Assignment 2 Report

Kushal Kumar Jain
2019111001

October 1, 2022

1 Theory

1.1 How does ELMo differ from CoVe? Discuss and differentiate both the strategies used to obtain the contextualized representations with equations and illustrations as necessary.

There are 2 major differences between ELMo and CoVe :
CoVe trains on the supervised machine translation dataset while ELMo trains on the unsupervised language model data.
CoVe treats the output of the last layer machine translation LSTM as the context vectors while ELMo is a weighted combination of all layers in the language model LSTM.

1.2 The architecture described in the ELMo paper includes a character convolutional layer at its base. Find out more on this, and describe this layer. Why is it used? Is there any alternative to this?

Character-level tokens goes through convolutional layers with different kernel sizes. The original "small" ELMo model uses kernels of size 1, 2, 3, 4, 5, 6, 7 with 32, 32, 64, 128, 256, 512, 1024 channels, respectively. Outputs from each convolutional layers are then max-pooled and concatenated to yield $32 + 32 + 64 + 128 + 256 + 512 + 1024 = 2048$ -length vector. This concatenated vector can be used as a word embedding. Since convolutional layer is well known for its feature-extracting property, this can be regarded as a character-level context extraction process.

2 Training ELMo For Language Modelling

2.1 Hyperparameters

I have kept the sequence length fixed at 20, so that it is easier to load into dataloader and make batches. Learning rate was 0.003 and it turned out to be fine.
Here I have used the Cross Entropy loss (pytorch inbuilt), along with Adam optimizer. The whole network converged in 50 epochs.

2.2 Graphs

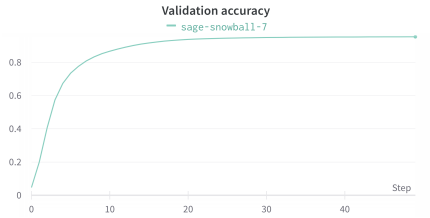

Figure 1: Validation Accuracy

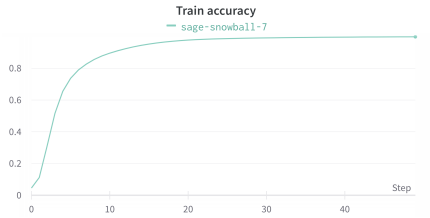

Figure 2: Training Accuracy

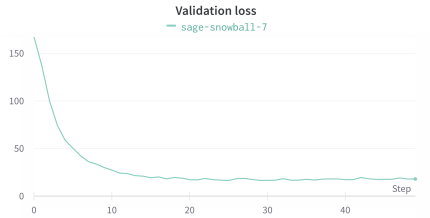

Figure 3: Validation Loss

Figure 4: Training Loss

2.3 Metrics

Training Loss for epoch 50 : 13.950760135427117
Training Accuracy for epoch 50 : 99.6690034866333 %
Val Loss for epoch 50 : 23.64764827489853
Val Accuracy for epoch 50 : 95.12194991111755 %

Test Accuracy : 94 %

3 Training The Classifier

3.1 Architecture

I used the weighted sum of bi-lstm outputs as input to a unidirectional LSTM cell in model2. After which I applied a linear layer to narrow down the output to 50 and another linear layer to further classify the sentence in 5 categories.

3.2 Graphs

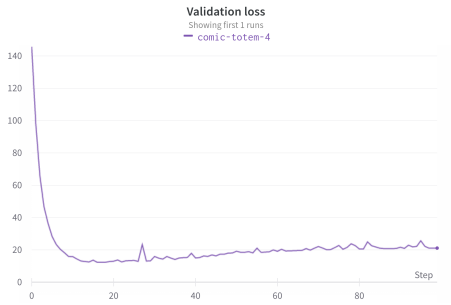

Figure 5: Validation Accuracy

Figure 6: Training Accuracy

3.3 Metrics

Training Loss for epoch 50 : 165.950760135427117
Training Accuracy for epoch 50 : 79.6690034866333 %
Val Loss for epoch 50 : 343.64764827489853
Val Accuracy for epoch 50 : 66.12194991111755 %

References

[1] Weights and biases.