Neural Network Language Generation (CL3.407)

Spring 2023, IIIT Hyderabad
Assignment 2 Report

Abhinav S Menon

Report

The main observations were as follows:

- **Range.** The perplexities of both models were in widely different ranges – the bayesian model had perplexities between 10 and 90, while the LSTM model stayed within 1 to 7 in all examples analysed. Thus, overall, the LSTM "models" language better, but there are specific scenarios in which the bayesian method displays more desirable behaviour.

```
Bayes
21.51: This is a sentence .
39.06: The quick brown fox jumps over the lazy dog .
44.48: I shot an elephant in my pyjamas .
76.74: Colourless green ideas sleep furiously .

LSTM
2.21: This is a sentence .
1.75: The quick brown fox jumps over the lazy dog .
2.07: I shot an elephant in my pyjamas .
2.32: Colourless green ideas sleep furiously .
```

- **Style.** The bayesian model "learns" style better than the LSTM model. That is, it scores a sentence written in the style of Henry James as more likely than one not, while the LSTM showed the opposite tendency.

```
Bayes
27.40: His presence at my home , while not wholly of an unpleasant nature , was
nevertheless rather disconcerting .
33.30: I can't say I didn't want him at my place , but I wasn't happy to see him
there either .

LSTM
1.40: His presence at my home , while not wholly of an unpleasant nature , was
```

```
nevertheless rather disconcerting .
1.32: I can't say I didn't want him at my place , but I wasn't happy to see him
there either .
```

- **Contractions.** At first, it appears that the LSTM model is very good at identifying the equivalence of contractions and their full forms, and that the bayesian model is bad.

```
Bayes
21.31: he can't go
19.68: he cannot go
35.46: there's a mouse
16.63: there is a mouse
26.76: what're you doing
12.04: what are you doing

LSTM
4.86: he can't go
4.86: he cannot go
3.62: there's a mouse
3.07: there is a mose
3.93: there is a mouse
3.73: what're you doing
3.48: what are you doing
```

However, it turns out that this is an artifact of the contractions not being present in the data (and therefore being identified as `<unk>` tokens). Thus the above results can be explained as specific cases in the models' general treatment of OOV words (see next point).

```
Bayes
21.31: he qwerty go
35.46: qwerty a mouse
26.76: qwerty you doing

LSTM
4.86: he qwerty go
3.62: qwerty a mouse
3.73: qwerty you doing
```

- **OOV Tokens.** `<unk>` tokens blow up bayesian perplexities, but have almost no effect (sometimes a slight decrease) on LSTMs.

```
Bayes
19.90: she saw a horse .
23.82: she saw a gfhsk .

LSTM
3.19: she saw a horse .
```

```
2.68: she saw a gfhsk .
```

This may be an artefact of the differing vocabulary sizes (10k in the case of the LSTM, vs. 20k in the case of the bayesian model). * **Long-distance dependencies.** The LSTM can be "fooled" into forgetting LDDs by adding intervening words. Interestingly, the bayesian model doesn't change perplexities very much (it slightly increase) even when intervening words are added.

```
Bayes
23.20: the man closed .
25.50: the man who went to the store on the corner closed .
32.31: the cat broke .
35.30: the cat hiding under the wooden table broke .

LSTM
3.34: the man closed .
1.77: the man who went to the store on the corner closed .
3.79: the cat broke .
2.12: the cat hiding under the wooden table broke .
```

- **Fullstops.** The bayesian model finds . very unlikely, but LSTMs lose perplexity significantly when it is added.

```
Bayes
28.04: the cat ran
39.11: the cat ran .
18.12: the man screamed
27.59: the man screamed .

LSTM
6.82: the cat ran
4.07: the cat ran .
3.58: the man screamed
2.50: the man screamed .

- lstms can be fooled by putting quotes. small reduction then.
- bayes continues to mark it much higher
```

- **Quotes.** Quotes always reduce the LSTM's perp, while they hugely increase the bayesian model's.

```
Bayes
63.31: " the cat ran . "
49.34: " the man screamed . "

LSTM
2.25: " the cat ran . "
1.74: " the man screamed . "
```

Reproduction

Downloading

The bayesian model can be downloaded from **here** and the LSTM model from **here**. They need to be in the same directory as all the code files.

Running

To run the bayesian or LSTM model, simply run the `bayes.py` or `lstm.py` Python scripts and enter the sentence to be scored at the prompt.

```
$ python3 bayes.py
Enter a sentence to score: The quick brown fox jumps over the lazy dog .
perp: 39.06245698819558
```