Michael Pollack: LING 574 HW7

Due 11PM on May 16, 2024

In this assignment, you will

- Develop understanding of recurrent neural networks, especially as used for language modeling
- Implement components of data processing
- Implement masking of losses for an RNN language model

All files referenced herein may be found in `/dropbox/23-24/574/hw7/` on patas.

1 Recurrent Neural Network Decoders/Taggers [35 pts]

Q1: Understanding Masking [15 pts] Suppose that we want to train a (word-level) language model on the following two sentences:

<s> the cat sits </s>
<s> the model reads the sentence </s>

We saw in HW6 that padding is necessary to make these sentences have the same length so that they can be batched together, as:

<s> the cat sits </s> PAD PAD
<s> the model reads the sentence </s>

Please answer the following questions about these sequences:

- In a recurrent language model, what would the input batch be? What would the target labels be? [4 pts]

  Answer:
  The input batch would be:
  `[[<s>, the, cat, sits, </s>, PAD],`
  `[<s>, the, model, reads, the, sentence]]`

  The target labels would be the next word in each sequence, meaning:
  `[[the, cat, sits, </s>, PAD, PAD],`
  `[the, model, reads, the, sentence, <s>]]`

- Recurrent language models use a *mask* of ones and zeros to 'eliminate' the loss for PAD tokens. What would the mask be for this batch? [3 pts]

  Answer:
  `[[1, 1, 1, 1, 1, 0],`
  `[1, 1, 1, 1, 1, 1]]`

- Suppose that we have the following per-token losses:

$$\begin{bmatrix} 0.1 & 0.3 & 0.2 & 0.4 & 0.7 & 0.5 \\ 0.2 & 0.6 & 0.1 & 0.8 & 0.9 & 0.4 \end{bmatrix}$$

What is the *masked* loss matrix? [3 pts]

Answer:
```
[[0.1, 0.3, 0.2, 0.4, 0, 0],
[0.2, 0.6, 0.1, 0.8, 0.9, 0.4]]
```

- Why is it important to mask losses in this way? What might a model learn to do if the loss is not masked? [5 pts]

Answer:
Masking losses in this way ensures that these losses do not contribute to the overall loss calculation during training. This ensures that the model focuses on meaningful tokens rather than padding. Without masking, the model develop a bias towards padding tokens, generate incorrect loss calculations, or in general start treating padding tokens as if they had actual meaning, which will hinder the model's effectiveness.

Q2: Evaluating Language Models [20 pts] Given a corpus $W = w_1 w_2 \dots w_N$ (so N is the number of tokens in the corpus), a common (intrinsic) evaluation metric for language models is *perplexity*, defined as

$$PP(W) = P(w_1 \dots w_N)^{-\frac{1}{N}}$$

This can be thought of as the inverse probability that the model assigns to the corpus, normalized by the size of the corpus.

- Is a lower or higher perplexity better? [2 pts]

Answer:
It is better to have lower perplexity, as this indicates better predictive performance.

- For a recurrent language model, write an expression for $P(w_1 \dots w_N)$ using the chain rule of probability. How is this different from the expression for a feed-forward language model? [5 pts]

Answer:
The chain rule would dictate that the joint probability of a sequence is the product of the probabilities of each of its events, given the previous events. That would mean something like:

$P(w_1 ... w_N) = P(w_1) \times P(w_2|w_1) \times P(w_3|w_1, w_2) \times ... \times P(w_N|w_1, ..., w_{N-1})$

In a cleaner and more condensed notation, this could be:

$P(w_1 ... w_N) = \prod_{i=1}^{N} P(w_i|w_1, ..., w_{i-1})$

Or perhaps (in a clever bit of foreshadowing):

$P(w_1 ... w_N) = \prod_{i=1}^{N} P(w_i|w_{<i})$

- Show that

$$PP(W) = e^{-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})}$$

where $w_{<i} = w_1 w_2 \dots w_{i-1}$ and log is the natural (base e) logarithm. [5 pts]

[Note: using base e measures perplexity in a unit known as *nats*. Using base 2 would measure it in bits.]

