MT Exercise 2: RNNs and Language Modelling

Mikal Ermias Tadesse, 21-724-927
Sudehsna Sivakumar, 21-721-634

1 Training a recurrent neural network language model

Chosen Data:

We first started training the model with the book "Nathan the Wise; a dramatic poem in five acts by Gotthold Ephraim Lessing" but when we generated a sample text, but the sample was almost unintelligible and we thought this was due to some unconventional formatting which wouldn't allow proper preprocessing. Then we also noticed that it was only around 3000 segments which was too short anyways. We then chose another dataset: "Moby Dick; Or, The Whale by Herman Melville". Moby Dick, being a long and better formatted prose novel, provided better training data.

Special Attributes:

Moby Dick contains distinctive linguistic and thematic features, including 19th-century literary English. These stylistic elements may influence our model to generate longer, more descriptive sentences, potentially with hints of marine language patterns.

Sample Generation:

With the default settings the generated sample exhibits a strong stylistic match to Moby Dick, with references to nautical elements (e.g., “sea,” “harpoon,” “prow”) and dramatic tone. The presence of names like “Ahab” suggests the model successfully internalized important tokens. While semantic coherence is for the most part lacking and <unk> tokens remain present.

2 Parameter tuning: Experimenting with dropout

Tables and line-plots:

Epoch Perplexity

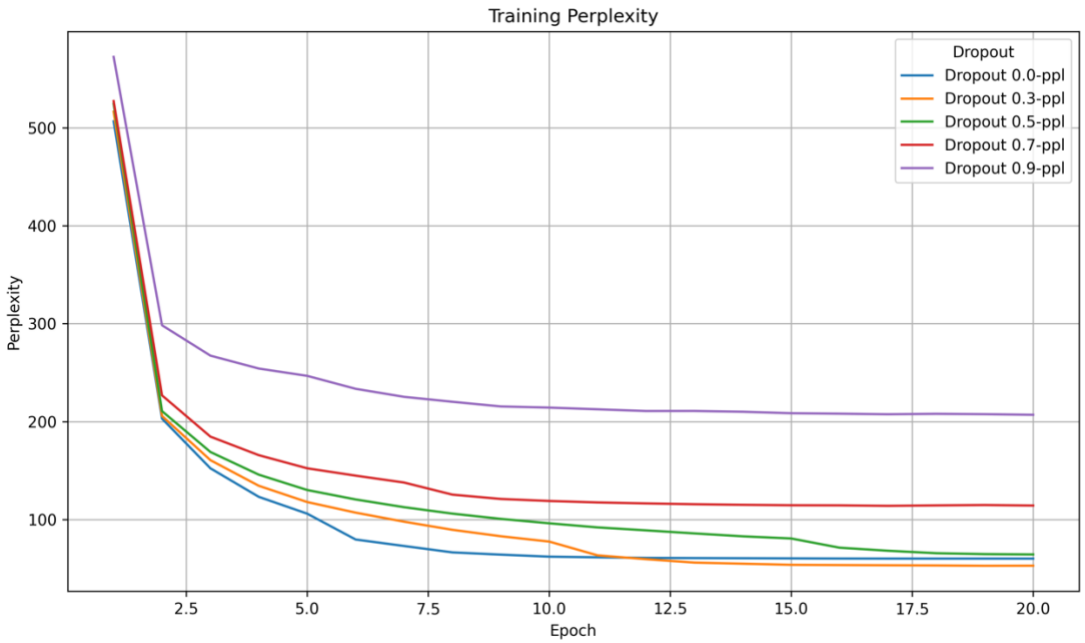
Epoch	Dropout 0.0-ppl	Dropout 0.3-ppl	Dropout 0.5-ppl	Dropout 0.7-ppl	Dropout 0.9-ppl
1.0	506.47	516.34	524.29	527.13	572.21
2.0	202.99	205.72	210.66	226.73	298.3
3.0	152.3	160.62	168.95	184.63	267.31
4.0	123.17	134.45	145.8	165.7	254.13
5.0	106.01	117.88	130.15	152.32	246.67
6.0	79.65	107.08	120.54	144.84	233.46
7.0	73.04	97.84	112.65	137.8	225.26
8.0	66.51	89.55	106.07	125.45	220.22
9.0	64.22	82.98	100.71	121.04	215.45
10.0	62.16	77.55	96.19	118.99	214.28
11.0	61.48	63.48	92.01	117.48	212.54
12.0	60.89	59.59	89.0	116.5	210.8
13.0	60.7	56.09	85.88	115.66	210.87
14.0	60.55	54.97	82.9	115.06	210.03
15.0	60.4	53.83	80.71	114.62	208.54
16.0	60.25	53.57	71.3	114.52	208.08
17.0	60.2	53.36	68.18	114.0	207.47
18.0	60.19	53.18	65.67	114.47	207.94
19.0	60.19	52.85	64.76	114.84	207.56
20.0	60.18	52.93	64.42	114.32	206.94

