Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет компьютерных систем и сетей

Кафедра информатики

Машинное обучение

РЕФЕРАТ

на тему

# ОСОБЕННОСТИ БОЛЬШИХ ДАННЫХ И ГЛУБОКОГО ОБУЧЕНИЯ

Магистрант:

Н.Р. Ровдо

858641

Проверил:

М.В.Стержанов

МИНСК 2020

# СОДЕРЖАНИЕ

# 1 ВВЕДЕНИЕ

Новая наука - наука о данных - занимающаяся работой над сбором, пред-обработкой и анализом данных получает большую популярность. Одним из перспективных результатов работы науки являются разработанные программистами (англ. data scientists) машинные методы обучения. Среди множества способов анализа данных используются те, что применяются в основном математике и статистике.

В настоящее время вычислительная техника способна обрабатывать большие данные в реальном времени.

Интернет-ресурсы очень часто используют большие данные для продвижения своих услуг. Большие данные помогают выявить закономерность действий пользователя и предложить ему купить книгу или фильм, имея при этом высокий шанс на удачу, т.к это предложение основано на индивидуальных особенностях покупателя.

Эта технология используются не только в сфере IT. Ею пользуются в милиции для выявления преступных элементов, в медицине для определения болезни по симптоматике пациента, в военной сфере для анализа боевой обстановки, в банковской сфере для решения о выдаче кредита, в сфере изучения социума и многого другого.

Большие данные являются новым этапом в развитии технологий, а в будущем будут иметь место в каждой жизнедеятельности человека, что позволит оптимально использовать ресурсы и улучшить качество жизни.

# 2 БОЛЬШИЕ ДАННЫЕ

Большие данные - это постоянно увеличивающийся набор неструктурированных и структурированных данных, которые

обрабатываются компьютером. В современном мире существует чёткое различие между большими данными и структурированными базами данных, которые имеют схему разметки и описанную структуру для каждой сущности. Большие данные, хоть и вначале не имеющие столь большой объём, обладают крайне полезным свойством, они могут подстраиваться под решаемую задачу, если использовать их совместно с методами машинного обучения. Три V: скорость(velocity), разнообразие (variety) и объём (volume) - являются главными характеризующими определениями больших данных.

Обычно рассматривается объём данных от 100 терабайт, но в последнее время эта цифра доходит до эксабайта(10^18 байт). Именно с такого размера компьютер способен чётко определять закономерности для дальнейшей обработки аналитическими средствами и приводить к понятному человеку агрегированном виде.

Вариативность позволяет машине видеть закономерности там, где человек их не выявит. Например, зависимость роста заработных плат от магнитных аномалий или скорость развития эпидемии от ухудшения погоды.

Как правило, все данные, поступающие на обработку, агрегируются в облачных хранилищах и в обязательном порядке дублируются на независимое хранилище для исключения возможности потери этих данных.

Данные могут быть различного вида: это может быть аудио-файл, текстовый документ, графическое изображение, программный код и другое. В обычной ситуации процесс обработки информации следующий: информация проходит через специальный алгоритм контролируемый программистом, далее информация приводится в структурированный вид удобный для понимания людьми, а после людьми же и анализируется, после чего принимается решение. Всё больше и больше внедряется использование методов машинного обучения, где весь процесс обработки данных, принятия решения контролируется машиной, следовательно, меньше вклад человека.

Как правило, главными источниками больших данных принято считать социальные сети и интернет вещей, ещё они могут приходить из больших организаций и компаний, внутренняя информация которой собиралась и не анализировалась долгое время. Также сфера астрономии и биологии имеет большие данные для анализа.

Инструментарий работы с большими данными сейчас крайне широк. Если посмотреть на тот объём информации, который необходимо не только хранить, но и обрабатывать, то использование классических жёстких дисков - это крайне плохой вариант. Для обработки данных следует использовать современные твёрдотельные накопители SSD, а если их использование невозможно, то можно их арендовать, однако перенос данных на арендованные носители по сети может оказаться не только дорогим, а ещё и крайне долгим процессом.