Answer:

$e^{\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})} = (e^{\sum_{i=1}^{N}\log P(w_i|w_{<i})})^{-\frac{1}{N}}$

$= (e^{\log P(w_1)} \times e^{\log P(w_2|w_1)} \times ... \times e^{\log P(w_N|w_{<N})})^{-\frac{1}{N}}$

$= (P(w_1) \times P(w_2|w_1) \times ... \times P(w_N|w_1, ..., w_{N-1}))^{-\frac{1}{N}}$

$= (\prod_{i=1}^{N} P(w_i|w_{<i}))^{-\frac{1}{N}}$

Using our definition from earlier:

$(\prod_{i=1}^{N} P(w_i|w_{<i}))^{-\frac{1}{N}} = P(w_1...w_N)^{-\frac{1}{N}}$

Therefore:

$PP(W) = P(w_1...w_N)^{-\frac{1}{N}} = e^{-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})}$

- What is another name for the exponent $-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i \mid w_{<i})$ in the above expression? [Hint: it appears in training as well.] [3 pts]

Answer:
This is known as the cross-entropy loss or the negative log-likelihood.

- Suppose that the same text corpus were tokenized with two different vocabularies of different sizes (perhaps, e.g., one replaces infrequent tokens with an UNK token) and two language models were trained on the resulting tokenized text. All else being equal, would you expect perplexity to be lower or higher for the model with a smaller vocabulary? What consequences does this have for comparing different language models? [5 pts]

Answer:
The model trained with a smaller vocabulary would have higher perplexity. This is because it would be exposed to less words and thus be less able to perform predictions than the model trained with a larger vocabulary. This would mean that, when comparing different language models, we can generally expect those trained on larger vocabularies to have smaller perplexities.

2 Implementing an RNN Character Language Model [25 pts]

See hw7.tar.gz

3 Running the Language Model [15 pts]

`run.py` contains a basic training loop for SST language modeling. It will record the training and dev loss (and perplexity) at each epoch, and save the best model according to dev loss. Periodically (as specified

by a command-line flag), it also outputs generated text from the best model.

Q1: Default parameters Execute `run.py` with its default arguments. Paste below the texts that are generated every 4 epochs, as well as the epoch with the best dev loss and the dev perplexity from that epoch. In 2-3 sentences, describe any trends that you see. [Note that generated text will not necessarily be completely coherent: recall that this is a *character-level* language model.] [5 pts]

Answer:
While the text doesn't necessarily become more coherent, you do see clearer words starting to form towards the end. You especially see more use of the word "the" paired with nouns, and what seems to be the beginnings of more complex noun phrases.

Output:

```
["<s>the the from sto senting and 's with the the the t", '<s>the for and more the porting
a whan of the conters', '<s>the film , the farlating , be and the and pore is ', '<s>the actor
, the the remod sore film a come the the', '<s>the pary , a the plontion and a rest , as the
the ', '<s>the mod and the sperer and and , the the pille and', '<s>the senter of the move
least of the the the deanti', '<s>the fare it comat the film the be coming of in it ', '<s>and
the be film a comect a a mant sure a mover and', '<s>the the band the more and the perpite
a comating s']

['<s>it the sent and in the the and be make and the the', '<s>sore a film .</s>ary the movie
.</s>were the could the c', '<s>the somes the film , a make the mories and sears a', '<s>the
sever .</s>and comedy the the perferfing that the', '<s>a the be a sumedy is lith of be a
so a dessere .</s>a', '<s>the preating and in a movie and the strome the fil', '<s>the has
the the seat and in the sexer and a story ', '<s>the is and the winte is and it sellic to
a stist a', '<s>a for the sill , the comere and a decicomet , a fi', '<s>the the lent the
so coming and a sing entere the c']

['<s>a film the a film a spines comenting film and stor', '<s>a him .</s>art to the movie
the perfict and a movie a', '<s>the intere that the the fars mand actor to surpric', '<s>that
be some and the have in the sunding strile an', "<s>it 's and and have to the senting that
what the mo", '<s>the we the comedy beap .</s>enge , the still of the m', '<s>the for the
movie , and movie and hore and on the ', '<s>a doings and the film the move in the comedy
the p', "<s>it 's enterters , the provie .</s>and .</s>ar and the mo", '<s>a comate , the
movie fun the to sust the real and ']

['<s>the pretion of the hard it and real movie the mate', '<s>story with the director to part
.</s>and could and th', '<s>conderthing and to desiging , and the cent and the', '<s>the movie
the film in the bele and the movie .</s>and', '<s>the movie the remant and the incere the
comesting ', "<s>it 's but the the film and be is should the maning", '<s>the film .</s>and
is a film , comedy a stare and and ', '<s>a donely for a destion with a desents and the the
', '<s>a the care .</s>and not the completter and work the f', '<s>the characction and the
chare in the comedy is a re']

['<s>the port to a and the love of the the is in the pe', '<s>the plot and much the movie
the movie the word .</s>a', "<s>a compless it 's suth all sonse and menting the mo", '<s>the
sees in the movie .</s>and recoming , the feen co', '<s>a presting and with the performance
, a story of t', '<s>the starn to the the and remach of the master .</s>an', '<s>like the
sented .</s>aker .</s>..  be instare is and the ', '<s>a character of come in the film , and
```