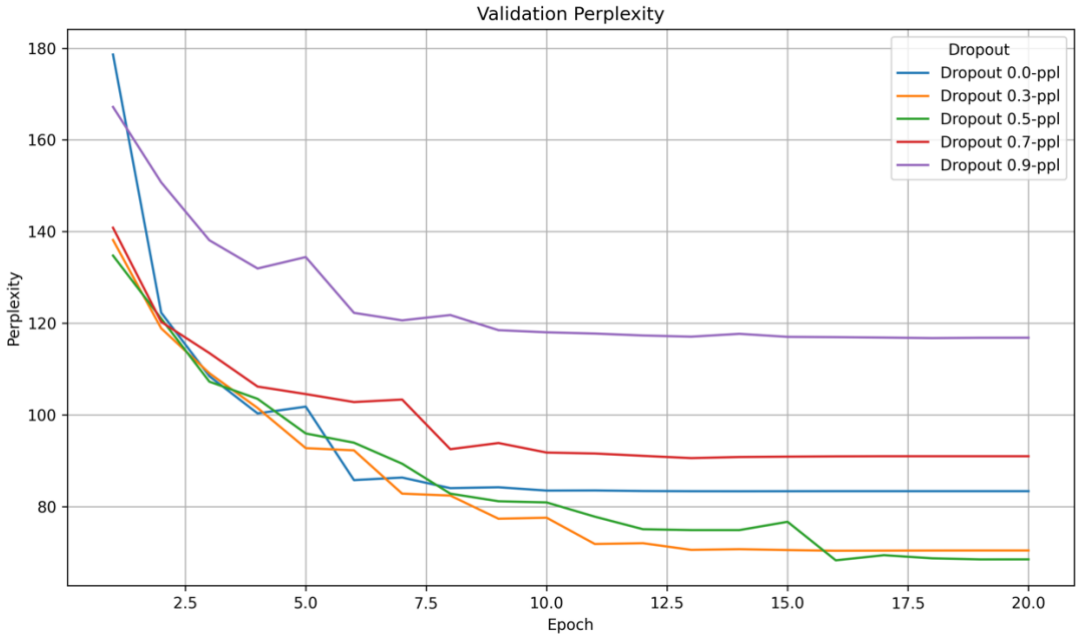
Valid Perplexity

Epoch	Dropout 0.0-ppl	Dropout 0.3-ppl	Dropout 0.5-ppl	Dropout 0.7-ppl	Dropout 0.9-ppl
1.0	178.61	138.16	134.76	140.83	167.18
2.0	122.31	118.9	121.02	120.31	150.73
3.0	108.46	109.11	107.27	113.5	138.1
4.0	100.35	101.53	103.52	106.2	131.95
5.0	101.82	92.77	95.97	104.56	134.44
6.0	85.8	92.29	93.96	102.82	122.28
7.0	86.36	82.85	89.38	103.37	120.64
8.0	84.05	82.43	82.85	92.54	121.8
9.0	84.24	77.38	81.19	93.89	118.5
10.0	83.52	77.6	80.93	91.81	118.03
11.0	83.55	71.87	77.83	91.6	117.76
12.0	83.42	72.04	75.09	91.09	117.34
13.0	83.38	70.59	74.9	90.6	117.08
14.0	83.37	70.74	74.9	90.84	117.7
15.0	83.38	70.55	76.71	90.92	117.04
16.0	83.39	70.4	68.31	90.98	116.98
17.0	83.39	70.43	69.43	91.01	116.88
18.0	83.39	70.45	68.75	91.01	116.77
19.0	83.39	70.46	68.51	91.01	116.84
20.0	83.39	70.46	68.52	91.01	116.85

Test Perplexity

Model	Dropout 0.0	Dropout 0.3	Dropout 0.5	Dropout 0.7	Dropout 0.9
0.0	139.09	132.47	127.83	143.15	213.86





Connection between perplexities:

There is a clear positive correlation between training, validation, and test perplexity across all models. As training perplexity decreases, validation and test perplexities generally follow the same trend, indicating that the models generalize consistently.

Among the dropout settings tested, dropout = 0.5 yielded the best performance, achieving the lowest validation and test perplexity.

This suggests that a 50% dropout rate performs best under these circumstances providing the right balance between underfitting and overfitting for our dataset and model configuration.

Best and Worst samples:

While the lower-perplexity model produced slightly more structured and thematically consistent text, featuring phrases related to the sea, boats, and characters like Ahab - the output still lacked grammatical correctness and coherence. The overall fluency was poor, and <unk> tokens appeared frequently. In contrast, the model with the highest perplexity produced more chaotic and fragmented text with even less structure and weaker thematic relevance. In both cases, the outputs did not convincingly resemble natural or fluent language. Overall, the model with dropout 0.5 generates more coherent, and stylistically appropriate text (e.g., “whale-boat,” “my boat,” “thou”). In contrast, dropout 0.9 leads to weird phrasing and less intelligible content which reflects its poor generalization.

Repository: https://github.com/Sudehsna/mt-exercise-02