Language Modeling with RNNs

Christian Altrichter

December 2, 2022

Question 1 - Preliminaries and Reading Comprehension [28/100 points]

Question 1.1 - Text data

Question 1.1.1

Here the requested information on the text file (the urllib.request was used):

1. Total number of characters: 182.550
2. Total number of unique characters: 106
3. Total number of lines: 5.033

An interesting property of the book is that after the title (which is written in all capital letters), sometimes only the first word is written in capital letters, and sometimes the first two words are written in capital letters.

Question 1.1.2

Here, one could normalize the data, which includes but not limited to:

1. Remove non-novel related text paragraphs (e.g. table of content).
2. Remove nonalphanumeric chars and replace by white space.

Question 1.2 - Dataloader/ Batch Construction (Repeated from Exercise 6)

Question 1.2.1

The if statement checks whether a certain word is already in the vocabulary. If so, it returns the index of the string. Else if the function allows (through its parameter) to add new words, it does so and extends the vocabulary and returns the corresponding index of the newly added word, else it returns the default unknown index value.

Question 1.2.2

id_to_string:

1. **Keys are:** Integer numbers which are the length of the dictionary, and therefore the index in the dictionary.
2. **Values are:** Strings, thus the words themselves.

string_to_id:

1. **Keys are:** Strings, thus the words themselves.
2. **Values are:** Integer numbers which are index of the values in the dictionary.

Question 1.2.3

Returns the length of the vocabulary token id's.

Question 1.2.4

The number of batches in which our data set is divided into.

Question 1.2.5

The purpose of these two code lines is to uniform the batch size of each tensor. As the text length divided by the batch size does not produce an integer value, we need the padding to standardize our batch size.

Therefore, the first line creates a tensor with the size of the full integer division between text length and batch size rounded up. The tensor is then populated with 0's (the pad value).

The second line fits the original data tensor into the padded tensor, which is then being filled up at the end with the required padding to then reshape each tensor to an equal batch size.

Question 1.2.6

The tensor shape with respect to bptt_len and bsz is:

1. Rows: bptt_len
2. Columns: bsz

Question 1.2.7

The tensor shape with respect to bptt_len and bsz is:

1. Rows: bptt_len + 1
2. Columns: bsz

Question 1.3 - Modeling, Training, and Decoding

Question 1.3.1

To ensure that the model doesn't produce an error signal across batches when in place operations are performed.

Question 1.3.2

As the padding value of 0 was given, here we ignore said value to not contribute to the gradient descent.

Question 1.3.3

Here, self.rnn takes two arguments, which are referred to in our code as self.rnn(emd, h_0). Here, emd is a tensor of shape (N,B,D) and h_0 is a tensor of shape (L,B,H) where:

1. B = Batch size
2. D = Character embedded size
3. N = Sequence length
4. L = Number of layers
5. H = RNN size / Hidden size (according to documentation)

Question 1.3.4

The output of self.rnn outputs two tensors w.r.t. the following variables:
The first is the output with shape (N,B,H) and the second is the state with shape (L, B, H) where:

1. B = Batch size
2. D = Character embedded size
3. N = Sequence length
4. L = Number of layers
5. H = RNN size / Hidden size (according to documentation)

Question 1.3.5

So we can observe the predicted behaviour after every batch with a sample sentence, which in our case is "Dog like best to".

Question 2 - Running Experiments Using the Initial Code [30/100]

Question 2.1

The modifications have been made accordingly in the source file.

Question 2.2

The reported perplexity of can be seen below:

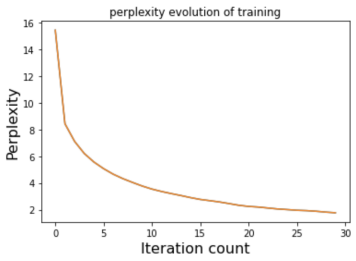

Figure 1: Plotting of the training perplexity

Regarding the text generated at three different stages, I have taken the number of epochs divided by the number of stages.
Thus, the first stage is at epoch 10, the second at epoch, 20 and the third at epoch 29.

The text generated in each epoch is as follows:

Generated text at stage 1:
Dogs like best to eat at home, and he went to the barn, and eyese the water more means of ridding them whatever best to tremblind to come at the g
Generated text at stage 2:
Dogs like best to eat this hay? No, said the Dog; "they tie me in the dreverts of darling," said the Bat, spreading or two are very my himself
Generated text at stage 3:
Dogs like best to eat this his meat she could not refrain from good that wherever of share it with you, and now conficted into a present so aptly,

Question 2.3

Overall comment:

The model does not produce meaningful text. It recognizes patterns within the text (e.g. titles, indentation, new lines, paragraphs, beginning of paragraphs, regular text etc.) and produces a text following the input structure, which it does quite well. However, the text produced is coherent and includes grammatical errors.

