Общее собрание Отделения математических наук РАН

Фундаментальные проблемы технологий искусственного интеллекта

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН,
рук. лаб. Машинного обучения и семантического анализа Института ИИ МГУ,
зав. кафедрой Математических методов прогнозирования ВМК МГУ,
зав. кафедрой Машинного обучения и цифровой гуманитаристики МФТИ,
г.н.с. ФИЦ ИУ РАН

voron@mlsa-iai.ru

Содержание

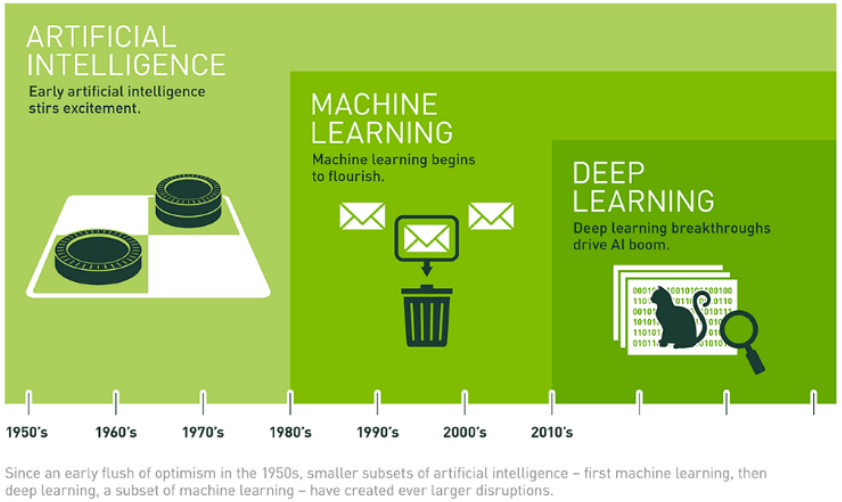
1. Фундаментальные основы искусственного интеллекта
   - От экспертных систем к машинному обучению
   - Задачи оптимизации в машинном обучении
   - Обучаемая векторизация и глубокие нейронные сети
2. Проблески общего искусственного интеллекта
   - Модели внимания, трансформеры и генеративные модели
   - Большие языковые модели (LLM)
   - Свойство эмерджентности
3. Фундаментальные проблемы технологий ИИ
   - Математические проблемы
   - Технологические проблемы
   - Социо-гуманитарные и организационные проблемы

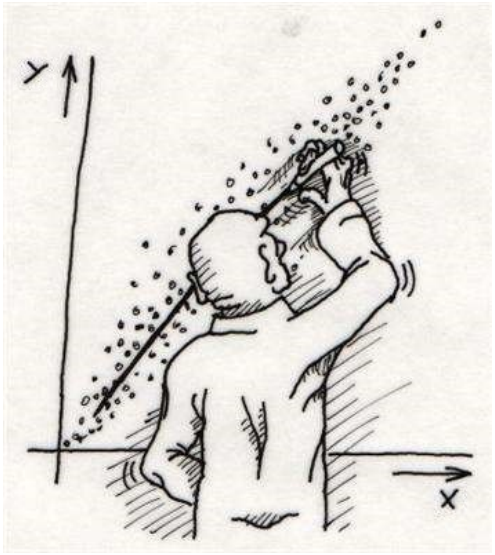
Бум искусственного интеллекта

1997: IBM Deep Blue обыграл чемпиона мира по шахматам
2005: Беспилотный автомобиль: DARPA Grand Challenge
2006: Google Translate – статистический машинный перевод
2011: 40 лет DARPA CALO привели к созданию Apple Siri
2011: IBM Watson победил в ТВ-игре «Jeopardy!»
2011–2018: ImageNet: 25% → 2,5% ошибок против 5% у людей
2015: Фонд OpenAI в $1 млрд. Илона Маска и Сэма Альтмана
2016: DeepMind, OpenAI: динамическое обучение играм Atari
2016: Google DeepMind обыграл чемпиона мира по игре го
2017: OpenAI обыграл чемпиона мира по компьютерной игре Dota 2
2020: Модель GPT-3 синтезирует тексты, неотличимые от человеческих
2023: GPT-4 демонстрирует «проблески общего искусственного интеллекта»

Машинное обучение (Machine Learning, ML)

- одна из ключевых информационных технологий будущего
- наиболее успешное направление ИИ, вытеснившее экспертные системы и инженерию знаний





- проведение функции через заданные точки в сложно устроенных пространствах
- математическое моделирование в условиях, когда знаний мало, данных много
- тысячи различных методов и алгоритмов
- около 100 000 научных публикаций в год

Задачи машинного обучения с учителем

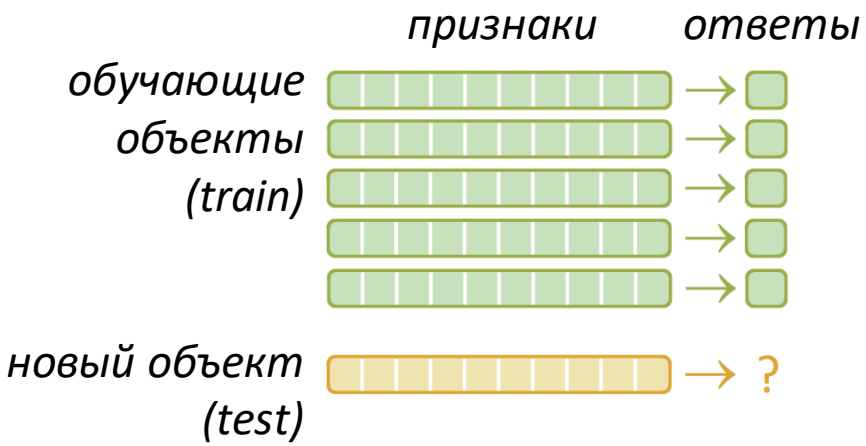
Этап №1 – обучение с учителем

- **На входе:**
  данные – выборка прецедентов «*объект* → *ответ*»,
  каждый объект описывается набором *признаков*
- **На выходе:**
  модель, предсказывающая ответ по объекту

Если нет данных,
то нет
и машинного
обучения

Этап №2 – применение

- **На входе:**
  данные – новый объект
- **На выходе:**
  предсказание ответа на новом объекте