Приведём пример, допустим компания намерена арендовать удалённые сервера для обработки 50 петабайтов информации. При передачи данных по сети со скоростью 1 гигабит в секунду при идеальных условиях вам бы понадобилось 10 лет, чтобы загрузить данные. В данном случае, наиболее приемлемым способом передачи является физическое копирование и перевозка данных с помощью сервиса AWS Snowmobile. В таком случае понадобиться несколько недель на весь процесс.

Наиболее популярные инструменты для структурирования, анализа и параллельного доступа к данным:

- Hadoop
- MapReduce
- Язык программирования R
- NoSQL

Очень важна возможность использования параллельных вычислений в работе с большими данными, что позволит в разы сократить время на её обработку.

# 3 НЕЙРОННЫЕ СЕТИ

Искусственная нейронная сеть - это математическая система или модель, имеющая программную реализацию, моделирующая поведение биологических нейронов живого организма. Нейронные сети находят своё применения в различных областях. Стоит упомянуть, что всеобщее мнение о том, что биологические и математические нейроны имеют схожую структуру, имеет всё больше и больше противников. Отличий может быть гораздо больше. Тем не менее, искусственные нейронные сети обучаются схожими методами, как и биологические.

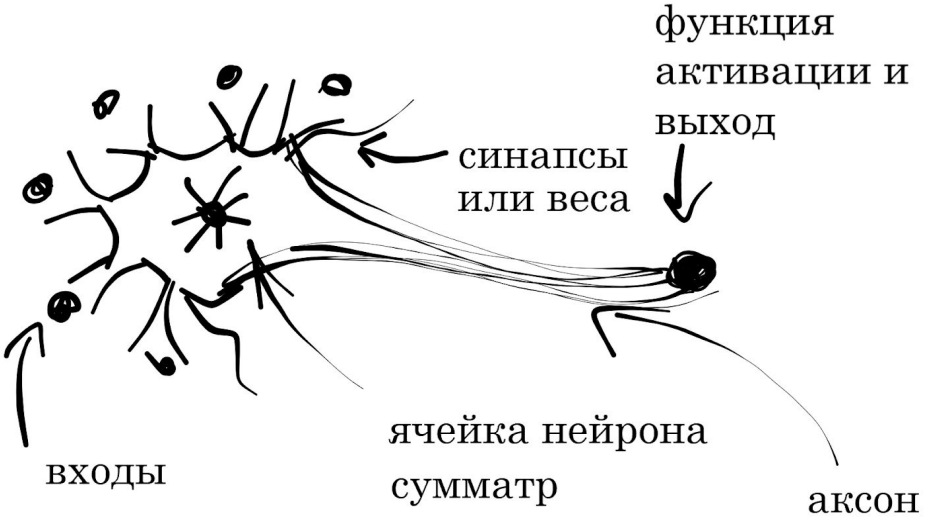


рис.1 Модель искусственного нейрона

Следует отметить плюсы использования нейронных сетей по сравнению с методами машинного обучения: а именно при применении

последних необходимо вручную выбирать ключевые признаки, которые могут оказаться ошибочными, с другой стороны нейронные сети требуют больших вычислительных мощностей для эффективной работы, т.к каждая связь нейронов, порядок которых может доходить до миллионов, требует неоднократного вычисления, т.к необходимо обрабатывать большой объём информации. Нейронные сети требуют достаточно большой объем информации для корректной работы, однако немаловажную роль является структура сети, параметры связи нейронов, количество слоёв, функции активации и т.д. Грамотный выбор каждого из параметров сделает нейронную сеть эффективной.

