Федеральное государственное бюджетное образовательное учреждение
высшего образования «Санкт-Петербургский государственный
университет»

Кафедра Системного Программирования

Захаров Роман Вадимович

Машинное обучение в медицине. Автоматическое распознавание легких на флюорографических снимках.

Курсовая работа

Научный руководитель:
ведущий разработчик ООО «НМТ» Невоструев К.Н.

Санкт-Петербург
2016

Оглавление

Введение

Актуальность работы

В наше время все чаще возникают задачи, собранные названием Big Data. Развитие деятельности человека подразумевает накопление большого количества данных. Анализ этих данных может быть проведен вручную, а может быть проделанным с помощью машинного обучения, то есть с помощью алгоритмов автоматического нахождения закономерностей в эмпирических данных. Одна из областей применения средств машинного обучения – медицина.

Среди алгоритмов машинного обучения выделяют алгоритмы глубокого машинного обучения – сложных многослойных нейронных сетей. В рамках этой работы будет исследована возможность применения алгоритмов глубокого машинного обучения к флюорографическим снимкам человека, а именно для автоматического распознавания легких человека, что может быть использовано для последующего обнаружения туберкулеза.

Доступные программные средства

Обработка снимков происходила с помощью языка программирования Python 2.7 и библиотек numpy, scikit-learn, panda, caffe и matplotlib в среде программирования PyCharm. Эти технологии свободно доступны для исследовательских задач.

Предметная область

Фундаментальная проблема данного исследования – анализ в задачах распознавания отдельных частей снимков легких человека. Работа нацелена на создание основы практической реализации методов поиска анатомических примитивов (легкие, кости, органы) с использованием алгоритмов глубокого обучения.

1. Постановка задачи

1.1. Цели

Целью данной работы является осуществление и сравнение нескольких методов для флюорографических снимков с целью обнаружения на них легких.

1.2. Задачи

В рамках данной курсовой работы были поставлены перечисленные ниже критерии. Требования к конечному продукту:

- Получение результатов анализа обнаружения легких на снимках.
- Высокая точность итогового применения алгоритма.

Тестирование и наличие тестов:

- Оценивающих реальную точность алгоритма, то есть наличие отдельных выборок для тренировки машинного алгоритма и для его тестирования.
- Сравнивающих результаты с имеющимися данными.

2. Методы обнаружения легких

В целях поиска различных паталогий легких на флюорографических снимках (например, туберкулеза) нам следует научиться выделять на снимках сами легкие. Это важно, так как в дальнейшем для распознавания болезней будут применяться алгоритмы машинного обучения. Если не определять легкие, а использовать для обучения снимки целиком, то мы рискуем отдать на вход алгоритма слишком много лишних данных, среди которых классификатор будет искать признаки для обучения. Также выделение легких позволит сократить количество требуемых аппаратных ресурсов для обучения и количество затраченного на обучение времени.

Разумеется, на небольшой обучающей выборке проще всего выделить легкие вручную. Однако, когда мы работаем с большой выборкой снимков, нам необходимо использовать алгоритмы компьютерного распознавания.

В ходе работы было реализовано и применено несколько алгоритмов для автоматического обнаружения легких.

2.1. Простое обнаружение

Первый метод является экспертной системой, работающей по простому алгоритму (без использования каких-либо алгоритмов машинного обучения).

Его описание можно задать с помощью следующих последовательных шагов:

- Выделение контура легких с помощью цвета. На этом этапе так же выделяются контуры тела.
- Отсечение границ тела с помощью анализирования гистограмм для строк и столбцов изображения.
- Повторный анализ гистограмм для более точного удаления ненужных частей.

- Вписание оставшихся выделенных частей снимка в прямоугольник. Полагается, что к этому моменту выделены только легкие.

Этот алгоритм хорошо себя показывает на большинстве изображений. Однако, дает сбои для некоторых снимков, например, когда пациенту снизили мощность флюорографического аппарата с целью снижения радиации по причине частого прохождения процедуры флюорографии. Также, прямоугольная форма недостаточно хороша для дальнейшего анализа – ввиду того, что в прямоугольник включаются не только легкие, и часто в прямоугольник включается больше не легких, чем легких, ввиду анатомической формы органов дыхания человека.

