LING 575K HW7

David Luongo
Due 11PM on May 19, 2022

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]

  The input batch would be a tensor with 2 sequences of one-hots corresponding to the tokens in the sentence. The dimensions would be [2, 6, vocab size]

  The target label for each input word would be the next token in the sequence.

- 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]

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

- 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]

$$\begin{bmatrix} 0.1 & 0.3 & 0.2 & 0.4 & 0 & 0 \\ 0.2 & 0.6 & 0.1 & 0.8 & 0.9 & 0.4 \end{bmatrix}$$

- 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]
  It would not zero out the losses corresponding to the PAD tokens. It might learn to to put PAD tokens in at the end of sentences where we don't want to see them in the output of the model.

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]
  Lower is better because it means the model is less "confused" by the corpus.

- 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]

  $\Pi_{t=1}^{N} P(w_t | w_1 ... w_{t-1})$

  t=1 is a special case, because there is no previous token. For t=1, the right side of the conditional: w_1 ... w_{t-1} in the above formula is equal to 1.

  This is different than feed-forward language models because feed-forward language models only calculate probability based on the previous n tokens (where n is the n-gram hyperparameter for the FFLM) whereas recurrent models calculate probability based on the entire previous sequence (vanishing gradient problems aside).

- 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]

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

- 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]

  per-word cross entropy

- 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]

  Perplexity would be lower for the model with the smaller vocabulary because the lower probability tokens have had their probabilities consolidated in a sum which is the probability of the UNK token. Without the UNK token, these probabilities are combined with a product which means the probability term in the perplexity calculation would be smaller. Since perplexity takes the inverse of that term, a product combination would result in higher perplexity. This means the lower perplexity UNK model would perform better since it would be less confused by the corpus.

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]

```
'<s>the when of sout and and compesting and sopting an', '<s>is it here
   .</s>ere of so comes in his sentere .</s>ell ', '<s>and and the
   deentanter and and the a the come sunc', '<s>a ward in ard the the the
    and and the is some the ', '<s>and the that the in the sonts a the
   comanter the s', '<s>a stort the boring the and beand and and , ho of
   t', '<s>this a dout , stort a dout , .</s>ad not a the the th', '<s>a
   mover with and wan comes and the crementing an a', '<s>the mest is and
    in the the movere an pave and be t', '<s>a .</s>ray , the sout
   sunting , an in and the for the'

'<s>should and surbeter the cane in and the soming in ', "<s>enthat the
   sade .</s>it 's leal the move the stor .</s> ", '<s>a care a movie a
   film of the sturition the sime .</s>', "<s>and a chored .</s>o one a
   story .</s>er makes it hat 's ", '<s>a to sand the movie , of the some
    stirk comen .</s>a ', '<s>the movie story , the movie .</s>reas .</s>
   reas and a th', "<s>a fant the film the plon 's movie , in the more co
   ", '<s>the movie .</s>ar not the this .</s> , story and gener ,', '<s>
   a this comenting the sire and a that the come and ', '<s>a no scrare
   .</s>as with and and the move .</s> so ming '

'<s>a for pore .</s> and a movie all with make appaniting', '<s>a offer ,
    and the seel of the the male the movie s', '<s>the some in a story to
    be for really movie the sho', '<s>a but has enteres , and be the
   still of chare and ', '<s>a for the not in the film a matter a the
   seen the ', '<s>the more in seen movie has stirch the not is sere ',
   "<s>a some is it 's a love and story and the some and ", '<s>the send
   the the film in the film , the mane the a', "<s>as n't .</s> the
   strant , the the senter film the fil", '<s>the film a slale in the
   some a film it a some is a'

'<s>a funny the litter the experition and love .</s> and ', '<s>a mane
   and the somately to story movie of the film', '<s>the superic .</s> to
    the simple , and the sime , the ', '<s>the faring .</s> movie of the
   with the shean , the ca', '<s>the surperic .</s> the sill , in a lot
   and a that mes', '<s>the movie that the sile strange , and the long of
    ', '<s>a strally for the to some the propaning .</s> more .</s>', '<s
   >a real and the simp film the comening the semen th', "<s>a the
```

```
   propering of the good the lear .</s>. mind 's f", '<s>a like the sick
   to chared .</s> movie prosting and be'

'<s>the sees and the surplay .</s> the the story for the ', '<s>the movie
    set and film .</s> ... a really for a some ', '<s>a life , and many
   stard to story the porting .</s> mo', '<s>the subsite about explige to
    be whe film and be is', '<s>the interelition to the deal and should
   spire , wh', "<s>the fare does it 's a strong and the fantion of th",
   "<s>a mart the some .</s> it 's speen it 's simple .</s> goo", '<s>a
   chare .</s> who is an the story is a film what with', '<s>a movie the
   some .</s> movie the movie .</s> disal and t', '<s>the that a the than
    film and a sure to some .</s> ...'
```

Epoch 19 dev loss: 1.732304573059082; perplexity (nats): 5.653668403625488

In the generated text, we see many non-words, and other tokens that are nonsensical. We see EOS and BOS tokens that aren't necessarily paired. We also see repeated words such as "and" that are repeated back to back which we would probably never see in a genuine natural language context.

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]

```
'<s>a monater .</s> the and a a the film the and story an', '<s>the rear
   and and the sting in the make and and on ', '<s>a comportion of the
   sear the with it , and the som', '<s>the not that the movie and and
   see that the little', '<s>it of the neth and mand the star , and story
    is an', '<s>a shene the film , a movie the deally and medion t', '<s>
   should that the film and the dumbing and sense mov', "<s>a a a comping
    's the film the .</s> about in a loy an", '<s>it in a the stire and
   the make .</s> the movie whe th', '<s>a movie and not , the love .</s>
    and to a ever the mo'
```

"<s>the film 's the better .</s> hord so story the movie ", "<s>it 's comper movee and see that the reall and the ", "<s>the sunder and the film .</s> it 's and bar good with", '<s>the stire and a care and the film , a finn , but a', '<s>for the staring stoly .</s> bitting of so story to an', '<s>the film and the film be is and make and the film ', '<s>the stranged .</s> and so sere for the heart , and a ', '<s>even the the movie , so sheated it at like the mov', '<s>the in seem .</s> .</s> a movie .</s> and stoon , sonding .', '<s>whith a director the directure .</s> the film , but t'

'<s>the rear and movie of has his the way with the con', '<s>the does both the such and a movie that in a broo ', '<s>a film is the more and she the remas at shot that ', '<s>the stroving portory mide of the to a with the fil', '<s>a movie in the with a the work .</s> .</s> and a a be re', '<s>a with a conture of the there and the well a movie', '<s>the movie in a film and the matter of sees and be ', '<s>an and partion .</s> and a peative for a with a rome ', '<s>the a movie of the film with a movie and a less fr', '<s>a do the the movie , but expleal to the porting .</s>'

'<s>a despection .</s> that the rean , the senter work .</s>', '<s>the roment and a movie and a famation the fare for', '<s>the story , but in good and a tirecting the scene ', "<s>a with a look is n't the stract , something of the", '<s>the movie the there to the film .</s> in the ther hol', '<s>the direction .</s> .</s> .</s> and a story of a more it se', '<s>the senter shew and bad film and a the movie of th', '<s>a compictic is somether and the plot and the film ', "<s>it 's matter and character and disture and bere th", '<s>the with a sent make the movie and the start of th'

"<s>the movie , it 's show the comedy .</s> and the comed", '<s>the fan of the real and not the distare the the se', '<s>the to the film that work and a comedy .</s> for the ', '<s>a film the and movie .</s> of the laugh .</s> .</s> and in ', "<s>it 's a some the respectable the sees and a movie ", '<s>the virtion .</s> in into a be and so a film .</s> to se', '<s>in a the the take that a to a will the from recent', '<s>the wart .</s> fatent and spection the film .</s> and ev', '<s>the movie .</s> .</s> but a for stranged that a divertio', '<s>a film of one of a the film of the to story and a '

"<s>it 's a film .</s> performance and the tire the distr", '<s>an and a live and the straction , and the tragin ,', "<s>the film of it 's the real the senter .</s> and the l", '<s>the chance that .</s> .</s> , and a man are the film and', '<s>the stend of the to the coment .</s> and a resome cre', '<s>that the character and start .</s> manding the film t', '<s>a movie and the film and a to film to interest and', '<s>a become of the movie that show and portright is t', '<s>the contright and a there and a real the movie and', "<s>it 's a culture of concance not the start and an i"

’<s>a comes .</s> and and so a compell that a stally comp’, ’<s>the character and both the film .</s> .</s> .</s> is the st’, ’<s>a movie to see on the see .</s> .</s> and the the film a’, ’<s>a for a movie .</s> and stand .</s> a starding the comed’, "<s>the story and a performance , and it ’s so his fal", ’<s>a movie and the film .</s> and the the some start .</s>.’, ’<s>the stror shelf of the charmally and movie .</s> and ’, ’<s>the compleal to white , deness of the film to stor’, ’<s>the tale .</s> and the strange and so the surprise .</s>’, ’<s>a film .</s> .</s> and unconfict that it seen , shell wa’

’<s>funny and staller that shell .</s>. the movie with it’, "<s>it ’s deere a the end , many as the and the care a", ’<s>the movie with the movie .</s> .</s>.</s>, and a movie .</s> .’, ’<s>a comedy film .</s> an onese .</s>ne a film comedy .</s>ne ’, ’<s>a senter contriction .</s>ne film lame and interestin’, ’<s>whether .</s>. but a film and a movie that the movie ’, ’<s>a this love with the the character and something t’, "<s>a with film .</s>g’700001? intensable .</s> an and a wit", ’<s>a fanger .</s> a film comedy .</s> is the matter in it .’, "<s>but it ’s a film .</s> of the little and its time of "

’<s>the movie , the sacker is surprising , but the sel’, ’<s>an and a the film that on a shaper , the story , t’, "<s>the end and so work the suspert ’s stand the be se", ’<s>it make the see in its the story and denession and’, ’<s>the love and the start .</s>. comes of a film charact’, ’<s>a character that to see something the to chers for’, ’<s>the movie and an and a completer .</s>bother the come’, ’<s>the starger that the movie that the unterstate and’, ’<s>a film of the story and story .</s>bit is suce and a ’, "<s>a lation famility time and the film ’s are a story"

’<s>the charmance and the you like the comedy of the m’, ’<s>a movie of the a the the story and story and see t’, ’<s>a film .</s>nest that a the movie and the make a be t’, "<s>it ’s a see the film , in the interester the movie", "<s>it ’s an action .</s>’0‘ the sure in its some the sta", "<s>the sered with the fans .</s>nel .</s>’ steration have t", ’<s>the senter the feel to intertating and the than .</s>’, ’<s>the the lark , in a comedy and of the movie .</s>noth’, ’<s>a movie and a time , and characters , in the there’, ’<s>of see performances that many with the interesting’

’<s>a much a lange .</s>. the and a not better but it the’, ’<s>screen the and a beant the sintation of how the di’, "<s>a stranger .</s>g’’ of the tragal and cast .</s>. the ch", ’<s>the movie and the sease .</s>. the startic laugh of t’, ’<s>the from the sturt and the some many and gar , and’, ’<s>the some comes and intertating .</s>be in with the fa’, ’<s>not interested propert the director sered and make’, "<s>the comedy .</s>’s some in intertainly the the sumpic", ’<s>the comedy , movie and the faming and be a film ma’, "<s>a documentally be is the respection .</s>’s the so be"

```
'<s>director seem and be a mest a genoler is a villing', '<s>the proves
   and the so .</s> .</s> and a comes to a serst', '<s>the movie that the
    some it work of its are a movie', "<s>it 's the the can in surprane
   .</s>. , in the movie t", "<s>a coming is a tale , and it 's delight
   .</s>. the sto", '<s>a movie .</s>never of a parts that a make a with
   a th', '<s>a prediction of contentions and this that somethin', '<s>
   the respection .</s>. should the movie comanic and an', '<s>the
   sensing and startion .</s>s the script .</s> .</s> a pe', '<s>the
   piece .</s> .</s>. neife and the the entertain of th'
```

Epoch 42 dev loss: 1.6598708629608154; perplexity (nats): 5.258631706237793

For my hyperparameters, I increased the number of epochs to 50, the hidden dimension to 50, and the learning rate to .004. It seems like the words are longer (there are fewer spaces). It seems like some sentences are being completed with a BOS and EOS. Overall, it seems like the intelligibility of it has not improved. It does seem that in both runs, the text conforms to English orthographic patterns in general.

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]

HW5 model had no BOS tokens at the beginning of sentences. This is because the source code did not include it. In the feed forward model, we see more valid in-vocabulary English words. This could be because of the difference in "memory" of the two types of models. N-gram FFNN would likely generate shorter words, which is what we saw in HW5. However, since the RNN has memory further back, it might be likely to generate longer words based on its previous memory. However, producing longer words may be more likely to cause errors due to the vanishing gradient problem and the lower frequency of longer words compared to short words in English.