Использование нейронных сетей позволяет решать задачи классификации, предсказания, обнаружения лучше и быстрее человека, т.к эффективность мозга человека может быть снижена различными внешними естественными факторами. Это позволяет нейронной сети избежать большой доли человеческих ошибок при обработке данных и принятии решений в местах, где эта самая ошибка может оказаться критической, например в медицине или в системе безопасности.

Однако, в современном мире невозможно автоматизировать любые процессы нейронными сетями, это связано с тем, что существуют так называемые нестандартные ситуации, последовательность действий при которых нейронными сетями может быть не заложена. Ситуацию облегчает то, что ситуация перестаёт быть нестандартной до тех пор пока человек не научится в ней принимать правильные решения.

Нейронные сети способны выполнять широкий круг задач, в рамки которого входят:

1. Прогнозирование
2. Распознавание

3. Моделирование
4. Классификация
5. Кластеризация
6. Выявление аномалий
7. Ранжирование и т.д

Стоит сказать, что не только нейронные сети способны наиболее эффективно решать эти задачи, есть случаи, когда простая линейная регрессия способна выдавать лучший результат.

# 4 ГЛУБОКОЕ ОБУЧЕНИЕ

Глубокое обучение представляет собой комплекс методом машинного обучения использующий многоуровневые нейронные сети для работы с абстрактными моделями обученными на больших данных.

Глубокое обучение является подмножеством более общего класса - методов машинного обучения, который в свою очередь является подмножеством методов искусственного интеллекта.

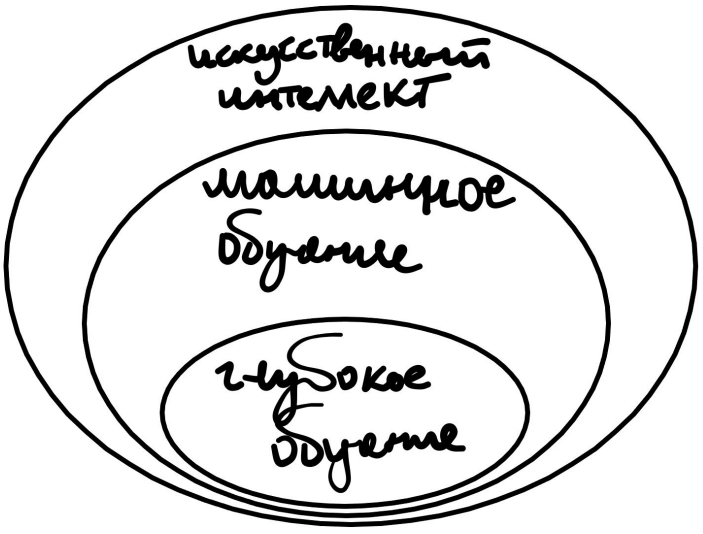


рис. 2. Концепция подмножества методов искусственного интеллекта

Представление данных может иметь вид таблицы заполненного записями, запись может состоять из набора ячеек, каждая из которых имеет значение в определённом диапазоне, например, температура за каждый час в сутках, или что в более упрощённом для машины виде значения от нуля до единицы, т.е представления температуры не на шкале Цельсия, скажем от -100 до 100 для решения задач по прогнозированию температуры. Одни представления являются более удобными для обработки, чем другие. Возьмём, например задачу компьютерного зрения, для её решения понадобится изображения и местонахождение объекта на изображении. Местонахождение можно закодировать четырьмя координатами: левая верхняя точка и правая нижняя точка прямоугольника в рамках которого находится объект.


рис. 3. Пример кодирования местоположения объекта по целочисленным координатам

Очень важно правильно выбрать обучающую выборку для нейронной сети. Допустим, в распоряжении есть набор данных распределённых по

регионам и было выбрано соотношение 75% для разграничения обучающей и тестовой выборки. Будет неправильно сортировать регионы по алфавиту и взять первые 75% в качестве обучающей выборки и остаток на тестовую. В данном случае будет потеряна информация о пропущенных регионах. Правильно будет взять случайные 75% данных от каждого региона. В таком случае выборка будет равномерной.

