Assignment11

August 12, 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))
```

```
Downloading data from https://s3.amazonaws.com/text-datasets/nietzsche.txt
606208/600901 [==============================] - 0s 1us/step
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 = []
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.

[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:

[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 [==============================] - 190s 120ms/step - loss: 2.2394
--- Generating with seed: "es to solve). on the part of
pious, or merely church-going p"
------ temperature: 0.2
es to solve). on the part of
pious, or merely church-going promes of the disting the spirit in the present
and the sense of the prome and think of the present and the distory and thinks
of the promes of the self-concenting things and sense of the condection and in
the man to be thinks in the self-and think of the individual and instinction of
the experies in the been and think of the great of the present of the different
in the present in the promes and th
------ temperature: 0.5
present of the different in the present in the promes and things of the must of
the lears of the present are in the but think as in the specious in the believe
in a such a bene in the many into exeration and all he one of the methous of
must his interrication of this and pose to in the believe in the great at the
spirit and deed may interplience to things a featting to be think of thinks of
this one of self--a think and his such at every senvenow of the ex
------ temperature: 1.0
ne of self--a think and his such at every senvenow of the extepter to alsayd
finded sty hadophou, in think bedicction which their, himolg intend"ursing,
in his condcient of yy advocie leous its hesours of can ourcesse belic- are has
greatt who had
oright
ourable undiefinalgen of do thinkore it thr a concencioually nor the meacly.
rowed we trode: sing
 being he jungerary. and of slank of the fuith he
owh only the good, but a pance of for astolay consequent,
------ temperature: 1.2
he
owh only the good, but a pance of for astolay consequent, to creat which hip
sayppence; meneing latirining them: man,
cormordwithousfm.

b foind. he
whomresthe
to suf,
the indecimitudredare for nem is bne! it a compack that pentive find and only
mush, in b4d athire intile liy tilalia.ute justly til raite godsurex; : this an
aw heurs to ob[jy! itsom(eoutinely a end atele! dovex , a
yoon my himself which
it, hnoral
are mude pariow--orod id
intirive a, do

epoch 2
1565/1565 [==============================] - 187s 120ms/step - loss: 1.6100
--- Generating with seed: "od, who is the sentinel and witness of every act,
```

```
every mome"
------ temperature: 0.2
od, who is the sentinel and witness of every act, every momerned to the
suffering the spirit in the spirit to the sublight to the spirit is a perhaps an
and the spirit in the soul it is the subelest the subeled to the decise the soul
the spirit in the spirit to the spirit to the spirit of the strange of the
spirit and the spirit to the spirit and the strange to the spirit to the spirit
to the spirit to the subeled to the spirit in the spirit in the substio
------ temperature: 0.5
it to the subeled to the spirit in the spirit in the substions, it is the most
prestans the
greater and chorce of him a suffering to be therefore the berefore its an and
with the honomenness to the same
present the great should not supposity
and submit of the lack of the distooler and their something the speak of the
should any them is the belief for preduction of subligned and far to the science
for the
fact themselves to one modern there one is man to th
------ temperature: 1.0
for the
fact themselves to one modern there one is man to the same
itself gow.--tyman eight of the peris the areter of one hand.--austo instincts
of the
frind experied to dilace, would ofor that in
the himmer, by belimation there
indicts are for the  what is the periodue and juits in the necessity--growtimate
spiritual anger if of a wared, hulk us an viegorful very of a man there and that
does the
exists theoself--must be nor he actirates a "too them
ormen
------ temperature: 1.2

exists theoself--must be nor he actirates a "too them
ormentd
will nawlest theosemanant afting, sure such barvance, to degreekg; but
ideag, an pranfingroses, windrehister duferestity, a new
fanicrita.----hutorches mobeler. fulttrodanged, ane
super-itman, like lince him: has sole every growtumeed
percois, we, thereffect comed to as rewaring:e, kfo--there istonaesed
trifey ios, no becoces? ther hentements perialed, hupenced? intragonation, lect
implitions),
epoch 3
1565/1565 [==============================] - 194s 124ms/step - loss: 1.5231
--- Generating with seed: "uper-abundance, the protection are there lacking
under which"
------ temperature: 0.2
uper-abundance, the protection are there lacking under which is the strength and
the the states and the same the present and the sublight of the sublight to the
```

```
state and the spirit and intermition of the sense of the state of the same the
same and the same that is a spirit and the state of the same the sense of the
enthures of the state of the same the man and the the the state of the states
the stand the subjegation of the state of the same as a sublight
------ temperature: 0.5
stand the subjegation of the state of the same as a sublight, the staller there
is a mader, as a soul, the helication of the individual signification of the
serpors, and all fautation and
your destruition of the sublight, in the consequent and the emolity and some
passion, which is the decourse and the the charm of the spirits of the master of
the life of the present which is present which is the destructed have station
that and the soul in the individual
------ temperature: 1.0
destructed have station that and the soul in the individual intersention of
prible, but, that their wother of her-is which, compassion, hraog? though, who
will the wat moralsting
and waw never howretoing that we also the prewarity, learnee tzedelf for a
possess is necessaprespressing that was
human flithe
how have he wimable main there may is all the requirestr of man"--your instincts
it from ran distancian cipal of the sensibility clamible, that he philos
------ temperature: 1.2
distancian cipal of the sensibility clamible, that he philosophazing are in the
valuality- and egoiset, the m: vilove
pry
to the fuith this
in ourtboolozd, who
 vali-mentifoue, that womaticigatic of human moral loys other cladundivess" even
the facition
who
thost: was the table betwould iotherderary.qioge vigwex alwabner
re divid.t"?
lwhyrace: hild, brreamitverath, as that physiows rrbe:tur of
command, sufferity for lixe towaowsted
an cumfelus contemm
that
epoch 4
1565/1565 [==============================] - 200s 128ms/step - loss: 1.4767
--- Generating with seed: "express
accurately what all these masters of new modes of sp"
------ temperature: 0.2
express
accurately what all these masters of new modes of spirit and the subjuctive of
the supersious the also and the supersious the more the more also the world the
contemption of the contemption of the supersious the supersious for the
supersious and some more the supersious more the supersious of the contemption
of the supersious the contemption of the supersious in the supersious the
faculture of the supersious and the supersious and the supersious an
```

```
------ temperature: 0.5
e of the supersious and the supersious and the supersious and supersation and
feeling of a something and all the called. if the world and really the bensised
the same instance, and because the sublime must be the new sthing and who has
the conception of the same order the inclined one more curression of all the
profounded and man long and all its conception of concess and is the starser the
stronger and free spirit which be longer who has a religious and a
------ temperature: 1.0
er and free spirit which be longer who has a religious and among is the feelium,
world, as the guelt gradions, headang, what explaed, the europead, in our
science for the sectede end
grastic word how them astempt to be uponly considerately, the trufferent,
afcorationfound. longer clight of a
falsouses: (as well from a sunes not would
be all hrilly man be before the cormonance of -the
whole op
dingle in mentide," as they appcedvors its when animally to the
------ temperature: 1.2
le in mentide," as they appcedvors its when animally to the turneson, i de bill
the valuability.
suchon evinally if
ye to swhole who strementher, though byleffeliness, this, his tritane open meful
been imbely "threutionfore of curtise).--here: as unmeto and shhupering and
shysimen's dion,
or woman without here berad the maj commened
rotheled advidian magrify schopainly
proupt
supershomlun, that
bencine pleasured to the uniuted in a confisifued."
and othep
```

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.

[ ]: