

CSE490 Homework 2

ayp2, chunz2

November 17, 2018

1 9.1 Plots

The following plots were based on the model using the default hyperparameters, as restated below.

SEQUENCE LENGTH = 100
BATCH SIZE = 256
FEATURE SIZE = 512
FEATURE SIZE = 512
TEST BATCH SIZE = 256
EPOCHS = 20
LEARNING RATE = 0.002
WEIGHT DECAY = 0.0005

Loss by epoch

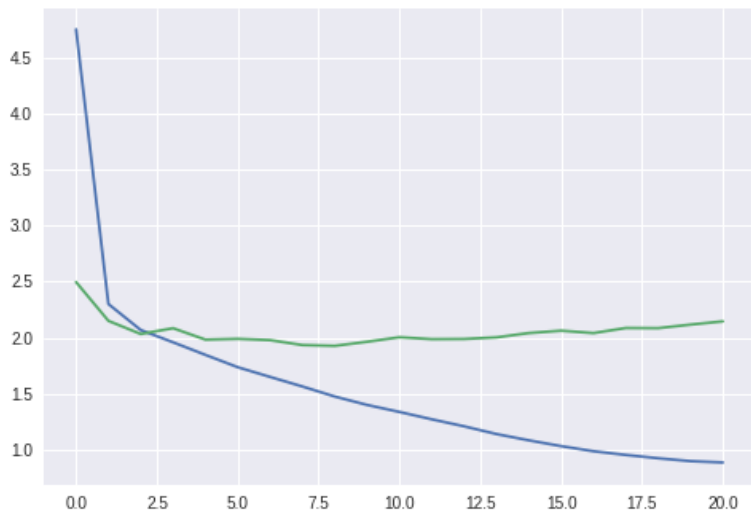


Figure 1: Blue = training loss, Green = test loss

Loss by epoch

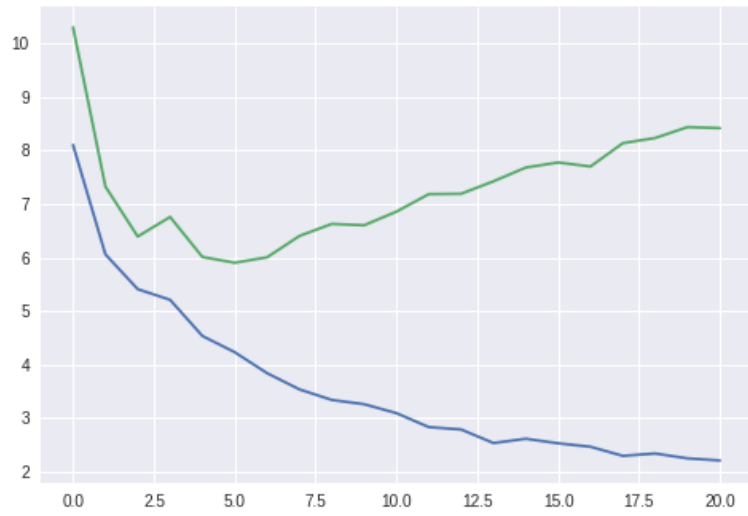


Figure 2: Blue = training perplexity, Green = test perplexity

Loss by epoch

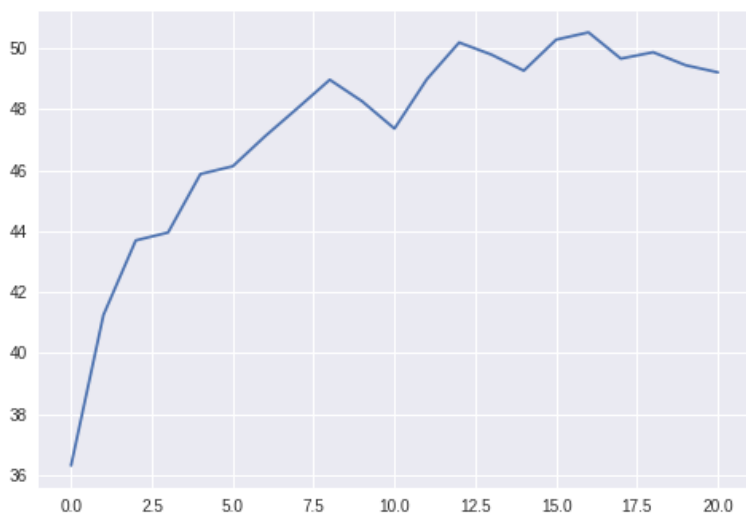


Figure 3: Blue = test accuracy

2 9.2 Final stats

Final test loss: 2.1461

Final test accuracy: 3086856/6272000 (49 percent)

Final test perplexity: 8.415808763377878

Best test loss: 1.9902

Best test accuracy: 51 percent

Clearly our model is overfitting our data, visible just by looking at our graphs. Our attempts at changing the hyperparameters didn't result in any notable gains or trends in accuracy or loss, so we ended up using results from the default hyperparameters, even though it's obviously not optimal.

3 9.3 Favorite Sentences (in terms of humor)

Max - "I never wanted" as seed words :

"I never wanted the told mery took. He's it. He's it..." "Hermiones. "Hermiones.
"Hermiones. "Hermiones. "Hermiones. "Hermiones. "Hermiones. "Hermiones.
"Hermiones. "Hermiones. "Hermiones. "Hermiones. "Hermiones. "H"

Sample - "Harry Potter" as seed words :

"Harry Potter ? Lee Durmberce was not on the moment. "My felt there Dumbleder, feec a feet Pofes? 1ately meally of all room Sent dement. "Hally, der's cont ree compent flugh. . ."

Beam - "Did you" as seed words :

"Did you. Harry. ... them the little front of him. They'd have the pod's complaining. "Hermione. "They're the door into the deading, Harry. ..." He's it..." He headed down in his recently," said Harry. ... "

4 9.4 Creating Sentences Sampling methods

Beam search generated the 'best' sentences that we observed, where 'best' is based on the reasonableness of words generated. Regular sampling had a notable amount of almost -but-slightly-incorrect english words. Using the max strategy produced the most consistent (repeating/redundant phrases). Beam search provided some middle ground with consistency and variance - more variability than max, but rarely non-English words.

5 9.5 Temperatures

0 temperature produced the saddest outputs (forgot that we had changed temperature to 0 when checking the results and thought there was a bug) and higher temperature (1.5) produced more gibberish than usual. We ended up trying temperature of 0.5, 1, 1.5. If we define best as most reasonable words that had both good variety and validity, then temperature 1 was the best.

Temperature 0 produced duplicates of the same sentences generated after a given epoch. This is because dividing by 0 (actually an extremely small temperature) really exaggerates and skews the current distribution so that already probable things become incredibly more probable when the normalization happens soon after. Temperatures greater than 0 but less than one also do skew the distribution, but it becomes less extreme the closer temperature approaches 1.

Temperature 1 should just divide the distribution by 1, so no change to distribution before it is normalized.

Temperature greater than 1 should make our distribution flatter, because we're making all our probability values smaller by dividing by a number greater than 1. Then if we normalize afterwards, we'll get a more even spread = it'll be easier for our sampling / beam algorithms to choose more unusual characters next, because the distribution is less skewed.

If we allowed for negative temperature, that would be weird because we couldn't normalize the distribution to 1 if all the values in our distribution are negative after we divide by temperature.

6 New Corpus Questions

1. We initially wanted to do the bible, thinking that is easy to find, large, and interesting. Comparing raw file sizes though, HarryPotter.txt was 6.6M bytes and the bible was only 4.5M. So we switched doing Harry Potter fanfiction, thinking it'd be interesting to see how different JK Rowling vs. internet writing would be on the same subject (more or less). We went through archiveofourown.org, filtering from most popular to least, grabbing any fanfiction with $\geq 50,000$ words so it wouldn't take forever. We added about 7M bytes worth of fanfiction. The final character count was 7,173,923.

2. The Harry Potter (canon) corpus generated lots of stuff about doors. It otherwise chose a lot of filler text like 'he', 'was', and 'and'. The Harry Potter fanfiction corpus generated plenty of filler text as well, but it also kept landing on some pretty funny terms like 'sucked', 'arrested', and 'ded'.

3.

Outputs for each sampling method

[illegible]

Figure 4: At default temperatures:

7 Words Instead of Characters Questions

1. New difficulties: highly probably words largely dominated output. I guess words that are common (ex. 'the') were just significantly higher probability than anything else. We were able to mitigate this a little bit by jacking up the temperature. It was interesting, the temperature had to be much higher than the character based sequencing to see it making a difference. At 1.3 we would still produce 'the the the ...' a ton. At 4.2 it became more varied.

2. Our new vocabulary size was 9,913 (including a token representing unknown words, aka occurrence less than 5 times)

3. No, hyperparameter tuning didn't seem needed - batch size, sequence length, feature size, and most other hyperparameters didn't make a difference for us at the word level. The word level was very stubborn. Until we moved temperature to a high value, it would generally just repeat a single word. And once it produced different words, they were very random words. Fiddling with hyperparameters didn't help us achieve a better accuracy.