МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №8

по дисциплине «Искусственные нейронные сети»

Тема: Генерация текста на основе “Алисы в стране чудес”

Студентка гр. 7382 _____ Головина Е.С.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы.

Реализовать рекуррентную нейронную сеть для генерации текста на основе “Алисы в стране чудес”.

Задачи.

- Ознакомиться с генерацией текста
- Ознакомиться с системой Callback в Keras

Требования.

- Реализовать модель ИНС, которая будет генерировать текст
- Написать собственный CallBack, который будет показывать то как генерируется текст во время обучения (то есть раз в какое-то количество эпох генирировать и выводить текст у необученной модели)
- Отследить процесс обучения при помощи TensorFlowCallBack, в отчете привести результаты и их анализ

Ход работы.

1. Подготовка датасета

В качестве данных используем текст книги «Приключения Алисы в Стране Чудес» Льюиса Кэрролла. В начале загрузим текст в память и преобразуем все символы в символы нижнего регистра.

```
filename = "wonderland.txt"
raw_text = open(filename).read()
raw_text = raw_text.lower()
```

Преобразуем символы в целые числа. Для этого создадим набор всех отдельных символов в книге, а затем создадим карту каждого символа с уникальным целым числом.

```
chars = sorted(list(set(raw_text)))
char_to_int = dict((c, i) for i, c in enumerate(chars))
```

Суммируем набор данных:

```
n_chars = len(raw_text)
n_vocab = len(chars)
print "Total Characters: ", n_chars
print "Total Vocab: ", n_vocab
```

Разделим текст книги на подпоследовательности с фиксированной длиной в 100 символов произвольной длины.

```
seq_length = 100
dataX = []
dataY = []
for i in range(0, n_chars - seq_length, 1):
        seq_in = raw_text[i:i + seq_length]
        seq_out = raw_text[i + seq_length]
        dataX.append([char_to_int[char] for char in seq_in])
        dataY.append(char_to_int[seq_out])
n_patterns = len(dataX)
print ("Total Patterns: ", n_patterns)
```

Преобразуем список входных последовательностей в форму [образцы, временные шаги, особенности], которую ожидает сеть LSTM. Затем нам нужно изменить масштаб целых чисел в диапазоне от 0 до 1, чтобы облегчить изучение шаблонов сетью LSTM. Далее преобразуем выходные шаблоны (отдельные символы, преобразованные в целые числа) в одну кодировку.

```
# reshape X to be [samples, time steps, features]
X = numpy.reshape(dataX, (n_patterns, seq_length, 1))
# normalize
X = X / float(n_vocab)
# one hot encode the output variable
y = np_utils.to_categorical(dataY)
```

2. Создание модели

Создаем один скрытый слой LSTM с 256 единицами памяти, слой Dropout с вероятностью 0.2. Выходной уровень – слой Dense, использующий функцию активации softmax для вывода прогнозирования вероятности для каждого из 47 символов в диапазоне от 0 до 1. Так как решаем задачу классификации отдельных символов по 47 классам, в качестве функции потерь используем перекрестную энтропию, а алгоритм оптимизации - adam.

```
model = Sequential()
model.add(LSTM(256, input_shape=(X.shape[1], X.shape[2])))
model.add(Dropout(0.2))
model.add(Dense(y.shape[1], activation='softmax'))
model.compile(loss='categorical_crossentropy', optimizer='adam')
```

Так как обучение проходит довольно медленно, используем контрольные точки для сохранения всех весов сети после прохождения каждой эпохи.

```
# define the checkpoint
filepath="weights-improvement-{epoch:02d}-{loss:.4f}.hdf5"
checkpoint = ModelCheckpoint(filepath, monitor='loss', verbose=1,
save_best_only=True, mode='min')
callbacks_list = [checkpoint]
```

Обучение:

```
model.fit(X, y, epochs=20, batch_size=128, callbacks=callbacks_list)
```