Использование глубокого обучения упрощает процесс выбора признаков, делает его автономным. Метод сам решает, как признак является более значимым, а какой - менее значимым, при этом проводя обучение как с учителем, так и без.

Немаловажно, что нейронная сеть с одним скрытым слоем имеет ограничения по входным шаблонам данных, в то время как сеть с более, чем одним скрытым слоем, например свёрточная нейронная сеть, такого недостатка не имеет. Она способна эффективно классифицировать изображения.

Наиболее известным методом обучения нейронной сети является метод обратного распространения ошибки. Данный метод работает, как глубокими, так и не с глубокими нейронными сетями. Основной смысл в том, чтобы настроить веса нейронов таким образом, чтобы выходные с наибольшей вероятностью выдавали ожидаемый результат. В данном методе очень важна обратная связь, т.к она является важнейшей частью метода. Главной целью метода является баланс правильных ответов нейронной сети как на данные, которые использовались при обучении, так и на данные, которые в обучении не участвовали.

Метод состоит из трёх основных шагов: получение данных на вход нейронной сети с последующим расчетом весов в направлении выхода, после получения ответа сети происходит вычисление ошибки и перерасчёт весов в обратном направление и в конце - корректировка весов (соединений между

соответствующими нейронами). После того, как обучение закончено, получение результата предполагает только первый этап метода, а именно, только получение ответов сети исходя из начальных данных и их последовательного расчета между нейронами. Стоит отметить, что первый этап является самым быстрым по времени, а оставшиеся два занимают основную часть обучения.

# 5 ВИДЫ ГЛУБОКОГО ОБУЧЕНИЯ

## 5.1 Обучение с учителем

Обучение с учителем - это самый известный и самый понятный тип глубокого обучения. Данный тип обучения характерен для решения двух типов задач: прогнозирование и классификация. Тип предполагает наличия пар объект-ответ, по которому сеть будет проводить обучение. Объект может состоять из одного или нескольких значений именуемых признаками. Набор таких пар называется обучающей выборкой. Сеть обучается по размеченным данным и подстраивает выходные значения под обучающую выборку. В обучении с учителем вводят такие характерные элементы как оценка точности. Она представляет собой такой же набор пар объект-ответ, только отличие в том, что этот набор не участвует в обучении. С помощь данного набора с помощью различных метрик(например f1 мера) оценивается эффективность обученной модели. Далее, в зависимости от результата, применяется дообучение модели большим количеством данных, либо же, корректировка архитектуры сети, либо добавляются, либо убираются признаки.

## 5.2 Обучение без учителя

Обучение без учителя является противоположностью обучению с учителем. Обучение без учителя используется для построения моделей в которых невозможно чётко классифицировать или предсказать объект. Задача модели при таком обучении - найти закономерности между объектами или их группировка. Задача кластеризации - яркий представитель задачи, для которой необходимо обучения без учителя. При обучении без учителя не ставится задача принятия решения исходя из данных, задача состоит в понимании данных.

