LTAT.01.001 Homework 5

Elizaveta Korotkova

0. +
1. +
2. +
3. See Figures 1-6.
4. I used `evaluate` on the trained models, the results are in question 7. A sample output of `evaluate` would be (only the last lines which contain metrics):

```
2019-04-25 00:06:50,122 - INFO - allennlp.commands.evaluate - Finished evaluating.

2019-04-25 00:06:50,124 - INFO - allennlp.commands.evaluate - Metrics:

2019-04-25 00:06:50,125 - INFO - allennlp.commands.evaluate - accuracy: 0.42036199095022625

2019-04-25 00:06:50,131 - INFO - allennlp.commands.evaluate - loss: 1.3652190634182522
```

5. Config file is in `bert_config.jsonnet`.
6. Trained the models.
7. Evaluated on test data. Baseline accuracy 0.4204, ELMo 0.4336, BERT 0.5014.
8. The result for the three short reviews was the following:

```
input:  {"sentence": "An incredibly well filmed movie!"}

prediction:  {"logits": [1.8116538524627686, -2.4207756519317627, -0.7414960265159607,
3.5211236476898193, -3.149897813796997]}

input:  {"sentence": "Terrible film, did not like at all."}

prediction:  {"logits": [-1.2918798923492432, 2.0291922092437744, 0.8780262470245361,
-3.0022523403167725, 1.4492517709732056]}

input:  {"sentence": "Loved the acting and the beautiful Scottish scenery."}

prediction:  {"logits": [2.6428189277648926, -2.372917652130127, -0.6464166641235352,
3.351029396057129, -4.014708518981934]}
```

We can see that the model understands at least generally what is positive and what is negative.

9. The best performing mode was BERT. The baseline model's best epoch was the second one (epoch 1 by AllenNLP numbering – starting from 0), after that it started overfitting very badly (at the 10th epoch, training accuracy was 0.98 and validation accuracy 0.35). The ELMo model's best epoch was the 6th. The BERT model's best was, again, the second, but it does not overfit as much as the baseline. I would assume that the speed of convergence depends in part on model regularization.
10. See curves in Figures 1-6. We can see that the baseline and BERT models' validation accuracies do not grow after the 2nd epoch, the models start to overfit (BERT not as badly as baseline). ELMo's validation accuracy fluctuates longer. BERT's training loss is eventually lower than ELMo's, while the baseline's loss shows a typical curve, but overfits too much by the end of training.
11. I completed all of the steps. Training the ELMo and BERT models took quite a while. For me, dealing with AllenNLP is quite hard in general; I see why it can be helpful and convenient once one gets used to it, but many things are not entirely clear to me for now.

