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. 8382 _____ Кулачкова М.К.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2021

Цель работы

Создать генеративную модель на основе рекуррентной нейронной сети.

Задачи

- Ознакомиться с задачей генерации текста;
- Ознакомиться с системой CallBack в Keras.

Требования

1. Реализовать модель ИНС, которая будет генерировать текст;
2. Написать собственный CallBack, который будет показывать то, как генерируется текст во время обучения (т.е. раз в какое-то количество эпох генерировать и выводить текст у модели в процессе обучения);
3. Отследить процесс обучения при помощи TensorBoard, в отчете привести результаты и их анализ.

Подготовка данных

Текст «Алисы в стране чудес» необходимо представить в виде, который подходит для использования в нейросети. Так как текст будет генерироваться посимвольно, для сокращения словаря сначала все символы текста преобразовываются в нижний регистр. Затем символы представляются в виде целых чисел.

Исходный текст делится на подпоследовательности одинаковой длины. Было решено выбрать последовательности длиной 200 символов. На основании этих 200 символов нейронной сетью будет предсказываться следующий символ. Из последовательностей и следующих за ними символов формируются тензоры входных и выходных данных. Исходные данные преобразуются в форму, подходящую для слоя LSTM. Затем данные нормализуются. Выходные данные представляются в виде векторов длины, совпадающей с размером словаря, с нулями на всех позициях, кроме соответствующей индексу символа в

словаре. На выходе модели будем ожидать вектор вероятностей того, что за последовательностью, переданной на вход, будет следовать каждый из символов словаря.

Функция для генерации текста

Реализована функция, генерирующая символьную последовательность. Для генерации текста сначала генерируется случайное число от 0 до len(dataX)-1. Здесь len(dataX)-1 – индекс последней последовательности в массиве последовательностей, полученном из исходного текста. Сгенерированное число – это индекс последовательности, на основе которой будет сгенерирован первый символ.

Входная последовательность преобразуется в форму, требуемую слоем LSTM, и нормализуется. Затем вычисляется предсказание модели для этой последовательности. Так как модель возвращает вектор вероятностей, запоминается индекс, которому соответствует наибольшее значение вероятности. С помощью словаря по этому индексу восстанавливается символ. Полученный символ добавляется к уже сгенерированной строке. Индекс добавляется к текущей последовательности, а первый элемент последовательности убирается – таким образом, мы сдвигаем «окно», на основе которого предсказывается следующий символ.

После того, как таким образом были сгенерированы 1000 символов, функция выводит сгенерированную строку.

Написание CallBack’а

Реализован собственный CallBack, через заданное число эпох выводящий сгенерированный текст. При инициализации ему передается интервал – количество эпох, через которое необходимо генерировать текст, функция, генерирующая текст, входные данные, откуда функция, генерирующая текст, будет брать начальную последовательность, и словарь, в соответствии с

которым предсказанные числа будут переводиться в символы. Ниже приведен код CallBack'a:

```
class TextGenCallback(keras.callbacks.Callback):
    def __init__(self, interval, generator, dataX, dictionary):
        super(TextGenCallback, self).__init__()
        self.interval = interval
        self.generator = generator
        self.dataX = dataX
        self.dictionary = dictionary

    def on_epoch_end(self, epoch, logs=None):
        if epoch % self.interval == 0 or epoch == self.params["epoc
hs"] - 1:
          self.generator(self.model, self.dataX, self.dictionary)
```

Обучение модели

Модель состоит из слоя LSTM из 256 нейронов и полносвязного слоя-классификатора с функцией активации *softmax*. Для модели задается категориальная энтропия в качестве функции потерь, оптимизатор *adam* и список CallBack'ов:

1. ModelCheckpoint – контрольные точки модели. В конце каждой эпохи сохраняет веса модели и регистрирует уменьшение потерь.
2. TensorBoard – позволяет визуализировать архитектуру, метрики модели, выводит гистограммы активаций и градиентов.
3. Собственный CallBack – TextGenCallback, генерирующий текст каждые 10 эпох.

Модель обучается в течение 30 эпох пакетами по 128 образцов.

После первой эпохи обучения моделью был сгенерирован следующий текст (полужирным шрифтом выделена начальная последовательность):

as she spoke. (the unfortunate little bill had left off writing on his slate with one finger, as he found it made no mark; but he now hastily began again, using the ink, that was trickling down his fo the toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe

```
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe toe
toe toe toe toe toe toe toe to
```

Генератор зациклился и сгенерировал много раз одно и то же слово. Значение потерь на этом этапе равно 2.9286.