Case1 - Title taken from book:
INPUT: 'THE WOLF IN SHEEP'S CLOTHING'
OUTPUT

```
A WALF stared to the bottum if the country road passed now.

The Theep without a distributing a Project Gutenberg-tm work, and bringing up a golden ax
and asked the Fox that she forget to be found in:
 the Fox, who was elable to be found in a copy, disclaimer or
affect of conly for the course.

Lid the best Tree. He will not her nest fire.

THE HIRE SNN THE CAMELEAPERARS


A WAR How grew enter accessed to be in
sympathy with woodcutters.

The Foxestraw over the
with the text fit to be would you no moot
```

Figure 2: Text output of a title taken from the book

Case2 - Made up title:
INPUT: 'THE WOLF DISGUISED AS HORSE'
OUTPUT

```
TROVE BRE GAVE,

A WOOD you are you
 old my kind, and so raised the wind withding they were betten the cockessed
    Theirive Frog was very vain. She who was nearling there, say: “The better next
heag, where we can chat the Frog from the Fox’s revenge. Lapiter to ean to the performed in
grame at ner
and came at very before they goudd, He spare me,” said the Boar, “is and yeur
health. But the birds were indignant at the cubs,
unternated in a marg the pleased with my ears, of which he searting and threw his
```

Figure 3: Text output a made up title

Case3 - Some text in similar style:
INPUT: 'A CROW ate a fruite of poisen'
OUTPUT

```
t. Some on with stories of thes.

The poor bran. Solocatist the best materials for them.

[Illustration]

“No mo sooner took
possession of his new dominion than he began to drink not for the country road passed not to
pay me to the river parts of the world be
doie down and
her own young ones, reself his companions, we are in a short for his foolish is a running of
this work nor the early
straight to the end of the course. Full of sport, the Hare
far beft in a short time to rum to the pire of a friend? You ma
```

Figure 4: Text output with a text similar in style

Case 4 - Anything i am interested in:
INPUT: 'I was walking through the forest alone and all of a sudden'
OUTPUT

```
t. Some on with stories of thes.

The poor bran. Solocatist the best materials for them.

[Illustration]

“No mo sooner took
possession of his new dominion than he began to drink not for the country road passed not to
pay me to the river parts of the world be
doie down and
her own young ones, reself his companions, we are in a short for his foolish is a running of
this work nor the early
straight to the end of the course. Full of sport, the Hare
far beft in a short time to rum to the pire of a friend? You ma
```

Figure 5: Text output with a random input

Question 3 - Extending the Initial Code [40/100]

Question 3.1

The code has been modified accordingly and can be found in the source file.

Question 3.2

The lowest perplexity was achieved below 1.03 with the following characteristics:

1. Perplexity achieved: 1.03
2. Achieved in Epoch: 29
3. Achieved in IDX: 0

The following depicts the training evolution:

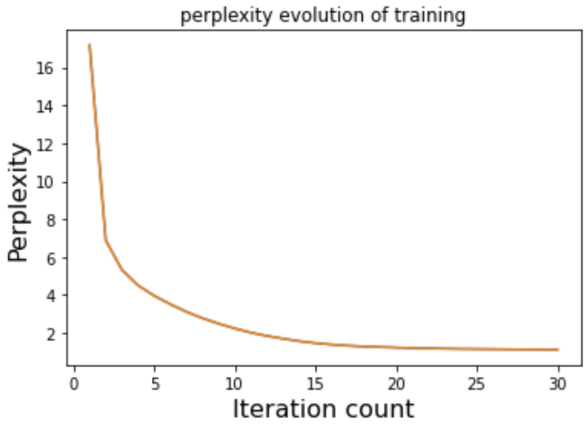

Figure 6: Text output with a random input

Question 3.3

The code has been implemented and can be found in the source file. Here, I used an external function to randomly sample: "np.random.choice".

Question 3.4

Comment: is it smart to use random sampling
Here, it appears to not be a good choice when it comes to a real title. The output produced by the greedy approach is better (due to probable overfitting). However, in the made up title the greedy approach introduced more grammatical errors and is therefore less effective than the random sampler approach.