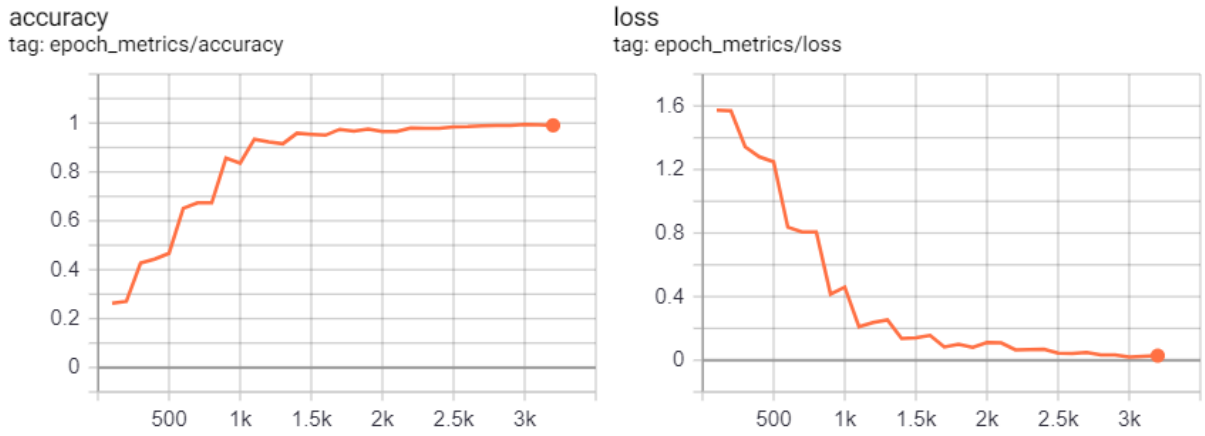

Figure 1: Baseline model, training metrics

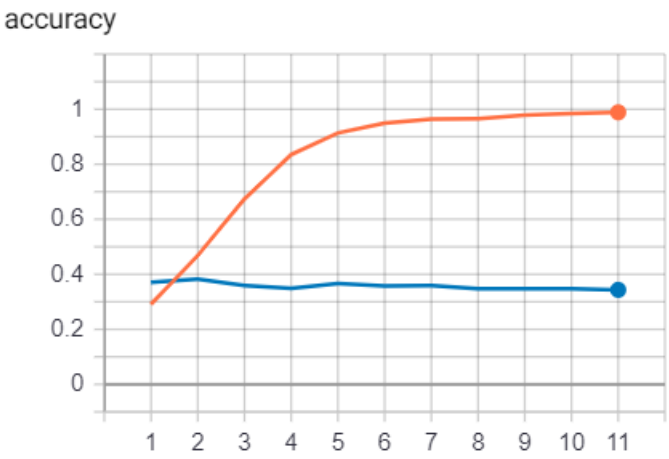

Figure 2: Baseline model, training (orange) and validation (blue) accuracy

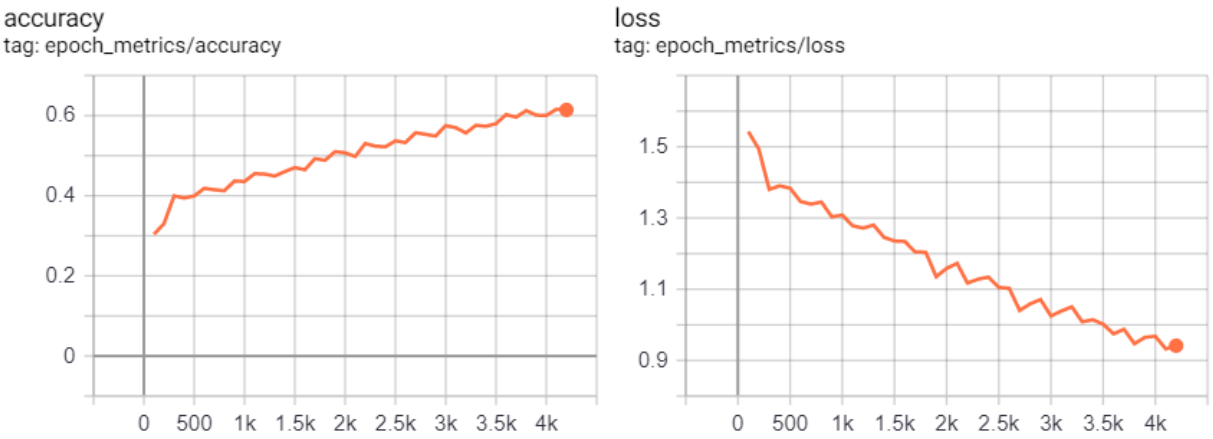

Figure 3: ELMo model, training metrics

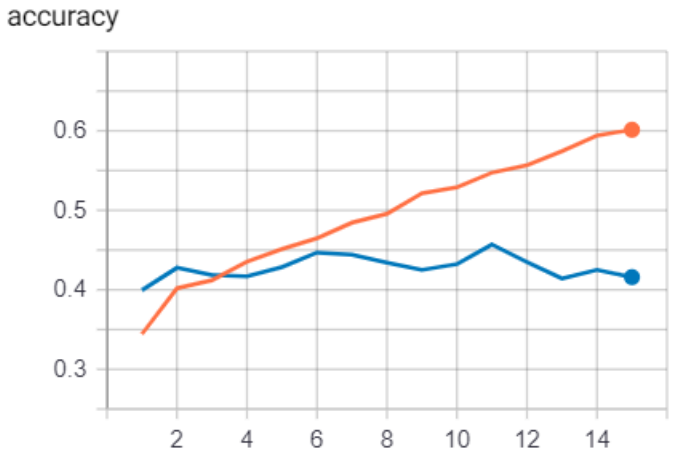

Figure 4: ELMo model, training (orange) and validation (blue) accuracy

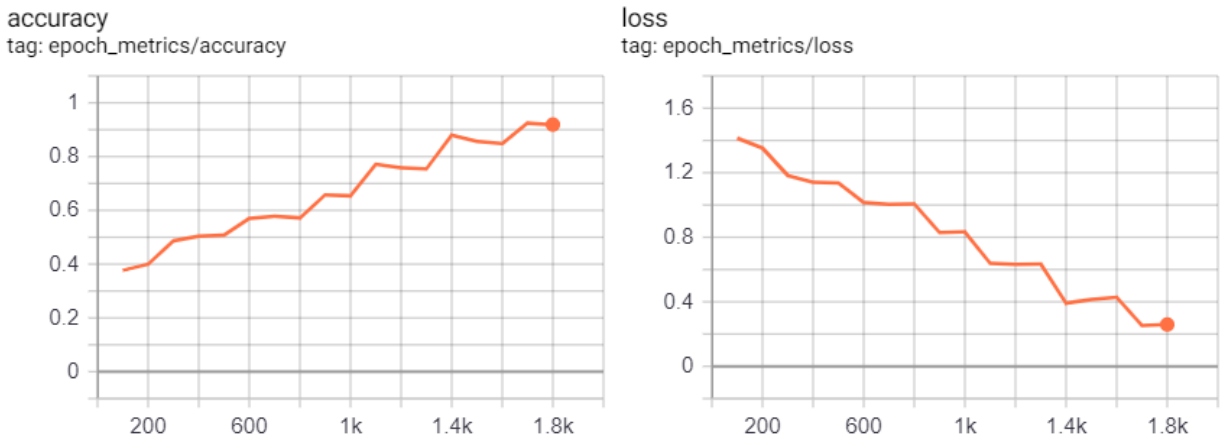

Figure 5: BERT model, training metrics

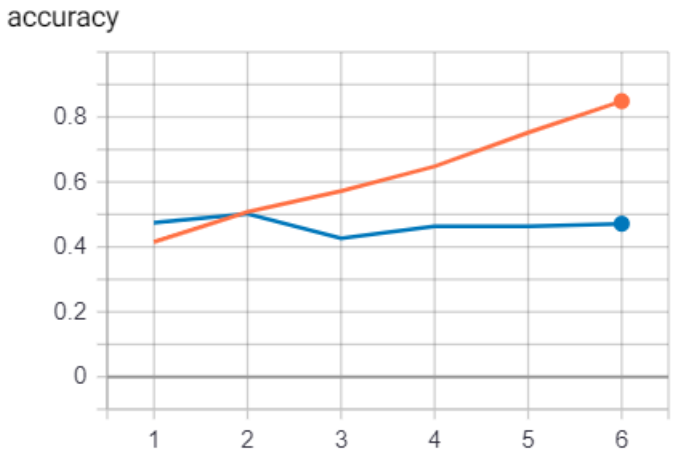

Figure 6: BERT model, training (orange) and validation (blue) accuracy