**Введение**

На сегодняшний день машинное обучение и искусственный интеллект считаются одними из самых перспективных направлений как в ИТ-отрасли, так и во многих других. Технологии на основе нейронных сетей все активнее проникают в повседневную жизнь. Даже в, казалось бы, неочевидных вещах, таких как новостная лента в социальных сетях, новости сформированы искусственным интеллектом на основе предпочтений пользователя.

Основными задачами подобных моделей являются в первую очередь управление, классификация и прогнозирование. Главным преимуществом нейронных сетей является возможность находить и воспроизводить сложные зависимости с разным количеством переменных, а так же способность к обучению, в ходе которого происходит определение связей и зависимостей между исходными данными.

Модели машинного обучения используются в самых разных сферах деятельности человека. С помощью нейронных сетей осуществляются разные задачи такие как управление качеством изделий на производствах, эффективный поиск месторождений в нефтедобыче, обработка изображений, распознавание лиц, отпечатков, голоса, прогнозирование в экономике и многие другие. Одним из самых востребованных применений является медицина. Работа в медицинских учреждениях всегда связана со сбором, обработкой и анализом различных визуальных данных, таких как рентген, КТ и многие другие. Здесь требуются алгоритмы, способные анализировать и обрабатывать изображения. Направление искусственного интеллекта, которое занимается работой с изображениями называется Computer vision или компьютерное зрение. Данное направление считается наиболее полезным в области медицинской диагностики.

Целью данной дипломной работы является проектирование системы, на основе нейронной сети, способной распознавать признаки пневмонии на снимках рентгенограммы. Распознавание данного заболевания методами машинного обучения является актуальной задачей на данный момент. События 2020 года показали, что во многих странах диагностика данного заболевания стала проблемой, так как из-за огромного количества заболевших появилась острая нехватка специалистов.

Предметом исследования является изучение и применение актуальных библиотек для создания и обучения нейронной сети

Объектом исследования является процесс создания нейронной сети для распознавания признаков пневмонии на изображениях.

**Постановка задачи**

Целью данной выпускной квалификационной работы является

**Анализ предметной области**

**Пневмония. Диагностика заболевания**

Пневмония — воспаление одного или двух легких, возникшее на фоне вирусной, бактериальной или грибковой инфекции, это одно из самых опасных и тяжелых заболеваний дыхательной системы. Характеризуется выбросом жидкости в альвеолы (Альвеола — пузырьковидное образование в лёгких, оплетённое сетью капилляров. Через стенки альвеол (в лёгких человека их свыше 700 млн) происходит газообмен при дыхании), а также такими симптомами как: температура, одышка, кашель (в большинстве случаев с обильным выделением мокроты), учащение дыхания, снижение насыщения крови кислородом, что приводит к серьезным последствиям для организма. Осложнения пневмонии в форме дыхательной недостаточности, острой гипоксии грозят летальным исходом

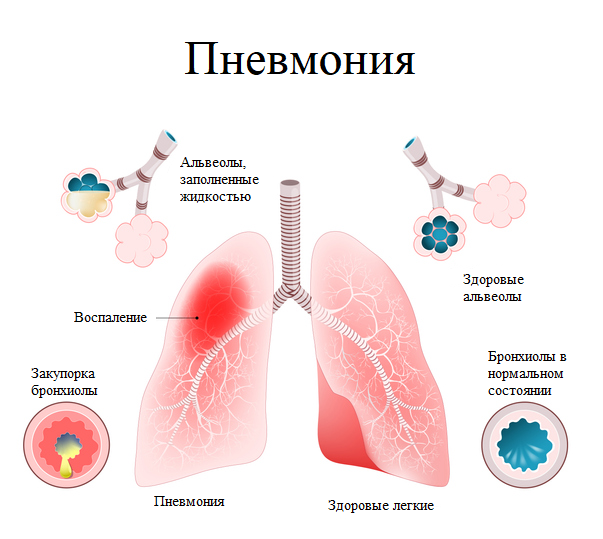


Рис.1 Пневмония

Пневмония может быть диагностирована разными инструментальными методами. Рассмотрим некоторые из них.

**1.Флюорография** – рентгенологическое исследование, заключающееся в фотографировании видимого изображения на флюоресцентном экране, которое образуется в результате прохождения рентгеновских лучей через тело человека и неравномерного поглощения органами и тканями организма.

С помощью флюорографии могут быть обнаружены:

* опухолевые образования и другие патологии;
* присутствие очагов воспаления;
* инородные предметы.

Преимуществами метода являются:

* минимальные затраты времени;
* минимальная стоимость исследования;
* низкое радиационное воздействие;
* широкая доступность метода и простота проведения исследований;
* не требует специальной подготовки пациента.

Недостатки метода:

* Менее точная диагностика, по сравнению с другими методами.

**2.Рентгенография грудной клетки** – это метод диагностики, позволяющий получить изображение органов грудной полости с помощью облучения рентгеновскими лучами. Различные ткани организма в зависимости от своей плотности по-разному пропускают рентгеновские лучи, а значит, по-разному отображаются на снимке (рентгенограмме).

Рентген грудной клетки используется для выявления таких патологий как:

* присутствие в легких очагов воспаления, зачастую характерных для пневмонии и туберкулеза;
* наличие опухолевых образований и отеков;
* наличие патологических скоплений газов и жидкостей;
* инородные предметы в легких, пищеводе и дыхательных путях.

Среди преимуществ данного метода можно назвать следующие факторы:

* широкая доступность метода и простота проведения исследований;
* не требует специальной подготовки пациента;
* низкая стоимость исследования.

Основные недостатки метода:

* информативность метода ниже, чем у КТ и МРТ;
* статичность изображения, невозможность проводить исследование в разных проекциях;
* большие затраты времени по сравнению с флюорографией.

**3.Компьютерная томография (КТ)** – это метод лучевой диагностики, дающий возможность послойного исследования организма, в основе которого лежит метод рентгенографии. Аппарат вращается вокруг пациента и делает серию снимков, которые затем обрабатываются и расшифровываются врачом.

С помощью КТ могут быть обнаружены:

* воспаления тканей;
* различные новообразования и отеки;
* онкологические заболевания.

Преимущества метода:

* высокая точность детализации исследуемых органов;
* результат проведения КТ-исследования – трехмерный снимок;
* полученная после КТ информация проста и понятна для восприятия.

Недостатки метода:

* относительно более высокая стоимость исследования.

**4.Магнитно-резонансная томография (МРТ)** – это диагностика внутренних органов и тканей человека с помощью ядерного магнитного резонанса. МРТ позволяет получить трехмерное изображение внутренних органов в реальном времени, что обеспечивает точную визуализацию различных структур организма.

При проведении МРТ могут быть обнаружены:

* наличие воспалительных процессов в организме;
* новообразования;
* патология структур тканей.

Преимущества МРТ:

* полученные результаты исследования характеризуются высокой точностью;
* нет ограничений по количеству МРТ процедур, в отличии от радиологических методов;
* результаты исследования в виде трехмерного снимка.

Недостатки метода:

* отсутствует возможность комплексного исследования полых органов, таких как легкие, желчный пузырь, мочевой пузырь;
* ограничения на проведение МРТ пациентам с металлическими предметами в организме и кардиостимуляторами;
* высокая стоимость проведения исследования;
* сравнительно большие временные затраты.

Хоть такие методы как КТ и МРТ достаточно точны и информативны, такие исследования все же не так доступны на данный момент. Среди рассмотренных методов диагностики оптимальными являются методы радиологических исследований, такие как рентгенография и флюорография. Эти методы широко распространены, не требуют больших затрат времени и средств. Пусть полученные с помощью них изображения сравнительно хуже, чем на МРТ и КТ, но они обладают достаточным количеством информативных признаков для классификации.

**Теоретические основы нейронных сетей**

**Устройство нейронных сетей**

Нейронная сеть или искусственная нейронная сеть – это математическая модель, а так же ее программная или аппаратная реализация, построенная по принципу организации и функционирования биологических нейронных сетей. Нейроны являются основным типом клеток в нервной системе, которые генерируют и передают электрохимические сигналы. Они в основном общаются друг с другом с помощью нейротрансмиттеров на конкретных узлах, называемых синапсами. Большинство из них имеют две основные структуры: аксон и дендриты. Биологический нейрон состоит из ядра и отростков нервных волокон двух типов:

* дендриты – разветвленный отросток нейрона, который получает информацию;
* аксон – отросток, по которому нейрон может передавать импульс.

Аксон контактирует с дендритами других нейронов с помощью синапсов, способных влиять на силу передаваемого сигнала. Структура, состоящая из большого количества нейронов получила название биологической нейронной сети.

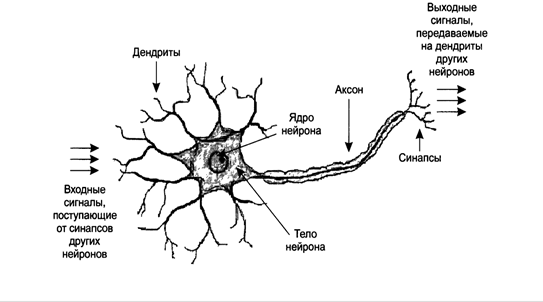


Рис. 2 биологический нейрон

Такой термин как «нейронная сеть» впервые появился в середине двадцатого века. Прогресс был связан с развитием нейробиологии. Первые работы, в которых был получены результаты в данном направлении, были проделаны американскими учеными У. Маккалаком и У.Питтсом. в 1943 году ими была предложена математическая модель биологического нейрона.

Подобно своему биологическому прототипу, искусственный нейрон является основным элементом любой искусственной нейронной сети.

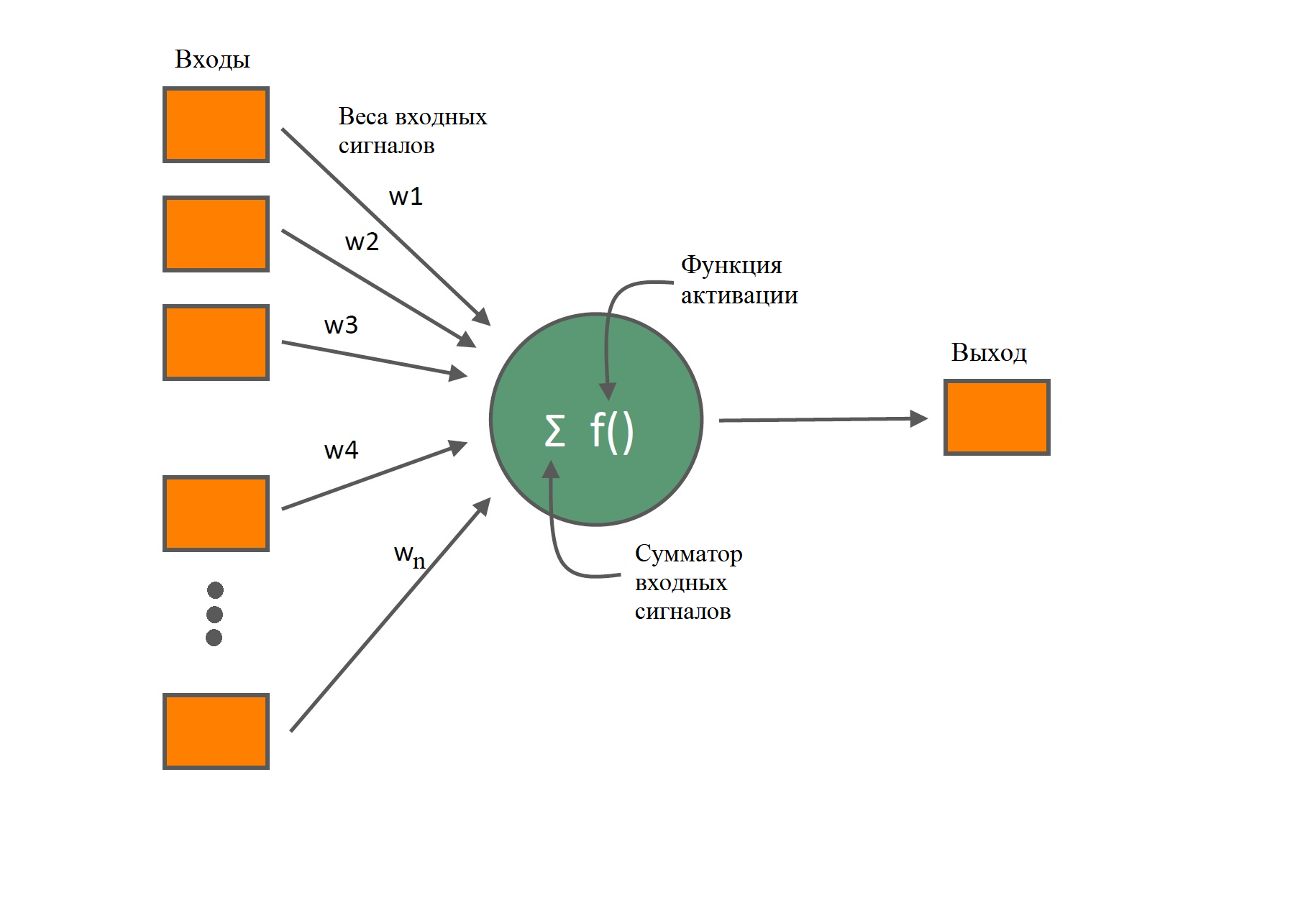


Рис. 3 искусственный нейрон

Математическая модель искусственного нейрона представляет собой сумматор всех входящих сигналов, применяющий к полученной сумме некоторую функцию, называемую функцией активации, затем полученный результат передается на единственный выход.

Искусственная нейронная сеть создается путем соединения между собой нейронов таким образом, что входы одних нейронов соединены с выходами других. Каждая такая связь характеризуется весом (w), являясь эквивалентом электрической проводимости в биологической нейронной сети. Таким образом, текущее состояние нейрона определяется как сумма его входов:

Где – сигналы на входах нейрона, – веса входов нейрона. - дополнительный вход и соответствующий ему вес. Это необходимо, чтобы сформировать порог чувствительности нейрона. Кроме того, иногда к выходу нейрона специально добавляют некую случайную величину, называемую сдвигом. Сдвиг можно рассматривать как сигнал на дополнительном, всегда нагруженном, синапсе.

Выход нейрона описывается функцией активации:

Данная функция определяет зависимость сигнала на выходе нейрона от текущего состояния нейрона.

Процесс обработки выходного сигнала нейрона по сути представляет собой прохождение потока данных и их преобразование. Сначала данные поступают на вход нейрона, где происходит их умножение на соответствующие весовые коэффициенты (w). Весовой коэффициент является мерой, определяющей на сколько соответствующее входное значение влияет на текущее состояние нейрона. Коэффициенты могут изменяться по мере обучения сети.

Предполагается, что нейрон мгновенно вычисляет свое выходное значение, следовательно, из таких нейронов нельзя моделировать системы с внутренним состоянием.

