NLU+ Coursework 1: Recurrent Neural Networks

Introductory Remarks

Something about our project the markers should be aware of.

Question 2: Language Modelling

a) Finetuning of hyperparameters

At this step, a hyperparameter grid search was conducted, considering the following values for each hyperparameter:

- Hidden units: 25, 50
- Learning rate: 0.01, 0.05, 0.5
- Lookback steps: 0, 2, 5.

Table 1 shows the results of the hyperparameter search.

Table 1: Results of the hyperparameter search, with 10 epochs of learning.

Hidden units	Learning rate	Lookback steps	Adjusted loss
25	0.05	0	5.67
25	0.05	2	5.669
25	0.05	5	5.647
25	0.1	0	5.472
25	0.1	2	5.469
25	0.1	5	5.46
25	0.5	0	5.403
25	0.5	2	5.295
25	0.5	5	5.288
50	0.05	0	5.583
50	0.05	2	5.555
50	0.05	5	5.537
50	0.1	0	5.417
50	0.1	2	5.424
50	0.1	5	5.436
50	0.5	0	5.303
50	0.5	2	5.357
50	0.5	5	5.342

The best adjusted loss was found to be 5.288 with the following hyperparameters: hidden units 25, learning rate 0.5, and lookback steps 5.

Interpreting the results, it is noticable that the adjusted loss improves with more lookback steps. This is expected as backpropagation through time (BPTT) allows the model to unfold, i.e. propagate errors further back in time [@guoBackPropagationTime13]. This is particularly useful for long-term dependencies, which are common in natural language, the domain of our task.

Furthermore, it was noticable that a smaller learning rate did not result in better performance. This is likely due to the fact that the model was trained for a fixed number of epochs, and a smaller learning rate would require more epochs to converge to the minimum. This is supported by the fact that the best learning rate was the largest one, 0.5. For infinite training time we would expect larger learning rates to perform worse, as they would likely overshoot the minimum more often.

In terms of the number of hidden units, there is no clear pattern in the results. Normally, we would expect more hidden units to outperform the smaller number of hidden units. Our most viable explanation is that the training data was too small for the larger number of hidden dimensions to be beneficial, and the results would be different in this regard when trained on more data.