Michael Pollack: LING 575K HW5

May 2, 2024

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.]

  Answer:
  You can calculate this by looking at the connections between the layers of the model:

  First, between the input and the embedding layer. The parameters here would contain the size of the vocabulary times the embedding dimension. Therefore: $d_e \times |V|$

  Next, between the embedding and the hidden layers. This would constitute the hidden layer dimension times amount of projection layer embeddings, which itself is dictated by the n-gram size and the size of the dimension of the embedding layer. Therefore, this would be: $d_h \times nd_e$

  Finally, between the hidden and the output layers. This would contain the size of the vocabulary times the size of the hidden layer's dimension. Therefore: $|V| \times d_h$

  Altogether, this gives us $(d_e \times |V|) + (d_h \times nd_e) + (|V| \times d_h)$ parameters in our model.

- 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.

  Answer:
  This probability is computed by inputing the sequence $w_{t-1}, ..., w_{t-n}$ as embeddings, which are fed through hidden layers in a neural network. The final output of this network apples a softmax function that produces a probability distribution over the vocabulary for w_t.

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

  Answer:
  They have significantly lower parameters due to "low"-dimensional embeddings and embeddings enable generalizing to similar words.

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.

  Answer:
  Alrighty, let's do this.

  First, let's look at our goal. By definition:

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

  Now let's get there. By our knowledge of a sigmoid function, we know that:

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

  Therefore:

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

  $= \frac{2}{1+e^{-2x}} - \frac{1+e^{-2x}}{1+e^{-2x}}$

  $= \frac{1-e^{-2x}}{1+e^{-2x}}$

  $= \frac{1-e^{-2x}}{1+e^{-2x}} \times \frac{e^x}{e^x}$

  $= \frac{e^x - e^{-x}}{e^x + e^{-x}}$

  Therefore:

  $2\sigma(2x) - 1 = \tanh(x)$

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

  Answer:
  Ok, so by the quotient rule, we know that:

  $\frac{d}{dx}(\frac{u}{v}) = \frac{u'v - uv'}{v^2}$

  In this case:

  $u = e^x - e^{-x}$

  $v = e^x + e^{-x}$

  Therefore:

  $u' = e^x + e^{-x}$

  $v' = e^x - e^{-x}$

Applying this to the definition of $\tanh x$ shown earlier would mean:

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

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

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

$= \frac{4}{(e^x+e^{-x})^2}$

We know that:

$\cosh(x) = \frac{e^x+e^{-x}}{2}$

Therefore:

$\frac{4}{(e^x+e^{-x})^2} = \cosh^{-2}(x)$

$= \frac{4}{4\cosh^2(x)}$

$= \frac{1}{\cosh^2(x)}$

Ok, now let's walk backwards from our desired answer a bit.

$1 - \tanh^2(x) = 1 - \frac{\sinh^2(x)}{\cosh^2(x)}$

$= \frac{\cosh^2(x)}{\cosh^2(x)} - \frac{\sinh^2(x)}{\cosh^2(x)}$

$= \frac{\cosh^2(x)-\sinh^2(x)}{\cosh^2(x)}$

$= \frac{(\cosh(x)-\sinh(x))^2}{\cosh^2(x)}$

Hooo boy, ok. Now let's look at that numerator:
$(cosh(x) - \sinh(x))^2 = (\frac{e^x+e^{-x}}{2})^2 - (\frac{e^x-e^{-x}}{2})^2$

$= \frac{e^{2x}+2+e^{-2x}}{4} - \frac{e^{2x}-2+e^{-2x}}{4}$

$= \frac{4}{4}$

$= 1$

Therefore:

$\frac{(\cosh(x)-\sinh(x))^2}{\cosh^2(x)} = \frac{1}{\cosh^2(x)}$

Connecting this with our previous calculations, this shows that:

$\frac{d}{dx}\tanh(x) = \frac{1}{\cosh^2(x)}$

$= 1 - \tanh^2(x)$

2 Implementing the Feed-Forward Language Model [40 pts]

See code.

3 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 flag `--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]

Answer:
Interestingly, it doesn't seem as though the outputs are becoming any more coherent. You can see what looks to be longer sentences, but they still will contain some nonsense words interspersed throughout them.