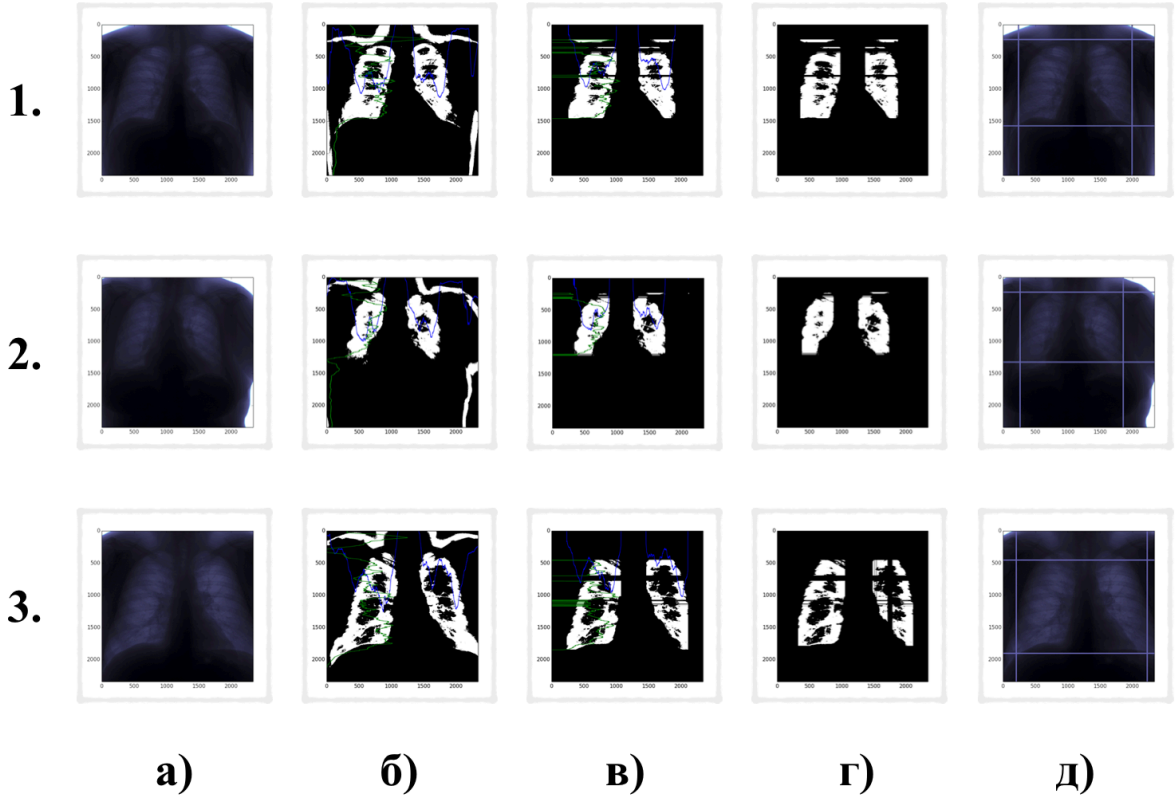

Рис. 1: Пример работы алгоритма простого обнаружения. а) - необработанное изображение, б) - выделение границ легких и тела человека, в) - отсечение ненужного выделения границ тела человека, г) - повторное отсечение, д) - выделенный прямоугольник с легкими (для наглядности показан внутри снимка, а не вырезан).

2.2. Обнаружение с использованием машинного обучения

Помимо простого алгоритма обнаружения легких, были созданы экспертные системы на основе алгоритмов машинного обучения.

2.2.1. Технология измерения точности алгоритмов

Чтобы использовать алгоритмы машинного обучения необходимо иметь большой специальный набор данных для обучения.

Для того, чтобы измерять точность алгоритмов машинного обучения, были размечены легкие на 40-а снимках и создана контрольная функция, точно определяющая, содержится ли квадратный кусочек снимка хотя бы на 30% внутри легких.