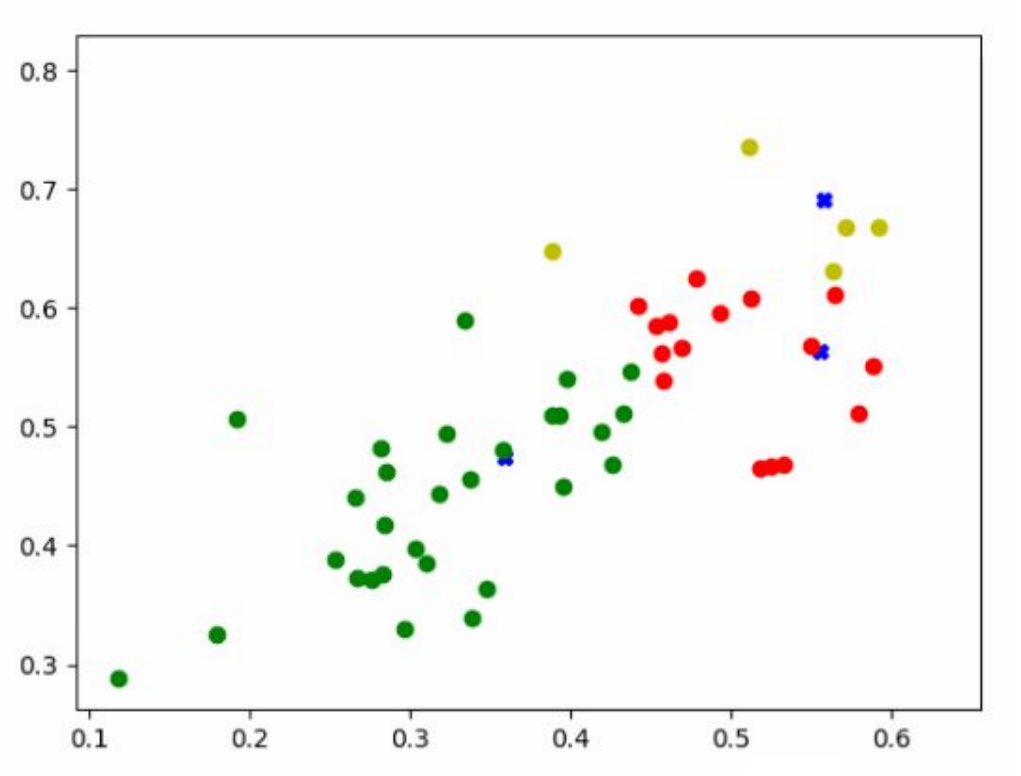


рис.4. Результат кластеризации нейронной сетью после обучения без учителя

Ещё одна задача, которую можно решить обучением без учителя - задача снижения размерности, она позволяет наглядно представить данные убрав незначительные признаки из выборки и позволить представить многомерные данные в понятном человеку виде, например привести их к двумерному виду.

## 5.3 Обучение с подкреплением

Обучение с подкреплением позволяет “направлять” обучение сети в нужном направлении. Данный тип обучения использует систему наград и штрафов для тренировки сети. Такой тип обучения применяется для обучения сети достигать цели при взаимодействии с определённым окружением, например победить в шахматной партии. Таким образом через множество итераций модель поймёт каким образом необходимо действовать в той или иной ситуации для того, чтобы получить желаемый результат. В таком случае данный тип обучения можно считать частным случаем обучения с учителем, где учителем является взаимодействующая среда (учитель неявный). Среда наказывает или поощряет модель в зависимости от результата, таким образом через множество итераций система сведёт количество наказаний к минимуму и получит жеалемый результат, что будет решением задачи.

Примеры реализаций:

- PyBrain
- OpenAI

## 5.4 Обучение с частичным привлечением учителя

Обучение с частичным привлечением учителя применяют, когда есть небольшая часть размеченных данных и достаточно большое количество неразмеченных. Отмечается, что при таком полуавтоматическом обучении возможно значительно улучшить точность обучения. Зачастую бывает так, что разметить данные - является очень трудозатратной частью, по сравнению с получением этих самых данных. В таком случае данный метод занимает

идеальную позицию. Данный вид обучения занимает промежуточное положение между обучением с учителем и без. При таком обучении модель способна предположить, что неразмеченный блок данных возможно кластеризовать, что впоследствии поможет получить лучший результат.

Такой подход применяется в таких задачах как обработка естественного языка.

# 6 АРХИТЕКТУРЫ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

Архитектура нейронных сетей варьируется от задачи, поставленной перед ней. Одни лучше подходят, для решения задачи классификации, другие, для распознавания образов. Далее будет рассмотрено две различных архитектуры глубоких нейронных сетей.

## 6.1 Рекуррентные сети: LTSM

Рекуррентные нейронные сети - это сети, которые позволяют сохранять информацию и содержать в себе обратные связи. Такие сети применяют при программировании роботов, для работы с временными данными, для распознавания голоса и реча, для распознавания рукописного текста, для генерации аудио-произведений.

Рекуррентную сеть можно как несколько обычных нейронных сетей связанных друг с другом, но каждая из которых передаёт информацию последующей. Получается цепочка из сетевых блоков.

За время существования рекуррентных сетей было предложено множество архитектур, как простых, так и сложных. Частный случай эффективной рекуррентной сети - LTSM сеть - сеть с долгосрочной и краткосрочной памятью. Она может значительно превосходить

оригинальную рекуррентную сеть. Особенность данной сети в том, что она универсальна. Это значит, что при достаточном количестве элементов в данной сети, она способна выполнить любую числовую операцию.

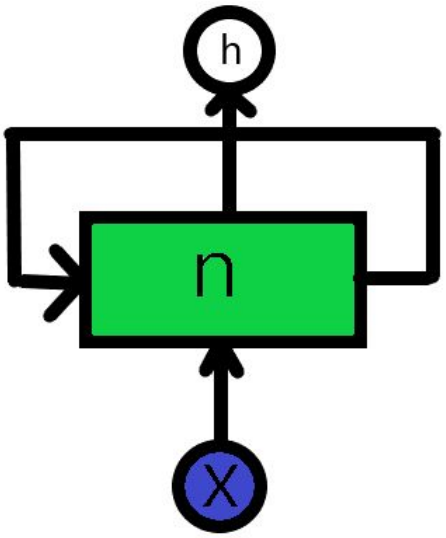


рис. 5. Общее представление рекуррентной сети

Внутри LTSM-сети, находятся не обычные рекуррентные блоки, а так называемые LTSM-модули. Этот модуль способен запоминать на какое-то время. Ключевой особенностью данной сети является отсутствие функции активации внутри своих компонентов, что позволяет эффективно обучать сеть с помощью описанного выше метода обратного распространения ошибки.

Главный компонент блока сети LTSM - горизонтальная линия сверху, проходящая через весь блок. Данные в этой цепи проходят почти без изменений. Данные в этой цепи участвуют в небольшом количестве преобразований.

Кроме этого, в блоке присутствуют структуры, которые могут удалять содержимое данных, проходящих по основной линии. Данный блок также может частично передать информацию, т.к использует сигмоидальный слой и функцию поточечного умножения. В блоке таких фильтров три. Все эти блоки влияют на конечный результат. Рассмотрим каждый фильтр в

отдельности. Первый фильтр - фильтр забывания. Он определяет, какую информацию необходимо забыть. Этот фильтр работает по принципу короткой памяти. Второй фильтр - входной. Он решает, какую информацию необходимо запомнить в ячейке LTSM-блока, а какие данные необходимо обновить.