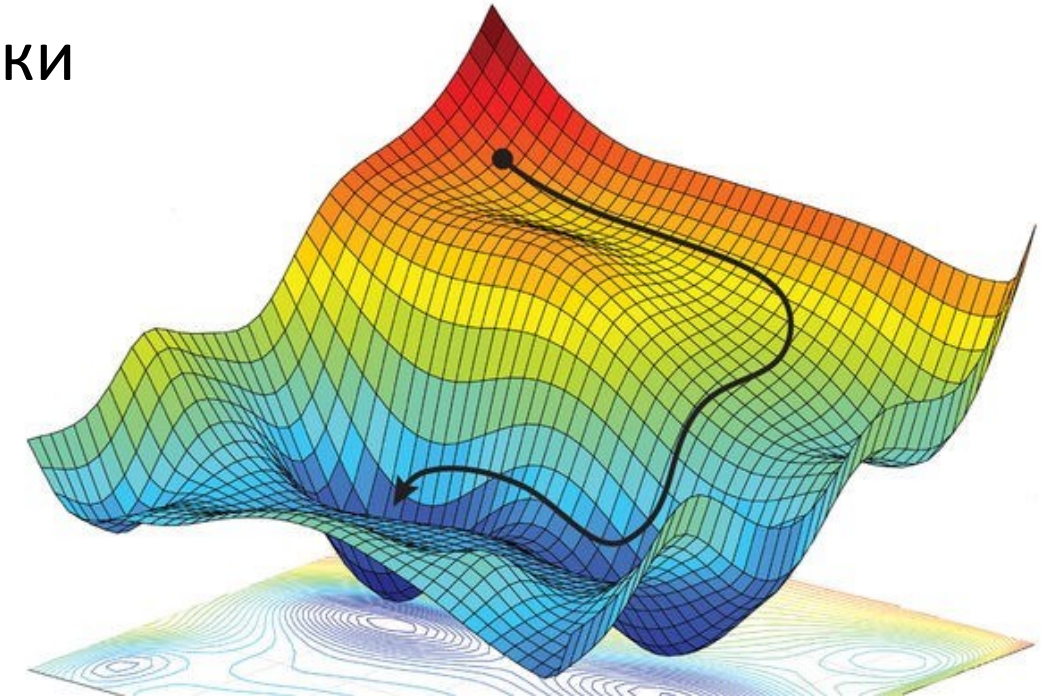
Машинное обучение – это оптимизация

x – вектор объекта обучающей выборки

w – вектор параметров модели

$\mathrm{Loss}(x, w)$ – функция потерь

$Q(w)$ – критерий качества модели



Задача на этапе обучения модели:

$$Q(w) = \sum_{x} \mathrm{Loss}(x, w) \to \min$$

Способ решения – численные методы оптимизации

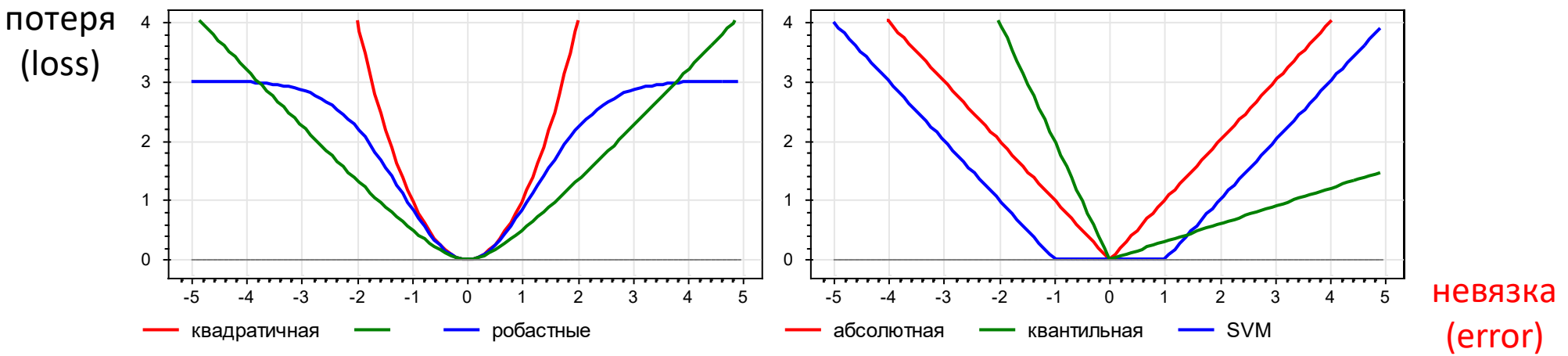
Восстановление регрессии (regression)

x – вектор объекта обучающей выборки, y – числовой ответ

$a(x, w)$ – модель регрессии с параметрами w

Например, $a(x, w) = \sum_j w_j x_j$ — линейная модель регрессии

$\text{Loss}(x, w) = (a(x, w) - y)^2$ – квадратичная функция потерь



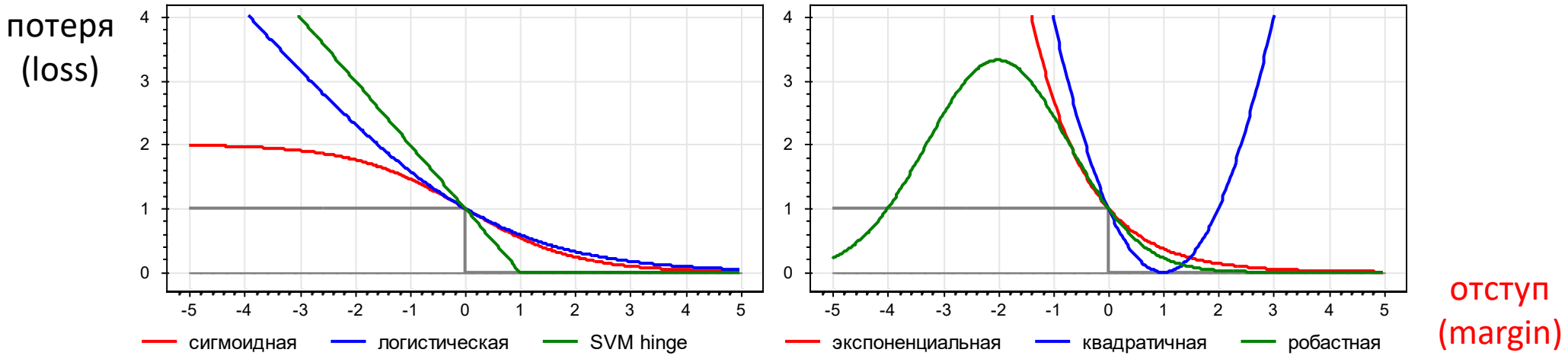
Классификация (classification)

x – вектор объекта обучающей выборки, y – ответ ($+1$ или -1)

$a(x, w)$ – модель классификации с параметрами w

Например, $a(x, w) = \text{sign}\left(\sum_j w_j x_j\right)$ — линейная модель

$\text{Loss}(x, w) = \max\left(0, 1 - y\sum_j w_j x_j\right)$ – функция потерь hinge



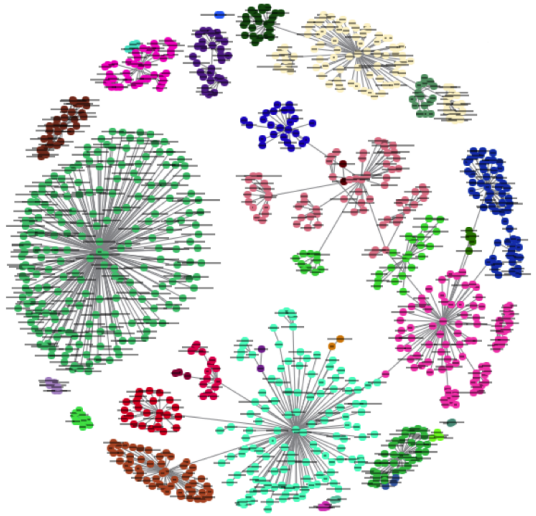
Кластеризация (clustering)

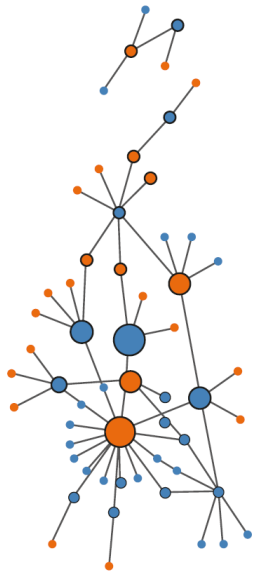
x – вектор объекта обучающей выборки, ответов не дано