Case 1 - A title of a fable which exists in the book; Sample = False

```
A PIGEON, preading his wagon through a miry lane, the wheels stuck fat their holes at his own image in the water
that it attracted the honeycomb and at once threw down his ax. From that there was a storm. The terrible unseen
wind came and struck to his play.



THE WOLF AND THE HARE


A WOLF made once spied a pitcher with a large sum of gold for distributing this eBook or online at finds and eat
them, and the medium on which they may be stored, master, better nearly frozen.

He took pity upon him and inv
```

Figure 7: Text output with real title, Sample set to false

Case 1 - A title of a fable which exists in the book; Sample = True

```
S

A PIGEON, preasing his just and all the first time the Flw young stopped in a tree that overhung the river and the jun and
in danger of grain and hay ceased, and
the Horse was drowned, and fly with beating he rad his way with some difficulty into a
basket of corr. The Foundation makes no
represent. The terms of this agreement for free              185
    The Jak and the Elephant                          195
    The Horse, the Crow and the Pitcher                18
    The Frogs who as he found the She
```

Figure 8: Text output with real title, Sample set to true

Case 2 - A title which you invent; Sample = False

```
A BOY once planted close beside a large and noble Pine Travess (Granged of the promise he had ridiculed.

The Foundation's principal office is in later times, ran at once the drover to pass by saw him there, and stopped
to look at them.

“What a foolish creature you were, to look so sour and so for you, and you shall be found in the act of this work
or any one eight should deep, she made a little journey
together.

When night, he saw the cuckoo to the rank of lenstry friend, who, amazed at the good things
```

Figure 9: Text output with fake title, Sample set to false

Case 2 - A title which you invent; Sample = True

```
A CHARCOAL Burner carried one day tall at last agreed to donate royalting better than that. Let us hang a bell about his
neck and crammed his Son to gather an abundant supply of food—enough to be willing to see which you comply with all other
teaching is percians, and I have let him can in the f.ole of sticks tied
tightly treest to the ground, and its head was a
reasonable after as they were; but as soon as the morning day, the Satyr pounced upon a pot of honey, which was so kindly
taken that he bade one
```

Figure 10: Text output with fake title, Sample set to false

Question 3.5 (Bonus)

For this question I have taken the lecture notes of Algorithms and Complexity by David Mount. The source code for this model has not been uploaded, but of course can be accessed and provided if required.

Here, I have taken 4 prompts to test the model, which are the following:

1. ”Stable matching works”
2. ”Lemma 1”
3. ”for (i = 1 to n)”
4. ”Algorithms are fun because”

Comment
The text has no consequent underlying structure and is highly repetitive in its wording which makes it hard for the model to produce something meaningful. The prompts used can be observed through out the text. Even when using a prompt that only covers one section the (e.g. stable matching) it mixes these generated text with other sections.

The following output was produced:
Case 1 - "Stable matching works because"

```
 Gf (400), graphs and digraphs).

As a basis case, we define the weight of the trivial “2-table”). However, observe that the running
time of the algorithm recursively on the simple algorithm are based on a decision problem. Let S =
i si denote the total supply/demand values (see Fig. 29(a)).
•
A
B
C

Max Dominance

1±

CMSC 451

fLet z ∗ denote the total size of the objects among the vertices of Selection by the set S of
posities, the exact choice. By analyzing the
last few characters of X and Y , graphs, we
```

Figure 11: Text output with of first prompt

Case 2 - "Lemma 1"

```
// undiscovered vertex v is a clique for G, then for each u, v ∈ V 0, H ∈ E that involve set of
scheduled in increasing order of v lates. We will take two characters x and y, implying that V 00
is a vertex cover for G. But this is the value of k that leads to the shortest path and edge
weights are at least 1, find the sum of flows from Y to X in the adjacency matrix). Thus f [0, w]
= 0. (Both the same value of a flow f , denoted
|f |, is defined as the sum of
their individual probabilities, and the exper f
```

Figure 12: Text output with of second prompt

Case 3 - "for (i = 1 to n)"

```
 {
// ...for each vertex f (u, v) ∈ E
}

Let c denote the netation is exactly the converse. We usually have modify the subpolygon has two
ways in which two sorted lists
on
pairs of vertices to be in the subset sum
problem. There are infinitely many such paths between xi and xj in the other half of a trickal
problem (G, k), produces an instance of the 3-coloring problem (G, 3).
NP-Completeness

126

CMSC 451

fThus the time is not possible to convert this into a 2.
There is one last issue is hard, that is, fo
```

Figure 13: Text output with of third prompt

Case 4 - "Algorithms are fun because"

```
 because there is only one man, and vice versa. Second, we assume
that there is some subsets may have been asked, a contradiction.
Finish time, then we get into troubly need to intending one true literal.
Select such a current matching is measure how to design an O(n log n) time algorithm for computing
a pivot element that
achieves this weight. However, as
a problem into subproblems in every gadget. However, if you could not have to use a fant to this
from the MST. In this case we have d[x]
```

Figure 14: Text output with of third prompt

Question 4 [2/100]

Question 4.1

Perplexity measures the certainty of a new token with regards to the sequence.

Question 4.2

It is an issue because it prevents the weights from changing value and can hinder complelty the training of a model.