Deep Learning - Midterm

Brendon Boldt

Oct/23/2017

1 Varying Hidden Unit Count

The code for the following experiement can be found in `mt.py`. When we vary only the number of hidden units in the LSTM model, we get the log perplexities shown in Table 1.

Table 1: Log perplexities given by varying the number of hidden units

# of Units	Log Perplexity
32	1.820
64	1.629
128	1.421
256	1.222
512	0.887

It is the case that as the number of hidden units increases from 32 to 512, the training log perplexity goes down by a consistent amount every time hidden unit count is doubled. From this, we can infer that the model is better able to make predictions on the training data, yet since no test data was used in the experiment, we cannot say whether or not this lower log perplexity corresponds to a better generalization. That is, the model might be overfitting to the training data which becomes more likely as the hidden unit count is increased.

Looking at the actual output of the model, the utterances in Table 2 were produced by each of the models within the last few epochs of training.

Table 2: Examples of utterances produced by each of the models

# of Units	Utterance
32	The will'st dooths, that/ She cordom, merrinds-fatie./ M
64	The Edwird' promst you Edjus of my content/ A gaze did;
128	The of Georging/ To the voices,–/ Shenkery name, this;
256	The crook on thee, I am sure: my lord?/ FROTH:/ I thoug
512	The hands are hore./ LUCENTIO:/ It love merchance sent

In general, as the number of hidden units increased, the number of non-words produced by the model decreased; that is, the higher-count models generally produced a larger portion of correct English words. This, again, could either be explained by a better overall modelling ability or by overfitting to the data. In the most extreme version of overfitting, the model would only produce exactly the words it has seen in the training data.

Looking at both perplexity and model output, it is exptected that lower log perplexities would correspond to more English-like model output since a lower log perplexity means that model is better able to predict and/or replicate the training data.

2 Varying Sequence Length

When we vary the sequence length in the LSTM model, we get the log perplexities shown in Table 3.

Table 3: Log perplexities given by varying the number of hidden units

Seq Length	Log Perplexity
25	1.326
50	1.421
75	1.370

Given the scarcity of data, it is not clear whether there is a trend in the log perplexities with respect to sequence length. We would expect that the log perplexity would go down with longer sequences since the model would be able to "make decisions" (predict letters) based on a larger amount of preceding data. That being said, since a longer sequence length means the model has to account for more complex word interactions, it would likely require a larger number of epochs to train (this could explain the increase in perplexity with longer sequence lengths).

Looking again at the output of the model, the utterances in Table 2 were produced by each of the models within the last few epochs of training.

Table 4: Examples of utterances produced by each of the models

Seq Length	Utterance
25	The done my breast,/ I child! Come./ Third Serving tell
50	The Duke of Wilt mouthly dark. Here's to garments h
75	The skited by the queen or company./ LEONTES:/ Farewell

While the changes are not quite as noticeable as with changing the number of hidden units, utterances tended to have a more natural flow with longer sequence lengths. Individual words more or less seemed just as natural going from one model to another, but the way that the words fit together was better with an increased sequence length. This is an expected result because a longer sequence length means that the model is learning patterns across a larger number of words. Since sentence flow in the English language is greatly affected by the words surrounding other words, taking account of larger number of words would greatly aid in this task.

Note: Since I did not have access to the GPU computers in Hancock (since the building was closed) for the duration of the break (and did not have time the week before), I could not run the models required and instead, have used data from my peers. The presentation and analysis of the data, though, is solely my own work.