Sentiment analysis, IMDB dataset

Baseline model

Preprocessing steps

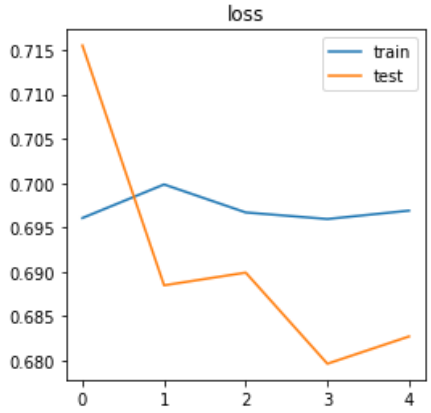
1. Loading data (using PyTorch DataSet module)
2. Tokenization (*spacy* tokenizer)
3. Removing stop-words
4. Building vocabulary from train data (size = 25000)

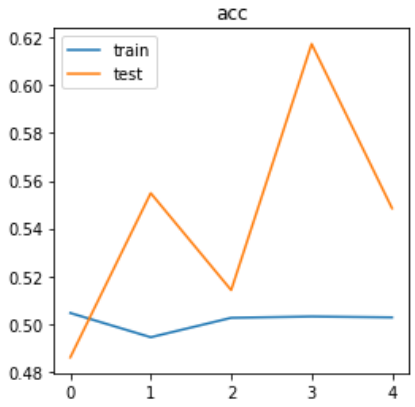
Model configuration

Optimizer – Adam, learn rate = 1e-3, loss function – binary cross entropy with logits

1. Embedding (random)
2. RNN
3. Linear layer

Results (5 epochs): train loss 0.70, train acc 0.50, test loss 0.68, test acc 0.55





Common problem – vanishing gradient

LSTM (Long Short-Term Memory) model

Model configuration

1. Embedding – pre-trained “glove.6B.100d”
2. LSTM
3. Dropout (0.5)
4. Linear layer

Results (5 epochs): train loss 0.24, train acc 0.91, test loss 0.26, test acc 0.90