В целях использования результатов работы алгоритмов машинного обучения была создана аналогичная функция, которая также для кусочка снимка отвечает содержится ли он в легких хотя бы на 30%, но уже на основе обученной компьютерной модели. С помощью этой функции можно увеличивать текущую выборку для обучения более сложных нейронных сетей, распознающих туберкулез или другие заболевания легких.

2.2.2. Создание выборок для обучения и тестов

С помощью 40-а уже размеченных файлов была создана обучающая выборка из 30 000 кусочков снимков квадратной формы и тестирующая выборка из 10 000 аналогичных кусочков.

С помощью контрольной функции эти изображения были разделены на два класса: класс кусочков снимков, содержащихся в легких хотя бы на 30% и класс кусочков снимков, содержащихся в легких менее, чем на 30% или не содержащиеся в легких совсем.

Все изображения были подвергнуты нормализации (стоит заметить, что снимок не цветной, а состоящий из 65 536 оттенков серого цвета; самый темный пиксель становился черным, самый светлый – белым,

остальные – отображались на соответствующий отрезок), что важно при использовании алгоритмов машинного обучения.

2.2.3. Обнаружение с помощью K-means и Random Forest

Если попробовать анализировать небольшие квадратные кусочки снимков, то, построив для каждого из них гистограммы цвета, можно обратить внимание, что гистрограммы кусочков, содержащихся в легких похожи друг на друга, гистограммы кусочков с рисунком костей тоже, и так для всех анатомических примитивов снимка. С помощью алгоритма кластеризации K-means можно разделить изображения на наборы на основе особенностей форм их гистограмм.

Для использования K-means были построены и нормализованы гистограммы для обучающей и тестовой выборок. Также в качестве признаков были выделены отношение между минимальным и максимальным цветом для кусочка, а также отношения минимального и максимального цветов к среднему цвету всего снимка.

Гистограммы изображений обучаюшей выборки были переданы алгоритму K-means, результаты кластеризации номера классов и нормализованные вышеупомянутые признаки были переданы на вход алгоритма машинного обучения Random Forest для создания компьютерной модели медицинской экспертной системы. Модель была протестирована на тестирующей выборке.

Итоговая точность модели – 88%. Можно заметить, что в оставшиеся 12% включаются преимущественно ложноположительные результаты, то есть части снимка, автоматически распознанные как легкие, не являющиеся таковыми на деле. Ложноположительные результаты «лучше» ложноотрицательных тем, что нам важнее не исключить настоящие легкие из дальнейшего рассмотрения, чем включить части снимка, не входящие в легкие.

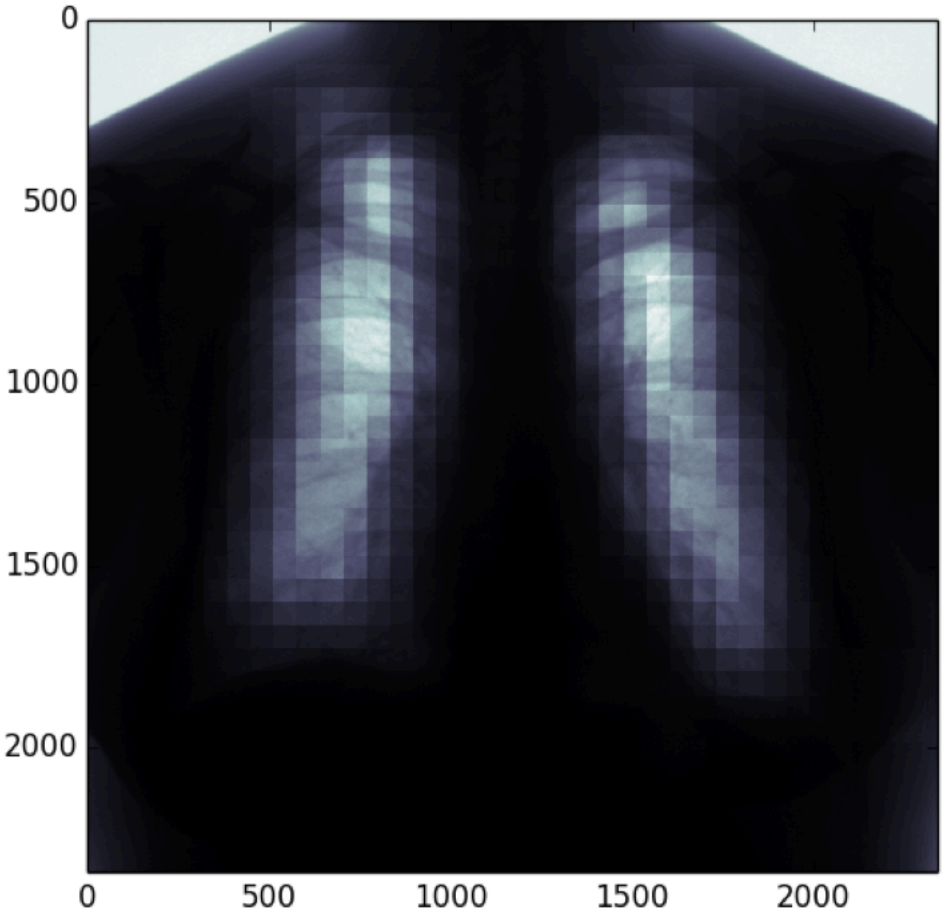

Рис. 2: Пример работы алгоритма на основе K-means и Random Forest. Предполагаемая область легких осветлена по отношению к другим частям изображения.

2.2.4. Обнаружение с помощью CNN

Сверточные нейронные сети (англ. CNN) широко применяются для задач, связанных с поиском и классификацией данных на изображении. В ходе работы был обучен бинарный классификатор на той же выборке, что и для Random Forest, однако модель на основе сверточной нейронной сети работает с изображениями, а не с их признаками.

В качестве структуры нейронной сети были использованы 5 пар сверточных (англ. convolution layers) и субдискретизирующих слоев (англ. subsampling layers), а также 3 полносвязных слоя (англ. fully connected layers).

Использование пар сверточных и субдискретизирующих слов позволяет создать распознающий алгоритм, не чувствительный к шумам

на исходном изображении, а также создать достаточно хорошую модель засчет сложной многослойной нейронной логики системы. Использующийся метод обучения – метод обратного распространения ошибки. [1, 4]

Достигнутая точность – 95%.

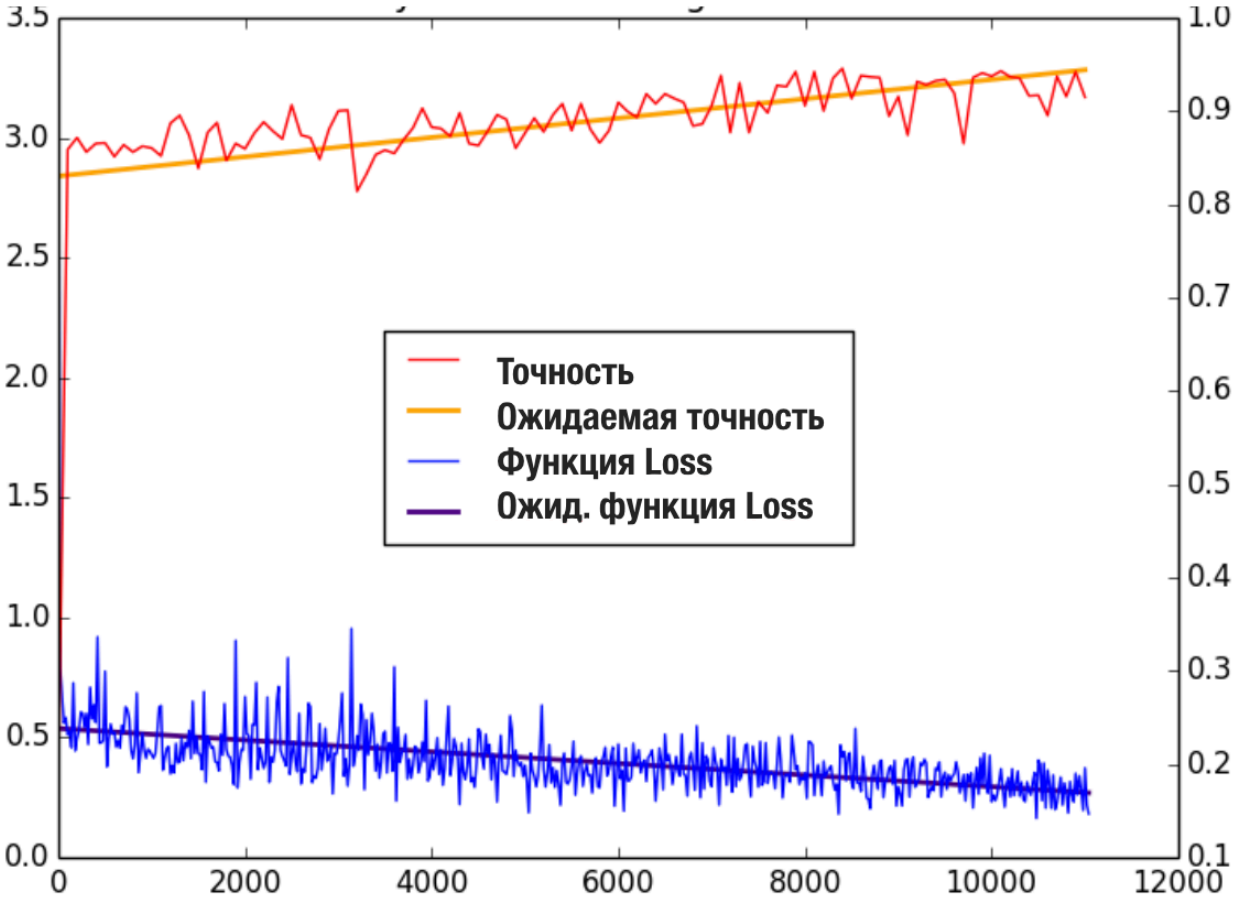

Рис. 3: График зависимости точности работы модели CNN на тестовой выборке от количества итераций обучения.

3. Результаты

3.1. Сравнение методов обнаружения

При одинаковой выборке точность модели на основе сверточной нейронной сети оказалась точнее, чем метод, основанный на K-means и Random Forest. Однако, обучение последнего происходит быстрее за счет использования меньшего числа признаков.

Сравнение методов представлено в следующей таблице.

Экспертная система	Простой метод	K-means + Random Forest	Сверточная нейронная сеть
Размер выборки	Отсутствует	30 000 изобр. для обучения, 10 000 изобр. для теста	30 000 изобр. для обучения, 10 000 изобр. для теста
Обучение	Отсутствует	Быстрое	Очень долгое
Достигнутая точность распознавания	Низкая	88%	95%
Применимость к другим задачам	Низкая	Средняя	Высокая

3.2. Универсальность CNN

Важная особенность способа поиска легких с помощью CNN – это то, что подобный метод можно применить к другим задачам медицины, в том числе и не для флюорографических снимков, и полученный опыт создания экспертной системы может пригодиться в работах над решением аналогичных проблем. [1]

Заключение

В рамках данной работы были достигнуты следующие результаты:

- Рассмотрены и применены методы машинного обучения (K-means, Random Forest, CNN). [2]
- Изучены литература и статьи об алгоритмах машинного обучения. [3, 5]
- Проведено сравнение точности алгоритмов.
- Апробированы различные готовые библиотеки и решения для алгоритмов глубокого обучения.
- Достигнута большая точность распознавания легких (95%).

Список литературы

[1] Caffe: Convolutional Architecture for Fast Feature Embedding / Yangqing Jia, Evan Shelhamer, Jeff Donahue et al. // arXiv preprint arXiv:1408.5093, 2014.

[2] Polyak B.T., Tsybakov. Optimal Orders of Accuracy for Search Algorithms of Stochastic Optimization. Probl. Peredachi Inf., 1990.

[3] Robbins H., Monro S. Stochastic Approximation Method. — Ann. Math. Statist, 1951.

[4] Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort et al. // Journal of Machine Learning Research. — Vol. 12. — P. 2825–2830, 2011.

[5] Я.З. Цыпкин. Адаптация и обучение в автоматических системах. — Москва: Наука, 1968.