; recall:  92.13%; FB1:  91.04  1886
```

To get this, run perl conll03eval.txt < dev2.out

Extra Credit

I utilized Task 2's BiLSTM in order to add the CNN feature. I looked over 30 characters and fed that into the BiLSTM _CNN, to have performance gain shown at the bottom.

```
processed 51577 tokens with 5942 phrases; found: 6108 phrases; correct: 5300.
accuracy:  97.94%; precision:  86.77%; recall:  89.20%; FB1:  87.97
              LOC: precision:  90.45%; recall:  94.34%; FB1:  92.35  1916
             MISC: precision:  83.52%; recall:  82.97%; FB1:  83.24  916
              ORG: precision:  81.49%; recall:  83.37%; FB1:  82.42  1372
              PER: precision:  88.45%; recall:  91.42%; FB1:  89.91  1904
```

To get this, run perl conll03eval.txt < dev3.out