LING 575K HW5

David Luongo
Due 11PM on May 5, 2022

1 Understanding the Feed-Forward Language Model [20 pts]

Q1: Architecture You can find a description of the model in the second half of the slides from lecture #6. [12 pts]

- How many parameters are there? Please write your answer in terms of the following quantities: d_e, the token embedding dimension; $|V|$, the size of the vocabulary; d_h: the dimension of the hidden layer; n: the n-gram size, i.e. how many previous tokens are used as input to the model. [Note: you may assume that there are no "direct connections" between the embeddings and the final layer.]

  $|V| + n * d_e * d_h * |V|$

- A traditional n-gram language model estimates probabilities $p(w_t|w_{t-1}, \ldots, w_{t-n})$ using counts from a corpus. How does the feed-forward language model compute this probability? Answer with a sentence or two describing the overall computation.

  Words are input as a one-hot vector which is the size of the vocabulary into a shared lookup table to get word embeddings. These embeddings are concatenated before being passed through a linear (according to the weights between the embedding and hidden layer) and tanh function in the hidden layer. Finally, the tanh layer is passed into a linear (according to the weights between the hidden and output layer) softmax function which is the size of the vocabulary

- What is a major advantage of the feed-forward language model over traditional n-gram models?

  The use of embeddings instead of simple vocabulary indices allow the model to generalize what it learns to words that are similar to words in the training data.

  There are also a significantly lower number of parameters.

Q2: tanh The model uses the hyperbolic tangent (tanh) activation function, defined as: [8 pts]

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

- Show that $\tanh(x) = 2\sigma(2x) - 1$, where $\sigma(x)$ is the sigmoid function.

  $\frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^x * (1-e^{-2x})}{e^x * (1+e^{-2x})} = \frac{1-e^{-2x}}{1+e^{-2x}} = \frac{1}{1+e^{-2x}} - \frac{e^{-2x}}{1+e^{-2x}} = \frac{1}{1+e^{-2x}} - \frac{e^{-2x}+1-1}{1+e^{-2x}} = \frac{1}{1+e^{-2x}} - (\frac{e^{-2x}+1}{1+e^{-2x}} - \frac{1}{1+e^{-2x}})$
  $= 2(\frac{1}{1+e^{-2x}}) - \frac{e^{-2x}+1}{e^{-2x}+1} = 2\sigma(2x) - 1$

- Show that $\frac{d}{dx}\tanh(x) = 1 - \tanh^2(x)$.

  $\frac{d}{dx}\tanh(x) = \frac{(e^x+e^{-x})(e^x+e^{-x})-(e^x-e^{-x})(e^x-e^{-x})}{(e^x+e^{-x})(e^x+e^{-x})} = 1 - \frac{(e^x-e^{-x})(e^x-e^{-x})}{(e^x+e^{-x})(e^x+e^{-x})} = 1 - \tanh^2(x)$

2 Running the Language Model [15 pts]

`run.py` contains a basic training loop for a feed-forward language model, which will record the training loss and generate text every N epochs (controlled by the flat `--generate_every`, set to 4 by default).

Q1: Basic parameters Execute `run.py` with its default arguments. Paste below the texts that are generated every 4 epochs. 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]

"it 's sting and and in is a sure that the film the", 'the movie the rest of the the movie of the port an', "it 's of the seart , and a how be a surping the pr", "last could better and the say and back the every '", 'a would have a care .</s>make a movie some the pros t', "it 's a belies .</s> the plistration with the story w", 'the a look a poot the some of the ereated .</s>mone o', "so has shere 's freth the movie work the start of ", 'share of the content and of the most of the expict', 'a movie is of the there spectice of the sore .< $/s¿$men'

'a story .</s>l movie , and the that the story of the ', 'a senter as a poles the baristic constical proces ', 'the self it is stire and of the some of the real t', 'a film that makes a come .</s>d a film that is a poti', 'a movie that and some the suse to the movie and in', 'the film is a seally apper the melication and and ', "a conter and lomentic andly and poot 's and the fi", "a doind of the story and it 's lome that mast the ", 'the screeming and the movie in it a stary .</s>ster ,', "it 's a work and comedy and the movie seel to real"

