תרגיל 3 - NLP

25 בדצמבר 2022

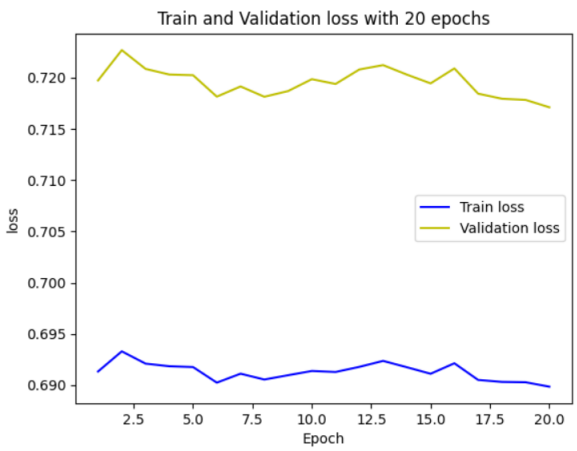
שגיאה ודיוק עבור *LogLinear – One hot average*

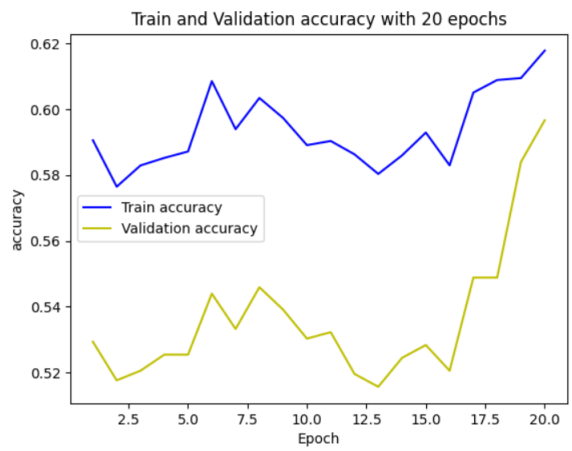
(א) שגיאה

For the below question we trained on 20 epochs with a learning rate of 0.01. Note that although the hyper-parameters stayed fixed, we received very different learning curves for the same training task. The random initialization of the weights may be a good explanantion for that since we are tying to converge to a local minima. (what do you think?).

We found it intriguing that the train and validation curves are quite offen very symmetric. (Although had some runs where this didnt happen).

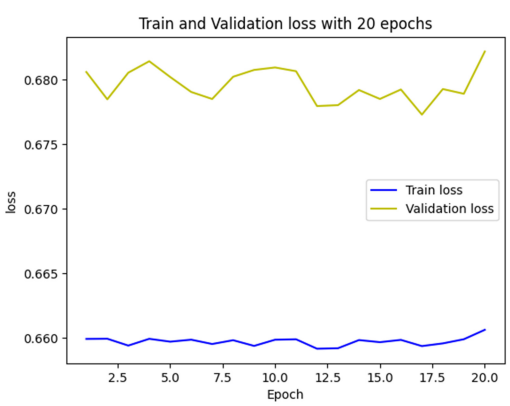
At first we used the predict method (in evaluate function) of the model in the training and validation:

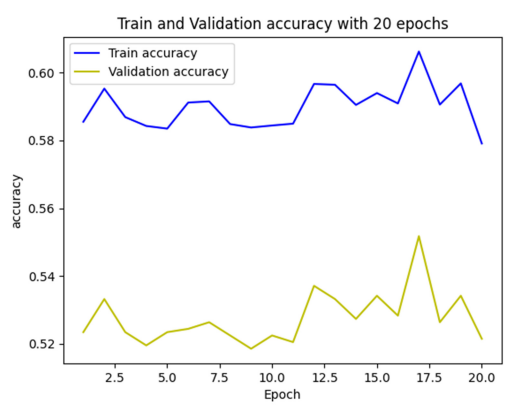




Using the predict method is wrong since the sigmoid layer is used twice (once in the predict method and once in the criterion loss BCEwithlogits).

Here are the results using the Forward method (in evaluate function) of the model in the training and validation. (which we will use in the subsequent questions)





