Assignment_11-text-generation-with-LSTM

May 26, 2021

[1]:
```
import keras
keras.__version__
```

[1]:
```
'2.4.3'
```

0.1 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.

0.1.1 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.

0.1.2 Preparing the data

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

[2]:
```
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.

[3]:
```
# 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 = []

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…
```

0.1.3 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.

[4]:
```
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:

[5]:
```
optimizer = keras.optimizers.RMSprop(lr=0.01)
model.compile(loss='categorical_crossentropy', optimizer=optimizer)
```

0.1.4 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"):

[6]:
```
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.

[7]:
```
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 [==============================] - 206s 130ms/step - loss: 2.2546
--- Generating with seed: "lieve that love can do everything--it is the
superstition
pe"
------ temperature: 0.2
lieve that love can do everything--it is the superstition
pertoring that is some one and some his some and are the sense and some and some
one and the power the religious that the some and in the such and for the power
of the some one some one some his one the some one such is the sense of the
power and the feel his even the some one some the presine in the present of the
religious the some and string and the sense of the sense and such and religious
the s
------ temperature: 0.5
ring and the sense of the sense and such and religious the some one of the some
onger is is the power--and the more never that even the charracted who
groditumed the senst and the such the great is not one are are reflect of the
form that a promory, this stronger one streth and the senscient do gen. the sone
one nother of the supermoral on the really conscience of the moral from one
severed the some be that in religious proble such proposition of the some
------ temperature: 1.0
me be that in religious proble such proposition of the some horebs sychicid,
hithertate--.
in ageaning ans comminially funcued coniele rave: as immearance who feel
perpoddured all inverous, as handnessid be mestasic the sont revenguraiin th.
byst, when that who meast must the considew becaeenedgatily our thinks
conmecours. such a poge nothers he givere
power thancioned flol nothers, one beataked, bodgment and develfixisip
```

```
restaryity, strance of "peinelly,
------ temperature: 1.2
bodgment and develfixisip restaryity, strance of "peinelly, reover
unlestantrress accigiodh
bdvenouble ourimist which ongurly,
whatred
fecess, if the midides for peture is
quical britreaway-hal? a cowe-- houghth, so lessule--as are
ductive one gre, yven to refate still erlured
abvoncy wed
for there doydic ducced complancic, ard" the greempatives at-langering, whenh
peroul, without begoanard gry feeler the-spooms,-njures; as ny rull, spirit,
excosment and f
epoch 2
1565/1565 [==============================] - 199s 127ms/step - loss: 1.6137
--- Generating with seed: "ry inception
is naturally ordered. yet everything evolved: t"
------ temperature: 0.2
ry inception
is naturally ordered. yet everything evolved: the most great the superficial and
desire the more and the spirit the superficial and the something and the fast
the far a man is the man is the same the man is which the for the fat its as in
the demonity of the man and the man the man and subsection of the superficial
the subtle the far and the subtle the fat in the fate the man and the
superficial to the same and the man is the such a consider t
------ temperature: 0.5
superficial to the same and the man is the such a consider there of the
respections the the profound of nature concepsioned of the freedom that he has
more that profound sperial of the perhaps the value and destriction that the
concernation and suberality and the man, it is the same above the many and
diffection and them whether the german should not the science of the short of
the orable the ancient in the are and desire the origin that the conteres this
------ temperature: 1.0
ent in the are and desire the origin that the conteres this as a are down, as if
scorn of magical shemitity of fiells, le exiction
whulh the make love humself a tomeding that must respects, with the shead, as a
heaute the socturitly one is a sapprificate at his age is venien a must spirit
of which being the tamicality of which their really there fat allot onceate
bould of a pertation. a consined and old for thie must be therenored that the
liftatic
vigh wi
------ temperature: 1.2
nd old for thie must be therenored that the liftatic
vigh with cfelsesfocourably it genionation. homburtic quity agaisters, in oreed
ceviliotically may, thragning much new
muspensious (th itsuacity-orneing, affell -grien, from the lite; motele? there
is: intermane's
scarific, hus belongings formsesly
```