the gire in ', '<s>the movie the conting the the film is a real the f', '<s>a film of a sean the make the film .</s>and character']

Best Dev Loss → Epoch 19
Epoch 19 train loss: 1.8475168546040852
Epoch 19 dev loss: 1.7222588062286377; perplexity (nats): 5.597157001495361

Q2: Modify hyper-parameter(s) Re-run the training loop, modifying some combination of the following hyper-parameters, which are specified by command-line flags:

- Hidden layer size
- Embedding size
- Learning rate
- Number of epochs [in particular: making it larger]
- Softmax temperature.
- L_2 regularization coefficient.
- Dropout (probability with which neurons are dropped from the input and to the output during training)

Include your model's generated texts here. In 2-3 sentences, state exactly what hyper-parameter change(s) you made, and what effects (if any) you see in terms of the dev set perplexity and text that the model generated. [5 pts]

Answer:
I decided to go a mildly psychotic route and increase my epoch count to 60. The outputs are below. Interestingly, the trend mentioned earlier with more complete noun phrases seems to continue with the addition of more epochs. By the last output, most words seem to be real and formed into coherent nounphrases (albeit not complete sentences). It is also interesting to see that the perplexity continually decreased (at a decreasing rate) across epochs.

Output:
["<s>the the from sto senting and 's with the the the t", '<s>the for and more the porting a whan of the conters', '<s>the film , the farlating , be and the and pore is ', '<s>the actor , the the remod sore film a come the the', '<s>the pary , a the plontion and a rest , as the the ', '<s>the mod and the sperer and and , the the pille and', '<s>the senter of the move least of the the the deanti', '<s>the fare it comat the film the be coming of in it ', '<s>and the be film a comect a a mant sure a mover and', '<s>the the band the more and the perpite a comating s']

['<s>it the sent and in the the and be make and the the', '<s>sore a film .</s>ary the movie .</s>were the could the c', '<s>the somes the film , a make the mories and sears a', '<s>the sever .</s>and comedy the the perferfing that the', '<s>a the be a sumedy is lith of be a so a dessere .</s>a', '<s>the preating and in a movie and the strome the fil', '<s>the has the the seat and in the sexer and a story ', '<s>the is and the winte is and it sellic to a stist a', '<s>a for the sill , the comere and a decicomet , a fi', '<s>the the lent the so coming and a sing entere the c']

['<s>a film the a film a spines comenting film and stor', '<s>a him .</s>art to the movie the perfict and a movie a', '<s>the intere that the the fars mand actor to surpric', '<s>that be some and the have in the sunding strile an', "<s>it 's and and have to the senting that what the mo", '<s>the we the comedy beap .</s>enge , the still of the m', '<s>the for the movie , and movie and hore and on the ', '<s>a doings and the film the move in the comedy the p', "<s>it 's enterters , the provie .</s>and .</s>ar and the mo", '<s>a comate , the movie fun the to sust the real and ']

['<s>the pretion of the hard it and real movie the mate', '<s>story with the director to part .</s>and could and th', '<s>conderthing and to desiging , and the cent and the', '<s>the movie the film in the bele and the movie .</s>and', '<s>the movie the remant and the incere the comesting ', "<s>it 's but the the film and be is should the maning", '<s>the film .</s>and is a film , comedy a stare and and ', '<s>a donely for a destion with a desents and the the ', '<s>a the care .</s>and not the completter and work the f', '<s>the characction and the chare in the comedy is a re']

['<s>the port to a and the love of the the is in the pe', '<s>the plot and much the movie the movie the word .</s>a', "<s>a compless it 's suth all sonse and menting the mo", '<s>the sees in the movie .</s>and recoming , the feen co', '<s>a presting and with the performance , a story of t', '<s>the starn to the the and remach of the master .</s>an', '<s>like the sented .</s>aker .</s>.. be instare is and the ', '<s>a character of come in the film , and the gire in ', '<s>the movie the conting the the film is a real the f', '<s>a film of a sean the make the film .</s>and character']

['<s>a movie .</s>an and an and the film the the film .</s>ag', "<s>it 's the restic to a like it 's decase .</s>. in the", '<s>the tare and the time of strack to comedy , and an', '<s>the make some the staral with the complict .</s>an we', '<s>but the masters .</s>and .</s>.</s and and movie and dark', "<s>it 's mender , but the sear not not explice is sur", '<s>the manical .</s>and is its bad the slight , and come', '<s>a director .</s>and make .</s>... a provication and and ', "<s>it 's a will this movie mary comedy the concender ", "<s>it 's reart and reciter and prade .</s>act and carabl"]

['<s>the a suppition .</s>.</s>and of the the proved and an a', "<s>it 's dening that the to a some , the had a film .", '<s>the comerent is a caming .</s>.. like the prostent of', "<s>it 's a a susting the the film , and in it 's a co", '<s>the starts , but the movie the plot comes and a fi', '<s>a film .</s>.. has a care .</s>... some and one .</s>... be', "<s>a sing to film .</s>.. and it 's a film , and the fal", '<s>the mater that he consister and a film fast to fil', '<s>the still and the more that with a movie a suberic', '<s>a comedy for the film .</s>and .</s>.. show .</s>... be a g']

['<s>a character and and do some and destintly of the f', '<s>the family site the prosess , and the mast and an ', '<s>a movie is a bad is with the makes piction present', '<s>a film .</s>... and a disting that with a story and t', '<s>a most and the consentic .</s>.. had the film story a', '<s>a dienness .</s>... seen with a his in the disting .</s>', "<s>a bast amour and like the stricks , it 's some , t", '<s>the sever .</s>.. chare .</s>. is a contertain and the t', "<s>the be a film 's a with a seric and an ender and t", "<s>it 's a the predect and her funny protes .</s>.. and "]

['<s>the film of the portor probed of the movie .</s>... a', '<s>the performance of the pave

the increent have movi', '<s>a the familial to the deliver .</s>... some makes a m', "<s>it 's should story that with the character and the", "<s>a movie 's maning .</s>... a we 's a and see the to t", '<s>the movie and a and and the movie of the lack is a', '<s>the way the parts and distally look .</s>. the movie ', "<s>it 's lame .</s>.</s>. a complain of a the film and exce", '<s>a compers to recall and the for the the picture an', "<s>it 's a compentieles .</s>.</s>. a performance and the c"]

['<s>sond of the started and in the movie .</s>. and movie', '<s>the movie that the portory that singers the perfor', "<s>it 's deliver .</s>. see a sual it some and disting a", "<s>the film that 's a deserise to script .</s>.</s>. the fi", "<s>it 's a an and it 's a consistent that it 's event", '<s>concerent visual the rare be distally film and a s', '<s>and and has comedy of for the comperses and intere', "<s>the characters in a the distering movie , it 's be", "<s>the probly .</s>. so movie .</s>. n't the most to sees a", "<s>it 's a story , and a film .</s>. 's no this does be "]

['<s>a consistent .</s>. and every the well the some the l', '<s>the susts that should the the pretical .</s>. plot a ', "<s>it 's a contomating and startic that it 's a dilli", '<s>the some the that the surpront film , but the film', '<s>a stand and a script .</s>.</s>. the sast .</s>.</s>. a spect ', "<s>it 's a be has a gener .</s>.</s>. and a can was a roman", '<s>the story the be pretent performances .</s>. definiti', '<s>the work and a look .</s>.</s>. the work to movie and th', '<s>the self-pleation of the movie .</s>.</s>. the movie .</s>.', '<s>a care .</s>. and a strearing and a does all the movi']

['<s>a carse .</s>.</s>. a set of the film of the story .</s>.</s>.', "<s>a movie .</s>... so has story .</s>. sart .</s>. '' bare an", "<s>it 's characters that a be a dark story .</s>. be a m", "<s>the senting , but the movie is a film .</s>... it 's ", '<s>the presising and the film .</s>... the plot has stor', '<s>the movie .</s>. and as the film and a starty , seet ', '<s>the interester in a conting , but an at the film .', "<s>a presently the script 's a comedy and over and th", "<s>an in the scare .</s>.</s>. it 's a an adady .</s>. .</s>. so ", "<s>it 's a the the surprent .</s>.</s>. a story for the com"]

['<s>a the film despective and movie , and a comedy of ', "<s>the point .</s>. seem , it 's a charm and the stant ,", '<s>a complessed of the story and story , but the feat', "<s>the they , it 's make of the movie is the tries as", "<s>the sental in the prosessed to astantic and it 's ", '<s>make an indection .</s>. it feel story of an and and ', '<s>the probed the movie the movie is a proves , the p', '<s>a what and should by an enough and even the movie ', '<s>the but the mand and the still in the movie and re', '<s>a movie .</s>. some and the plot a film is so wardar ']

['<s>an and and the still to comedy and attime and ever', "<s>call is a lot the movie to , and a some that it 's", '<s>a not site and the created like the cears .</s>. a st', '<s>the manical .</s>. a does as its invinces the present', "<s>it 's a film about the movie that with the vile wi", '<s>the movie and herrish the terrow to so the bad a b', '<s>the feen .</s>. could and the script .</s>. grat to an a', '<s>a see be characters of the fation .</s>. the start , ', "<s>for the film 's movie and reality .</s>. a still and ", '<s>a screen to stand , the pressiver , and it decome ']

['<s>the film the to from the the stand in a show of a ', '<s>the start the part that .</s>. a seem a compled , and', "<s>the most than the film that it 's a performance , ", '<s>the film of the restor to a mander .</s>. a be film s', '<s>story .</s>. a some and a complather

```
.</s>.  so prement o', '<s>a a simply see movie .</s>.  or beat all see a matter ', '<s>a
makes the film is a with the hour .</s>.</s>.  an inter', '<s>an about a strong to a supponce
.</s>.  is a start , h', '<s>a film .</s>.  and an ord a crome and a film , the scr', '<s>the
speces and the script .</s>.  beat .</s>.  a hard worl']
```

Q3: Comparison to feed-forward language model In 2-3 sentences, please explain what differences you see in the text generated by this LSTM language model and the feed-forward language model that you trained in HW5. What do you think may be causing these effects (or lack thereof)? [5 pts]

Answer:
The Feed-Forward generation from HW5 appeared to have produced longer sentences that felt, at least to me, like they were forming a larger structure. In this LSTM model, we see much shorter sentences produced, but more accuracy on the word formation, with less nonsense words than those produced by the Feed-Forward model. I would attribute this to the LSTM's ability to look at larger contexts as opposed to the Feed-Forward model's strategy of simply producing the next most likely character.

4 Testing your code

In the dropbox folder for this assignment, you will find a file `test_all.py` with a few very simple unit tests for the methods that you need to implement. You can verify that your code passes the tests by running `pytest` from your code's directory, with the course's conda environment activated.

Submission Instructions

In your submission, include the following:

- readme.(txt|pdf) that includes your answers to §1 and §3.
- `hw7.tar.gz` containing:
  - run_hw7.sh. This should contain the code for activating the conda environment and your run commands for §3 above. You can use run_hw2.sh from the previous assignment as a template.
  - data.py
  - run.py