Output 1

```
['the film so for and funder .</s>n atter and have film', 'the movie the performances and
the .</s>stic .</s>le tha', "the plare comedy 's a film woo sumper it an are th", 'what is
a sting .</s>m it be whis in a mericaule and ', 'an into the soraytic here and the spection
the mav', 'in the movie is a more that in its on the parting ', "the and is a some of the
the perfor ' in the one o", 'the movie is of the comper is an in the make the s', 'a story
.</s>lite what is seep the to in the fire an ', "a film which and a port to decones to the
film 's "]
```

Output 2

```
['a completed to be sumply the post of the film , an', 'a comples the with a see the screen
and stries and', 'the film a story that in the some of the far the f', 'a detritic .</s>m
the strood sempling to in them .</s>le', 'the movie is a not and constrate the materest that',
"it 's a territy .</s>sting and for this the picture .", 'the film is a farting and of the
serion and but an', "it 's the film exprother way comper and some you s", "the story and stanneric
to not has it is n't a mov", "the movie , complete conver , it 's a film that 's"]
```

Output 3

```
["a prot 's word an are sed of the best in the movie", 'this a dinds , the shows when the
story of the pro', "a film 's so plays .</s>n the action the prome .</s> jus", "a director
's mentater to be what make and porth t", "a great and with a sten and the 's a completely
an", 'a compless film .</s>m the sense of the firment and h', "the film is the movie 's most
this is a direction ", 'a sure in the sure of the film with a completented', "it 's great
of the star the proms and came .</s> the ", "the film 's the mere and movie than ever stre
that"]
```

Output 4

```
['an and string that the prome of a something and wi', 'the film of a screen and can arten
be a real in li', "the film 's seen stand .</s>ming of a story and direc", 'a really and dear
the film can of the movie is a s', "whis a sunt and a more that it 's a satic , funny ", 'the
```

```
story .</s>lith the film that make and of the pro', "a far the film that have seen and here
's a seally", "an and the bear , but it 's an a dunny , it 's a c", 'a movie is a - to the
provies .</s>lity in the not yo', 'a performances .</s>m as on the presity on the film t']
```

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]

Explanation:
The hyperparameter that I chose to change was the number of previous characters. I increased this count from 16 to 32. It is hard to tell if this had any significant benefit. The outputs are different, but they seem to similarly feature nonsense words that don't discernibly decrease between epochs.

Output 1:
```
['the film its the far and parter .</s> the cast dive a', 'the work the secout in a far as
a for fal to a mos', 'a makes and the as a so the beal as a surping and ', "it 's not the
movie with some a preating to prout ", 'a see and the sinder and stre .</s> the director to
s', 'a ureal and and seart and a that is the respection', 'a they in the film and of the story
of a here and ', "it 's a but confere that of the film , and rean as", "it 's a stare and
the work not enter , and the mov", 'a searisully surplines and it the film a great .</s> ']
```

Output 2:
```
["the film hare is n't has and a coment .</s> the port ", "it 's the sarding and the comedy
.</s> a partic tome ", 'a film with a sterity , but the sting and the stal', 'the cast and
entert of the end poost the most of t', 'a fer and a fan as a beart to the film .</s> the
plot', 'screen of the film and movie is mast of the film b', 'a movie , and a comper to a
seart a cast and mile ', "it 's a delish , but it 's no has it 's for the mo", "it 's be a
direction to make the seems that is rea", 'the film , and and of the film shough has the seen']
```

Output 3:
```
["it 's not the film .</s> a good for the compless and ", "it 's not heart .</s> has as work
and sucting a start", "what 's so plivent the action in a well the cast a", 'a pleasing the
part of the and pretention .</s> the f', 'a film and the story of the film like a shough the',
'a movie is a from the same a makes a dial .</s> all t', 'the film be seem the story that
```

enter is strical a', "it 's serful the stuget that 's seem is a simmer w", "it 's a bad in the story .</s> see so perfing .</s> a st", 'a for a movie , and a trand , the that is a start ']

Output 4:

["it 's a sinding and plot 's stranger .</s> read lity ", 'a real and ressed formantic most this movie .</s>rang', 'an and despic is a simpless in the film to the eve', "it 's not the stand , with the good and and a chir", 'a kind of the most and suble .</s>. has to show the s', 'the film is a full see proble and story with in a ', "a real as a start of harricaris .</s> the film 's see", 'the enese of the story of the movie to rest the fi', "the 's a fast and ention to be a serong the movie ", 'the part and surprising stand , the real and star ']