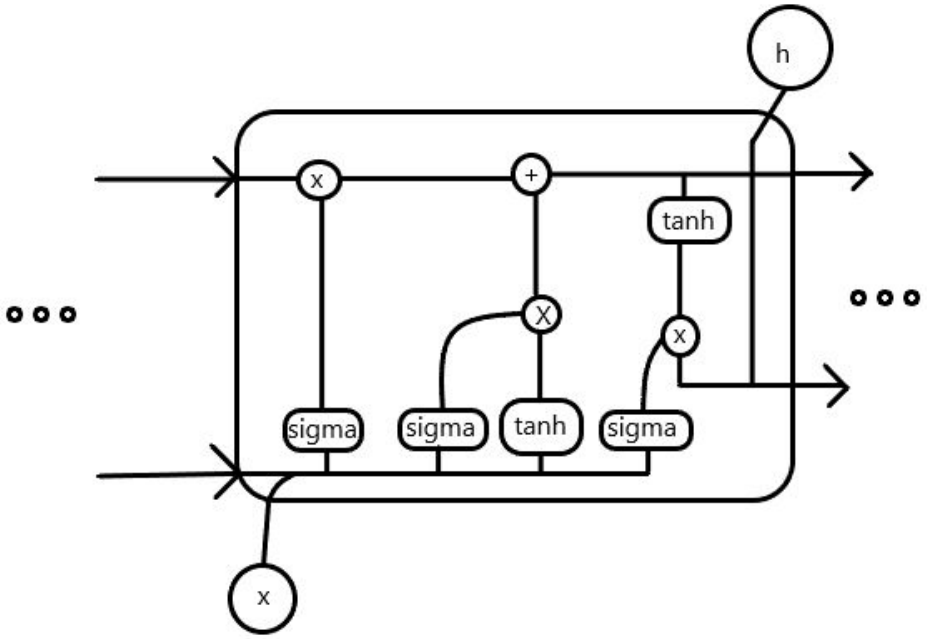


рис.6 Блок LTSM-сети

Таким образом LTSM-сеть способна последовательно обрабатывать данные и делать решения на основании прошлого временного ряда. Существуют и другие архитектуры LTSM. Они отличаются различными вариациями в работе фильтров и методов запоминания и очистки данных с главной линии.

### 6.2 Свёрточные нейронные сети: CNN

Сверточные нейронные сети - это такие структуры, которые предназначены для распознавания образов, а именно решающий задачи классификации, задача нахождения позиции объекта, задача переноса стиля, а также векторизации.

Как понятно из названия данные нейронные сети используют понятие свёртки для работы  с данными. Свёртка подразумевает работу с абстракцией данных, при этом данная структура сама определяет, с абстракциями какого порядка она будет работать. Последовательная послойная свёртка позволяет использовать одни части нейронных сетей решающих одну задачу для решения совершенно другой. Однако, такие слои, несмотря на свою эффективность, остаются слишком сложными для понимания и структуризации.

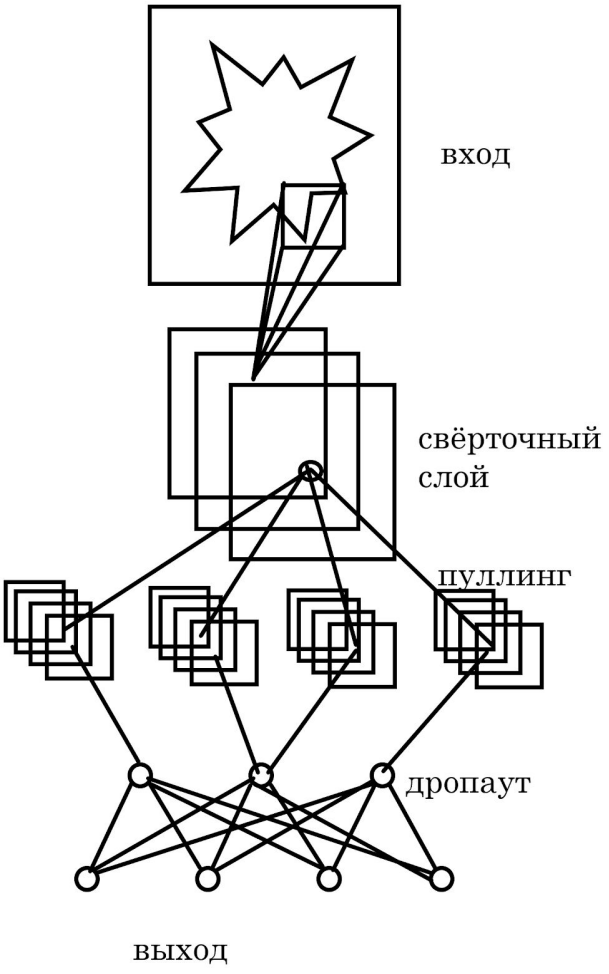


рис.7 Архитектура свёрточной нейронной сети

Основным компонентом сети как раз является свёрточный слой. Данный слой пофрагметно обрабатывает предыдущий, поэлементно суммируя результаты из каждого фрагмента, а также применяя фильтр.

