להוסיף c1 את הערך

C2 את הגרפים

D דוגמאות:

GRU

(c)

1) As shown directly from the maximum entropy value possible is

(2)

Average loss

The loss decreases gradually, eventually falls to zero (in comparison to the loss value before the training), the loss graph corresponds to both the logit histogram and the entropy graph that follows.

Average entropy

As thought in Intro to ML, the entropy is highest for the uniform distribution. As specified in the exercise, the initial weights were set to the uniform distribution, thus indeed we have the highest entropy there. It is also significant from the graph, how as the training goes on, the entropy decreases, which means the model is more certain of it’s predictions.

Logits histogram

It can be seen that the logits histogram corresponds to both the loss and entropy graphs. We know that the logit function equals to (when ) zero for the uniform distribution, and is varies for different distributions (exactly as the entropy is highest for uniform and lower for other distributions).

From the logit histogram we see that as the model reaches it’s final weights values, the logit range becomes “wider”, which means it’s certainty becomes higher, which makes the entropy lower.

(d)

By evaluating the model on the dev set, we can conclude the strong and weak parts of the model.  
While the model preforms good in evaluating the PER ORG and MISC labels, when evaluating the LOC label we have 9.7% error rate, 7.7% of them are while guessing the ORG label.  
While that might sound very high, we can see that while labeling right tokens, the certainty is above 97%, in labeling wrongly LOC as ORG, the certainty is around 60%, which means the model “feels” it predicts wrong. Moreover, labeling a LOC as a ORG is indeed a “hard” task, which can be hard for humans as well (As discussed in class).

Examples for that wrong classification and probability values can be shown in this examples from the test set.