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

ОТЧЕТ

по лабораторной работе №6

по дисциплине «Искусственные нейронные сети»

Тема: Прогноз успеха фильмов по обзорам

Студентка гр. 8382 _____ Кузина А.М.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2021

Цель работы

Прогноз успеха фильмов по обзорам (Predict Sentiment From Movie Reviews)

Задачи:

- Ознакомиться с задачей классификации
- Изучить способы представления текста для передачи в ИНС
- Достигнуть точность прогноза не менее 95%

Требования:

1. Построить и обучить нейронную сеть для обработки текста
2. Исследовать результаты при различном размере вектора представления текста
3. Написать функцию, которая позволяет ввести пользовательский текст (в отчете привести пример работы сети на пользовательском тексте)

Ход работы

В программу загружается датасет IMDb, при этом при загрузке сохраняются только заданное количество самых часто встречающихся слов. Каждый элемент датасета представлен как список индексов слов в словаре

Рассмотрим построенную модель сети:

```
model = models.Sequential()

model.add(layers.Dense(50, activation="relu", input_shape=(num, )))
model.add(layers.Dropout(0.5))
model.add(layers.Dense(50, activation="relu"))
model.add(layers.Dropout(0.35))
model.add(layers.Dense(50, activation="relu"))
model.add(layers.Dense(1, activation="sigmoid"))

model.compile(optimizer="adam", loss="binary_crossentropy",
metrics=["accuracy"])
```

Далее данная модель была обучена в течение 2х эпох с размером батча 500. С размером вектора 10000 достигнута точность предсказаний 89.5%.

Рассмотрим влияние длины вектора на точность работы сети:

Количество слов	Точность предсказаний

10000	0.8953
5000	0.8874
1000	0.8481
500	0.8084
100	0.666
50	0.5588

Чем меньше вектор, тем меньше точность предсказаний, однако отношение не линейное: переход от 5000 до 1000 тысяч дает небольшую прибавку к точности – около 1%, однако переход от 100 к 500 дает заметно большую прибавку к точности – около 14%.

Также был реализован функционал обработки пользовательского текста:

```
def read_txt(filepath, idim=10000):
    f = open(filepath, 'r')
    txt = f.read().lower()
    txt = re.sub(r"[^a-z0-9']", " ", txt)
    print(txt)
    txt = txt.split()

    index = imdb.get_word_index()

    coded = [-2]
    coded.extend([index.get(i, 0) for i in txt])
    for i in range(len(coded)):
        if coded[i]:
            coded[i] += 3
        if coded[i] >= idim:
            coded[i] = 2
    return coded
```

Файл открывается по названию, текст рецензии приводится к нижнему регистру, остаются только буквы и цифры из текста. После этого текст рецензии разделяется на массив слов. Затем каждому слову из рецензии задается частота появления из загруженного из датасета словаря. И после этого всем не самым частовстречаемым словам задается коэффициент 2, а остальные слова сдвигаются на 3 вперед, а коэффициент 1 означает начало рецензии.

По закодированному набору слов сеть делает предсказание и результат выводится на экран. Ниже приведены рассмотренные рецензии и предсказание сети по ним:

Отзыв	Оценка сети
Lindsay Lohan is accomplished in the dual role of twins separated at birth who meet when they're 11—and contrive to reunite their divorced parents. But the process takes more than two hours, during which seemingly inconsequential details—like the fact that the twins' mother designs bridal gowns—are obtrusively emphasized so their significance can be revealed later. A remake of the 1961 movie, which was based on a story by Erich Kästner, this 1998 romantic comedy is mostly boring with its cumbersome exposition and close-ups of trivial objects scattered throughout lackluster montage sequences.	0.1261512
This film remakes the classic movie very well with interesting additions and changes that keep the viewers engaged and entertained. This is a great movie to watch with the family and can make many viewers cry.	0.9672026
shit	0.54471844
I was praying that this film will be the best film in history. The best actors play in it, the people who created world famous masterpieces were engaged in its creation. It takes enormous talent to adapt such an amazing book so mediocre, I am discouraged	0.95842206

Первый два отзыва корректно определены сетью и не представляют особенного интереса – в них эмоция описана прямо. Однако оставшиеся два отзыва достаточно интересны. Третий отзыв имеет очевидно отрицательный окрас, но не входит в словарь датасета, что можно понять по его закодированному виду. А в случае, если ни одно из слов рецензии не вошло в список частовстречаемых, сеть всегда выдает примерно одинаковый результат – чуть выше 0.5. В случае с последним отзывом, очевидно, что автор отзыва сильно негодует, однако сеть не воспринимает такое противопоставление. Количество положительно окрашенных слов перевешивает отрицательно окрашенные слова и слова отрицания. Поэтому предсказание неверное.

Выводы

В данной лабораторной работе была создана и нейронная сеть, предсказывающая эмоциональный окрас рецензии на фильм. Сеть была обучена на датасете IMDB и достигла точности классификации в 89.5%. Также был реализован функционал, позволяющий загружать и классифицировать пользовательские рецензии.

Был рассмотрен один из возможных вариантов представления текста для работы с нейронной сетью, а также влияние числа загружаемых слов из отзыва на точность предсказаний сети.