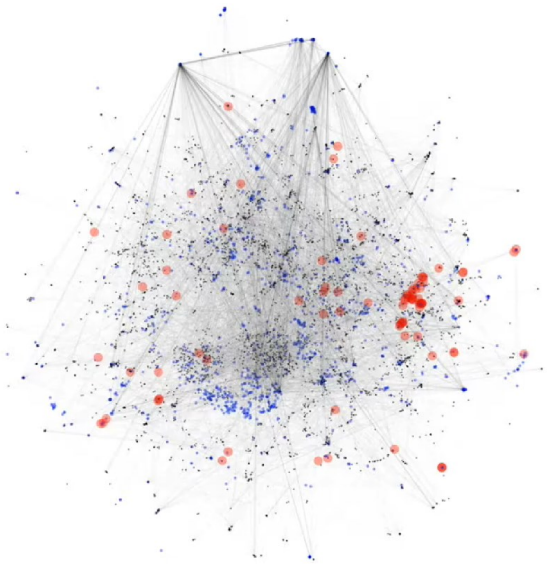
$a(x, w)$ – ближайший к x центр кластера

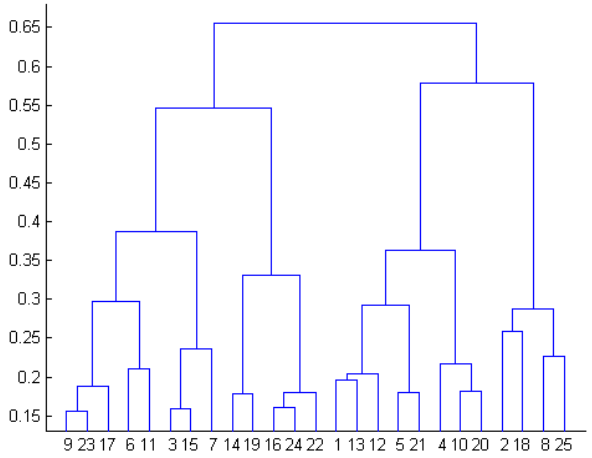
$w = \{c_1, \dots, c_K\}$ – векторы центров всех кластеров

$\text{Loss}(x, w) = \min_k \|x - c_k\|$ – расстояние до ближайшего кластера









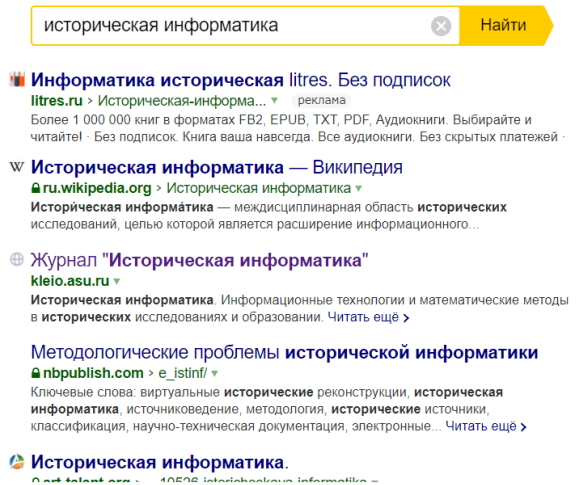
Ранжирование (learning to rank)

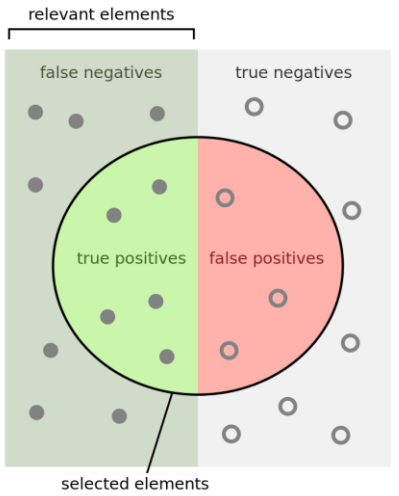
x – вектор пары «запрос-документ», y – оценка релевантности

$a(x, w)$ – модель ранжирования документов по запросу, параметр w

Например, $a(x, w) = \sum_j w_j x_j$ — линейная модель

$$\text{Loss}(x, x', w) = \max\Big(0, 1 - [y > y'](a(x, w) - a(x', w))\Big)$$





Precision = $\frac{\text{true positives}}{\text{selected elements}}$

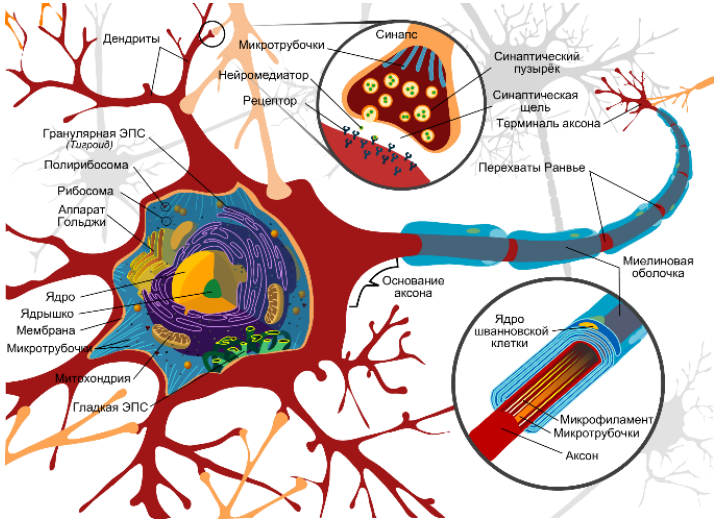
Recall = $\frac{\text{true positives}}{\text{relevant elements}}$