После каждого слоя свёртки следует слой активации, который скалярно складывает элементы свёрточного слоя и передаёт результат далее. Слой активации представляет собой любую нелинейную функцию, например ReLU. Также логически свёрточный слой и слой активации группируют в один.

Далее идёт слой уплотнения(пуллинга). Обычно данный слой имеет меньшую размерность и в неё пофрагментно, используя функцию максимума загружают значения. Данный слой вычленяет признаки из предыдущего слоя. К тому же, такая фильтрация признаков позволяет предотвратить переобучение. Кроме функции максимуму, ещё можно использовать и другие функции. Также применяется функция L2-нормирования.

После последовательности слоёв свёртки и пуллинга идёт полносвязную нейронную сеть. Она может также состоять из нескольких слоёв.

Обучать данную сеть надо с учителем, т.е на размеченных данных и использовать метод обратного распространения ошибки. Для увеличения эффективности свёрточной нейронной сети используют слои дропаута. Они позволяют удалять одиночный нейроны из сети, что позволит предотвратить переобучение.

Главным преимуществом данной архитектуры нейронной сети является лучшая производительность для задачи классификации изображений. По сравнению с полносвязной нейронной сетью в свёрточной сети снижено количество настраиваемых весов, что позволяет использовать обобщение фрагментов вместо попиксельной обработки изображения. Также данная сеть обладает возможностью парралельной работы. Так как для вычисления слоя можно параллельно высчитывать значения для разных фрагментов изображения. Этот метод устойчив к аугментации, т.е к поворотам и

изменениям изображения - свёрточная сеть будет продолжать выдавать правильный результат.

К недостаткам сети можно отнести большое количество параметров для настройки.

# 7 ВЫВОДЫ

В будущем будут разрабатываться новые архитектуры нейронных сетей, новые способы их обучения, что позволит максимально близко приблизить их к работе биологического мозга и создать искусственный интеллект.

Со временем роль информации будет только увеличиваться, что доказывает - все эти методы необходимы для сохранения конкуренции и тренда роста на мировом рынке.

Кроме того, очень важно сохранять целостность и защищённость информации, т.к халатное обращение с информацией способно негативно повлиять как на отдельного индивида, нарушив его право на приватность и конфиденциальность, так и на мировые процессы.

Благодаря развитию в сфере больших данных компании смогут лучше понимать желания клиентов, а потребители смогут получать нужные услуги, которые лучше всего им подходят.

# 8 СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

Введение в машинное обучение [Электронный ресурс]. – Электронные данные. – Режим доступа: https://habr.com/ru/post/427867/ – Дата доступа: 05.04.20.

Алгоритм обучения многослойной нейронной сети методом обратного распространения ошибки (Backpropagation) [Электронный ресурс]. – Электронные данные. – Режим доступа: https://habr.com/ru/post/198268/ – Дата доступа: 04.04.20.

Леонид Черняк. Большие Данные — новая теория и практика (рус.) // Открытые системы. СУБД. — 2011. — № 10. — ISSN 1028-7493.

Виктор Майер-Шенбергер, Кеннет Кукьер. Большие данные. Революция, которая изменит то, как мы живём, работаем и мыслим = Big Data. A Revolution That Will Transform How We Live, Work, and Think / пер. с англ. Инны Гайдюк. — М.: Манн, Иванов, Фербер, 2014. — 240 с. — ISBN 987-5-91657-936-9.

Большие данные и глубокое машинное обучение в искусственных нейронных сетях - Тихонов А.А. - Институт математики, физики и информатики. Красноярский государственный педагогический университет им. В.П. Астафьева. Красноярск -2019