3. Генерация текста

Загружаем полученные веса на итерации с наименьшим значением потери. В нашем случае – это получилась 20 итерация с потерей 1.9508.

```
# load the network weights
filename = "weights-improvement-20-1.9508.hdf5"
model.load_weights(filename)
model.compile(loss='categorical_crossentropy', optimizer='adam')
```

Функция для генерации текста:

```
def gen_text(model):
        # pick a random seed
        start = numpy.random.randint(0, len(dataX)-1)
        pattern = dataX[start]
        print ("Seed: ")
        print ("\"", ''.join([int_to_char[value] for value in pattern]),
"\"")
        # generate characters
        for i in range(1000):
                x = numpy.reshape(pattern, (1, len(pattern), 1))
                x = x / float(n_vocab)
                prediction = model.predict(x, verbose=0)
                index = numpy.argmax(prediction)
                result = int_to_char[index]
                seq_in = [int_to_char[value] for value in pattern]
                sys.stdout.write(result)
                pattern.append(index)
                pattern = pattern[1:len(pattern)]
```

4. Результаты работы

Посмотрим какой текст сгенерирует наша сеть после последней итерации.

Seed:

" turtle in

a tone of great relief. 'now at ours they had at the end of the bill,

"french, music, and "

teat s see seat the was io a letel caar wat oo sore and the whst on whrh the baterp laad oo he the hirte the was so the baterpillar seat the was so the worle the was soenting an inrele

thit was she farter why on tee bate of the brurte of the bourte of the war io a lergl caar wf the that sas the was soe tiitt and the whrl soe want on the whrle thene was soenking an incerinten oa the goure, and the whrte tabbit was soenking an inceling tonne aedute an the fareer

aaald the had been her head het hend the hatter and the whst hord to be a lott oh the whnle gareen whit sn the whst oo the whnle badut on the gart.

'io tes thet ' said the mock turtle, "thel i mo het i te beten a sittle ' said the mock turtle, "thet i moow the seaten saad to tey ' said the mock turtle, "thet i moow the seaten saad

to tey ' said the mock turtle, "thet i moow the seaten saad to tey ' said the mock turtle, "thet i moow the seaten saad to tey ' said the mock turtle, "thet i moow the seaten saad to tey '

Заметно, что много слов, которые невозможно понять, но примерно половина получилась нормально. Есть некоторые повторения, например, «said the mock turtle, "thet i moow the seaten saad to tey '» повторяется несколько раз в конце. Каких-то осмысленных предложений также не получилось.

5. Собственный CallBack

Напишем собственный callback, который будет показывать, как генерируется текст на эпохе 2, 6, 11, 16 и 19.

```
class CustomCallback(Callback):
        def __init__(self, stops):
                super(CustomCallback, self).__init__()
                self.stops = stops
        def on_epoch_end(self, epoch, logs={}):
                if epoch in self.stops:
                        gen_text(model)
```

Добавим в коллбэки, запускающиеся во время обучения сети – наш CustomCallback после 2, 6, 11, 16 и 19 эпох:

```
callbacks_list = [checkpoint, CustomCallback([2, 6, 11, 16, 19])]
```

Посмотрим какой текст будет сгенерирован.

После 2 эпохи:

Seed:

" he king.

'then it ought to be number one,' said alice.

the king turned pale, and shut his note-boo "

the and the

and the and the and the and the and the and the and the and the and the and the and

the and the and the and the and the and the and the and the and the and the and the

and the and

Весь текст состоит из повторений «the and», что не имеет смысла.

