

CSCI 544 HW 4 Report

Reading the data

The train, dev, and test data were taken and put into a dataframe. All None rows (empty lines from the files) were removed.

Converting the words to indices

Created a vocabulary of words and tags by extracting the unique words from the training dataset. Then, assigned each word and its corresponding tag to their indices from the vocabulary. Finally, created a new dataset where each row in the dataframe contains a sentence as a list of words, their corresponding word indices, and tag indices.

Handling the uneven length of sentences

The largest sentence length from the given batch was used, and padding was added as zero for the words and -1 to the labels for the length.

Dataloader

Created a data loader with the batch size of 32. The size of the each batch passing through the model is $32 * \text{Max length of the sentences in that batch}$

Task 1

The following is the architecture of the model with the layers as mentioned in the assignment

```
BLSTM(  
    (embedding): Embedding(23700, 100)  
    (lstm): LSTM(100, 256, dropout=0.33, bidirectional=True)  
    (linear): Linear(in_features=512, out_features=128, bias=True)  
    (elu): ELU(alpha=1.0)  
    (classifier): Linear(in_features=128, out_features=10, bias=True)  
)
```

- Optimizer used is the SGD with learning rate as 0.8 i.e., updating the model weights with larger steps. Assigned a scheduler to reduce the learning rate by 0.01 after every 20 steps
- Trained the model for 100 epochs

Handling the class imbalance

Assigned the class weights to have less weight for the most frequent class, while passing to the cross entropy loss.

Handling padded values

Didn't apply the cross entropy loss for the padded integers(-1) by ignoring index -1 in the loss.

Hyperparameters

```
BATCH_SIZE = 32
NUM_EPOCHS = 100
lr = 0.8
```

Results

```
processed 51578 tokens with 5942 phrases; found: 5637 phrases; correct: 3772.
accuracy: 93.65%; precision: 66.92%; recall: 63.48%; FB1: 65.15
      LOC: precision: 82.02%; recall: 77.95%; FB1: 79.93 1746
      MISC: precision: 69.36%; recall: 68.00%; FB1: 68.67 904
      ORG: precision: 57.07%; recall: 60.78%; FB1: 58.87 1428
      PER: precision: 57.60%; recall: 48.75%; FB1: 52.81 1559
```

Task 2

```
BLSTM_Glove(
  (embedding): Embedding(400000, 100)
  (lstm): LSTM(100, 256, dropout=0.33, bidirectional=True)
  (linear): Linear(in_features=512, out_features=128, bias=True)
  (elu): ELU(alpha=1.0)
  (classifier): Linear(in_features=128, out_features=9, bias=True)
)
```

Handling the case sensitivity

Added a flag as 1 if the entire word is in the capitals while passing to the model. Rest all the steps related to data loading are similar as the task 1.

Results

```
processed 51578 tokens with 5942 phrases; found: 5661 phrases; correct: 3686.  
accuracy: 93.31%; precision: 65.11%; recall: 62.03%; FB1: 63.54  
    LOC: precision: 80.24%; recall: 78.93%; FB1: 79.58 1807  
    MISC: precision: 65.61%; recall: 67.25%; FB1: 66.42 945  
    ORG: precision: 54.99%; recall: 56.67%; FB1: 55.82 1382  
    PER: precision: 56.06%; recall: 46.47%; FB1: 50.82 1527
```