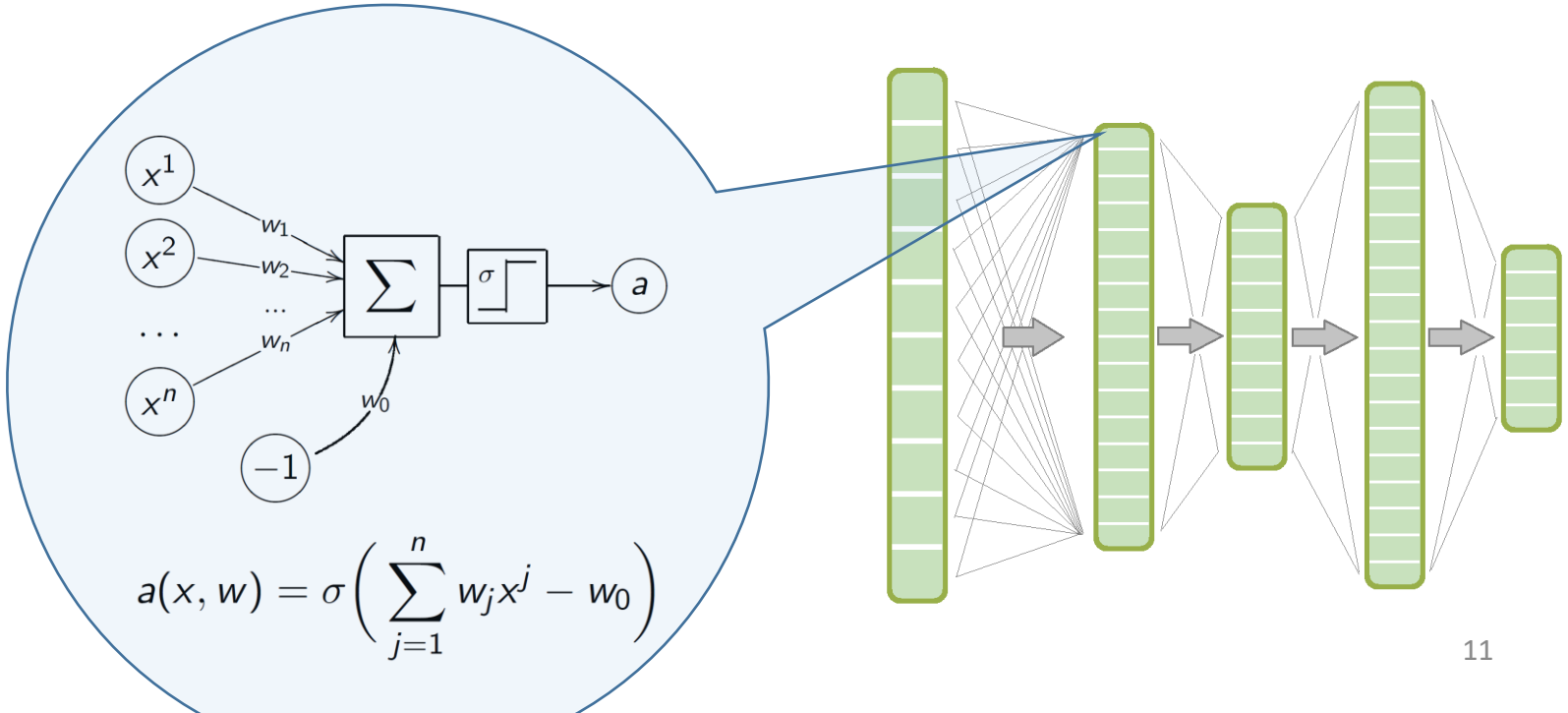
Искусственные нейронные сети

На каждом слое сети вектор объекта преобразуется в новый вектор

Каждое преобразование (нейрон) – взвешенная сумма признаков

Веса w являются обучаемыми параметрами модели

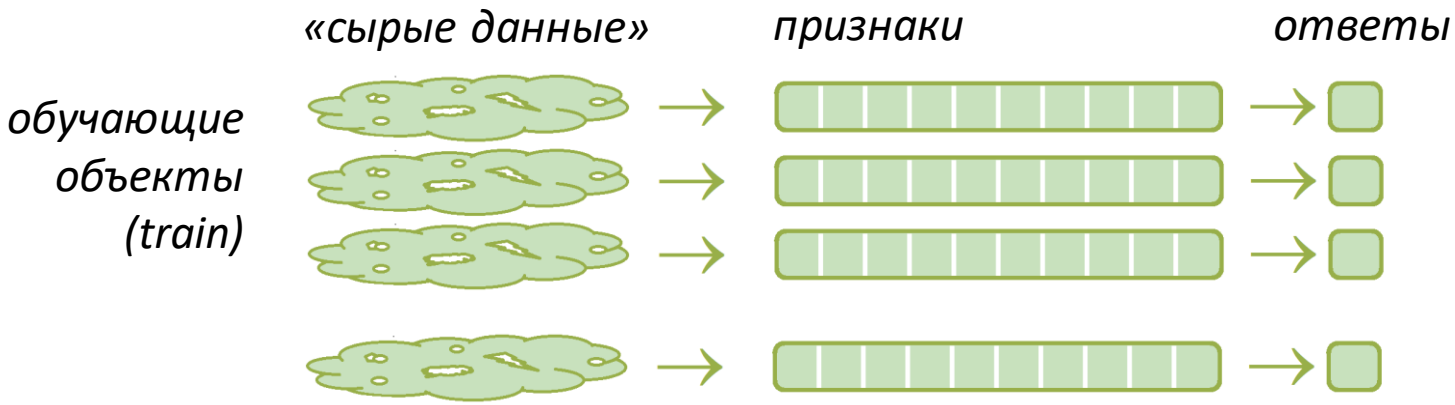




Глубокие нейронные сети

Вход: сложно структурированные «сырые» данные объектов

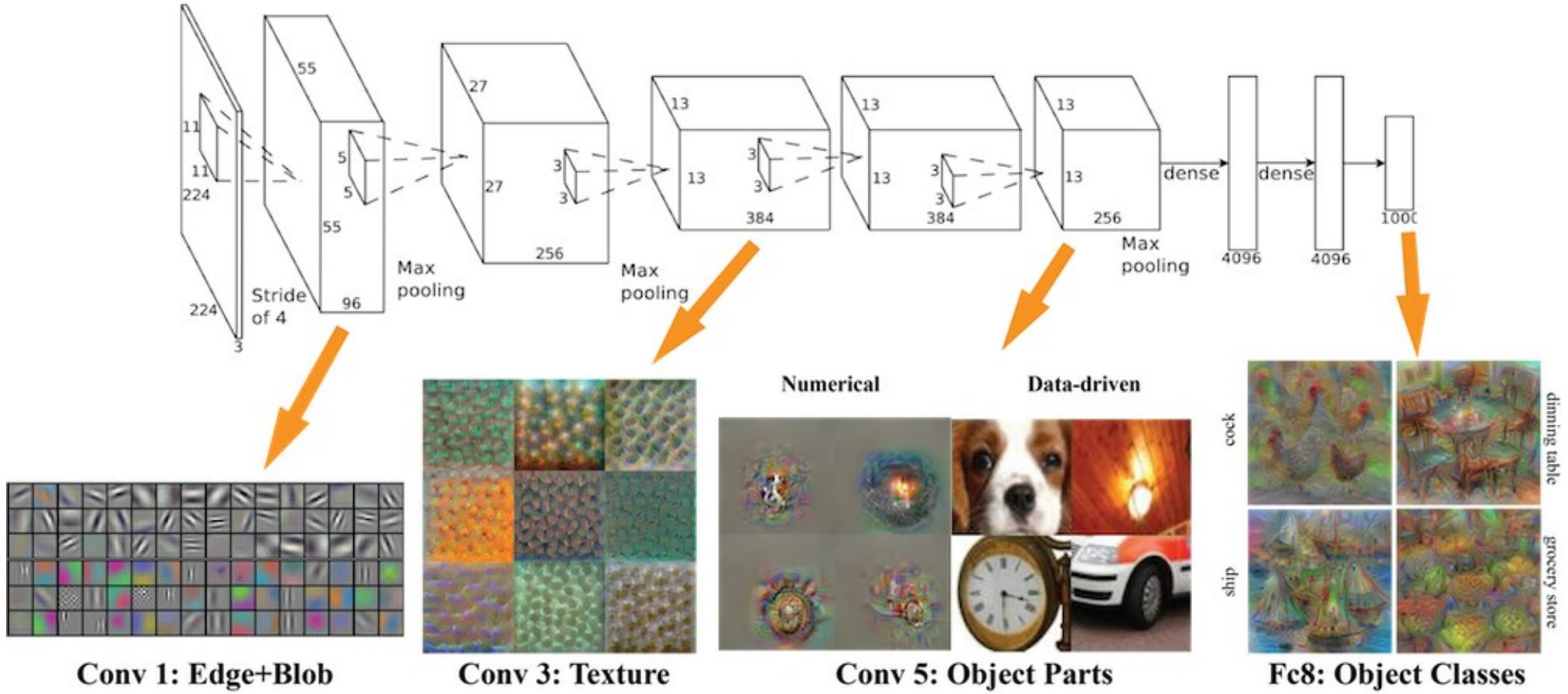
Выход: *векторные представления объектов*, затем ответы



Deep Learning – это всего лишь обучаемая векторизация сложных объектов

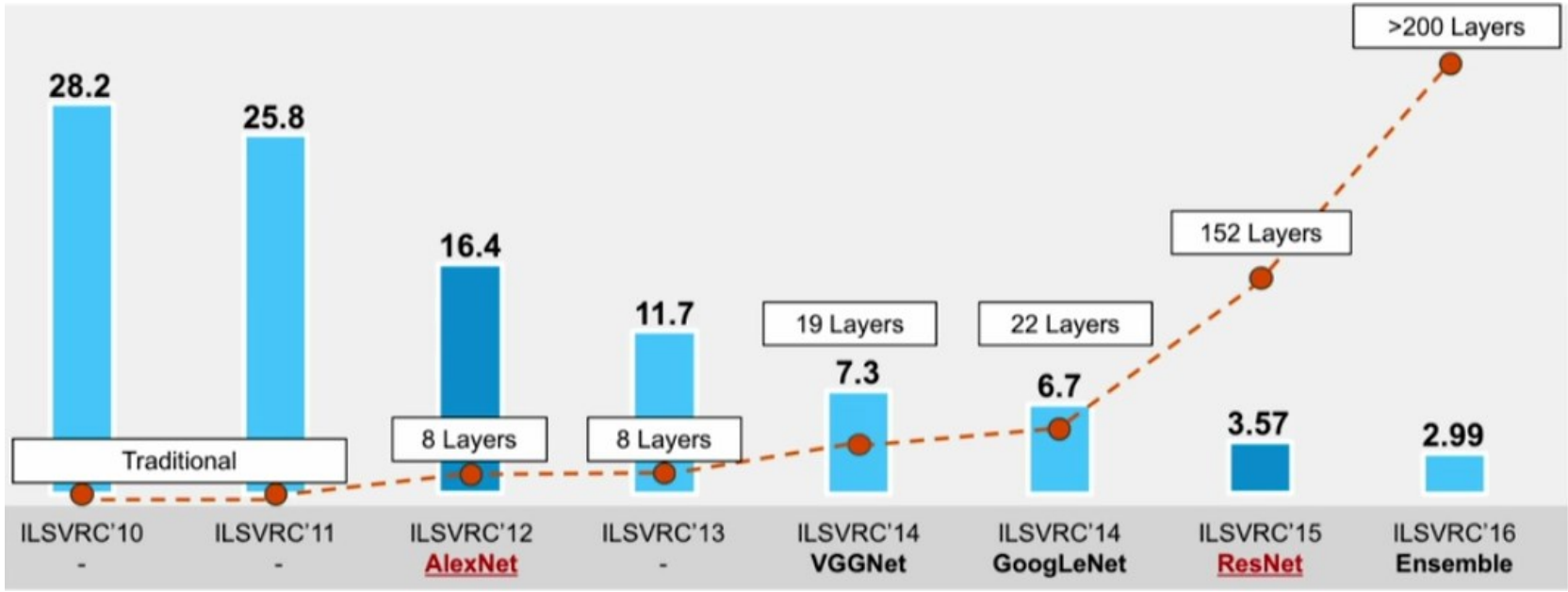
Примеры сложно структурированных объектов:
изображения, видео, временные ряды, тексты, транзакции, графы, ...

Глубокие свёрточные нейронные сети
для классификации изображений



Роль больших данных

ImageNet: открытая выборка 14M изображений, 20K категорий



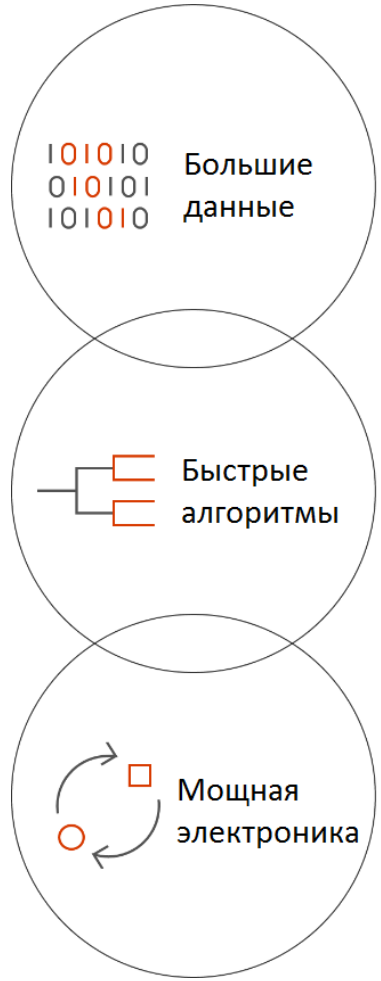
Старт в 2009 г. Человеческий уровень ошибок 5% пройден в 2015 г.

Три составляющих успеха Deep Learning

- Повсеместное применение компьютерных технологий
  → *накопление больших выборок данных*
  в частности, ImageNet

- Развитие математических методов и алгоритмов
  → *накопление критической массы опыта*
  методы оптимизации, контроль переобучения

- Достижения микроэлектроники
  → *рост вычислительных мощностей по закону Мура*
  в частности, GPU



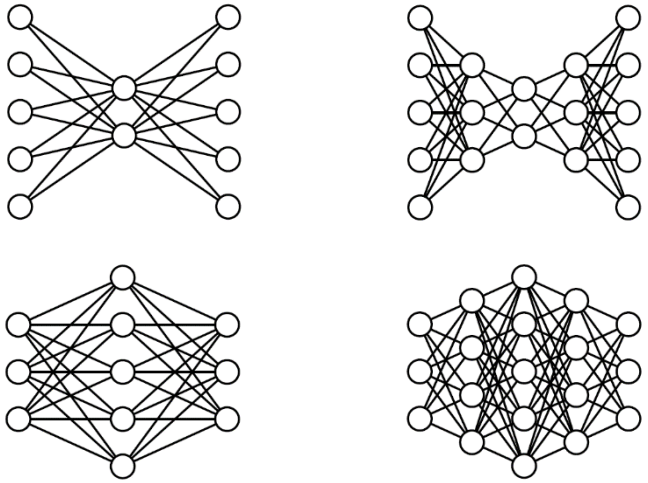
Обучаемая векторизация (autoencoders)

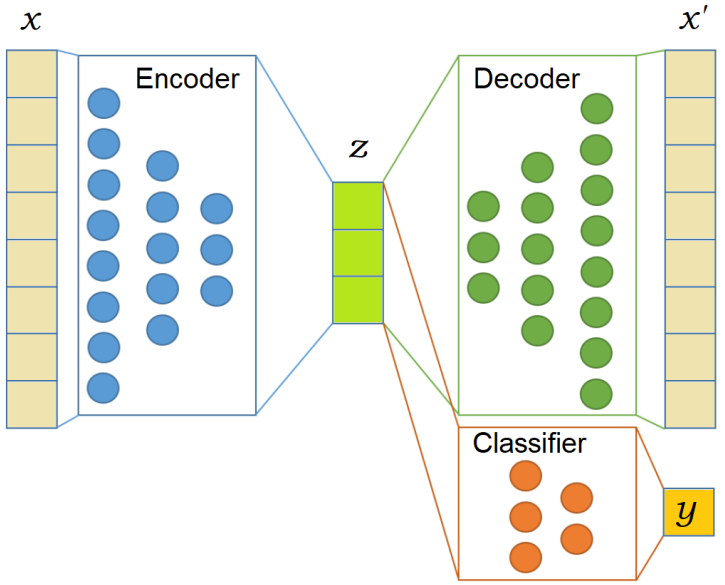
x – описание объекта обучающей выборки, ответов не дано

$z = f(x, w)$ – *модель кодирования* (векторизации) x в вектор z

$x' = g(z, w')$ – *модель декодирования* z в реконструкцию x'

$\text{Loss}(x, w) = \|g(f(x, w), w') - x\|$ – точность реконструкции объекта





обучаемая векторизация сложных объектов

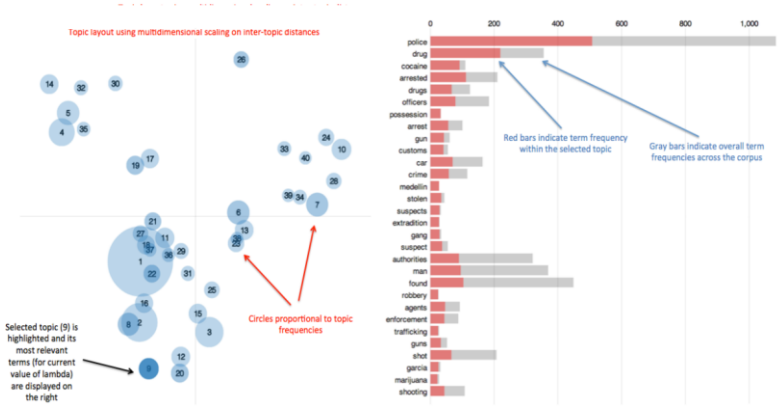
Тематизация текстов (topic modelling)

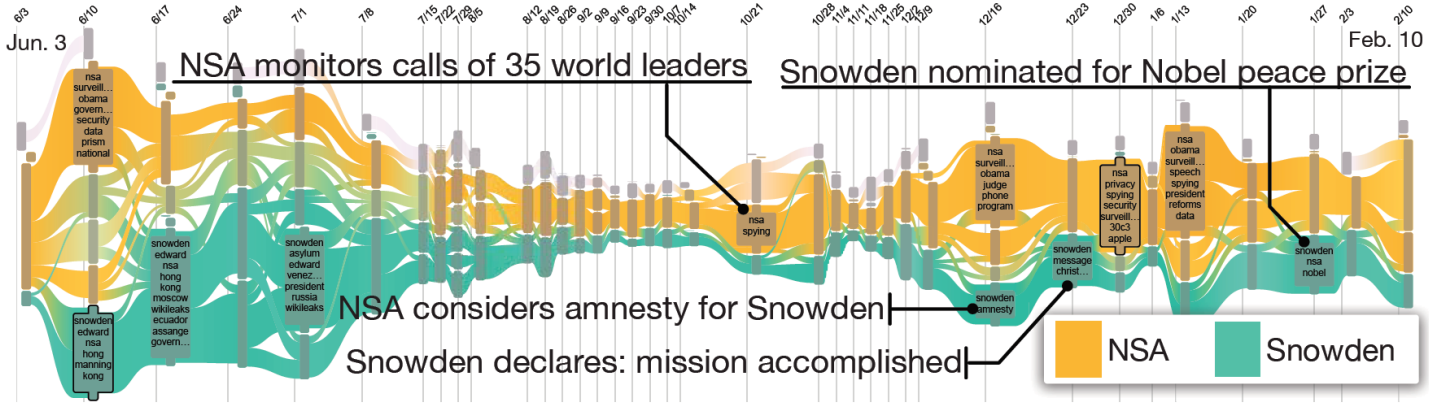
x – текст на естественном языке, «мешок слов» $p(\text{слово}|x)$

$z = f(x, w)$ – *модель кодирования* x в вектор тем $z = p(\text{тема}|x)$

$x' = g(z, w)$ – *модель декодирования* z в реконструкцию текста x'

$\text{Loss}(x, w) = \text{KL}\big(\, x \parallel g(f(x, w), w)\big)$ – точность реконструкции текста





Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. URSS, 2023. ISBN 978-5-9519-4345-3.

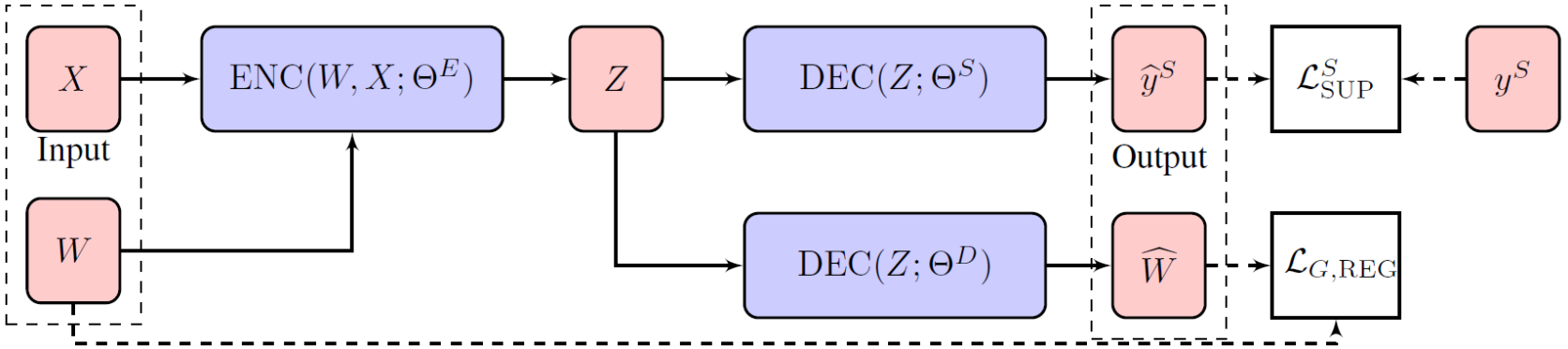
Векторизация графов (graph embeddings)

$x; (x, x')$ – данные об объектах и взаимодействиях между объектами

$z = f(x, w)$ – *модель кодирования* вершин графа x в векторы z

$x' = g(z, w')$ – *модель декодирования* z в реконструкцию x'

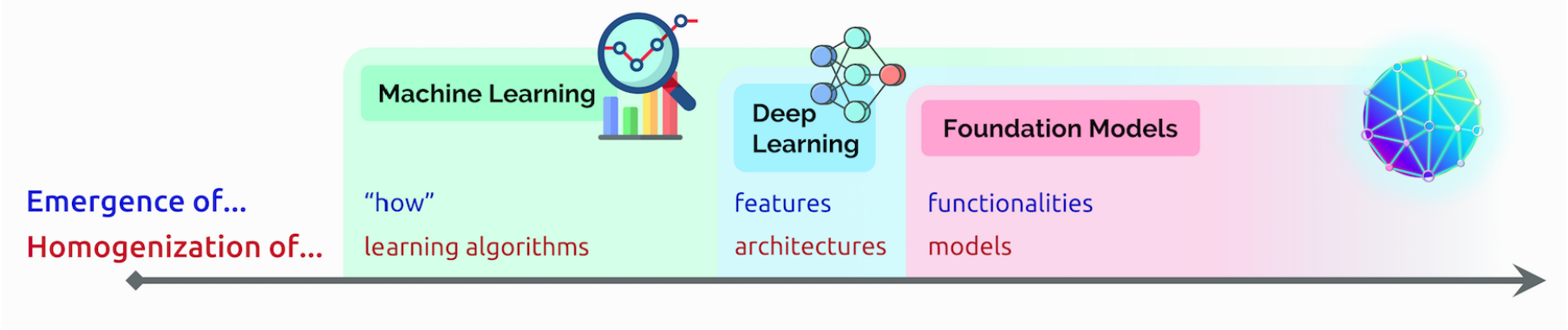
$\text{Loss}(x, w) = \|g(f(x, w), w') - x\| + \tau L_{\text{SUP}}(x, w_S)$ – сумма критериев

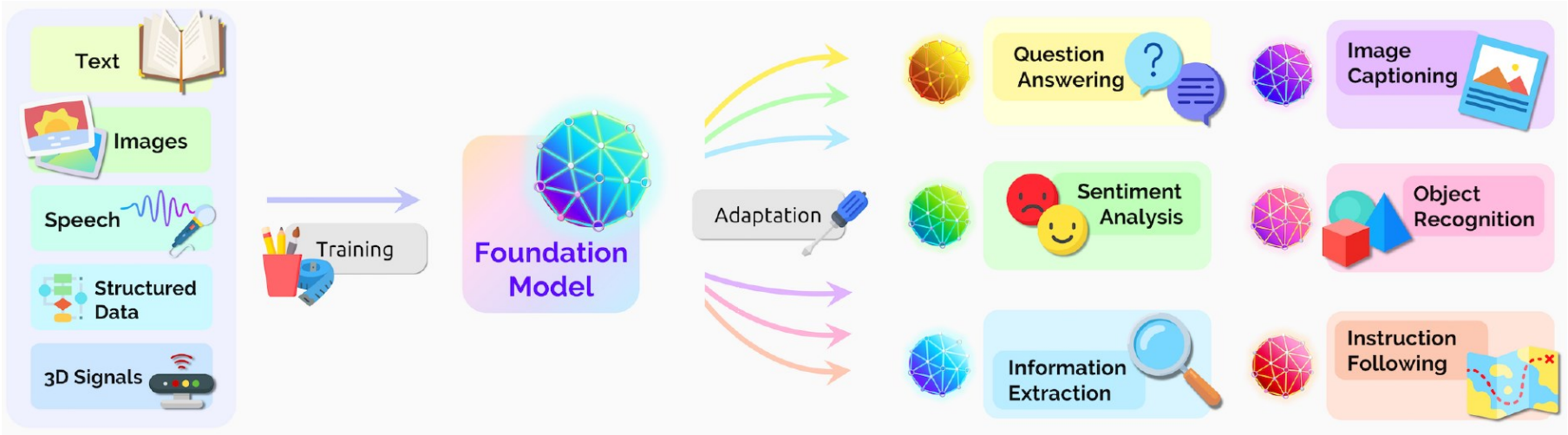


обучаемая векторизация сложных объектов по данным об их взаимодействиях

T.Mikolov et al. Efficient estimation of word representations in vector space, 2013.
I.Chami et al. Machine learning on graphs: a model and comprehensive taxonomy. 2020.

Фундаментальные модели (Foundation Models)





R.Bommasani et al. (CRFM, Stanford University). On the opportunities and risks of foundation models. 2021

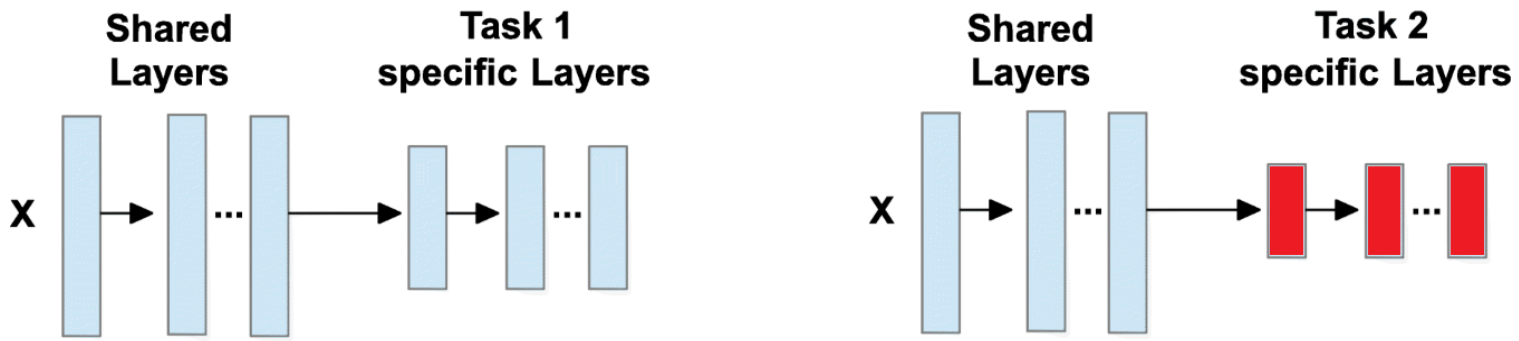
Предобучение (pre-training, transfer learning)

$z = f(x, w)$ – *модель векторизации*, универсальная для многих задач

$y = g(z, w')$ – часть модели, специфичная для своей задачи

$\min\limits_{w,\, w'}: \sum_x \text{Loss}_1(g_1(f(x, w), w'))$ – обучение по большим данным

$\min\limits_{w'}: \sum_{x'} \text{Loss}_2(g_2(f(x', w), w'))$ – обучение по своим данным

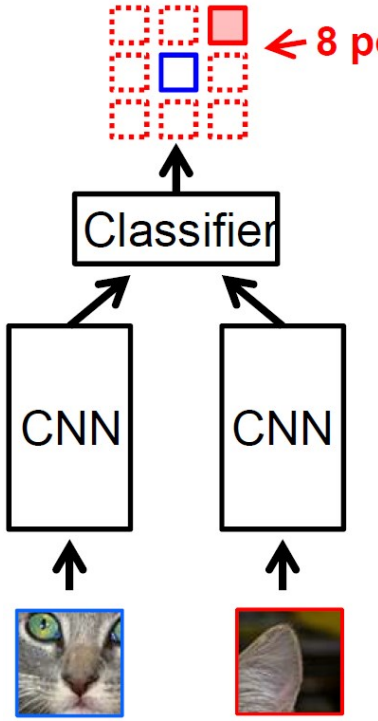


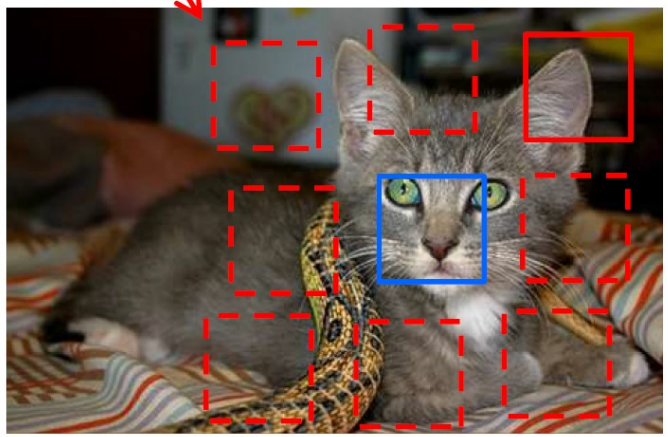
Sinno Jialin Pan, Qiang Yang. A Survey on Transfer Learning. 2009

Самостоятельное обучение (self-supervised)

Модель векторизации $z = f(x, w)$ обучается предсказывать взаимное расположение пар фрагментов одного изображения

Преимущество:
сеть выучивает векторные представления объектов без размеченной обучающей выборки





Unsupervised visual representation learning by context prediction, Carl Doersch, Abhinav Gupta, Alexei A. Efros, ICCV 2015