После 6 эпохи:

Seed:

" . 'i'm glad they don't give

birthday presents like that!' but she did not venture to say it out

loud "

the woeee to the whe woee

'nha ' said the mone and the har ant toee to the cane and the har an toe toee th the woue and the har and the toeee to the whe woe whe woeee to the whe woe toee

nhe whu ho whe woele wo the toee

'nha ' said the kact and the woued to the toue

nhe sou dnn to the toue

nhe she gotthe dad no the care and the har and the toeee to the whe woee

'nha ' said the kact and the woued to the toue

nhe sou dnn to the toue

nhe she gotthe dad no the care and the har and the toeee to the whe woee

'nha ' said the kact and the woued to the toue

nhe sou dnn to the toue

nhe she gotthe dad no the care and the har and the toeee to the whe woee

'nha ' said the kact and the woued to the toue

nhe sou dnn to the toue

nhe she gotthe dad no the care and the har and the toeee to the whe woee

'nha ' said the kact and the woued to the toue

nhe sou dnn to the toue

nhe she gotthe dad no the care and the har and the toeee to the whe woee

'nha ' sai

Тоже довольно зациклено, но слова уже получаются более разнообразные. Повторяется уже предложение, а не два слова, но довольно много несуществующих слов.

После 11 эпохи:

Seed:

" rude, so she

bore it as well as she could.

'the game's going on rather better now,' she said, by wa "

s an ier faad soe tait oo the catee.

'thet iad tot teen ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toate ' said the caterpillar.

'whlt iin toe toa

Очень зациклено, повторяется одна фраза, но больше половины слов реально существуют и их сочетание имеет смысл.

После 16 эпохи:

Seed:

" eating and drinking.

'they lived on treacle,' said the dormouse, after thinking a minute or two.

' "

the soett oave ' said the mock turtle an a loeke ou toore

andce tas iot a aet and the hors on the soeen of the coore and the woole tas aoonne the woile aadkn soe po the goost and the tese the foest of the boore '

'iec so the sorel ' said the mock turtle

'and the toell sald to the boose to tey,'

'toe mag hot she soale fad setee so the bin,e teme toe kagte' she mart ro the fory on the soos,'

'io yhu ao i can teke toe mage ' said the mock turtle. "whll you to tee to tel the sere to seye

'

'i mone the moes to the moot,' said the monk, and the qoeke suine tone to the tonke

'and toe toilg saadit, io a sore, said the mock turtle an a loeke ou toore

andce had not ao an once an anl the corrouse the woode

and eer aelen toeetinn the mors of the goush soe poeet of the goose ' and the whst hnr lont

the was oo the woole aadk to the courous, and the thite tabbit here to the boom an tee soeke

and the woole tat aoonne the woide she would bale whth the bormouse tf tee to sere the h

Заметно, что теперь гораздо большее разнообразие словах. Повторения есть, но они не настолько очевидны.

После 19 эпохи:

Seed:

" haven't,' said

alice)--'and perhaps you were never even introduced to a lobster--'

(alice began to s "

eay it was toe tiing of the soods so tee beone her

and she was sot then it whsl aerin the whsl would the was soenk ano oo her hn the tine, and she whstght it was soen a aoef aalen and she whstght the was sot a little berer see whet she was soeee ano ou head and whs wely dird and soeee to be an ou toone the whste thene whsh the bare and allnee shth the rase and the west ooce to ce lo here the while gad senerked to tee it

was the whstg the bade she past whsle the while and ier a mettle berer see if the was oo the tinee an her haae, and she test hnrdly to see the white rabbit was soe oint the was soenk ano oo her hnat, and she whstght the was sot a little bro of the sabdet was so tee thete whsh the was soeeking an incer oh the raste

and the whstght the whste tabdi she wound tee bean hn the siodse then so tee bt har on the thne, and she whstght in a lortee of the was soenking an inr lonk, and she thought the was sot a lute tu cear woted at she would her head an the whnte the taste

Получилось много непонятных слов, но, если их немного подкорректировать, получатся настоящие. Довольно много повторений, но и различных слов больше.

Таким образом, получили сгенерированный текст, который далек от идеала, но более менее похож на настоящий.

Вывод.

В ходе выполнения работы была реализована ИНС, генерирующая текст на основе книги «Приключения Алисы в Стране Чудес» Льюиса Кэрролла. Адекватного текста не получилось, но были получены базовые навыки в генерации текста с помощью нейронных сетей. Также был написан собственный callback для вывода сгенерированного текста после определенных эпох для наблюдения за прогрессом.