```
e
copasion look and the weas
the
alliosed almayed--intellectelity ou ; aunt is
new frled be
other the relicents uine deove plause to justs as one is the hcear
epoch 3
1565/1565 [==============================] - 199s 127ms/step - loss: 1.5277
--- Generating with seed: "he relapses of the convalescent! how it delights him,
suffer"
------ temperature: 0.2
he relapses of the convalescent! how it delights him,
suffering and responsibility of the superated to the discountable and so the
same of the one of the sense of the soul his conditions of the soul and souls
and as the most propers of the conditions of the soul and the soul and the
assumed and more to be a the discounter the sense of the condition of the soul
and soul has as the soul of the soul and the soul and the soul the spirits of
the spirits of the
------ temperature: 0.5
and the soul and the soul the spirits of the spirits of the propers of more
every been the existing and immenses itself will which the spirits of the habits
the philosopher the transcas its other need to more and indity. the any been its
he vasions of the ellow and could not the discontines of the ideal the
interpretation of the conditions of the every sour for the disinglishing which
as is the further product of the process and dielities of the dis in the
------ temperature: 1.0
rther product of the process and dielities of the dis in the
sincenced, whercer insentioned nature. "fhieder. "wholese under-eirlor only but
alnosise: at anyances for more find, at its artlies. on tell passed himself
nothing which uso
quigers, so fastical hourse that he state in
nee'sl.on: wished. princhor supless. that is impurses and
it, which only systeged like: the function alboetement,
ruls--
                roying is conduct this have? likewir an ama
------ temperature: 1.2

                roying is conduct this have? likewir an amatings
surpsed  fat in the triew
give as
then is cadefor, will let its of
noining to do ringeryed blaitisms: could tonate
its pase for indelse social , tepeath."


1tm

i lce, foot,". "earth, gwanked and some findstly:
```

```
all telf,
under of
scclinsars eitherthing imperity. probablies drulunatioc, generaly intruting
beoothed
mirsers, with every histin find co froe always man percoint ty.fur-sursenjs as
epoch 4
1565/1565 [==============================] - 199s 127ms/step - loss: 1.4824
--- Generating with seed: "s of pain
by the performance of acts of sympathy.--with the "
------ temperature: 0.2
s of pain
by the performance of acts of sympathy.--with the superiority and something that
the superiority of the superiority of the conditions of a distinguished and
superiority of the words and soul and soul and the supposing that they are have
the superiority of the superiority and soul and present the superiority of the
superiority of the superiority of the superiority and superiority and something
the superiority of the conditions of the conceptional
------ temperature: 0.5
thing the superiority of the conditions of the conceptional dispense for which
the false--for which has not sequence of the produced to many not the world one
was a fear the truth in the present and soul and art of its so it in the fear,
to his habit, is an and of the last in the characters, that the conceptional
being that is a man with the fair of something from the should of the prival to
upon the facts of present to the fundamentally are all the protec
------ temperature: 1.0
the facts of present to the fundamentally are all the protectional
juquest, to net really be asset partly individual.
theye
once the germans? no just real as astolutafled is samely refine, tide of teetiey
amantives."

1ppations, aghs, tlasi
same which i seableces, that he esse that we have age use we our reparting: in
eternal rest anything to alsooy the falser, he "these orourd to a svofting of
means that enity part
estempy in dispureness, or attuin lide
------ temperature: 1.2
eans that enity part
estempy in dispureness, or attuin lide at althorsch hims put there adicblats as
it is namms in past epht"--will
truth dur,, lace is with which it their any grean people whoen habity. as

factly,
the wholish, it and
sopocosm emectid
no ney, a doon
in which a time soon their extactryes in reproise; but its
of
```

```
dist the take to far which  "for who questike of their
seculty, that the farsage throse, "ie ent may comfestr as exgre better--on
```

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.

0.1.5 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.

[ ]: