Santosh_assignment11

May 31, 2021

Text generation with LSTM This notebook contains the code samples found in Chapter 8, Section 1 of Deep Learning with Python. Note that the original text features far more content, in particular further explanations and figures: in this notebook, you will only find source code and related comments.

Implementing character-level LSTM text generation Let's put these ideas in practice in a Keras implementation. The first thing we need is a lot of text data that we can use to learn a language model. You could use any sufficiently large text file or set of text files – Wikipedia, the Lord of the Rings, etc. In this example we will use some of the writings of Nietzsche, the late-19th century German philosopher (translated to English). The language model we will learn will thus be specifically a model of Nietzsche's writing style and topics of choice, rather than a more generic model of the English language.

Let's start by downloading the corpus and converting it to lowercase:

Preparing the data

```
[1]: import keras
     import numpy as np
     path = keras.utils.get_file(
         'nietzsche.txt',
         origin='https://s3.amazonaws.com/text-datasets/nietzsche.txt')
     text = open(path).read().lower()
     print('Corpus length:', len(text))
```

```
Corpus length: 600893
```

Next, we will extract partially-overlapping sequences of length maxlen, one-hot encode them and pack them in a 3D Numpy array x of shape (sequences, maxlen, unique_characters). Simultaneously, we prepare a array y containing the corresponding targets: the one-hot encoded characters that come right after each extracted sequence.

```
[2]: # Length of extracted character sequences
     maxlen = 60
     # We sample a new sequence every `step` characters
     step = 3
     # This holds our extracted sequences
     sentences = []
     # This holds the targets (the follow-up characters)
     next_chars = []
```

```python
for i in range(0, len(text) - maxlen, step):
    sentences.append(text[i: i + maxlen])
    next_chars.append(text[i + maxlen])
print('Number of sequences:', len(sentences))

# List of unique characters in the corpus
chars = sorted(list(set(text)))
print('Unique characters:', len(chars))

# Dictionary mapping unique characters to their index in `chars`
char_indices = dict((char, chars.index(char)) for char in chars)
# Next, one-hot encode the characters into binary arrays.
print('Vectorization...')
x = np.zeros((len(sentences), maxlen, len(chars)), dtype=np.bool)
y = np.zeros((len(sentences), len(chars)), dtype=np.bool)
for i, sentence in enumerate(sentences):
    for t, char in enumerate(sentence):
        x[i, t, char_indices[char]] = 1
    y[i, char_indices[next_chars[i]]] = 1
```

```
Number of sequences: 200278
Unique characters: 57
Vectorization…
```

Building the network   Our network is a single LSTM layer followed by a Dense classifier and softmax over all possible characters. But let us note that recurrent neural networks are not the only way to do sequence data generation; 1D convnets also have proven extremely successful at it in recent times.

```python
[3]: from keras import layers
     model = keras.models.Sequential()
     model.add(layers.LSTM(128, input_shape=(maxlen, len(chars))))
     model.add(layers.Dense(len(chars), activation='softmax'))
```

Since our targets are one-hot encoded, we will use categorical_crossentropy as the loss to train the model:

```python
[4]: optimizer = keras.optimizers.RMSprop(lr=0.01)
     model.compile(loss='categorical_crossentropy', optimizer=optimizer)
```

Training the language model and sampling from it   Given a trained model and a seed text snippet, we generate new text by repeatedly: 1) Drawing from the model a probability distribution over the next character given the text available so far 2) Reweighting the distribution to a certain "temperature" 3) Sampling the next character at random according to the reweighted distribution 4) Adding the new character at the end of the available text

This is the code we use to reweight the original probability distribution coming out of the model, and draw a character index from it (the "sampling function"):

```
[5]: def sample(preds, temperature=1.0):
         preds = np.asarray(preds).astype('float64')
         preds = np.log(preds) / temperature
         exp_preds = np.exp(preds)
         preds = exp_preds / np.sum(exp_preds)
         probas = np.random.multinomial(1, preds, 1)
         return np.argmax(probas)
```

