Санкт-Петербургский
Государственный
Политехнический
Университет

Институт прикладной
математики и механики

КАФЕДРА
ТЕЛЕМАТИКА

Методы исследовательской работы

вычислительная платформа поддержки технологий искусственного интеллекта

(занятие 12)

---

28 апреля
2022 г.

Что обсуждали на прошлой лекции

С точки зрения физики процессы вычислений происходят в 4-х мерном «пространстве-времени». При этом имеют место следующие свойства :
1. причинность – время отражает порядок следования «причина-следствие».
2. альтернативность – у события возможно несколько вариантов исхода.
3. цикличность – события в физическом пространстве-времени повторяются.

С точки зрения компьютерных наук пространство состояний событий расширяется (до 6-ти мерного многообразия), поэтому появляются новые возможности

4. управлять скоростью течения времени (скоростью транзакций)
5. использовать «прошлое» для «предсказания» будущего
6. изменять «настоящее», используя прогноз будущего.

Основной вопрос состоит в том, можно ли феномен интеллекта и свойства сознания объяснить, используя процессы вычисления

Проблемы реализации технологий ИИ

Компьютерные науки – область знаний, которая пересекается со многими другими областями науки, включая математику, статистику, теорию вероятностей, физику, обработку сигналов, а в последнее время машинное обучение, компьютерное зрение, психологию, лингвистику и науку о мозге.

Задачи, зависящие от множества переменных факторов, требуют очень сложных решений, которые трудны для понимания, объяснения и очень сложно алгоритмизируются.

Программирование алгоритмов, используемых для решения задач, связанных с большими объёмами данных, занимает у разработчиков очень много времени. Даже когда удаётся написать код, обрабатывающий большое количество разнообразных данных, этот код зачастую получается очень громоздки и тяжело проверяемым.

Однако, т.н. «натуральные вычисления» с такими проблемами не сталкиваются. Научение мозга как аналог процесса программирования происходит непрерывно благодаря феномену интеллекта и свойству сознания объяснить принимаемые решения. Можно ли всего этого добиться при создании систем ИИ ?

Как конструктивно определить понятие интеллект ?

Способность мышления, рационального познания, в отличие от таких, например, душевных способностей, как чувство, воля, интуиция, воображение и т. п.

БСЭ

Способность к познанию и решению проблем, которая объединяет познавательные способности: ощущение, восприятие, память, представление, мышление, воображение

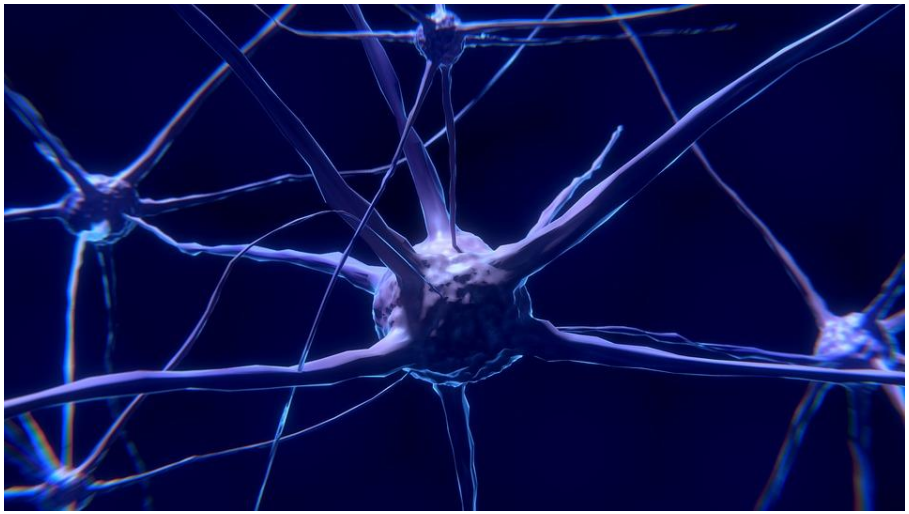
Википедия

Интеллект —умственное начало, основа целеполагания, планирования ресурсов и построение стратегии достижения цели.

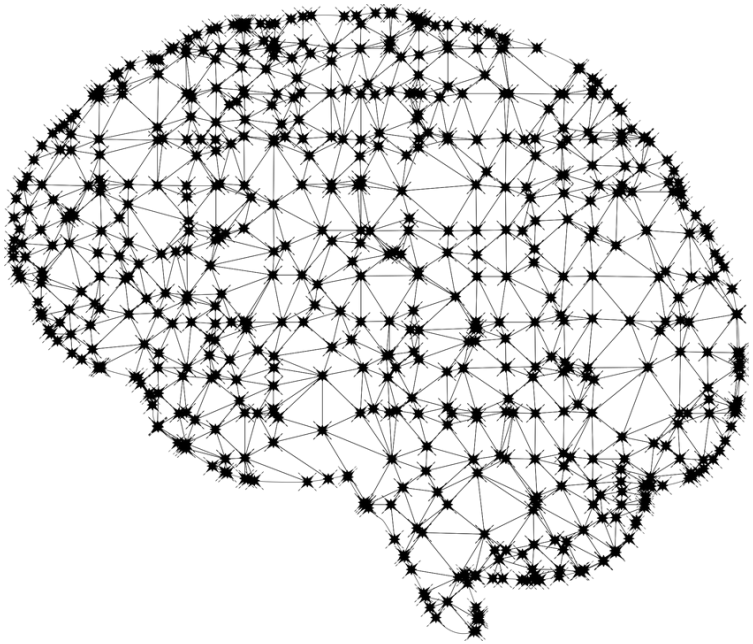
Толковый словарь Ожегова

Инструментарий интеллекта - человеческий мозг, который представляет собой углеродный компьютер, выполняющий миллиард миллиардов операций в секунду (1000 петафлопс =1 экзафлопс), потребляющий при этом 20 Ватт энергии. https://nplus1.ru/news/2019/08/28/nanotube-processor?ysclid=l2iv7hbnts

Итак, углеродный компьютер это



Элементная база



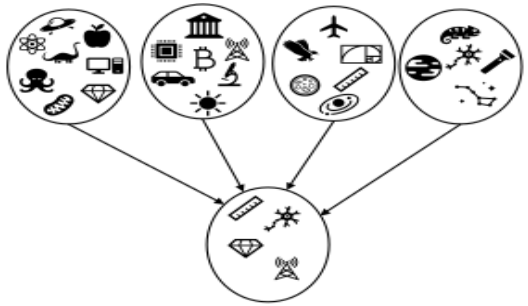
структура

Совокупность нейронов, которые взаимодействуют друг с другом с помощью специальных каналов, позволяющих им обмениваться информацией. Сигналы отдельных нейронов «**взвешиваются**» и комбинируются друг с другом перед тем, как активировать другие нейроны.

Искусственные Нейронные Сети (ИНС)

Искусственные Нейронные Сети — это математические модели, созданные по аналогии с биологическими нейронными сетями. ИНС способны моделировать и обрабатывать нелинейные отношения между входными и выходными сигналами. Адаптивное «**взвешивание**» сигналов между искусственными нейронами достигается благодаря алгоритму, считывающему наблюдаемые данные, изменяя при этом веса нейронов так, чтобы повысить точность «восприятия» данных….

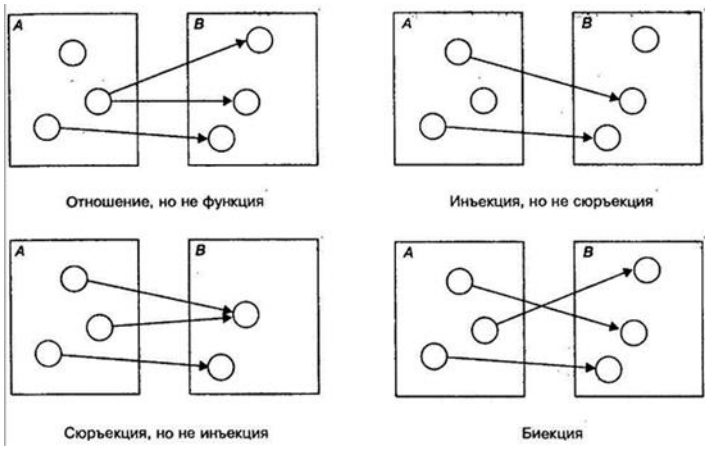
.



«Аксиома выбора»

Если «модель» воспринимаемого мира оказывается слишком «большой» её невозможно оптимизировать до уровня используемых понятий - «микрокоманд» процессов сознания .

Концепция «глубокого обучения»

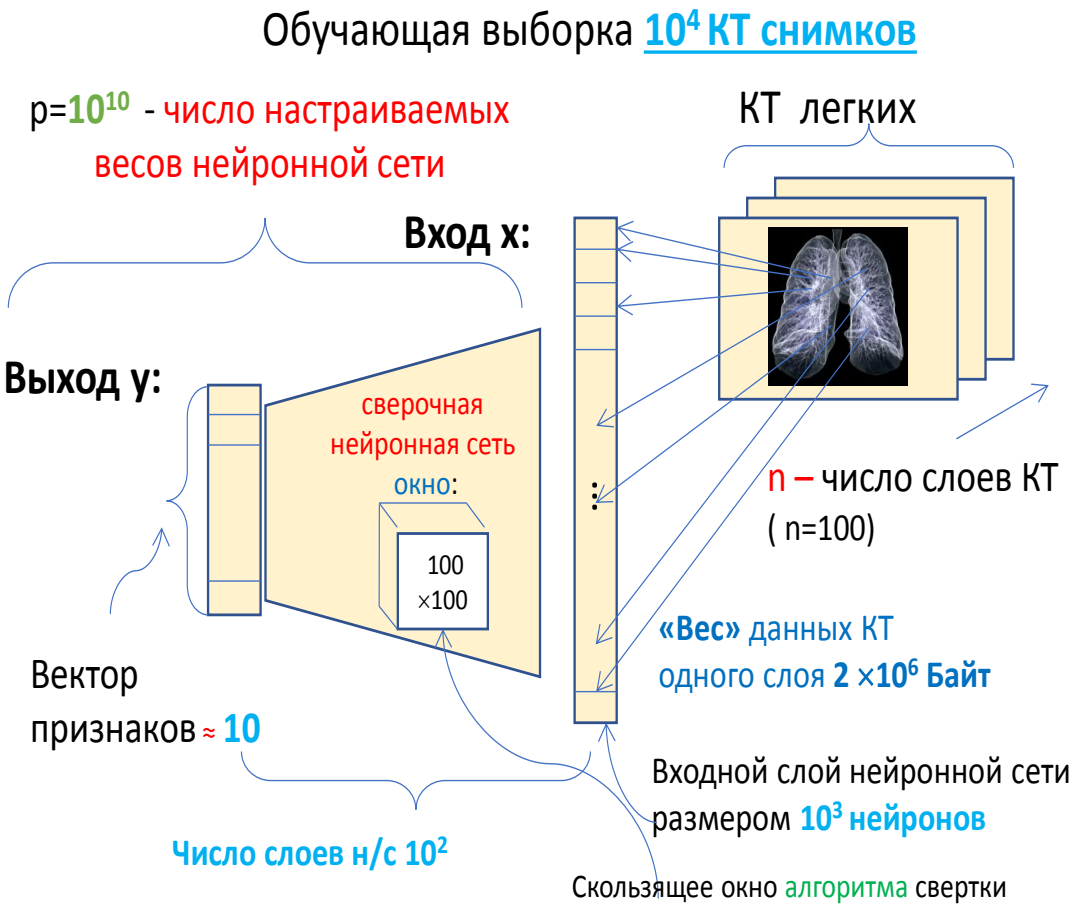


Виды отображений между различными множествам «данные» - «понятия»

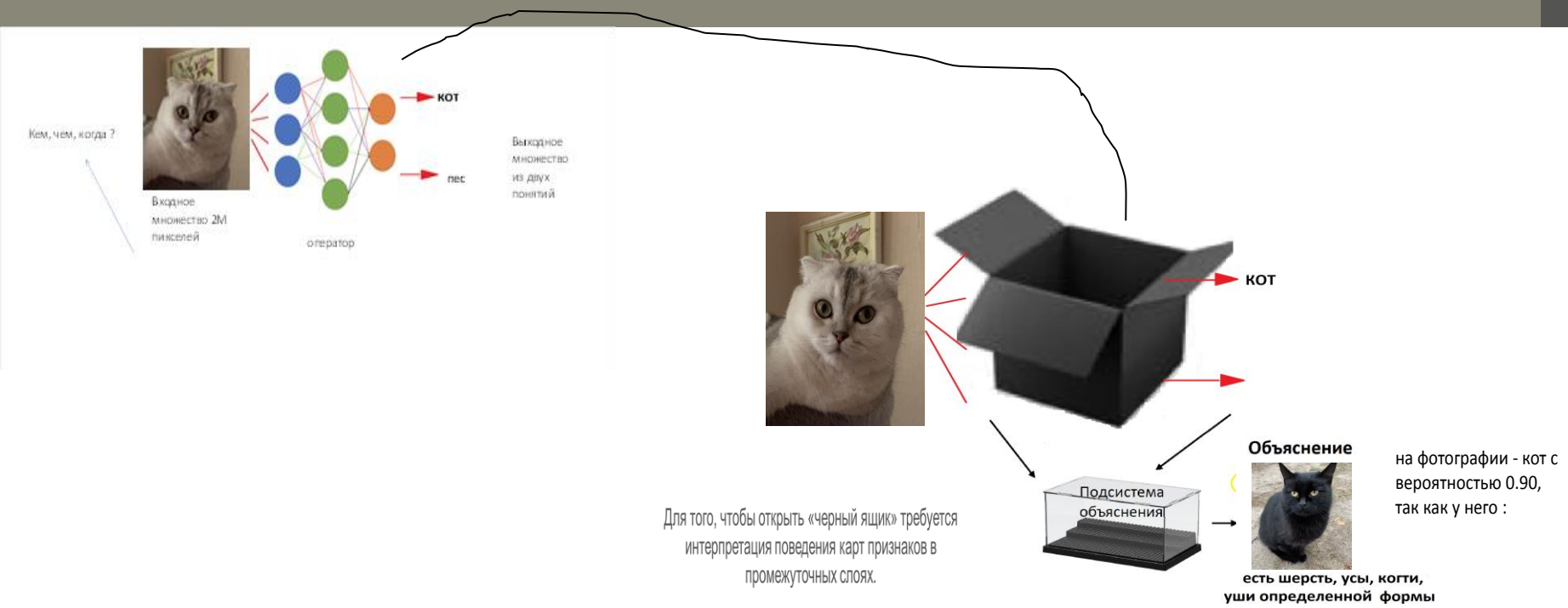
Термин глубокое обучение используется для описания нейронных сетей и используемых в них алгоритмах, принимающих «сырые» данные (из которых требуется извлечь некоторую полезную информацию). Эти данные обрабатываются, проходя через слои нейросети, для получения нужных выходных данных.

Соотношения понятий «символ – смысл»

Семантический треугольник Г. Фреге .



Вычислительная проблема интерпретации данных



Суть обучения ГНС: 2 М пикселей входного изображения кодируют с помощью hidden сигнатуры в выходном embedding vector позицию «кот" с вероятностью 0.98", а слово «пес» с вероятность 2%.

Пример ИИ vs машина «природы»

- Существующие методы машинного обучения «слабы», когда от них требуется обобщения за пределами обучающего распределения.
- В природе этой проблемы нет. Генетическая информация очень компактна, а биохимические машины организма - это очень эффективные системы, способные бороться со 2-м законом термодинамики.
- Геном человека в виде последовательного кода содержит всего около восьмисот миллионов байт информации (800 МБ), что эквивалентно размеру средней современной программы. Сам код ДНК выполняется кластером параллельных биохимических машин, которые переводят одномерный "код" ДНК в последовательности аминокислот, которые в свою очередь складываются в трехмерные белки, составляющие все живые существа от бактерий до человека.

Пример из области компьютерных наук

- Суть идеи Тьюринга: каждому алгоритму вычислений надо "подобрать" оптимальную машину.
- При этом возможности оптимизации платформы связать со статической гетерогенностью и динамической реконфигурируемостью архитектуры этой машины.
- Тело человека это не одна МТ, а кластер из различных МТ. В таком кластере есть как базовые вычислительные блоки (мультиядерные и мультредовые), так и специализированные ускорители.
- Саму платформу можно рассматривать на нескольких уровнях описания ее программно-аппаратных компонент, начиная от отдельных электронных блоков, до системно-программного обеспечения, реализуя при этом возможность изменение "под задачу" используемого "внутреннего" микрокода платформы, из которого на уровне архитектуры уже "собираются машинные команды" программы.
- Меняя «на лету» микрокод реконфигурируемой вычислительной платформы, процессы вычислений адаптируются в условиям и контексту «существования» алгоритма.

Что нам надо: Deep understanding vs deep learning

- Understanding: involve an ability not only to correlate and discern subtle patterns in complex data sets, but also has the capacity but also has the capacity to address questions such as
  - what, why, when, and how
- The only thing that matters in answer to these questions in the long run perspective is the leveraging of computation.

cognitive vs program based approach

Computer science is built on the concept of the cycle. The question is: can this concept be brought into the model of consciousness.

So, can we view cognition in terms of a kind of cycle:

- organisms (eg humans) take information from the outside, they build

and clarify internal cognitive models based on their perception of that information, etc.

- what computational features would we need in order to have systems that are capable of reasoning in a robust fashion about the world? And what it would take to bridge the worlds of deep learning (primarily focused on learning) and classical computer technology (which was more concerned with program as an internal cognitive models) ?

Интуиция и логический вывод: «Проблема Линды»

- Проблема Линды – ошибка коньюнкции или формальная ошибка, возникающая, когда предполагается, что коньюнкция конкретных условий более вероятна, чем одно общее условие:

  Thought Experiment:

  Which of the following events is most likely to occur, or are they equally likely?:

  - Vice President Kamala Harris will become the next president.
  - President Joe Biden will be removed from office **and** Vice President Kamala Harris will become the president.

- "conjunction» is a sentence of the form: "…..**and**….."
- For example, sentence: "Today is Saturday **and** the sun is shining" is a conjunction
- The probability of a conjunction is never greater than the probability of its conjuncts…. the probability of two things being true can never be greater than the probability of one of them being true

Информационный парадокс интуиции

- Переход из области осмысленного, где отношения формируются путем ответа на вопрос «почему ?», в область абстрактного, где не имеет смысла спрашивать “зачем?” – не имеет простой объяснительной (суррогатной) модели.
- Поиск смысла возможен только там, где может иметь место феномен предметной рефлексии. Для этого воспринимающий информацию субъект должен себя представить на месте другого субъекта.
- Например, выбирая вместо реального объекта, вероятность его обнаружить в эксперименте, человека интуитивно может приписать большую вероятность менее вероятному событию, **например** P(A&B) > P(A) ?, имеем P(A&B)=P(B|A)*P(A), а P(A) <1, следовательно P(A&B)<P(A)….

Однако, - log P(A&B) > -log P(A),

т.е.

информации в событии A&B больше, чем в событии A

I(A&B)>I(A)

Проблема, которую надо решать

Можно ли и как в объективные физические процессы «встроить» процессы вычислений ?

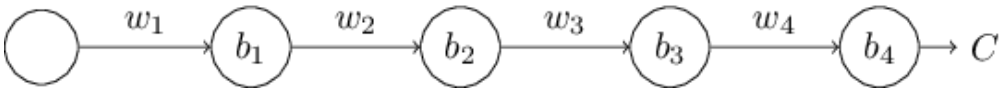
- Так как объективные законы сохранения – это следование тех или иных видов симметрии, то для решения сформулированной проблемы надо найти подходящую симметрию, например, симметрию обратимости состояний,
- одна из форм которой требует равноправности двух направлений «стрелы времени» и имеет вид P(A, t)=P(B, t-1),
- Чтобы этого добиться надо иметь ресурсы – например, «отрицательную корреляционную энергию» - суть которой рост энергетического потенциала (уменьшение энтропии) по мере усложнения системы.

Скорость обучения нейронных сетей

- Существуют фундаментальные причины для замедления обучения, которые связаны с использованием техник обучения на основе градиента.
- В многослойных глубоких нейронных сетях первые скрытые слои обучаются гораздо медленнее последних. Это явление известно под названием «проблемы исчезающего градиента», у которого есть «антипод» – взрывной рост градиента. Нестабильность является фундаментальной проблемой для градиентного обучения ГНС.
- Заметим, что случайная инициализация весов ГНС означает, что первый слой сети «выбрасывает» большую часть информации о входящем изображении.

пример

- Рассмотрим сеть с тремя скрытыми слоями:



Где w1, w2,… – это веса, b1, b2,… – смещения, C – некая функция стоимости, взвешенный вход нейрона равен:

$z_j = w_j a_{j-1} + b_j$

а выход нейрона это $f(z_i)$ - сигмойдная функция т.н. активации нейрона

- Градиент функции стоимости dC/db_i - связан с первым скрытным слоем ГНС (используется метод обратного распространения ошибки) $\frac{\partial C}{\partial b_1} = \sigma'(z_1) \times w_2 \times \sigma'(z_2) \times w_3 \times \sigma'(z_3) \times w_4 \times \sigma'(z_4) \times \frac{\partial C}{\partial a_4}$

Вариация параметров ГНС

- Если мы внесли изменение Δb_1 в смещение нейрона b_1. то это произведет серию каскадных изменений по всей сети. Прежде всего это повлияет на выход первого скрытого нейрона Δa_1 – что заставить измениться и Δz_2 во взвешенном входе на второй скрытый нейрон.
- Затем произойдёт изменение Δa_2 в выходе второго скрытого нейрона. И так далее, вплоть до изменения ΔC в стоимости выхода.

- Получается, что $$\frac{\partial C}{\partial b_1} \approx \frac{\Delta C}{\Delta b_1}$$

Уточнения

- Рассмотрим, как Δb_1 «заставляет» меняться выход a_1 первого скрытого нейрона.
- Имеем $a_1 = \sigma(z_1) = \sigma(w_1 a_0 + b_1)$, поэтому

$$\Delta a_1 \approx \frac{\partial \sigma(w_1 a_0 + b_1)}{\partial b_1} \Delta b_1$$

Компоненты выражения

- Член $\sigma'(z_1)$ - это первый член выражения для градиента $\partial C/\partial b_1$

- Интуитивно понятно, что этот член инициирует изменение смещения Δb_1 в изменение Δa_1 выходной активации нейрона. Изменение Δa_1 в свою очередь вызывает изменение взвешенного входа $z_2 = w_2 a_1 + b_2$ второго скрытого нейрона:

$$\Delta z_2 \approx \frac{\partial z_2}{\partial a_1} \Delta a_1$$

Распространение влияния смещения вдоль сети

- изменение смещения b_1 распространяется вдоль сети и влияет на z_2:

$$\Delta z_2 \approx \sigma'(z_1) w_2 \Delta b_1$$

Изменение функции выхода

- В итоге получается выражение, связывающее конечное изменение ΔC функции стоимости с начальным изменением Δb_1 смещения:

$$\Delta C \approx \sigma'(z_1) w_2 \sigma'(z_2) \ldots \sigma'(z_4) \frac{\partial C}{\partial a_4} \Delta b_1$$

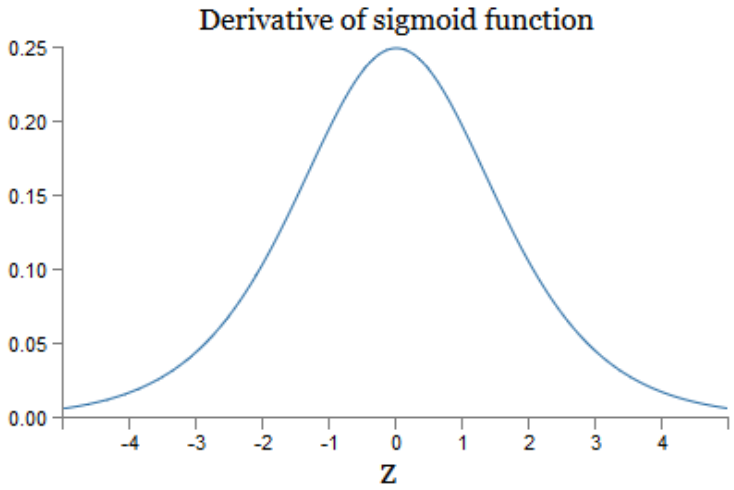
Исчезновение градиента

- Чтобы понять, почему возникает проблема исчезающего градиента подробно распишем выражение для градиента:

$$\frac{\partial C}{\partial b_1} = \sigma'(z_1)\ w_2\sigma'(z_2)\ w_3\sigma'(z_3)\, w_4\sigma'(z_4)\ \frac{\partial C}{\partial a_4}$$

Сигмойдная функция и ее производная

График стандартной сигмойдной функции активации достигает максимума в точке $\sigma'(0)=1/4$. При случайной инициализации весов сети, для выбора веса сети используется распределение Гаусса, у которого средне отклонение рано нулю нулём, а стандартное отклонением 1.



При такой инициализации веса ГНС будут удовлетворять неравенству $|w_j|<1$, а члены $w_j\sigma'(z_j)$ с высокой вероятностью будут удовлетворять неравенству $|w_j\sigma'(z_j)|<1/4$. Поэтому, если воять произведение множества таких членов, то оно будет экспоненциально уменьшаться, очевидно:
чем больше членов <1, тем меньше само произведение.

Уточнения

Сравним выражение для $\partial C/\partial b_1$ с выражением градиента относительно следующего смещения, допустим, $\partial C/\partial b_3$. Мы не записывали подробное выражение для $\partial C/\partial b_3$, но оно следует тем же закономерностям, что описаны выше для $\partial C/\partial b_1$. И вот сравнение двух выражений

$$\frac{\partial C}{\partial b_1} = \sigma'(z_1) \overbrace{w_2\sigma'(z_2)}^{<\frac{1}{4}} \overbrace{w_3\sigma'(z_3)}^{<\frac{1}{4}} \underbrace{w_4\sigma'(z_4)\frac{\partial C}{\partial a_4}}$$

common terms

$$\frac{\partial C}{\partial b_3} = \sigma'(z_3) \overbrace{w_4\sigma'(z_4)\frac{\partial C}{\partial a_4}}$$

в градиент $\partial C/\partial b_1$ входит два дополнительных члена, каждый из которых имеет вид $w_j\sigma'(z_j)$. Такие члены обычно <1 и не превышают величину 1/4. Поэтому градиент $\partial C/\partial b_1$ будет в 16 (или больше) раз меньше, чем $\partial C/\partial b_3$. И это основная причина возникновения проблемы исчезающего градиента. Однако, если время обучения расти веса w_j. то члены $w_j\sigma'(z_j)$ в произведении уже не будут удовлетворять неравенству $|w_j\sigma'(z_j)|<1/4$. Вместо этого градиент будет экспоненциально расти при обратном движении через слои. И вместо проблемы исчезающего градиента мы получим проблему взрывного роста градиента.

Формулировка сути рассматриваемой проблемы

- Суть рассматриваемой проблемы заключается не в исчезающих значениях градиента или его взрывном росте.
- Проблема в том, что при такой многослойной архитектуре НС и методах оптимизации параметров **градиент целевой функции в первых слоях является произведением членов из всех слоёв сети**. Поэтому, когда слоёв ГНС много, ситуация с оптимизацией параметров по сути становится неустойчивой и скорости настройки слоев существенно разные.

- Очевидно, что все слои ГНС смогут обучаться с примерно одной скоростью –если выбрать так члены произведения в выражении для градиента целевой функции, чтобы они балансировали друг друга.
- Итак, реальная проблема в том, что обучение ГНС «страдает» от проблемы «нестабильного градиента». Поэтому, если использовать стандартные обучающие техники на основе методов стохастического градиента, то разные слои сети обучаются с существенно разными скоростями.

Комментарии: Пропедевтика машинного обучения

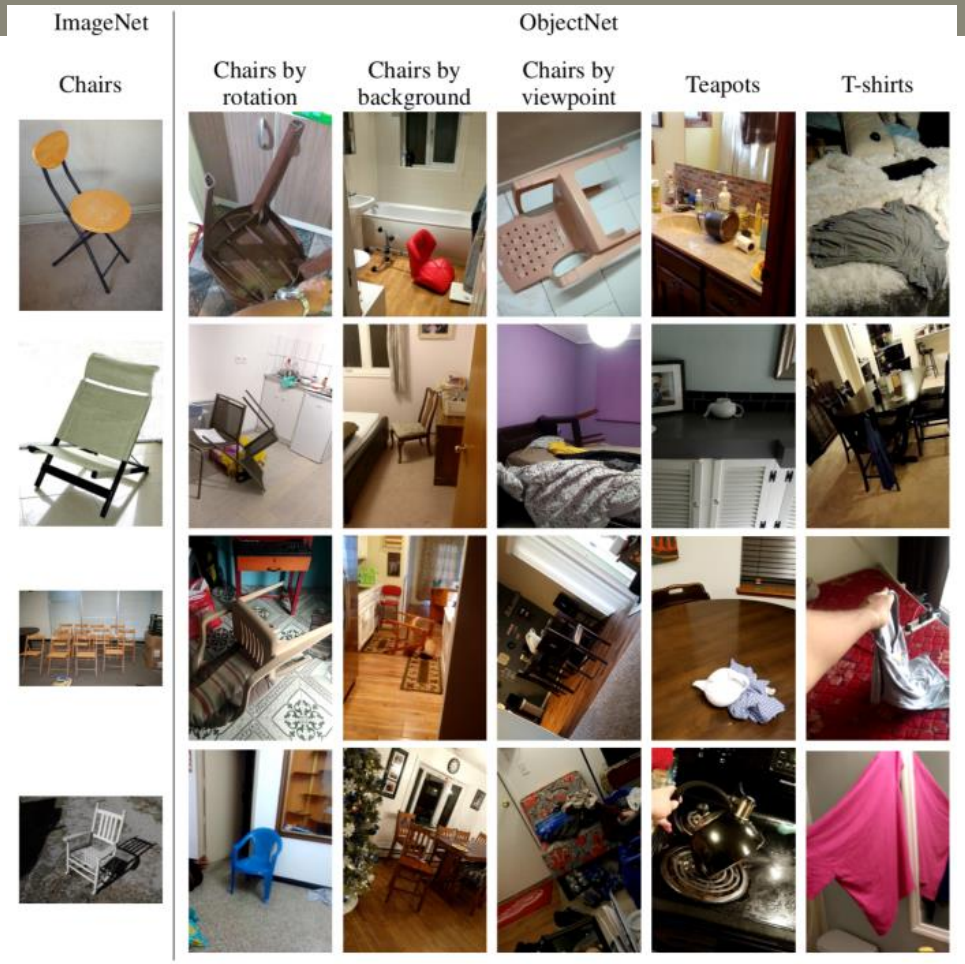
Надо учесть, что истинность любых формальных методов основана на непротиворечивости используемой аксиоматики. Аксиоматика машинного обучения основана на решении обратных алгоритмических задач или задач ретросинтеза.

- Для достижения практически значимого результата с помощью алгоритмов машинного обучения разработчики «интеллектуальных систем» должны использовать такие алгоритмы описания предметной области, чтобы обеспечивать условия сходимости методов оптимизации, в частности, градиентных методов.

- Существующие статистические модели обучения на основе градиентной оптимизации следует «пополнить» с учетом :
  - причинно-следственных связей между переменными
  - топологических инвариантов рассматриваемых структур

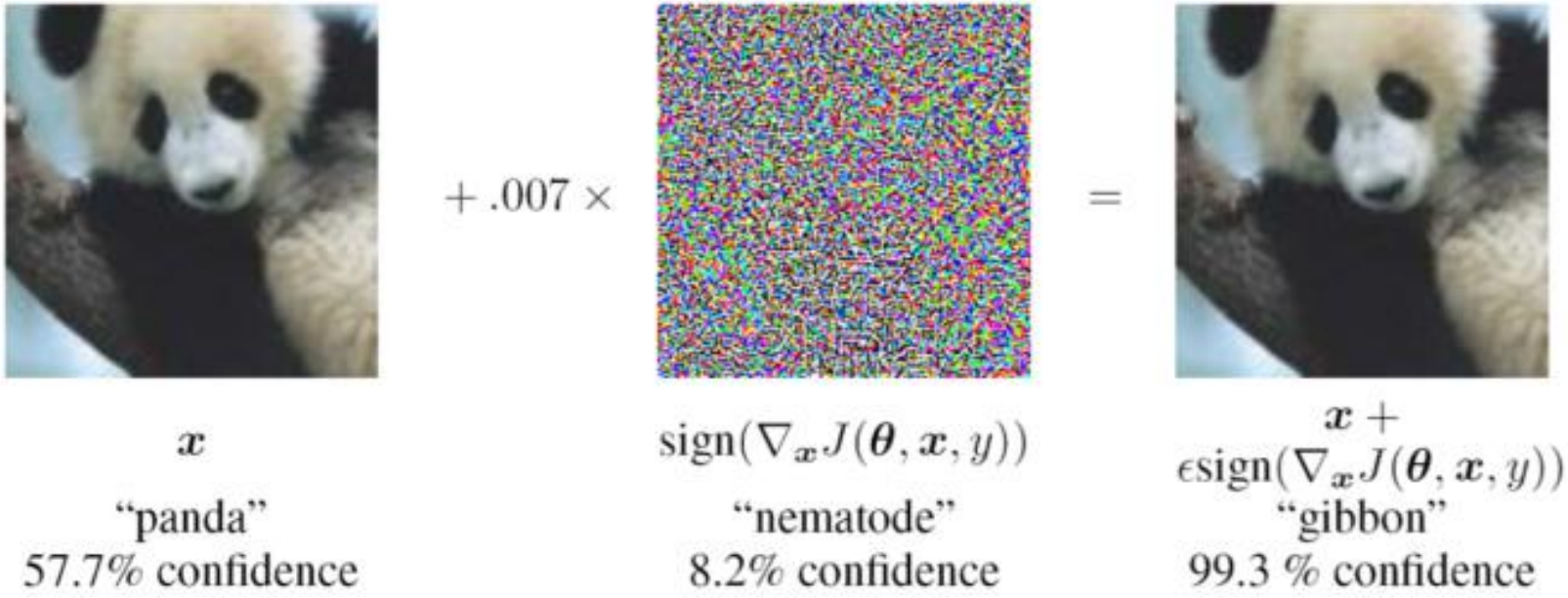
  а также
  - субъективного тезауруса воспринимаемых данных

«проклятие» отождествления



Объекты в обучающих наборах данных по сравнению с объектами в реальном мире отличаются . По мере того, как среда становится все более сложной, практически невозможно охватить все распределение посредством добавления большего количества обучающих примеров.

Пример семантической неустойчивости задач классификации



Так называемые состязательные атаки нацелены на чувствительность машинного обучения к условиям сходимости градиентных методов.

На этом фотографическом изображении панды добавление незаметного для наблюдателя слоя шума «заставляет» сверточную нейронную сеть принимать панду за гиббона….

машинному обучению надо учесть причинно-следственные связи

Просматривая видео последовательность, естественным образом можно сделать выводы о причинно-следственных связях между различными элементами:

- бита и рука бейсболиста движутся в синхронно,
- рука игрока вызывает движение биты, а не наоборот,
- бита вызывает резкое изменение траектории мяча.

Эти выводы о причинно-следственных связях в наблюдаемых явлениях люди делают интуитивно.

Для алгоритмов машинного обучения учет причинно-следственных связей (каузальность от causality) является сложной задачей. Современные статистические алгоритмы обучения хорошо работают в задачах выявлении сложных закономерностей в больших массивах данных - преобразовывать в реальном времени звук в текст, размечать тысячи изображений и видеокадров, проверять структуру снимков МРТ на предмет наличия в них патологических образований. Однако, этим алгоритмам сложно выводить объективные причинно-следственные связи.