test loss: 0.6746793450656696

test accuracy: 0.5595703125

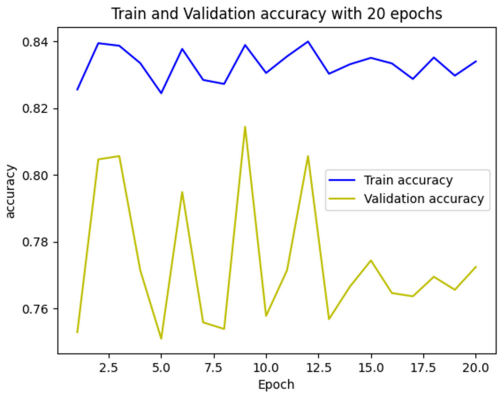
rare loss 0.7269768193364143

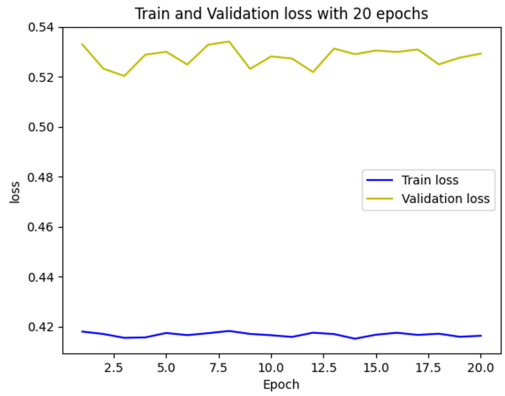
rare accuracy 0.3

polar loss 0.6962529244201798

polar accuracy 0.4838709677419355

שגיאה ודיוק עבור $LogLinear - w2v\ average$





test loss: 0.48748051115308044

test accuracy: 0.81640625

rare loss 0.6544381907209754

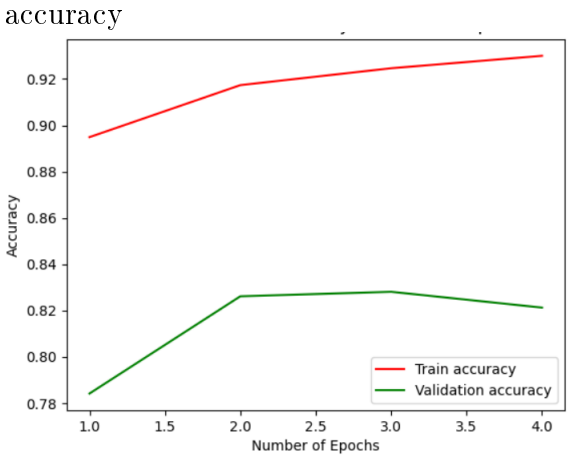
rare accuracy 0.66

polar loss 0.7204013147781934

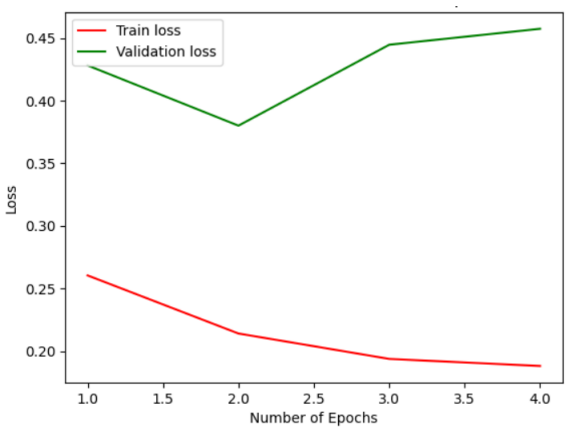
polar accuracy 0.4838709677419355

שגיאה ודיוק עבור $LogLinear - LSTM$

accuracy



loss



test loss: 0.326895251

test accuracy: 0.87250121

rare loss 0.521692719

rare accuracy 0.83421961

polar loss 0.9404013147781934

polar accuracy 0.54838709677419355

השוואת המודלים:

1. Let us note the w2v model outperformes the one_hot model. The words embeddings seems to be a more powerful feature than the one hot representation. This is expected as the Word2Vec embeddings are pre-trained.

2. The model comprising the LSTM units outperformes all the previous models. This is consistent with the literature.

Note that we did not have the problem of exploding/vanishing gradients as we trained with only 20 epochs. (However if we had this problem the LSTM model would have dealt with it)

The model better integrates the idea of a separation between the long-term-memory and the short memory. As a result, it is able to learn long term dependencies while still giving weight to the short-term-memory (hidden_states at each input to the unit). This is done through the use of gated units.

The result is a richer model that processes sequential information (in both direction in our case) allowing an overall better predictive performance. Note that we do not average the word vectors embeddings in the LSTM model as we do with the two other models. Thus, contextual information is better preserved.

3.

Negated sentences (sentences with double sentiment):

we get that the W2V model outperformed the one_hot model. This is consistent with the test/eval set. Note that both models have similar accuracy ranges as they both take an average of the vector embeddings to represent a sentence, restricting them quite badly.

The LSTM model performed the best. We surmise that processing sequential information bidirectionally as well as the overall better integration of the long memory is the reason for that. These sentences include two proses that we think need to be analysed and processed independently before assessing which is the one entailing meaning or sentiment. Through the use of forget, input, update gates for each word, the LSTM model is able to do just that.

Rare words:

We would expect both the W2V model and the one-hot model to perform similarly on unseen or rare words. However, we undstood already from above that the log linear model is richer when using w2v embeddings. In that sense, we learned a better approximation of the target distribution of sentiment in sentences. We therefore expect an overall better predicive power regardless of whether the word is rare or not.

The LSTM performance here is the highest. Moreover the performance on the rare words dataset is very close to the performanc on the test/eval set. This is consistent with the fact that this model processes sentences as a sequence (backward ands forward) leaving alot of the prediction of the rare word to its surrounding context.