МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧИ КЛАССИФИКАЦИИ РЕНТГЕНОГРАММ

А. А. Гулицкий

Белорусский государственный университет, г. Минск;
antongoulitski@gmail.com;
науч. рук. – Э. А. Чернявская, д-р физ.-мат. наук, проф.

В работе освещена тема больших объёмов данных, которая становится всё более актуальной в результате быстрого роста количества информации. На базе набора рентгенограмм произведён анализ некоторых методов обработки Big Data и описан метод, позволяющий диагностировать лёгочные заболевания. Для задачи сегментации области лёгких была использована нейронная сеть с архитектурой U-net, являющаяся наиболее популярной для аналогичных задач в области анализа биомедицинских изображений. Классификация произведена с помощью нейросети ResNet-50, итоговая точность которой составила 92 %.

Ключевые слова: Big Data; ResNet; U-net, mapreduce; рентгенограммы; глубокое машинное обучение.

ВВЕДЕНИЕ

За последние несколько лет своего существования человечество произвело больше данных, чем за всю историю до этого. Сегодня более пяти миллиардов пользователей взаимодействуют с данными ежедневно. К 2025 году их будет 6 миллиардов, и каждый из них будет взаимодействовать с данными как минимум раз в 18 секунд. Ежегодный отчёт IDC говорит о том, что объём общемировых данных вырастет с 33 зеттабайт в 2018 году до 175 в 2025 [1]. Сегодня проблема хранения, обработки и использования больших объёмов данных является одной из самой актуальных.

Анализ больших объёмов данных имеет особую значимость в области диагностики лёгочных заболеваний, т.к. в этой сфере существует потребность в постоянном анализе и мониторинге больших баз данных рентгеновских снимков для обнаружения заболевания, прежде чем оно сможет нанести серьёзный вред здоровью.

АНАЛИЗ МЕТОДОВ ГЛУБОКОГО МАШИННОГО ОБУЧЕНИЯ ДЛЯ ЗАДАЧИ КЛАССИФИКАЦИИ РЕНТГЕНОГРАММ

Среди методов анализа больших объёмов данных можно отдельно выделить методы класса Data Mining, краудсорсинг, смешение и интеграцию данных, машинное обучение, нейронные сети, алгоритмы mapreduce и т.д. Такое разнообразие методов может свидетельствовать о том, что постановка задачи в области Big Data может существенно различаться, единая методология отсутствует и в рамках каждой отдельной коллекции данных существует возможность анализа разными способами.

В контексте задачи классификации рентгенограмм было проведено несколько исследований повышения точности обнаружения аномалий. Так, авторы [2] и [10] провели исследование эффективности применения некоторых методов предобработки и добились итоговой повышения точности классификации с 40 % и 62 % соответственно до приблизительно 70 %. Следует отметить, что в этих исследованиях были использованы достаточно простые нейросети и итоговая точность классификации может повыситься благодаря использованию более актуальных архитектур нейронных сетей.

РЕАЛИЗАЦИЯ АЛГОРИТМА КЛАССИФИКАЦИИ РЕНТГЕНОГРАММ

В процессе работы с исходным набором данных были выделены следующие этапы анализа:

1. Усиление локального контраста.
2. Фильтрация коллекции данных.
3. Сегментация области лёгких.
4. Классификация.
5. Оценка.

Эти этапы были реализованы на Matlab и Python с Keras и бэкендом Tensorflow.

Одним из главных факторов, влияющих на качество изображения, а значит и на результативность обучения нейросети, является контрастность изображения. Для повышения контрастности изображения существует ряд методов, использующихся при разных начальных условиях. Постановка задачи в данной работе требует качественного повышения контраста в области лёгких, который сможет облегчить распознавание патологии, но при этом будет затрачивать минимальное количество ресурсов и времени для реализации. Метод выравнивания гистограммы [3] служит для получения возможности использования всего динамического диапазона изображения. В этом подходе исходная гистограмма изображения представляется в виде функции плотности вероятностей, в результате происходит оценка вероятности каждого уровня значения серого цвета, после чего происходит перераспределение уровней. Данный метод имеет высокую производительность и хорошо масштабируется при увеличении количества изображений, при работе с большими данными такие качества являются необходимыми.

После этапа повышения контрастности изображений последовал анализ имеющихся в коллекции данных рентгенограмм. В результате было выявлено, что многие снимки в датасете являются засвеченными и не подходят для дальнейшего анализа. Для отбора неудачных снимков в условиях большого размера набора рентгенограмм был выбран метод mapreduce, который используется для анализа данных большого объёма, имеющих специфику размещения в памяти компьютера. В данной рабо-

те в качестве функции map выступает функция, которая сохраняет данные изображения и среднее значение яркости в качестве промежуточных значений. После чего функция reduce получает список имён файлов изображений вместе с соответствующими значениями средней яркости и находит максимальные значения. В результате получилось отсеять более сотни изображений, которые могут негативно повлиять на результаты работы нейросети.

Для задачи сегментации области лёгких была выбрана обученная нейронная сеть на базе архитектуры U-net [4]. Данная архитектура наилучшим образом зарекомендовала себя в применении к задачам сегментации биомедицинских изображений [5]. В результате сегментации были получены маски более чем пяти тысяч изображений. Для оценки работы нейронной сети были вручную созданы маски для ста случайных изображений из датасета, которые будут считаться точной областью лёгких, далее было произведено сравнение между условным эталоном и результатами, полученными с помощью нейросети. Итогом сравнения стали значения F-меры $F = 0{,}91$ для набора данных, прошедшего предыдущие этапы обработки, и $F = 0{,}80$ для исходного набора данных. На основе результатов был сделан вывод о том, что повышение уровня контрастности оказало положительное влияние на качество полученных масок. Вызвано это тем, что границы между соответствующими областями становятся более различимы, а также более высокой средней контрастностью коллекции данных[6], на которых была обучена нейросеть по сравнению с исходным датасетом, использующемся в данной работе.

Финальным этапом анализа коллекции рентгенограмм будет классификация с помощью нейронной сети на базе ResNet. Данная архитектура была выбрана потому, что является одной из самых актуальных на сегодняшний день и хорошо показывает себя в задаче диагностирования пневмонии. Для сокращения времени обучения были использованы соответствующие веса [7]. В качестве метрики качества использована точность (accuracy), которая определяется как отношение количества верных решений классификатора ко всему количеству решений. Результат точности для обработанного датасета – 0,92. Аналогичные действия были произведены над датасетом без обработки и над датасетом с сегментацией без повышения контраста. Соответствующие результаты точности – 0,78 и 0,84, что доказывает целесообразность применения методов предобработки рентгеновских снимков при распознавании аномалий в легких. Для оценки эффективности используемых в работе методов было произведено сравнение с нейросетями [8] и [9], имеющими похожую архитектуру и ранее использовавшимися для классификации использованного в работе набора данных.

Таблица

Значения точности для различных нейросетей

Нейросеть	Точность
ResNet-50 с предобработкой	0,92
VGG-19 [8]	0,79
ResNet-18 [9]	0,85

Из полученных результатов видно, что использованные в работе методы обработки данных могут существенно увеличить точность в задаче диагностирования лёгочных заболеваний. Следует отметить, что нейросети, использованные для анализа, являются одними из самых актуальных и оптимизированных для задачи классификации биомедицинских изображений, и в случае более простых и устаревших архитектур методы, использованные в работе, могут дать значительно больший прирост к метрикам качества. Однако даже на примере этих нейронных сетей видны влияние этапов предварительной обработки на полученные результаты и важность индивидуального подхода к каждому конкретному набору больших данных.

Библиографические ссылки

1. IDC Regional reports [Электронный ресурс]. URL: https://www.seaaate.com/gb/en/our-story/data-age-2025 (дата обращения: 30.05.2019).
2. Deep Learning with Lung Segmentation and Bone Shadow Exclusion Techniques for Chest X-Ray Analysis of Lung Cancer [Электронный ресурс] / Yu. Gordienko [et al.] // ArXiv. URL: https://arxiv.org/abs/1712.07632 (дата обращения: 30.05.2019).
3. *Gonzalez R. C., Woods R. E., Eddins S. L.* Digital Image Processing Using MATLAB // Pearson Prentice-Hall. 2004. P. 81–84.
4. U-Net lung segmentation (Montgomery + Shenzhen) [Электронный ресурс]. URL: https://www.kaggle.com/eduardomineo/u-net-lung-segmentation-montgomery-shenzhen (дата обращения: 30.05.2019).
5. *Ronneberger O., Fischer P., Brox T.* U-Net: Convolutional Networks for Biomedical Image Segmentation // Medical Image Computing and Computer-Assisted Intervention (MICCAI). 2015. Vol. 9351. P. 234–241.
6. Tuberculosis Chest X-ray image data sets [Электронный ресурс]. URL: https://ceb.nlm.nih.gov/repositories/tuberculosis-chest-x-ray-image-data-sets (дата обращения: 30.05.2019).
7. ResNet-50 Pre-trained Model for Keras [Электронный ресурс]. URL: https://www.kaggle.com/keras/resnet50 (дата обращения: 30.05.2019).
8. Resnet18 Transfer Learning PyTorch [Электронный ресурс]. URL: https://www.kaggle.com/xanthate/resnet18-transfer-learning-pytorch (дата обращения: 30.05.2019).
9. Chest X-Ray – Keras VGG19 Transfer Learning [Электронный ресурс]. URL: https://www.kaggle.com/curiousprogrammer/chest-x-ray-keras-vgg19-transfer-learning (дата обращения: 30.05.2019).
10. Dimensionality Reduction in Deep Learning for Chest X-Ray Analysis of Lung Cancer [Электронный ресурс] / Yu. Gordienko [et al.] // ArXiv. URL: https://arxiv.org/abs/1801.06495 (дата обращения: 30.05.2019).