Finally, this is the loop where we repeatedly train and generated text. We start generating text using a range of different temperatures after every epoch. This allows us to see how the generated text evolves as the model starts converging, as well as the impact of temperature in the sampling strategy.

```
[6]: import random
     import sys
     for epoch in range(1, 5):
         print('epoch', epoch)
         # Fit the model for 1 epoch on the available training data
         model.fit(x, y,
                   batch_size=128,
                   epochs=1)

     # Select a text seed at random
         start_index = random.randint(0, len(text) - maxlen - 1)
         generated_text = text[start_index: start_index + maxlen]
         print('--- Generating with seed: "' + generated_text + '"')
         for temperature in [0.2, 0.5, 1.0, 1.2]:
             print('------ temperature:', temperature)
             sys.stdout.write(generated_text)
     # We generate 400 characters
             for i in range(400):
                 sampled = np.zeros((1, maxlen, len(chars)))
                 for t, char in enumerate(generated_text):
                     sampled[0, t, char_indices[char]] = 1.

                 preds = model.predict(sampled, verbose=0)[0]
                 next_index = sample(preds, temperature)
                 next_char = chars[next_index]
                 generated_text += next_char
                 generated_text = generated_text[1:]
                 sys.stdout.write(next_char)
                 sys.stdout.flush()
             print()
```

```
epoch 1
1565/1565 [==============================] - 201s 127ms/step - loss: 2.2686
--- Generating with seed: "he motley whirl of the senses--the
mob of the senses, as pla"
```

3

```
------ temperature: 0.2
he motley whirl of the senses--the
mob of the senses, as plated and the strenged and the stand and the strenged and
the something the presention of the something the strenged the stranger of the
something the stand of the some the still the something the strance of the
conceation, the strangeration, and the stand of the sense and the strenged and
the same the sense of the still the stand of the stand is is and the strenged
and the stand and all the somethi
------ temperature: 0.5
and is is and the strenged and the stand and all the something and the follning
the ighals in the discruesed to the greative except the same evinion of the
relacting, and who learn, the something the have the sume and form, and regoods
as is more of his of in the something, when a morality strong the
something the stunding the morality of the will stands the power he as to the
cate them the specied the also also not which the some of the
stand whe has whic
------ temperature: 1.0
d the also also not which the some of the
stand whe has which. precessine with eo-gide oftherspe, by the gakned by species
he name has religes but pact, whether--is to us indestoness. leart hremany of
which as nothin mosimenessnesty--if at mank of incallefalition, sensethed away
excepal
compartrally's; that nosore alove
"sperns in agains the gorthile, cturtinabit: what other ouring atnons, from
alove himself.!

        " nomsa: that frutm, of learasory cla
------ temperature: 1.2
ove himself.!

        " nomsa: that frutm, of learasory claps" ids of then itstled,
chuls.=--e:very more deperves, pyemofuress flamedustco of theyekne, which oncem.
gowl: and smculity dpenter
foolewhas to us ans ned noid, in issaliys's for doy rations worm nlthils dried
still highous whith views which
surdnd; and
sompthing-to thothhed, as at knowly ordelen dening hauscrom, the self-prese
knots, he presenties."-it and amend, cbresueveb)r, and germantlesily
epoch 2
1565/1565 [==============================] - 197s 126ms/step - loss: 1.6173
--- Generating with seed: "-charged with interrogative signs--and
often sick unto death"
------ temperature: 0.2
-charged with interrogative signs--and
often sick unto deather of the all the consend to the would be a present the
reason to the sense of the believed and all the believed and the common as the
suble to the been struggless of sufferent to the vistous and say as a power of
the would all the existence and the subter and the sense and say the would the
```

suble to the common to the world as a present the been the wart as the suble and
the same and the have the
------ temperature: 0.5
he been the wart as the suble and the same and the have the are now more ellows
and the states a compare of god intermoral makes and with every the best of the
same about the action of the us to the spirit and more viese the hourselves of a
distime in the more one self-astic and to spectiain the world respect of the
very retood and as a power--the origin, and commous age and vision of
exceplation of every religious intention of all the present himself to b
------ temperature: 1.0
of every religious intention of all the present himself to be kinds, in itself
who maist courdly. the soul us--which condiced
men in
also asciency as megnifulse soothers which ary in
ruled to the coursent should feety and the mothest tenn in
itself into fimure, them more as commance more stragge of respend of the betiess
entermital to himself to what is procauarisions.--no ward
present reself, is itself all comations and not skops, he the emver-yotheity co
------ temperature: 1.2
itself all comations and not skops, he the emver-yotheity condrations with it
loos,
should withing wath tregurely. affearome indyraed tyigh equal
our
wlouconspic votury, lack falh our with necess the gevory
futual juchs: as noer his psicors each on enlocation but one could ourselves.
the mots sinse cariy
chupur moleble (it so frothing
much as to great kisth formorable", as soomy over of onjoraus with ics amy
ondied and,-indiviate-miveds vory one goich ar
epoch 3
1565/1565 [=============================] - 197s 126ms/step - loss: 1.5284
--- Generating with seed: "annihilation, petrifaction, a hatred surmounting
love, perha"
------ temperature: 0.2
annihilation, petrifaction, a hatred surmounting love, perhaps and the self-
conscious and the most the conscious and the stranged and the most the most and
the most selfly the result of the most the rest as a preciation of the superie
and the stranged to the most only the self-conduced be the stranged to the most
and the self and the demistion of the futures of the self proved the rest as it
were the more the more of the most once is the souls and the most
------ temperature: 0.5
the more the more of the most once is the souls and the most also to a germans
it who distive
at mosing to its the duch as it would not itself and germans to be promite
itself--not without the standard and and the restruced the breaked and its of
the higher, the souls and the action is the possestically and heart of the
selfless to be so fact of
the doubt of the day the reals. he would not enthuling and the something that it

is the a destrustic of the supe
------ temperature: 1.0
ng and the something that it is the a destrustic of the supersotful
and intections in the world tib1, of whom even the be has
and amcondity of the wettent, an any.

