LTAT.01.001 Homework 6

Elizaveta Korotkova

Step 1

Validation accuracies are the following:

- Baseline: **0.3824**
- ELMo: **0.4414**
- BERT: **0.5023**

I will choose BERT as the baseline, because it shows the best accuracy (6% higher than ELMo). However, it is the slowest model to train.

BERT baseline config: `bert_baseline_config.jsonnet`

Step 2

The baseline BERT model showed validation accuracy 0.5023 and training accuracy 0.6007 at its best checkpoint. Loss progress shown in Figure 1.

1. First, I tried changing the hidden dimension of the LSTM. Originally it was 128, I tried also:

- `hidden_size = 64`: validation accuracy 0.4759 (the 2nd epoch was the best)
- `hidden_size = 32`: validation accuracy 0.4659 (2nd epoch also the best, accuracy-wise)

I was not able to get better validation metrics by making the hidden dimension smaller. After the second epoch, models keep overfitting (see Figures 2 and 3). Loss by the end of training is lower than with baseline, but that does not help with overfitting.

2. I also tried changing the batch size (originally 32). I made it bigger, which could lead to more stable training:

- `batch_size = 64`: validation accuracy 0.4696 (best validation loss on the 3rd epoch, best validation accuracy on the 5th; also starts overfitting after the first epochs, see Figure 4)

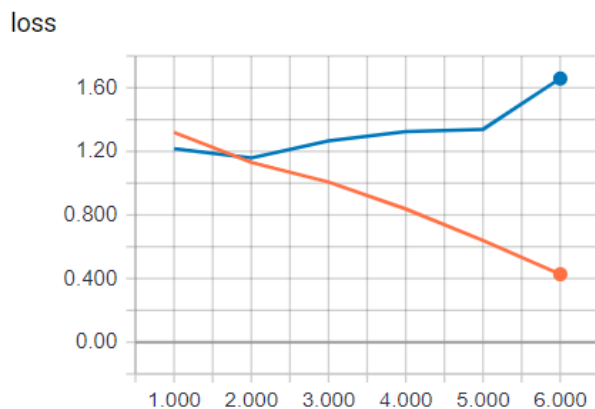3. As a next step, I simultaneously made batch size and hidden dimension smaller.

Figure 1: BERT baseline model, training and validation loss. Training loss shown in orange, validation blue
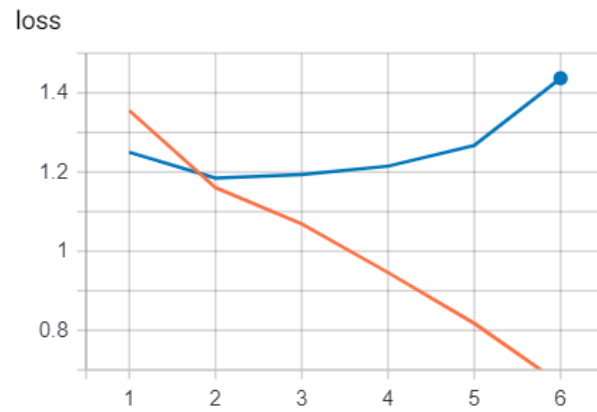
1

Figure 2: `hidden_size = 64`, training and validation loss. Training loss shown in orange, validation blue
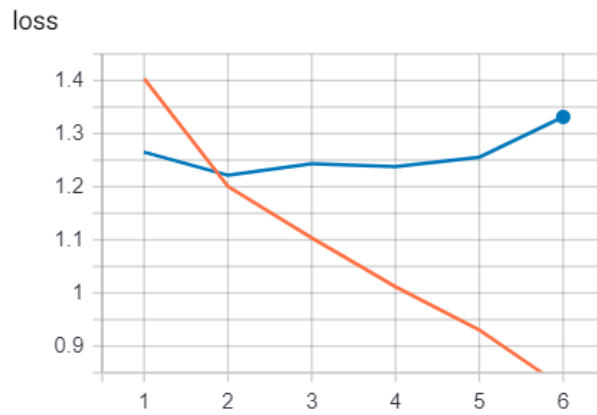

Figure 3: `hidden_size = 32`, training and validation loss. Training loss shown in orange, validation blue
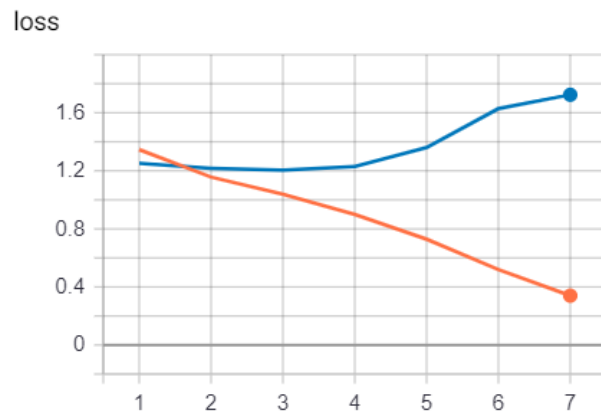

Figure 4: `batch_size = 64`, training and validation loss. Training loss shown in orange, validation blue
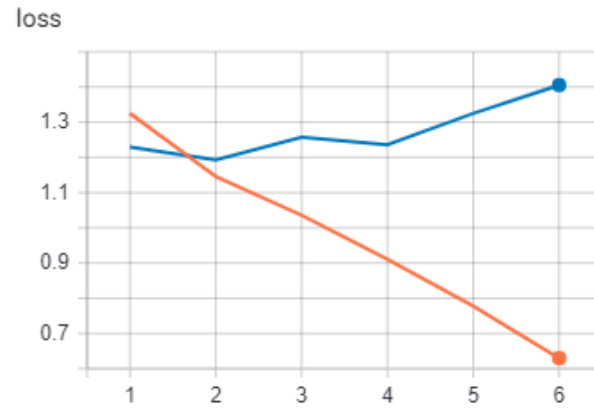
Figure 5: `batch_size = 16, hidden_size = 64`, training and validation loss. Training loss shown in orange, validation blue

- `batch_size = 16, hidden_size = 64`: validation accuracy 0.4759 (see loss progress in Figure 5)

So far, none of the hyperparameter tweaks I tried yielded better validation accuracy than the baseline.