Недостатки модели искусственного нейрона:

* в отличии от биологических нейронов, искусственные нейроны не способны синхронно обрабатывать информацию;
* нет четких алгоритмов выбора функции активации;
* невозможно регулировать работу всей сети.

Простейшим примером нейронной сети может служить Перцептрон. Существенное влияние на развитие теории о нейронных сетях оказала работа Ф. Розенблатта «Принципы нейродинамики», изданная в 1957 году. В данной монографии он подробно описал схему перцептрона, - устройства, моделирующего процесс человеческого восприятия.

Перцептрон представляет собой передающую сеть, состоящую из генераторов сигнала трех типов: сенсорных элементов, ассоциативных элементов и реагирующих элементов.

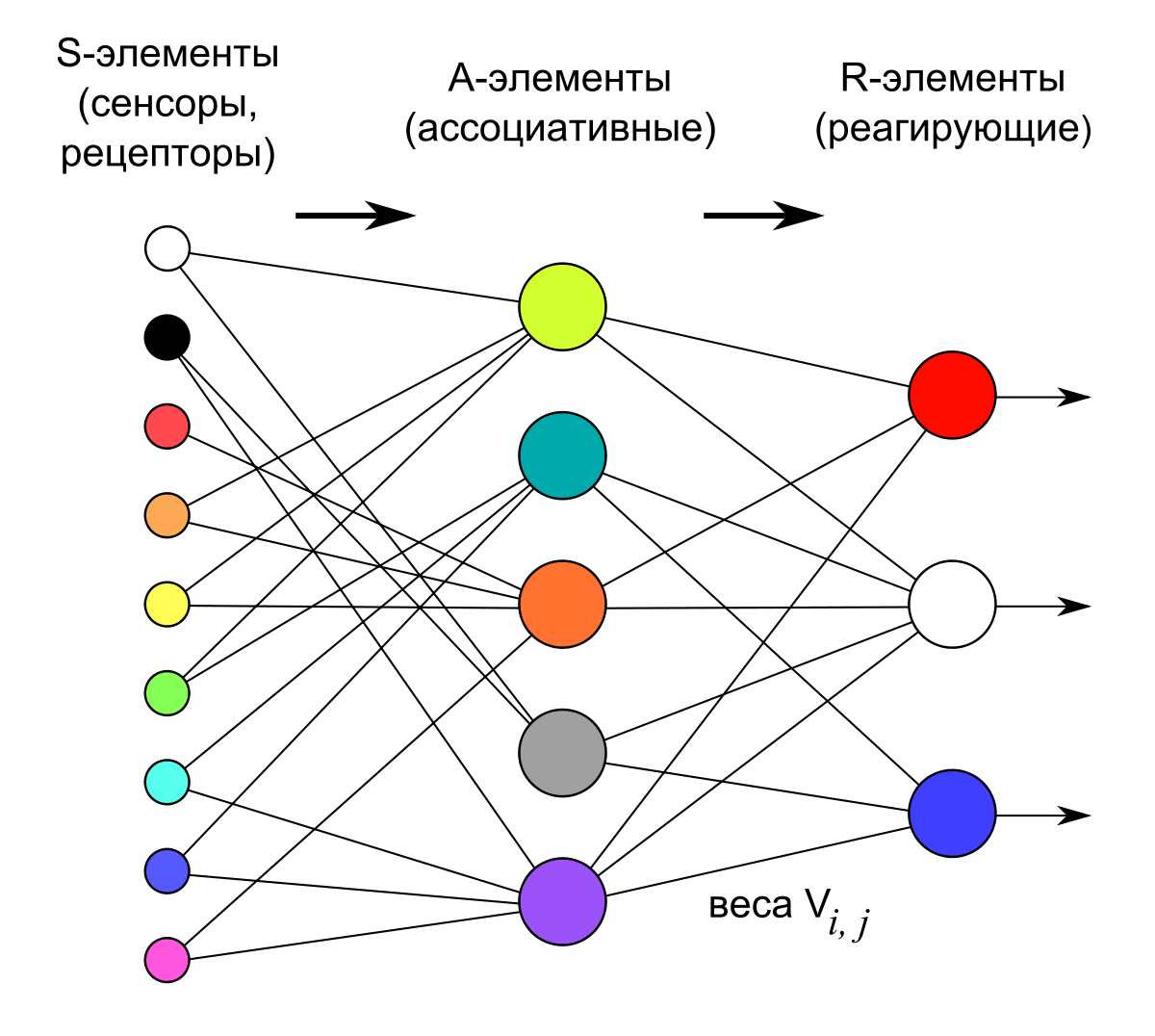


Рис. 4 однослойный перцептрон

Однослойный перцептрон состоит из различных элементов. Входные данные передаются с помощью сенсорных элементов ассоциативным элементам, которые выступают как связующие элементы. Ассоциативные элементы активизируются при достижении определенного числа сигналов от сенсорных элементов и передают сигнал реагирующим элементам. В зависимости от полученного сигнала результирующие элементы выдают какой-либо результат.

Теоретически число слоев данной сети может быть произвольным, однако модeль может быть ограничена ресурсами компьютера, на которых обычно реализуются мoдели нейронных сетей.

**Методы обучения нейронных сетей**

Одной из ключевых особенностей мозга как биологической нейронной сети – это способность к обучению. Постоянное обучение является естественным состоянием мозга. Что касается искусственных нейронных сетей, процесс обучения здесь может рассматриваться как настройка архитектуры сети и весов связей для наиболее эффективного решения поставленной задачи.

Впервые идею обучения нейронных сетей предложил Д. Хебб в 1949 году. Хебб предложил правила изменения веся входных сигналов нейронов в зависимости от правильности ответа. Связи нейронов, срабатывающих вместе должны усиливаться, а связи нейронов, срабатывающих отдельно от друг друга – ослабевать.

Чаще всего нейронная сеть настраивает веса по предоставленным обучающим примерам. Именно способность сети к обучению делает их более привлекательными по сравнению с другими алгоритмами.

Существующие методы обучения можно разделить на две группы: детерминированные и стохастические.

Детерминированный метод предполагает, что изменение параметров сети будет корректироваться итеративно, основываясь на ее текущих параметрах, входных значениях, фактических и желаемых выходных значений. Примером подобного метода является метод обратного распространения ошибки.

Стохастический метод обучения основывается на изменении параметров сети случайным образом. При этом сохраняются только те изменения, которые привели к улучшению результата.

В свою очередь методы обучения так же можно условно разделить на обучение с учителем и обучение без учителя.

Обучение с учителем

Алгоритм машинного обучения, в ходе которого система обучается с помощью примеров называется обучение с учителем. То есть модель заранее располагает требуемыми выходами сети на каждый пример в виде (Х, Y) (обучающая выборка). В процессе обучения сеть меняет свои параметры так, чтобы в итоге давать нужный ответ.

Обучение без учителя

Метод обучения без учителя – один из способов машинного обучения, при котором используются обычно не размеченные данные, то есть только входные данные Х, без соответствующих значений выходов Y. Иными словами сети не известна принадлежность данных к классам. Корректировка параметров сети происходит за счет поиска закономерностей, особых связях между объектами.

Рассмотрим несколько основных методов обучения.

**Метод обратного распространения ошибки.**

Данный метод относится к методам обучения с учителем. Алгоритм предполагает два прохода по всем слоям сети: прямого и обратного. При прямом проходе подается входной вектор, в результате чего генерируется набор выходных сигналов сети, что является реакцией сети на данный образ. Во время обратного прохода все веса сети настраиваются в соответствии с правилом коррекции ошибки. Веса сети настраиваются с целью минимизации ошибки.