                              in to bit out is exceptingly and in any respolses of
the souls
asso reeusic
talk, bedy inmay emos
of
the internomution of dipidences entirefulness through the cause of the maint
under to atsucs andisted whreavene his workstis of "welling errlely are
assemuate whoke th
------ temperature: 1.2
vene his workstis of "welling errlely are assemuate whoke they b:iedly but
entifury
with a tempole"--podsis. and
"truth, it comebjues of thot sprinens.
i mode-inived to venu
except with it, or alsampe.
tasted. it maskind, wholvanced w"rigining as bette. thinever eit he world you
of origin the
senses too which every ununempy if alnod to she thing dearly way of
diverseful, that cercoriliaris we ted to have distumbty and should a test
finsidedr, -it migutener
epoch 4
1565/1565 [==============================] - 197s 126ms/step - loss: 1.4811
--- Generating with seed: " gradually got so far,
that one no longer lets the consequen"
------ temperature: 0.2
 gradually got so far,
that one no longer lets the consequences of morality of the propided that the
conscience of its own distinct, and a soul, and a conscience and the present and
soul and a more that the souls and the spiritual and the find and the present
the concerning and the souls of the sense of the self a more pared and the
spiritual to a more man is the sense of the pride of the spiritual and the soul
and the sense and the be the person and a sou
------ temperature: 0.5
l and the soul and the sense and the be the person and a soul, and the spiritual
gain believe that is religion of its who has been that one man as its own soul
as the most preciselve to please of the seems
he is this only the most conssious and intereation of the morality and a thing
of we thouth the unimmose the seriat as an art of the senders of the make a
platour consciention of his was to so deserve of the acts to be acts some man,
the persons of the f
------ temperature: 1.0
eserve of the acts to be acts some man, the persons of the foodeftime fritter,

```
artistms to taste to the seads " that deex--long. this other futs that himself
borier of with -his solong,
the
shord
such ceet elsed bring evolve, inishers of those that lighted femitie, the higher
regard also expecien, a respicis,
in the sublencess is
im always attempte to bes
a
desires of the howel as self-impresspanad are self-converses thare in at harden
is as
science of cor
------ temperature: 1.2
d are self-converses thare in at harden is as
science of corcaer: allowing shurient turded "faidry thathis origine, far
xitable," power' is indicaltous,
nearly turdn andile!
it is an hroilizir of hoge sympaths."", superially;
and
today" of then denieation, efver abrest, to-day plys of ly": poseless-nr"
that have exease detain thar point musth eathing, that art then that is
parature: it is a tender inmores it, so of historical to tastic sparagy yous,
will r
```

As you can see, a low temperature results in extremely repetitive and predictable text, but where local structure is highly realistic: in particular, all words (a word being a local pattern of characters) are real English words. With higher temperatures, the generated text becomes more interesting, surprising, even creative; it may sometimes invent completely new words that sound somewhat plausible (such as "eterned" or "troveration"). With a high temperature, the local structure starts breaking down and most words look like semi-random strings of characters. Without a doubt, here 0.5 is the most interesting temperature for text generation in this specific setup. Always experiment with multiple sampling strategies! A clever balance between learned structure and randomness is what makes generation interesting.

Note that by training a bigger model, longer, on more data, you can achieve generated samples that will look much more coherent and realistic than ours. But of course, don't expect to ever generate any meaningful text, other than by random chance: all we are doing is sampling data from a statistical model of which characters come after which characters. Language is a communication channel, and there is a distinction between what communications are about, and the statistical structure of the messages in which communications are encoded. To evidence this distinction, here is a thought experiment: what if human language did a better job at compressing communications, much like our computers do with most of our digital communications? Then language would be no less meaningful, yet it would lack any intrinsic statistical structure, thus making it impossible to learn a language model like we just did.

Take aways

1) We can generate discrete sequence data by training a model to predict the next tokens(s) given previous tokens.

2) In the case of text, such a model is called a "language model" and could be based on either words or characters.
3) Sampling the next token requires balance between adhering to what the model judges likely, and introducing randomness.
4) One way to handle this is the notion of softmax temperature. Always experiment with different temperatures to find the "right" one.

[ ]: