Ofer Sabo
201511110

Part 2

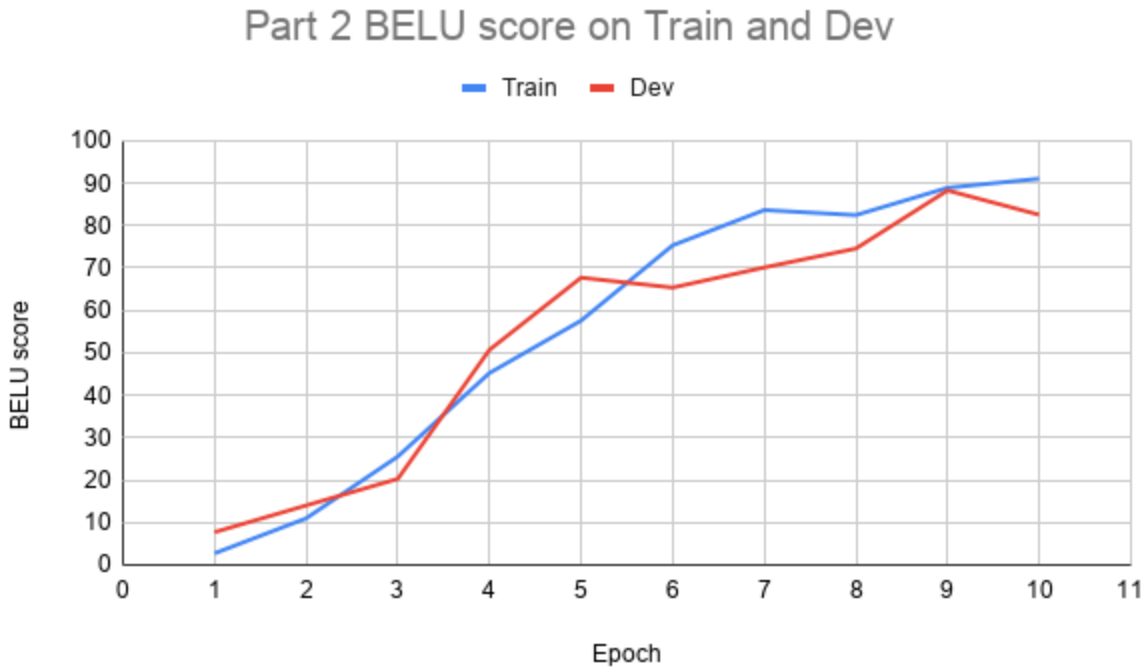
I used the Optuna package to find the hyperparameters which produce the best BELU score. As a result, these values are not “round”.

Hyperparameters:
"decoder_type": "gru",
"drop_out_rate": 0.054145200158513054,
“hidden_size": 250,
“embedding_dim": 50,
Learning rate: 0.0007355253529183.
Optimizer: adam.

Description of the model architecture:
Encoder - single direction and single-layer LSTM, with size hidden.
Decoder - single direction and single-layer GRU, with size hidden.
Source embedding: size “embedding_dim”.
Target embedding: size “embedding_dim”.
Max decoding steps is 100.
Also I clip the gradient into a fixed value of 1.

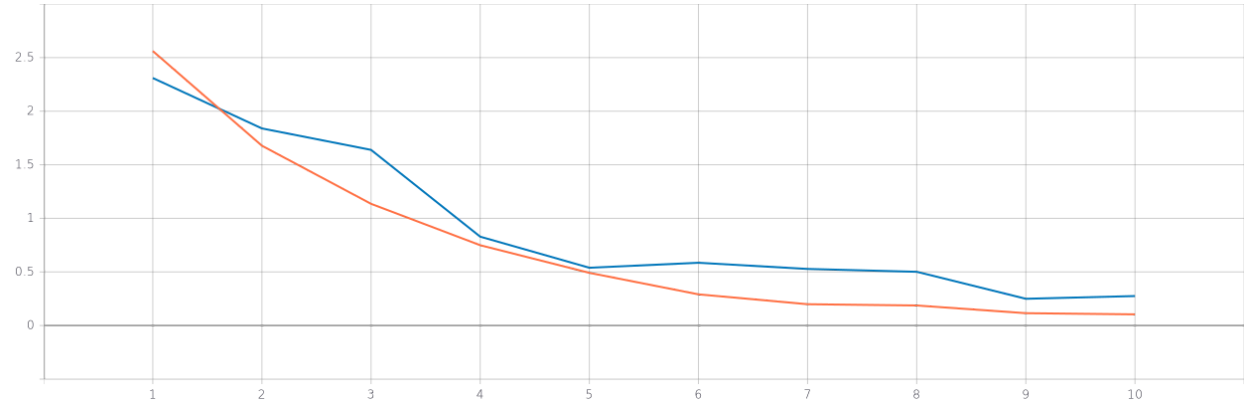
The data for the attention analysis is dumped into: “attention_data.jsonl”
Decoder attention function: BilinearAttention.

Ofer Sabo
201511110



Here is the loss value graph:

- Orange is the training loss curve.
- Blue is the Dev loss curve.



Also mention how the performance of this model compares to the performance of the model from part 1, and how it compares to the previous model in terms of training and evaluation *time*.

As can be seen in the BELU graph the performance of this model is by far better. The attention model boosts the BELU score by 26 points. Also, I observed that on average the attention model gets more stable BELU scores.

Ofer Sabo
201511110

In terms of wall time, i.e. training and evaluation time.
Part 1 model: training_duration is 0:02:01
Part 2 model: training_duration is 0:02:38

This reflects that the model 2 training duration lasts 1.3 times more.

In terms of evaluation, it is easier to count how many iterations the model completes in a single second.
Model 1: 285 iterations per second on average.
Model 2: 225 iterations per second on average.

This reflects that the model 1 evaluation time is 1.25 times faster.