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

ОТЧЕТ по лабораторной работе №7
по дисциплине «Искусственные нейронные сети»
Тема: Классификация обзоров фильмов

Студент гр. 7382 _____ Ленковский В.В.

Преподаватель _____ Жукова Н.А.

Санкт-Петербург
2020

Цель работы.

Классификация последовательностей - это проблема прогнозирующего моделирования, когда у вас есть некоторая последовательность входных данных в пространстве или времени, и задача состоит в том, чтобы предсказать категорию для последовательности.

Проблема усложняется тем, что последовательности могут различаться по длине, состоять из очень большого словарного запаса входных символов и могут потребовать от модели изучения долгосрочного контекста или зависимостей между символами во входной последовательности.

В данной лабораторной работе также будет использоваться датасет IMDb, однако обучение будет проводиться с помощью рекуррентной нейронной сети.

Основные теоретические положения.

Датасет IMDb состоит из 50 000 обзоров фильмов от пользователей, помеченных как положительные (1) и отрицательные (0). Это пример бинарной или двуклассовой классификации, важный и широко применяющийся тип задач машинного обучения.

1. Рецензии предварительно обрабатываются, и каждая из них кодируется последовательностью индексов слов в виде целых чисел.
2. Слова в обзорах индексируются по их общей частоте появления в датасете. Например, целое число «2» кодирует второе наиболее частое используемое слово.
3. 50 000 обзоров разделены на два набора: 25 000 для обучения и 25 000 для тестирования.

Ход работы.

1. Была построена и обучена нейронная сеть для обработки текста. Были использованы рекуррентные нейронные сети, они хорошо подходят для обработки текстов, т.к. могут хранить свое состояние и принимают текущее решение с учетом предыдущих. Также будем использовать одномерную свертку и пулинг. Данная архитектура дает точность: на тренировочных ~ 85%, на валидационных ~ 87%. Графики точности и ошибки предоставлены на рис. 1 и рис. 2 соответственно.

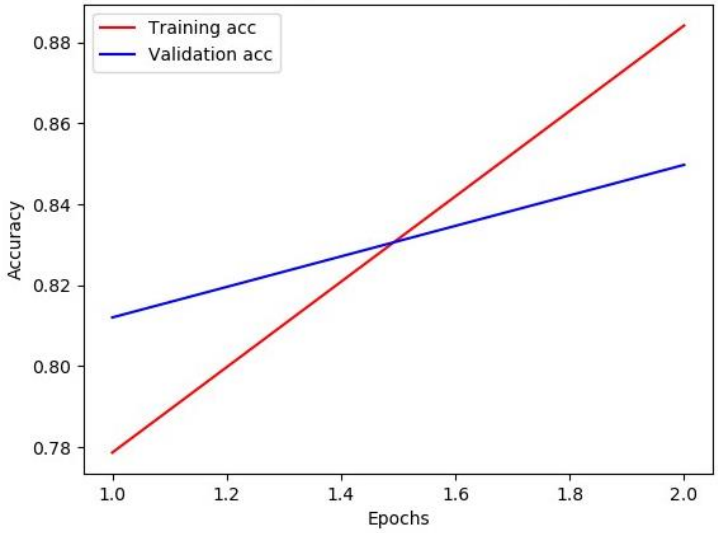

Рисунок 1 – График точности без dropout

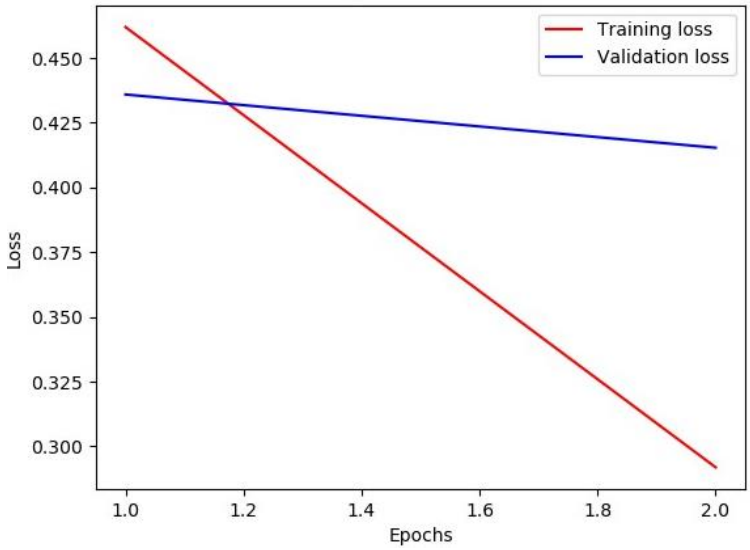

Рисунок 2 – График потерь без dropout

2. Рекуррентные нейронные сети, такие как LSTM, обычно имеют проблему переобучения. Решим эту проблему путем добавления в архитектуру сети слоев Dropout сделаем его около 0.3-0.5.