'a story .</s>ll be a bore that does for a convers to ', 'a surper some in the movie story .</s>stage portich t', 'a with a fance about the story to be a deles , and', "it 's a stare of the seen so be .</s>stage of the mos", 'the film is a film a cone that is made a frith the', 'a movie is a streen of the make a movie is a the h', 'the action to spious and a story the film in the c', 'a too proce to the movie is a screen to can are sc', 'a feel be wreling and here the into the batting th', 'the film and a movie in the story .</s>stered a film '

'some , but it is a not of the best in the film and', 'the film is a film the movie that year its film .</s>', 'a great the best for the spich the one , and the p', 'the movie is a screen the director comedy as a com', "it 's work the directed the film that is a with a ", 'a vight and not in the film it .</s>stentions that ma', 'the film is a stark is the movie of the film that ', 'not belies and doing , and offers that star film ,', 'the plot and soud , and the film and dear , and th', 'a bast and such the movie is a big cars of a melie'

Overall, this text has mostly recognizable words and some coherent chunks. The puntctuation seems a little strange in some places, such as beginning sentences with ', and extra spaces between commas as in 'the plot and soud , and the film and dear , and th'. There are some also non-words throughout. This could be expected, because while a prev chracter parameter of 16 allows for modeling whole words and space in-between words, sometimes there would lack a clear, coherent way to make a string of previously generated characters continue in a coherent clear way.

Q2: Modify one hyper-parameter Re-run the training loop, modifying one of the following hyper-parameters, which are specified by command-line flags:

- Hidden layer size
- Embedding size
- Number of previous characters (i.e. n-gram size; this is `--num_prev_chars`)
- Learning rate

- Number of epochs [in particular: making it larger]
- Softmax temperature. (We did not cover this in class: higher values of this temperature make the softmax probabilities more closely approximate arg max, while lower values make it look more and more like a uniform distribution. A value of 1 is the 'default' softmax value.)

Include your model's generated texts here. In 2-3 sentences, state exactly what hyper-parameter change you made, and what effects (if any) you see in terms of the text that the model generated. [10 pts]

For the following, I changed the number of previous chars from the default 16 to 5:

"and some hor the film , and the story .</s> it 's a p", 'a class the screen interesting to seen in the film', 'a stict is a not the prodience .</s> a stuck .</s> and t', 'the kill of the strong human of the film that a de', "it 's a film .</s>ser the beart with a sare in the wi", "the film of pert 's some of the film with a strati", "the movie that it 's there and feen intere .</s> the ", "the movie the better that it 's surider the prosty", 'a sunder care and some .</s> and the some .</s> the kire', 'a competer character and of the film the film , th'

"the bear with all the story .</s>solven or to it 's s", 'a stack as it is a haution .</s> the characters to th', 'a can and who stally story .</s>stance the film the s', 'the film .</s>st humbly , the fun in the film start a', "a strout film and here , but the film , but it 's ", 'a strecters and a more that it all the fincing as ', "it 's beat it a good not is a provision and and th", 'the movie the experies the movie that the movie th', 'the film .</s>ser that it a prove , film that has scr', 'the movie is a comedy of a stand a sture to contra'

'the film beludience and the story , and strous , t', 'the bears and the movie a have to chese and up tha', "it 's a performances to a movie of the least of th", 'the character the film as a down , the story , his', "it 's a have sare to the belation and the stary an", "the movie is n't the film the characters and fill ", 'the film the movie the film the movie the movie th', 'the movie that make a low the want shorking that t', "it 's a director the show man suck , but it a does", "it 's a constent , and the portrising the movie .</s>"

"a surprising , and that it 's a this movie , and t", 'a shork action .</s>ling that it the film the part an', 'the film and seems to show man with a great and in', 'a surprise is a movie the movie that it a can ente', "an every and the screen just that it 's a rear .</s>s", 'the end the work .</s> the enough .</s> the film .</s>sting', 'a movie film the film .</s>sell .</s>just enough the mov', 'a film and the film .</s>sting the promedican sit of ', 'a cripen .</s> and character and with the least the s', "it 's a low and that 's sentime to be a better the"

With only 5 previous characters, we see the coherence of the model suffer: There are even more uninterpretable words like "'a shork action." There are relatively few strings of words that even resemble a coherent sentence fragment. This makes sense, as generation based on the 5 previous characters would not be able to model larger words or even multiword sequences when the words in question are short.