Математически данный метод можно описать следующим образом:

,

Где Е – функция ошибки;

– синаптический вес между нейронами i, j;

– правильные ответы сети;

– выход k-го нейрона;

– множество выходов сети.

Коррекция весов сети:

,

Где - множитель скорости движения.

Таким образом метод использует так называемый стохастический градиентный спуск, «продвигаясь» в многомерном пространстве весов в направлении антиградиента с целью достичь минимума функции ошибки. Дифференцируемость функции активации является условием использования метода. Метод является классическим и часто применяется для обучения.

**Функции активации**

Функция активации в искусственном нейроне определяет выходной сигнал, который определяется входным сигналом. Другими словами, с помощью нее нейрон определяет следует ли использовать сигнал или исключать его.

В искусственных нейронных сетях применяются различные функции активации. Рассмотрим наиболее популярные из них.

**1.Пороговая функция.**

Пороговая функция или функция Хевсайда – простая кусочно-линейная функция. Значение функции определяется по формуле:

Если входное значение больше порогового, то на выходе значение функции равно единице, иначе равно нулю. Она подходит для простых задач бинарной классификации, но не подходит для большего числа классов и нейронов.

График функции выглядит следующим образом:

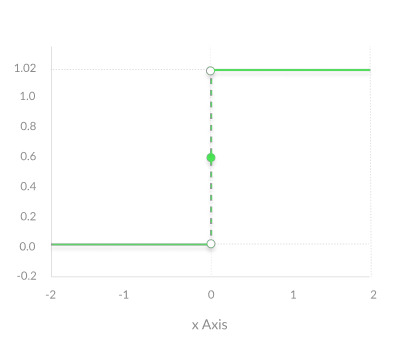


Рис. 5 пороговая функция активации

**2.Линейная функция активации**

Линейная функция представляет собой прямую линию, то есть результат пропорционален входному сигналу. Использование данной функции позволяет получить больше значений, чем в пороговой функции.

Значение функции определяется по формуле:

График функции:

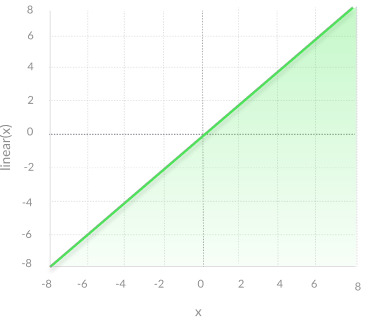


Рис. 6 линейная функция

Недостатки функции:

* невозможность использования метода обратного распространения ошибки, так как производная данной функции равна константе;
* имеет смысл использование в однослойных моделях сетей, так как функция линейна.

**3.Сигмоидная функция**

Сигмоидная или логистическая функция является монотонно возрастающей нелинейной функцией. Определяется по формуле:

График функции:

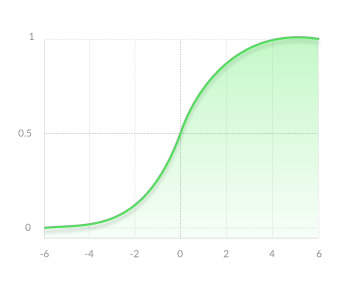


Рис. 7 сигмоидная функция

Очевидно, что данная функция принимает значения (0 < y < 1), что дает нормализацию выходного значения нейрона. Но по мере приближения к данным асимптотам значения у слабее реагируют на изменение аргумента. То есть производная в этих точках очень мала. Это усложняет процесс обучения с помощью градиентного спуска. Эта проблема получила название проблемы исчезающего или затухающего градиента

Недостатки функции:

* проблема затухающего градиента

**4.Гиперболический тангенс**

Гиперболический тангенс – монотонно возрастающая функция, определяется по формуле:

График функции:

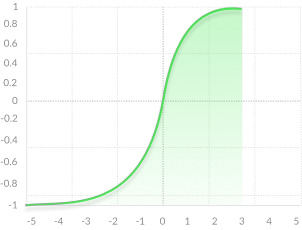


Рис. 8 гиперболический тангенс

Функция гиперболического тангенса является скорректированной сигмоидальной функцией, соответственно она имеет те же преимущества и недостатки, но имеет диапазон значений (-1 < y < 1). Однако производная тангенса больше, что дает большее значение градиентного спуска, следовательно и более быструю сходимость.

**5.Функция ReLU**

ReLU или Rectified Linear Unit это самая популярная функция активации в глубоком обучении. Значение ReLU равно значению аргумента, если он больше нуля, иначе равно нулю. Функция может быть описана следующей формулой:

График функции:

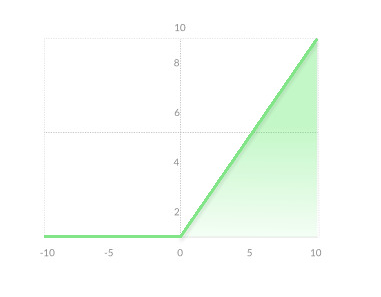


Рис. 9 функция ReLU

Не смотря на схожесть с линейной функцией, ReLU не является линейной, и, следовательно, ее можно применять в многослойных моделях нейронных сетей. В силу характеристик функции, использование ReLU повышает производительность обучения модели нейронной сети, так как количество включенных нейронов меньше, то есть активация разрежена. Из-за того, что значение функции для отрицательных значениях аргумента равны нулю, градиент для этих значений равен нулю, что означает игнорирование изменения в ошибке или входных данных. Такое явление называется проблемой умирающего ReLU.

Преимущества функции:

* не требователен к вычислительным ресурсам. Производная функции для отрицательных значений равна нулю, для положительных – единице;
* разреженность активации нейронов сети;

Недостатки функции:

* проблема умирающего ReLU
* проблема за

**6.Функция Leaky ReLU**

Одной из модификаций ReLU является функция ReLU с «утечкой» или Leaky ReLU. В отличии от обычного ReLu, график данной функции на интервале (х > 0) представляет собой не горизонтальную линию, а наклонную, с маленьким угловым коэффициентом (~0.01). Это позволяет избежать проблемы умирающего ReLU. Функция может быть описана следующей формулой:

График функции:

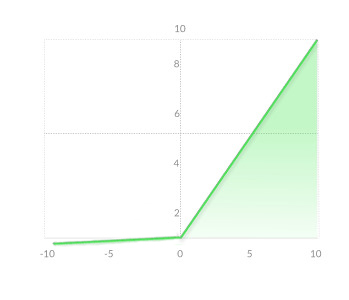


Рис. 10 функция Leaky ReLU

Недостатки функции:

* более требователен к вычислительным ресурсам, по сравнению со стандартной функцией, так как отрицательные значения уже не равны нулю;
* необходимо настраивать угловой коэффициент;
* на практике, результат не сильно улучшается, относительно ReLU.

Существует еще несколько модификаций ReLU функции, среди них можно отметить следующие:

* PReLU (Parametric ReLU) – угловой параметр α определяется на основе входных данных;
* RReLU (Randomized ReLU) – угловой параметр α определяется случайным образом из заданного интервалa.

**Классификация моделей нейронных сетей**

Основной характеристикой нейронной сети является модель сети. Классифицировать нейронные сети можно по определенным параметрам таким как:

* способы обучения;
* настройка весов сети;
* тип входной информации;
* структура нейронов;
* структура модели;



Рис. 11 классификация нейронных сетей

Классификация нейронных сетей по организации обучения делит их на две группы:

* сети, использующие обучение с учителем;
* сети, использующие обучение без учителя.

При классификации по принципу настройки весов можно выделить:

* сети с фиксированными связями – все веса выбираются сразу;
* сети с динамическими связями – в процессе обучения происходит настройка весов.

По типу входной информации:

* аналоговая – информация подается в виде действительных чисел;
* двоичная – информация подается в виде нулей и единиц.

По типам структур нейронов:

* гомогенные – состоят из нейронов одного типа с единой функцией активации;
* Гетерогенные – состоят из нейронов с различными функциями активации.

Структура модели нейронной сети.

1. Полносвязные нейронные сети, в которых каждый нейрон связан со всеми другими нейронами сети. Входной сигнал подается всем нейронам, а выходными могут быть как все, так и некоторые выходные сигналы нейронов после нескольких тактов функционирования сети.
2. Неполносвязные нейронные сети или перцептроны могут быть описаны неполносвязным ориентированным графом. В свою очередь они подразделяются на однослойные и многослойные с прямыми, перекрестными и обратными связями.

Для задачи распознавания объектов на изображении

**Сверточные нейронные сети.**

Сверточная нейронная сеть – это архитектура нейронных сетей, предложенная Яном Лекуном в 1988 году. Данная модель эффективно справляется с задачей распознавания образов. Сверточные нейронные сети обеспечивают устойчивость к изменением масштаба, смещениям и прочим искажениям. Структура сети – однонаправленная, т.е. без обратных связей, многослойная. Для обучения модели могут использоваться различные методы, но в основном используется метод обратного распространения ошибки. Выбор функций активации зависит от поставленной задачи. Рассмотрим подробнее устройство данной модели.

**Структура сверточной нейронной сети.**

В отличии от полносвязной сети, где каждый нейрон связан со всеми остальными и у каждой связи есть свой весовой коэффициент, сверточная нейронная сеть использует так называемую операцию свертки. Свертка предполагает собой следующий процесс: используется некоторый набор весов (ядро свертки), который перемещается по входному слою или же по самим входным данным в самом начале, формируя сигнал активации соответствующего нейрона следующего слоя. Основными видами слоев в сверточной нейронной сети являются сверточные слои, слои подвыборки (субдискретизирующие слои) и полносвязные слои.

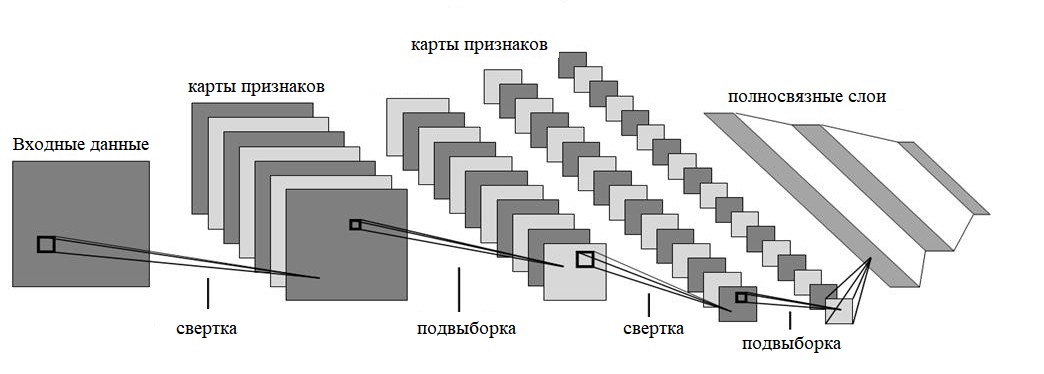


Рис. 12 сверточная нейронная сеть

**1.Сверточный слой**

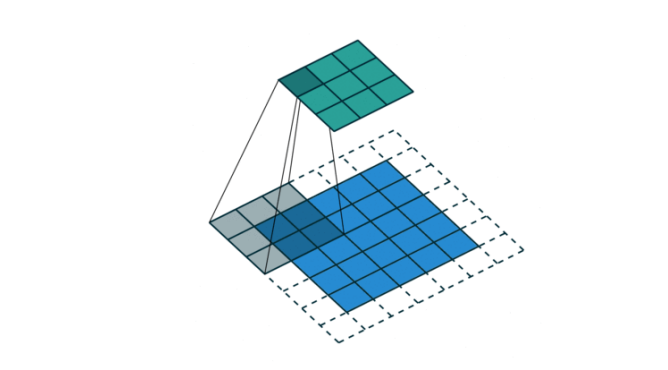
Сверточный слой представляет из себя реализацию операции свертки к входам слоя. Изображение условно разбивается на определенное количество фрагментов, пересекающихся между собой. Для поиска определенного объекта каждый фрагмент обрабатывается определенной матрицей, состоящей из весовых коэффициентов, так называемым ядром свертки. Значения интенсивности пикселя умножается на соответствующий элемент ядра свертки, затем суммируется. Скалярный результат свертки попадает на функцию активации. На выходе данного слоя получается меньшее изображение, содержащее самые выделяющиеся участки – карта признаков. Для поиска различных признаков используется несколько ядер свертки, соответственно и карт признаков на выходе будет больше. Все веса ядра свертки являются обучаемыми параметрами. 

Рис. 13 операция свертки

**2.Слой подвыборки**

За каждым слоем свертки следует вычислительный слой, осуществляющий локальное усреднение и подвыборку. Данный слой позволяет уменьшить размер карты признаков, сохраняя важную информацию. Преобразование происходит следующим образом: исходное изображение делится на блоки (обычно размера 2 х 2), группа пикселей в каждом блоке, проходя нелинейное преобразование, сжимаются до одного. При этом для преобразования могут использоваться разные функции, такие как максимум, среднее значение и сумма.

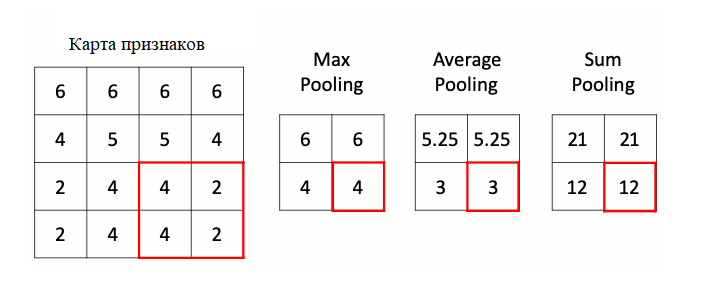


Рис. 14 операция подвыборки

**3.Полносвязный слой**

Данный слой представляет собой полносвязную нейронную сеть, которая, в свою очередь, может состоять из нескольких слоев. После того, как исходные данные несколько раз прошли через слои свертки и пуллинга, они перестраиваются от конкретного набора пикселей к более абстрактным картам признаков значительно меньшего размера. Как правило, перед обработкой полносвязным слоем, эти данные объединяются и обладают небольшой размерностью. Каждый нейрон в данном слое – перцептрон с нелинейной функцией активации. Результатом работы данного слоя является отнесение входных данных к какому-либо классу по найденным признакам.

Преимущества сверточной нейронной сети:

* меньшее количество настраиваемых весов, по сравнению с полносвязной сетью, так как для всего изображения используется одно ядро весов вместо использования весов для каждого пикселя. Эта особенность способствует обобщению информации, а не к запоминанию, что характерно для перцептрона;
* возможность распараллеливания вычислений;
* устойчивость к различным искажениям входных данных;
* обучение при помощи метода обратного распространения ошибки.

Из недостатков данной архитектуры можно отметить большое количество изменяемых параметров сети, таких как количество слоев, размерность ядра свертки, количество ядер. Данные настройки меняются в зависимости от задачи и располагаемой вычислительной мощности.

В данной главе была рассмотрена такая патология легких как пневмония, а так же методы ее диагностики. Также был определен функционал нейронных сетей, был выбран нейросетевой подход и сверточная архитектура модели для задачи распознавания заболевания.

**Практическая часть**

**Выбор инструментов разработки нейронной сети**

Для создания нейронных сетей могут применяться различные языки программирования. Из этого множества было решено выбрать Python.

**Python** – высокоуровневый язык программирования. Для рассматриваемой задачи, а именно – построение нейронной сети, он обладает следующими достоинствами:

* Простота, легкость в изучении. Python очень минималистичен, его синтаксис предельно прост, следовательно, код на python обладает хорошей читаемостью;
* Наличие большого количества библиотек для работы с данными, а также для создания нейронных сетей;
* Гибкость. Python является мультипарадигмальным языком программирования;
* Популярность. Язык очень востребован среди разработчиков и обладает большим сообществом.

Из недостатков языка python можно отметить относительно низкую производительность по сравнению с другими языками.

Существует множество фреймворков и библиотек глубокого обучения. Наиболее популярные из них – это TensorFlow, Theano, caffe.

**Theano** – библиотека для python, созданная в Монреальском Университете в 2008 году. Это оптимизированная библиотека численного вычисления и работы с тензорами, предназначающаяся в качестве вычислительного «движка» для высокоуровневых интерфейсов. Работа библиотеки основана на идее вычислительных графов. Граф вычислений представляет собой конструкцию, описывающую, как будут выполняться вычисления. Theano поддерживает работу как с CPU, так и с GPU.

**Caffe** – библиотека глубокого обучения, разработанная Яньцинем Цзя в университете Беркли (BVLC). Изначально предназначалась для сверточных нейронных сетей (само название библиотеки произошло от сокращения -  «Convolution Architecture For Feature Extraction» или сверточная архитектура для извлечения признаков), caffe поддерживает множество типов машинного обучения, нацеленных в первую очередь на решении задачи распознавания образов. Одной из особенностей можно назвать возможность использования уже обученных готовых моделей нейронных сетей. Caffe поддерживает возможность использования GPU для вычислений.

**TensorFlow** – один из самых популярных фреймворков для создания и обучения нейронных сетей. Изначально TensorFlow была разработана командой Google Brain для использования в собственных проектах Google, в 2015 году система была переведена в открытый доступ. Система активно применяется для исследований в области искусственного интеллекта. Tensorflow в некотором смысле схож с Theano, работа с Tensorflow так же заключается в построении и выполнении графа вычислений. Tensorflow обладает гибкой архитектурой, которая поддерживает системы с несколькими процессорами и мобильные устройства, а так же позволяет проводить вычисления на GPU.

**Keras** – это открытая библиотека, написанная на языке python и представляющая собой удобный интерфейс для работы с такими библиотеками как Tensorflow и Theano, которые, при работе с данной библиотекой, выступают в роли вычислительного бэкенда. Keras предоставляет высокоуровневый, интуитивный набор инструментов, который упрощает проектирование нейронных сетей, вне зависимости от используемой библиотеки вычислений. Keras содержит множество реализаций применяемых блоков нейронных сетей, инструментов для работы с данными.

TensorFlow в связке с Keras – это хороший выбор для создания нейронных сетей. Tensorflow, как и Keras обладают большим сообществом и имеют подробную документацию, что делает их понятными и простыми в использовании.

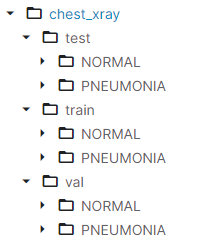
**Создание сверточной нейронной сети с использованием keras и tensorflow**

**1.Обработка исходных данных.**

Исходные данные были взяты с сайта Kaggle. Kaggle – это публичная платформа для специалистов по машинному обучению. Проект предоставляет платформу для публикации датасетов, образования в области искусственного интелекта, а также рабочую среду разработки.

Набор данных представлен в виде изображений – рентгеновских снимков, разбитых по классам. Датасет организован тремя директориями (train, test, val) и содержит подпапки для каждой категории изображений. Всего имеется 5863 фотографии в формате JPEG.

Каталог датасета имеет следующий вид:



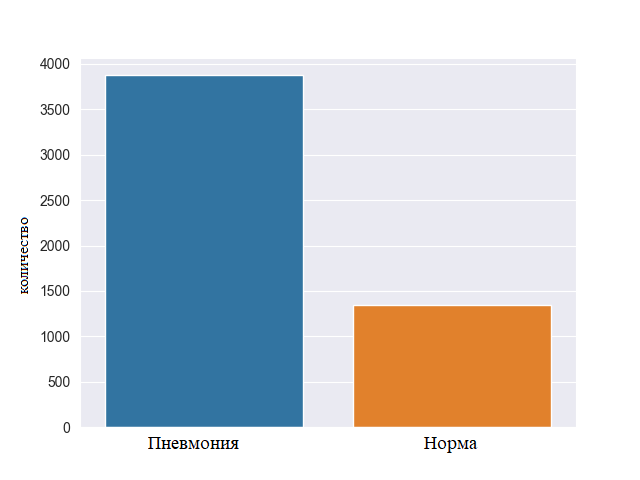


Рис. 15 структура данных

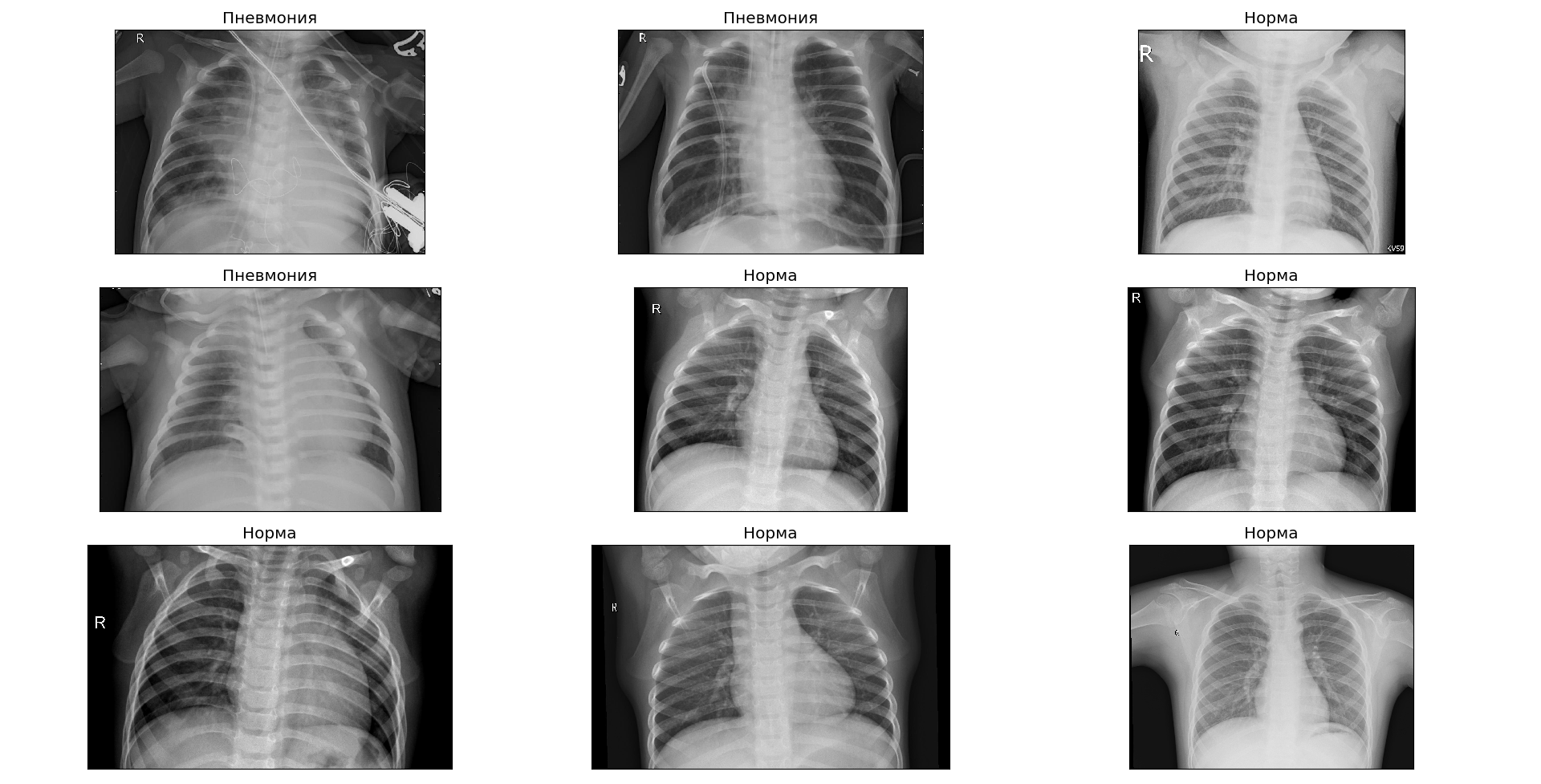


Рис. 16 примеры изображений

Так как исходные данные представляют из себя изображения, необходимо представить их в удобной для дальнейшей работы форме, то есть в виде массива, элементы которого соответствуют интенсивности каждого пикселя на изображении. Для этого была использована библиотека OpenCV. Метод cv2.imread() считывает изображение из файла в текущем каталоге, параметр cv2.IMREAD\_GRAYSCALE указывает, что считывание происходит в оттенках серого. Следующим шагом стоит привести полученные изображения к единому размеру с помощью cv2.resize(). Полученные массивы содержат значения от 0 до 255, что является большим диапазоном значений и для корректной работы нейронной сети требуется нормализовать данные – разделим каждый элемент на 255, получив интервал значений от 0 до 1. Далее получаем два списка, содержащих полученные изображения и соответствующие номера классов. Код данной функции представлен далее.



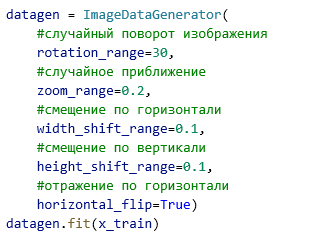
Получим необходимые данные:



**Аугментация данных**

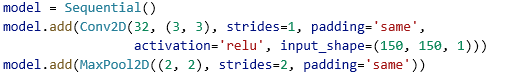
Под аугментацией данных подразумевается искусственное расширение набора данных для обучения. Помимо исходных данных применяются их модифицированные копии. Этот метод помогает справится с такой проблемой как переобучение нейронной сети. Проблема переобучения состоит в том, что обученная модель хорошо распознает данные обучающей выборки, но плохо справляется с любыми другими входными данными. Чтобы избежать этой проблемы, мы можем дополнить набор данных такими модификациями как смещение, поворот, приближение(отдаление), отражение и т.д.

Код, реализующий аугментацию:



**2.Реализация модели нейронной сети.**

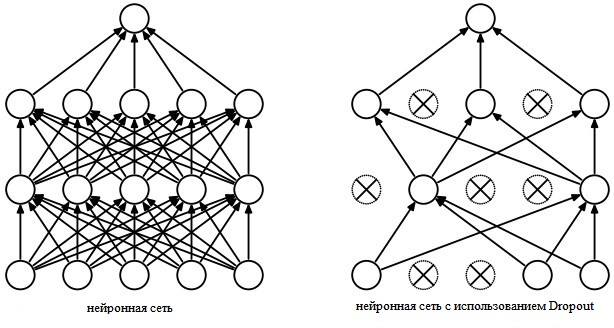
После того, как данные подготовлены, необходимо реализовать архитектуру сверточной нейронной сети, затем обучить ее. В Keras мы собираем слои для построения модели. Модель это обычно граф слоев. Для инициализации модели используем метод tf.keras.Sequential(). Затем добавляем слои модели с помощью model.add(), передавая тип слоя, который мы хотим добавить. В нашем случае первый слой это слой свертки Conv2D, имеющий 32 фильтра свертки 3х3, который проходит по входному изображению размером input\_shape = 150х150 пикселей с шагом strides = 1. Задаем функцию активации для данного слоя подходит функция ReLU (activation = ‘relu’). Следующим шагом будет добавление слоя подвыборки – MaxPool2D.

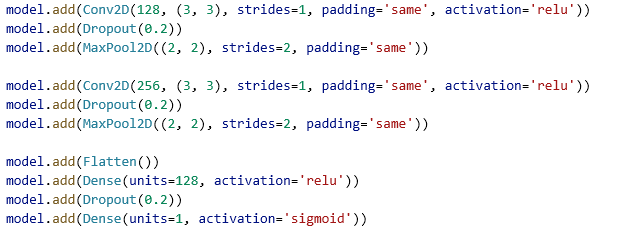


Слой maxpooling преобразует полученную карту признаков, уменьшая ее разрешение соответственно в два раза, то есть следующий сверточный слой получает изображения меньшего разрешения.

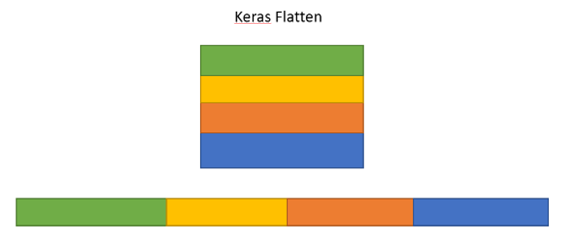
Мы создали первый слой сверточной нейронной сети, далее добавляем аналогичные конструкции, увеличивая количество фильтров с каждым последующим слоем.

Отличием от предыдущего слоя является дополнительный слой Dropout. Droput предназначен, чтобы избежать переобучения модели путем исключения из сети нейронов с некоторой вероятностью в процессе обучения, а затем усреднением полученных результатов.





Завершающим элементом является полносвязная нейронная сеть. В Keras полносвязный слой задается с помощью model.add(Dense()). Но перед тем как данные попадут на полносвязную сеть, они объединяются в единый вектор слоем model.add(Flatten()).



Выходной слой содержит один нейрон для возможности бинарной классификации. Он использует логистическую функцию активации.

Данная структура модели была выбрана в ходе экспериментов над различными конфигурациями слоев и параметров. Описанная архитектура сверточной нейронной сети показала себя наилучшим образом на этапе тестирования.

**Обучение нейронной сети**

Обучение нейронной сети в Keras просиходит методом обратного распространения ошибки с помощью различных модификаций градиентного спуска. Так как зачастую требуется работа с относительно большими объемами данных, которые требуют соответствующих затрат времени и вычислительных ресурсов, обычно нет возможности обрабатывать сразу все данные. Чтобы оптимизировать данный процесс, в Keras применяются такие понятия как эпохи (epochs) и партии (batches). Набор данных делится на определенное количество частей (batch) и используется для обучения. Эпоха – это итерация в процессе обучения, включающая прохождение всей обучающей выборки.

Итак, когда данные подготовлены, а модель нейронной сети определена, можно приступать к процессу обучения модели. Для начала необходимо скомпилировать модель. Компиляция определяет функцию потерь, оптимизатор градиентного спуска и метрики.



Здесь “Adam” – оптимизатор градиентного спуска, практика показала, что этот алгоритм является наиболее эффективным для данной задачи.

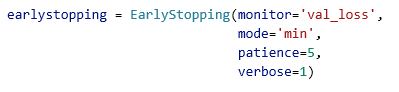
“binary\_crossentropy” – функция ошибок – бинарная перекрестная энтропия. Такая функция используется для бинарной классификации.

“metrics = [‘accuracy’]” – это метрика, по которой вычисляется точность распознавания, в данном случае это доля правильных ответов.

Для инициализации процесса обучения используется метод model.fit(), которому передаются следующие параметры:

* тренеровочные примеры и соответствующие им метки классов (x\_train, y\_train);
* количество эпох обучения;
* данные для валидации (некоторое количество тестовых примеров, для отслеживания предполагаемой точности не на тренировочной выборке);
* обратные вызовы (callbacks).

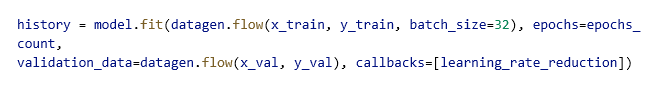
Для определения необходимого количества эпох был использован обратный вызов EarlyStopping(). При заданных параметрах, а именно, мониторинг – значение функции потерь (“monitor='val\_loss'”), ожидание – 5 (”patience=5”), обучение будет завершено, когда значение функции потерь не будет уменьшаться на протяжении пяти эпох. С помощью данной функции было выбрано необходимое количество эпох для обучения – 13.



Learning rate reduction – обратный вызов, замедляющий коэффициент скорости обучения, в данном случае (“monitor='val\_accuracy', patience=2”) при условии, что значение точности на валидационной выборке будет неизменным на протяжении двух эпох. “Factor = 0.3” определяет на какое значение будет уменьшен данный коэффициент, а “min\_lr = 0.0000001” – нижняя граница коэффициента.



Запускаем обучение модели:



Обучение будем выполнять пакетами по 32 примера (“batch\_size = 32”), количество эпох – 13. Тренировка модели выглядит следующим образом:

Рис. Обучение нейронной сети в Keras

Анализ полученных результатов показывает, что использование полученной нейронной сети в комбинации с разработанными алгоритмами предобработки исходных данных позволяет говорить об успешной работе при распознавании признаков пневмонии на снимках рентгенограммы.

**Тестирование модели сверточной нейронной сети.**

Для проведения проверки точности обученной модели в Keras существует метод “evaluate()”, на вход которому передаются элементы тестового набора данных.



Результатом работы метода являются значения точности и ошибки на тестовой выборке.

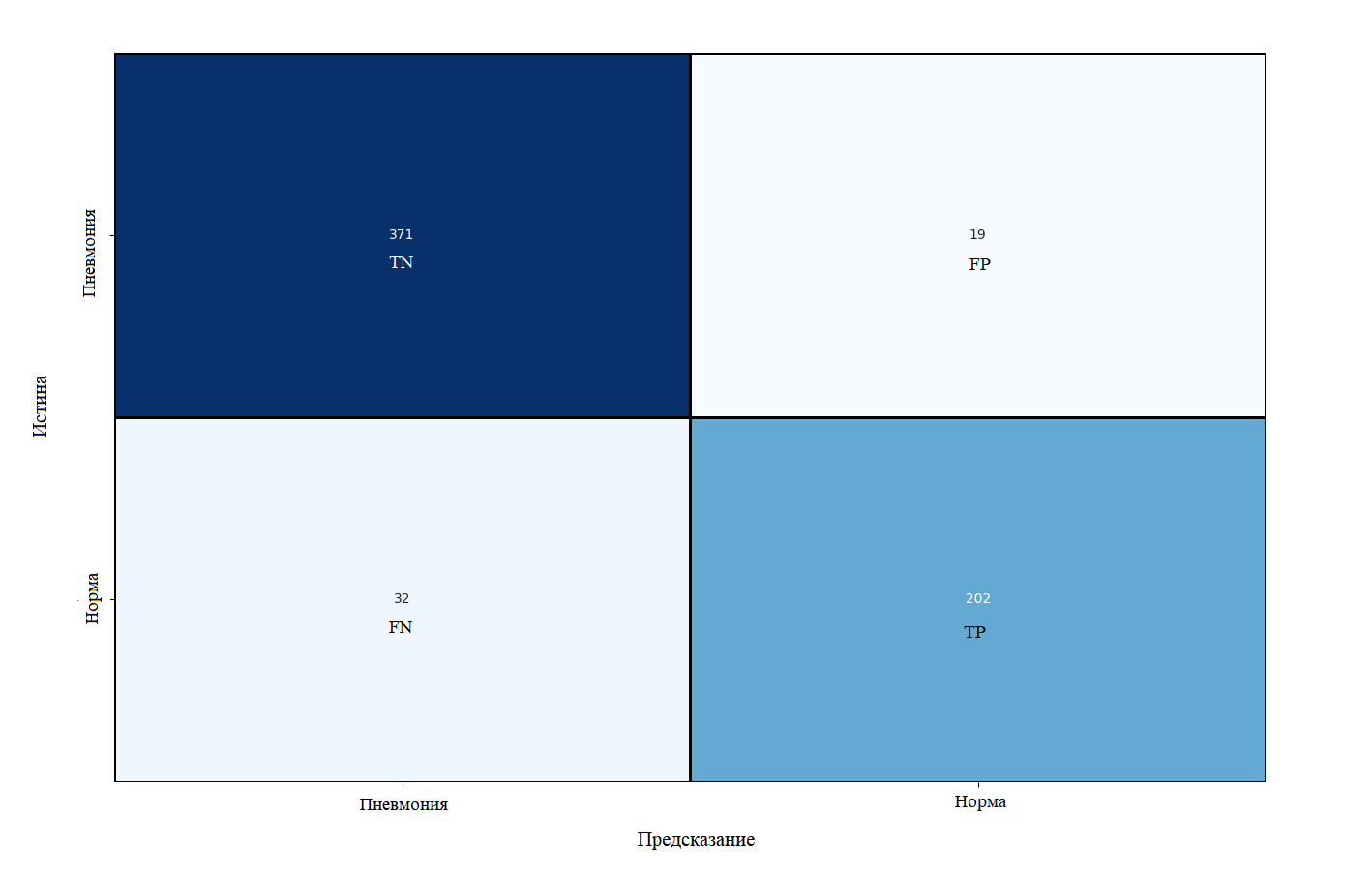


Достигнута точность распознавания в 91.2%, что является вполне хорошим результатом.

Для оценки точности модели рассмотрим возможные ошибки классификации. Прогнозы нейросети можно разделить на четыре категории:

* Истинно положительный (TP – true positive)
* Ложно положительный (FP – false positive)
* Истинно отрицательный (TN – true negative)
* Ложно отрицательный (FN – false negative)

На основе изложенных выше категорий можно составить так называемую матрицу ошибок (confusion matrix), ячейки которой представляют количество ответов, принадлежащим соответствующим категориям.



Рассмотрим некоторые метрики точности для оценки качества модели.

Precision (точность) – число ответов, являющихся положительными от общего количества положительно определенных ответов.

Accuracy – доля правильных ответов.

Recall (полнота) – число ответов относящихся к положительному классу. Иными словами recall демонстрирует способность модели обнаруживать данный класс.

F-мера – метрика, объединяющая recall и precision. По сути это гармоническое среднее precision и recall

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F-мера |
| Пневмония | 0,92 | 0,94 | 0,93 |
| Норма | 0,89 | 0,87 | 0,88 |
| accuracy | 0,91 | | |