Сравнивая рис. 1 и рис. 2 и рис. 3 и рис. 4 мы видим, что без Dropout ошибка на валидации пошла вверх, что говорит о переобучении. Dropout используется для устранения переобучения, путем случайного отключения связей или нейронов, таким образом, что либо связь выдает нулевой сигнал на выход, либо нейрон выдает на все свои выходы нулевой сигнал. Точность: на тренировочных ~ 86%, на валидационных ~ 89%. Графики точности и ошибки предоставлены на рис. 3 и рис. 4 соответственно.

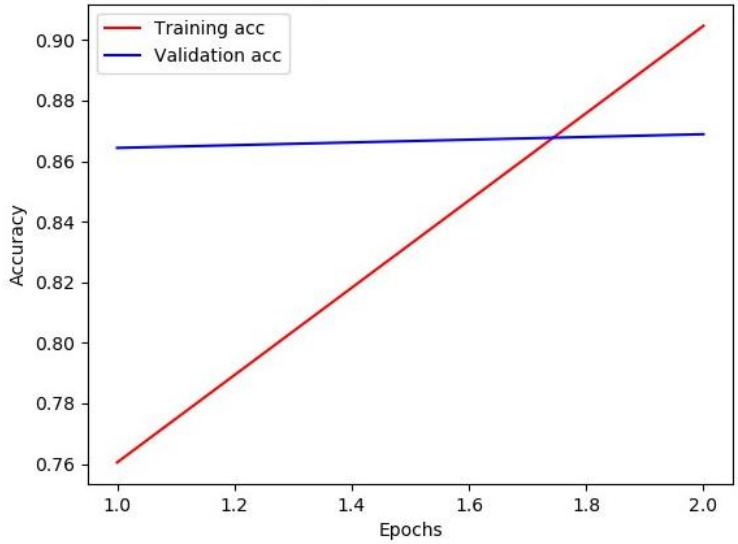

Рисунок 3 – График точности с dropout

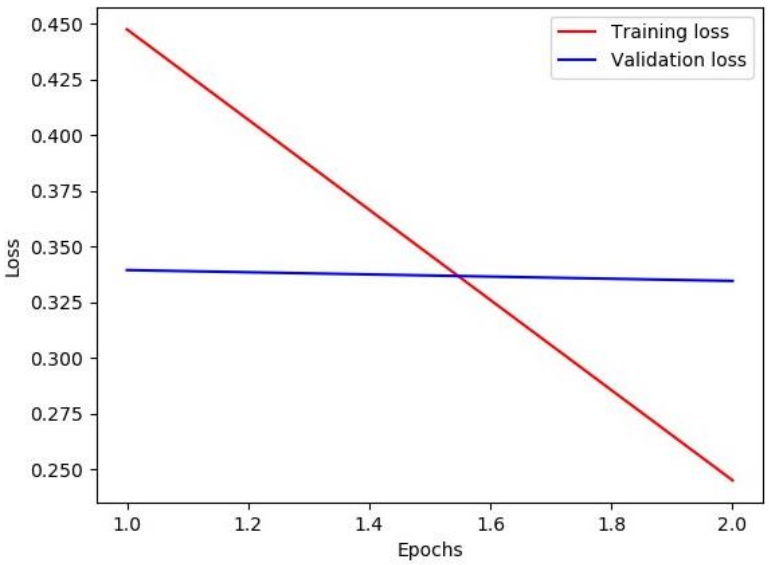

Рисунок 4 – График потерь с dropout

3. Написали функцию, которая позволяет ввести пользовательский текст.

При помощи данной функции можно получить из массива строк (обзоров) массив представлений в виде индексов слов в imdb датасете и подготовленные для прогона через модель.

Выводы.

В ходе работы была изучена задача классификация обзоров издатасета IMDB. Подобрана архитектура, дающая точность 89%. Проведя исследование, было выяснено, что при добавлении нескольких слоев свертки и пулинга увеличивается точность предсказания.