



Санкт-Петербургский
Государственный
Политехнический
Университет

Институт прикладной
математики и механики

КАФЕДРА
ТЕЛЕМАТИКА

**Семинар по специальности на английском
языке
(Workshop in English)**

**Hybrid artificial intelligence:
Distilling the Knowledge in a Neural Network**

(занятие 14)

11 мая
2022 г.

Что обсуждали на прошлой лекции: "propaedeutics" of machine learning

"propaedeutics" - (пропедевтика в медицине - заключение о сущности заболевания) это термин, обозначающий вводный курс, систематически изложенный в элементарной (легко объяснимой) форме. В нашем случае это вывод о том, что

- истинность результатов любых формальных методов основана на непротиворечивости используемой **аксиоматики**. Современная **аксиоматика машинного обучения** основана на решении обратных статистических задач (задач ретросинтеза), а эти задачи не имеют единственного решения
 - Поэтому для достижения практически значимого результата с использованием методов машинного обучения **применяются методы оптимизации**, в частности, градиентные методы поиска лучших с точки зрения **выбранного критерия** параметров (весов) нейронных сетей – инструментальных средств поиска решения.
 - Однако, у таких методов обучения обнаружено ряд «принципиальных» недостатков (в частности, катастрофическое «затухание» или «взрывном росте» градиента в процессе настройки параметров при «обратном распространении ошибки»), поэтому одна из актуальных задач - сделать методы обучения **robust**, то есть **повысить их «устойчивость»** к влиянию несущественных факторов. Это можно сделать, если «пополнить» методы машинного обучения новыми средствами, которые учитывают:
 - причинно-следственные связи между входными параметрами (переменными)
 - топологические инварианты обрабатываемых структур данных
- а также**
- субъективный тезаурус системы обучения (например, размерность выходного слоя нейронов используемой ИНС), характеризующий «разрешающую способность» системы. К сожалению, в процессе работы эту точность пока увеличить не удастся.

The essence of the problem is that the **gradient of the target function in the first layers** of the neural network is the **product of the gradients of all subsequent layers of the network.** (произведение градиентов всех последующих слоев сети)

- The case for hybrid artificial intelligence. By [Ben Dickson](#) - March 4, 2020
- Exploring the missing piece of artificial intelligence By [Ben Dickson](#) - December 9, 2019

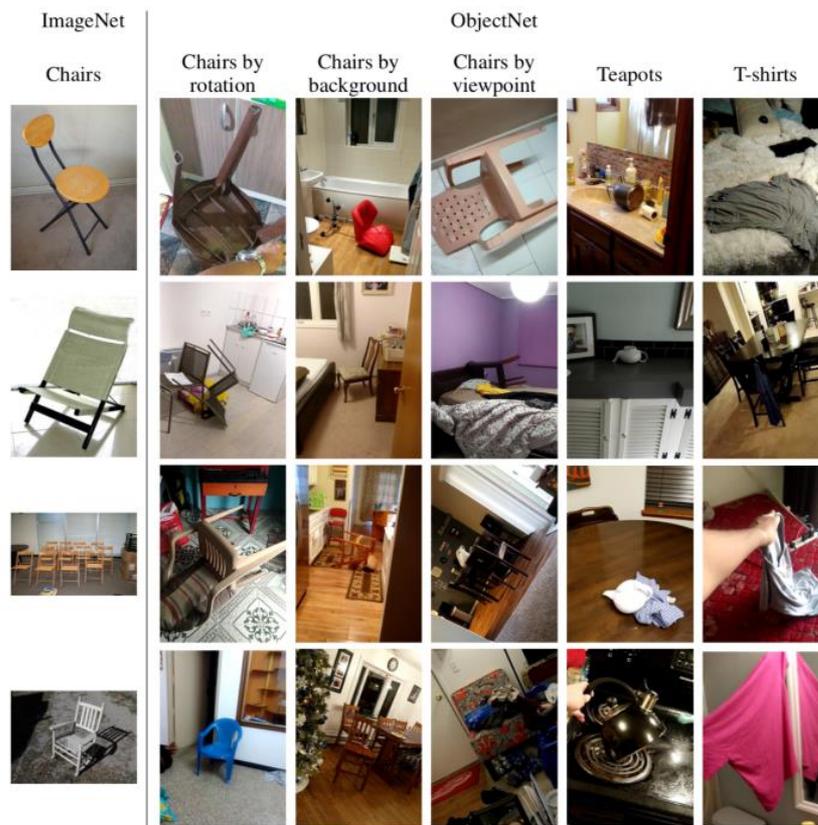
Почему машинному обучению с трудом дается причинно-следственная связь?

[MaxRokatansky](#) 18 мая 2021 в 18:39

Напоминание: Robust mean - ignores irrelevant factors

- Current machine learning methods seem weak when they are required to **generalize beyond** the **training distribution**, which is what is often needed in practice.
- What can we do to take AI to the next level of robust end effectiveness ?
- Understanding – or involve an ability not only to correlate (сопоставлять) and discern (различать) patterns in complex data sets, but also has the capacity to address questions such as
 - what, why, when, and how

Фундаментальная проблема машинного обучения - «проклятие» отождествления



Objects in training datasets compared to objects in the real world can be very different. По мере того, как среда в которой находится объект становится все более сложной, практически невозможно с помощью обучающей выборки охватить все возможные ситуации только посредством добавления большего количества примеров.

How to solve the information paradox of the identification abstraction (in particular Linda's paradox)

- Парадокс Линды - **ошибка**, возникающая, когда предполагается, что конъюнкция конкретных условий **более вероятна**, чем одно общее условие.
- Используя абстракцию вероятности, которая приписывается реальным объектам, человека интуитивно может присвоить большую вероятность менее вероятному событию, **например** $P(A\&B) > P(A)$?, имеем $P(A\&B)=P(B|A)*P(A)$, но всегда $P(A) < 1$, 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 **$I(A\&B) > I(A)$**

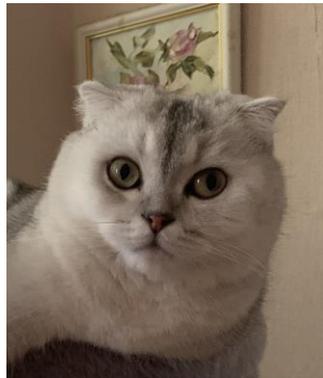
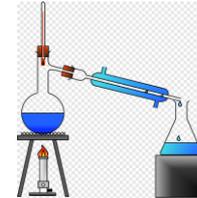
- It is the abstraction of identification that "controls" the transition from the realm of the **meaningful**, where relations between objective entities are formed by answering the question "why?", to the realm of the **abstract**, where the question "why?" **itself has no meaning** (the concept of "abstraction of identification" was introduced by A. A. Markov in 1954).

From the meaningful to the abstract - the problem of "knowledge distillation"

- Одним из наиболее актуальных вопросов в области машинного обучения является интерпретируемость результатов, полученных с помощью моделей на основе нейронных сетей.
- For models based on deep neural networks (DNN), it is quite difficult to get a clear rationale for decisions made with them.
- Для этого требуется алгебраизация зависимостей между признаками входных данных и рассматриваемыми целевыми классами, что может существенно прояснить интерпретацию полученных результатов. .

The Model of Explanation as a Result of the "Distillation of Knowledge"

Дистилляция - технологический процесс разделения и рафинирования многокомпонентных веществ



КОТ

Глубокая ИНС с $m \gg 100$ параметрами

Распознаваемый объект как совокупность 2 миллионами параметров - 2Mr пикселей, каждый из которых не характеризует объект



Объяснение



есть шерсть, усы, когти, уши определенной формы



«суррогатная модель», содержащая n – понятий, характеризующих «существенные» свойства «объекта»

How to compress the knowledge

The easiest way to improve the explainability of any machine learning algorithm's model is to follow the identity abstraction, that is, to **train many different models** on the same data, and then average their predictions.

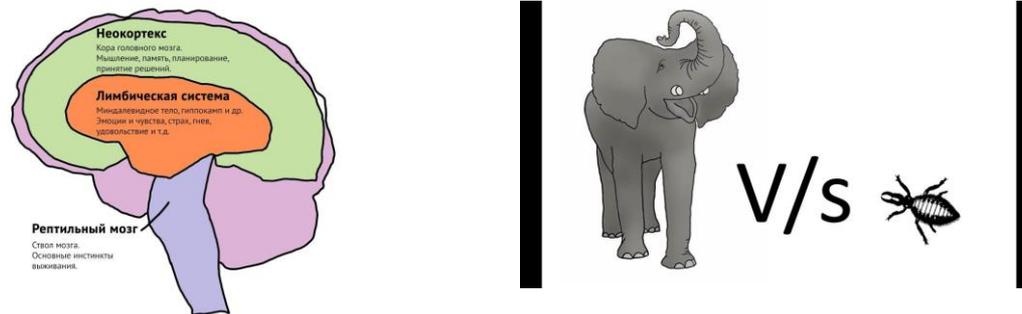
However, simple approach does not scale well

So, making predictions using a ensemble of models may be too computationally expensive, especially if the individual models are deep neural nets.

Main question: **is it possible to compress the knowledge that embedded in an ensemble of models, into a single model ?**

Гипотеза: Cognitive abilities depend on the mass of the brain

We know that many insects have a **specific and very efficient mechanism** that is optimized for extracting energy and information from the environment. This information insects use in real time mode for traveling, reproduction, or other life activities.



Elephant and insect have different complexity neural network, but both survive in the same conditions.... причина, как считается, заключается в количестве нейронов, а не в размерах мозга

In analogy, **large-scale machine learning system** can **extract** “stealth” structure from very large, highly **redundant data stream**. **Extract** in ML means **filtering** the input data stream with the selection of essential for this context features of the environment.

- What is mean essential for this context features
- How can estimate the amount of computation which is use to provide filtering process ?

The Elephant Brain Paradox - What are more important – structure of brain or data

Оказалось, что мозг африканского слона содержит больше нейронов, чем мозг человека. И это не просто незначительное различие: ровно в три раза больше нейронов, 257 миллиардов по сравнению с нашими 86 миллиардами.

При этом 98 процентов из этого количества нейронов мозга слона расположены в мозжечке, т.е. в задней части мозга.

Real computational resources of brain depends from neuron size but also from structure of their connection that provide **information exchange** between different parts of brain.

Итак, мозг человека не содержит больше нейронов, чем намного более крупный мозг слона, однако на кору головного мозга человека приходится почти **в три раза больше нейронов, чем на соответствующий отдел мозга слона**, хотя последний более чем в два раза крупнее.

That is why it is possible to train ..
simple models if it can easier to extract internal structure from the available data.....

Resume: in common sense - Knowledge “Distillation” is a way to transfer knowledge from **teacher to student**

In our case, **knowledge distillation** is to train a compact neural network using distilled knowledge **extrapolated** from a “deep learning” model or even ensemble of DL models. Using distilled knowledge, somebody can improve efficiency of train **compact** model without seriously compromising its performance and accuracy .

