

Лабораторная работа №4

«Распознавание последовательностей»

Цель работы:

Получить практические навыки обработки текстовой информации с помощью рекуррентных искусственных нейронных сетей при решении задачи определения тональности текста.

Подготовка к работе:

Ознакомиться с лекционным материалом на темы «Обработка последовательностей и рекуррентные нейронные сети» и «Развитие идей архитектуры LSTM».

Задание:

- 1) В среде Google Colab создать новый блокнот (notebook). Настроить блокнот для работы с аппаратным ускорителем GPU.
- 2) Загрузить набор данных IMDb, содержащий оцифрованные отзывы на фильмы, размеченные на два класса: позитивные и негативные. При загрузке набора данных параметр seed выбрать равным $(4k - 1)$, где k – порядковый номер студента по журналу. Вывести размеры полученных обучающих и тестовых массивов данных.
- 3) Вывести один отзыв из обучающего множества в виде списка индексов слов. Преобразовать список индексов в текст и вывести отзыв в виде текста. Вывести длину отзыва. Вывести метку класса данного отзыва и название класса (1 – Positive, 0 – Negative).
- 4) Вывести максимальную и минимальную длину отзыва в обучающем множестве.
- 5) Провести предобработку данных. Выбрать единую длину, к которой будут приведены все отзывы. Короткие отзывы дополнить спецсимволами, а длинные обрезать до выбранной длины.
- 6) Повторить п. 4.
- 7) Повторить п. 3. Сделать вывод о том, как отзыв преобразовался после предобработки.
- 8) Вывести предобработанные массивы обучающих и тестовых данных и их размерности.
- 9) Реализовать модель рекуррентной нейронной сети, состоящей из слоев Embedding, LSTM, Dropout, Dense, и обучить ее на обучающих данных с выделением части обучающих данных в качестве валидационных. Вывести информацию об архитектуре нейронной сети. Добиться качества обучения по метрике accuracy не менее 0.8.

- 10) Оценить качество обучения на тестовых данных:
 - вывести значение метрики качества классификации на тестовых данных,
 - вывести отчет о качестве классификации тестовой выборки.
 - построить ROC-кривую по результату обработки тестовой выборки и вычислить площадь под ROC-кривой (AUC ROC).
- 11) Сделать выводы по результатам применения рекуррентной нейронной сети для решения задачи определения тональности текста.
- 12) Составить отчет по лабораторной работе. Внести в отчет программный код и результат программного вывода (при наличии) в соответствии с пунктами задания. Внести в отчет выводы по результатам лабораторной работы.

Контрольные вопросы

- 1) Области применения рекуррентных нейронных сетей и задачи, решаемые с их помощью.
- 2) В чем заключается особенность данных, представляемых в виде последовательностей? Приведите примеры таких данных.
- 3) Опишите структуру простого рекуррентного нейрона.
- 4) Опишите структуру рекуррентной LSTM-ячейки.
- 5) Что означает термин «обратное распространение во времени» (backpropagation through time, BPTT)?
- 6) Какие типы слоев применяются при построении рекуррентных нейронных сетей в Keras?
- 7) Что такое ROC-кривая? Как использовать ее для оценки качества бинарной классификации? Что показывает метрика качества классификации AUC ROC?