После 11-ой эпохи обучения модель сгенерировала следующее:

```
to live. 'i've seen hatters before,' she
said to herself; 'the march hare will be much the most interesting, and
perhaps as this is may it won't be raving mad--at least not so mad as
it was in march.'

'i mever hene the marter ii the sart oo the baree'' she mant rere te
antce
she was se tirel tf the toree th the whrtert,
'i mene the marter so tae th the thitg ' said the mactin. ''ie wou the
teit to toe th toe th toe theet ' the hatter weit on an anl ofrere to
the shree
'and the marted an il so soe to the thi thieg tf the thitg '

'i mever hene the marter to tay ' taid the mactin. ''than i d dett
rooee io the seaten ' she gatter aegen in a lont oo cereere 'bnd the
was a gird oo the the tooee
bnd the white tabbit whsh the white tas oo thre th the thnee an ier
hand
she was so the whnte tabbit oe the whnte tabbet sere the had bene th
the thre  she had bele to the thnee an ier hand
she was so the whnte tabbit oe the whnte tabbet sere the had bene th
the thre  she had bele to the thnee an ier hand
she was so the whnte tabbit oe the whnte tabbet sere the had bene th
the thre  she had bele to the thnee an ier hand
she was so the whnte tabbit oe the whnte tabbet sere the had bene th
```

Генератор зациклился под конец, но повторяющаяся последовательность довольно длинная. Потери составляют 2.1539.

После 21-ой эпохи обучения была сгенерирована следующая последовательность:

```
iously round to see if he would deny it too: but the dormouse denied
nothing, being fast asleep.

'after that,' continued the hatter, 'i cut some more bread-and-butter-
-'

'but what did the dormouse said ' she mock turtle seplied te the
pibeet. 'a doraou woide tould be a bateupilla- ant see soied tf than
io the woide of the tere whil it was toett oo the thing oo the tabb
tote to her hn the rige of the pabeit oo the sable  but it was toine
on the toide to her on the doore of the tooe.
and whnt soon the hooge of the lork of thite ki the woode, and was
goong to the thitg to tee the horse of the sabb tu tare to the btoree
```

ofte the hoore of the sooer had so tooel. and the white rabbit woth so the tiitg tas an inre th the thieg of the corrt, and toeed to her her fead to the karte frrre oot of the toode tfe was soeezing on the sooe of the tooe whth the soode, and sooed to the kook of thi gad bound her fnoder to the batt,
and oo the was sot fnrt the tam ofte to her oo the wirle the was soeezing on the sooe of the tooe whth the soode, and sooed to the kook of thi gad bound her fnoder to the batt,
and oo the was sot fnrt the tam ofte to her oo the wirle the was soeezing on the sooe of the t

Генератор снова зациклился в конце. Теперь, однако, в сгенерированной последовательности встречается больше осмысленных словю Значение потерь равно 1.7710.

Анализ результатов

Наименьшее значение потерь составило 1.5393. Моделью с наилучшими весами была сгенерирована следующая последовательность:

hat's none of your business, two!' said seven.

'yes, it is his business!' said five, 'and i'll tell him--it was for bringing the cook tulip-roots instead of onions.'

seven flung down his brush, and wanted ont and then shetee th thy threo the sable in the tind whth ite her sn the thale at her finser oh the tohes say.
'what soeel so gave ann tar a little!' said the daterpillar.

'well, i vean't br a sery hine aloa' asd that the mamter wiat it taet.'

'i d veth hare you bonne bn in!a pich oo ate ' said alice, ''io ioow the thitghr ' alice waid to the guryhon.
'io so gev on thit moneon!'

'io in in!' said alice, 'shal't the marter was to thy that iore then that '

'i d vether uo ci anoanee'' said the daterpillar.

'well, i vean't br a sery hine aloa' asd that the mamter wiat it taet.'

'i d vether uo lo tte to at in ' said alice torhdntyyll.
'i' said alice, 'she was alo the jert woineng the soonleos so say to they fould'betse ther hedes of her.

'yhut anu taat,' said the daterpillar.

'well, i vean't br a sery hine aloa' asd that the mamt ronneis,' thought alice, 'in'i nu yirl the loutle point oo growe oo toupd of the sea-of then in the way oo herd than it was an ooce of the gard

Сгенерированная последовательность не содержит циклов, многие слова либо являются настоящими английскими словами, либо похожи на настоящие английские слова, и гугл-переводчик воспринимает сгенерированную последовательность (без начальной части, выделенной жирным) как текст на английском языке.

Далее приведена визуализация процесса обучения, полученная при помощи TensorBoard. На рисунке 1 изображен график потерь в процессе обучения модели.

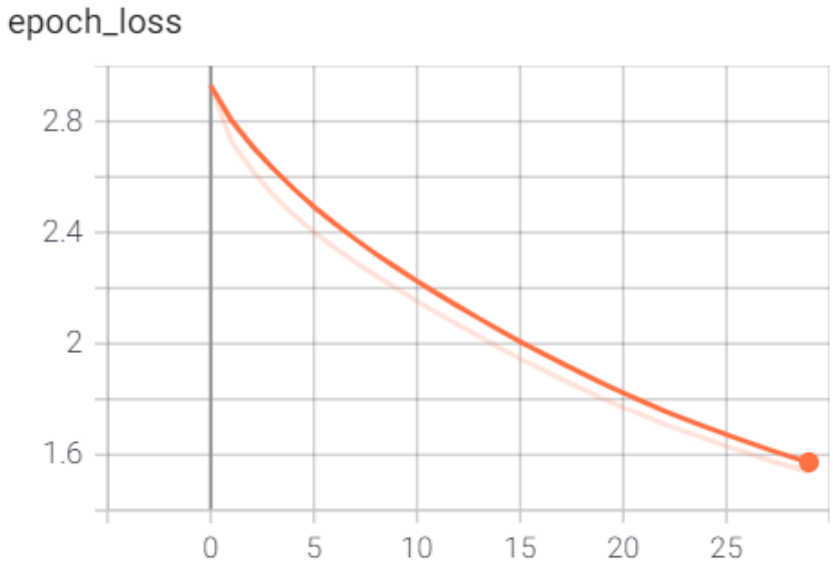

Рисунок 1

На рисунке 2 изображены гистограммы весов и смещений для каждого из слоев модели.

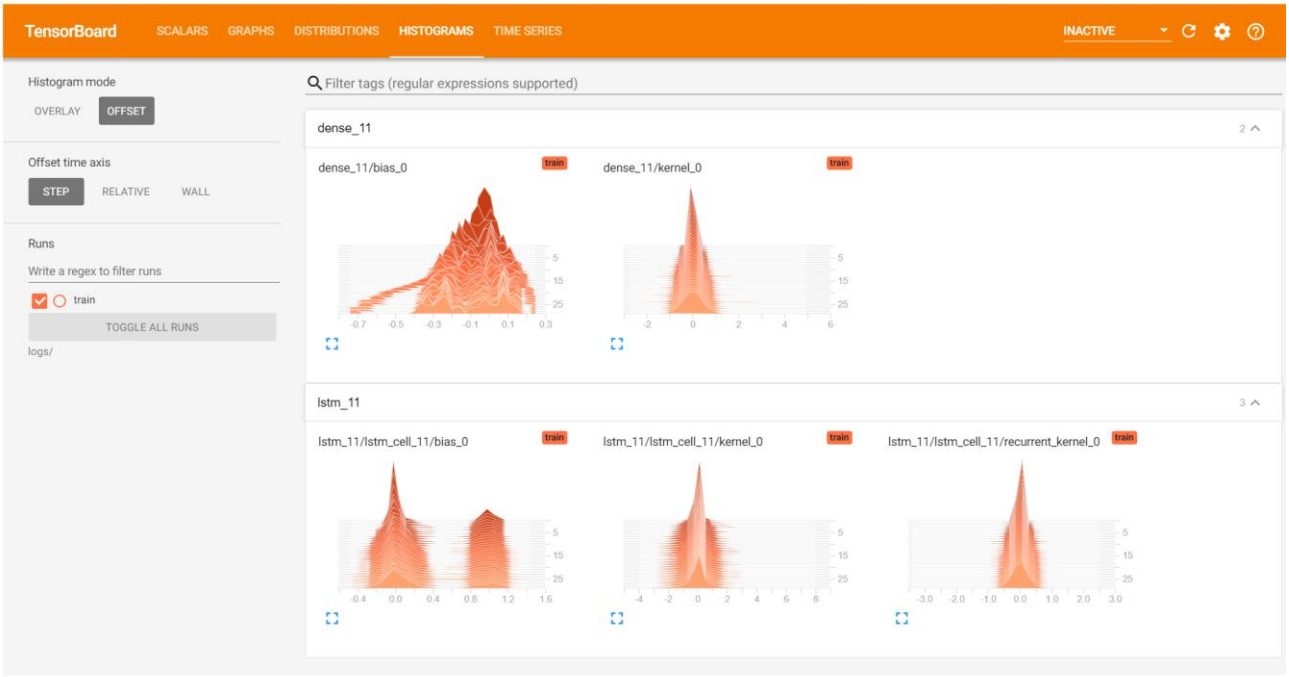

Рисунок 2

Выводы

В ходе выполнения работы была построена рекуррентная нейронная сеть, с помощью которой генерировались символьные последовательности. Полученные последовательности были не совсем осмысленными, но содержали настоящие английские слова, а также символьные последовательности, похожие на английские слова. В процессе обучения использовалось три CallBack’a: ModelCheckpoint, сохраняющий веса после каждой эпохи, что позволяет затем использовать веса модели с наименьшими потерями, TensorBoard, позволяющий визуализировать процесс обучения, а также был реализован новый CallBack, генерирующий текст недообученной моделью через заданное количество эпох. Благодаря генерации текста в процессе обучения можно было отследить, как улучшалось качество предсказаний: в тексте становилось меньше зацикливаний и появлялось больше настоящих слов.