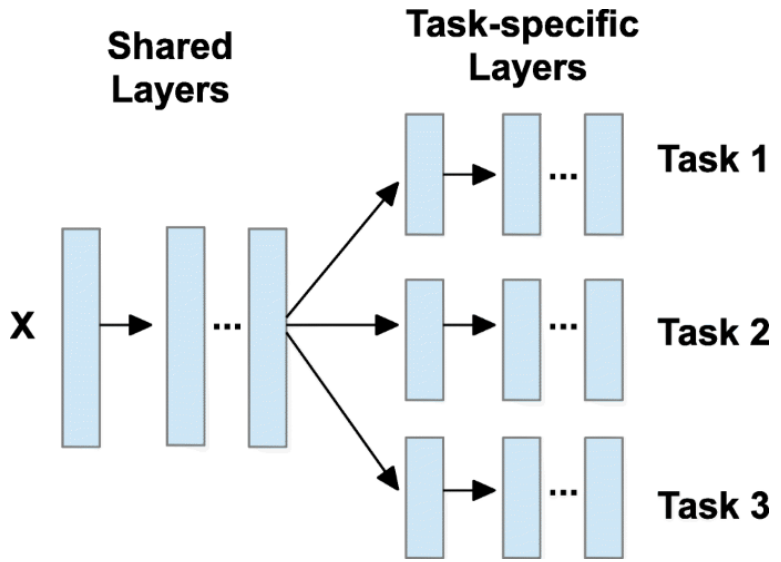
Многозадачное обучение (multi-task learning)

$z = f(x, w)$ – *модель векторизации*, универсальная для всех задач

$y = g_t(z, w'_t)$ – часть модели, специфичная для t-й задачи

$\min\limits_{w, w'_t}: \sum_t \sum_x \text{Loss}_t(g_t(f(x, w), w'_t))$ – обучение по всем задачам

few-shot learning – обучение по малому числу примеров



M.Crawshaw. Multi-task learning with deep neural networks: a survey. 2020

Y.Wang et al. Generalizing from a few examples: a survey on few-shot learning. 2020

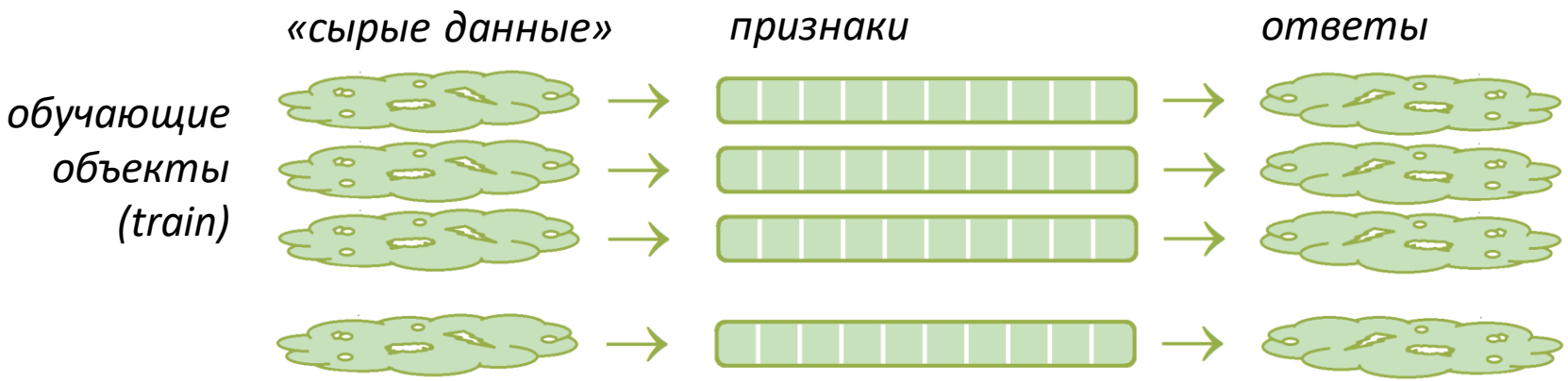
Содержание

1. Фундаментальные основы искусственного интеллекта
   - От экспертных систем к машинному обучению
   - Задачи оптимизации в машинном обучении
   - Обучаемая векторизация и глубокие нейронные сети
2. Проблески общего искусственного интеллекта
   - Модели внимания, трансформеры и генеративные модели
   - Большие языковые модели (LLM)
   - Свойство эмерджентности
3. Фундаментальные проблемы технологий ИИ
   - Математические проблемы
   - Технологические проблемы
   - Социо-гуманитарные и организационные проблемы

Нейронные сети для синтеза объектов

Вход: сложно структурированные объекты

Выход: сложно структурированные ответы



Примеры: синтез изображений, перенос стиля, распознавание речи, машинный перевод, суммаризация текстов, диалог с пользователем

Модели: seq2seq, CNN, RNN, LSTM, GAN, BERT, GPT и др.

Генеративная состязательная сеть (GAN)

$x = g(z, w)$ – модель генерации реалистичного объекты x из шума z

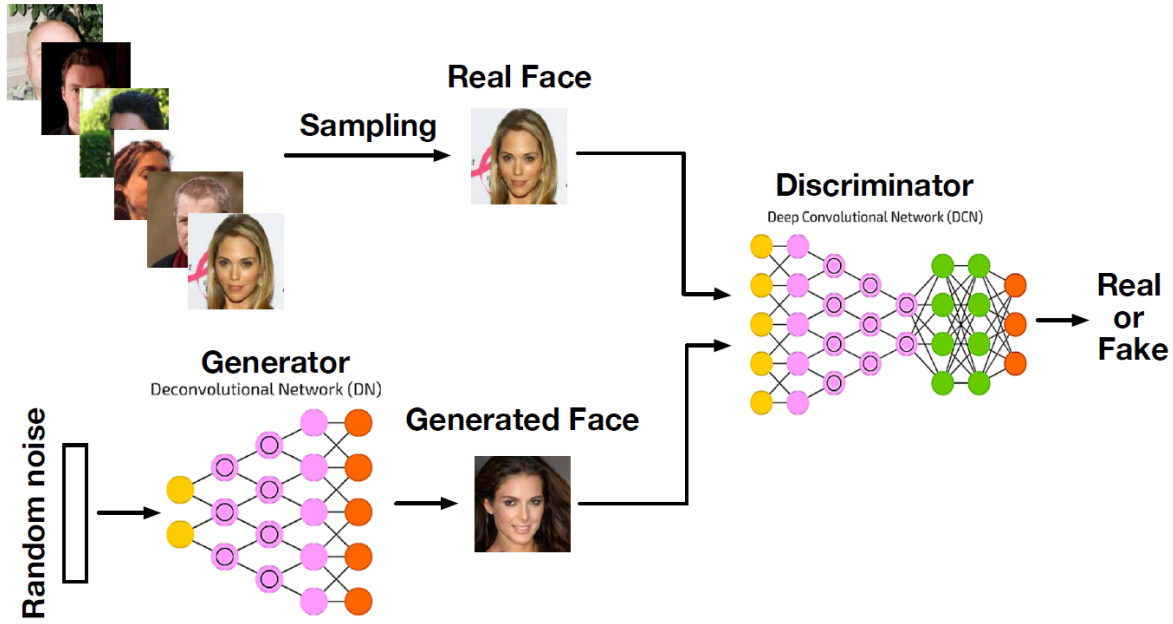
$f(x, w')$ – модель классификации x «реальный/сгенерированный»

$\min_{w} \max_{w'} \sum_x \ln f(x, w') + \ln\left(1 - f(g(z, w), w')\right)$ – совместное обучение

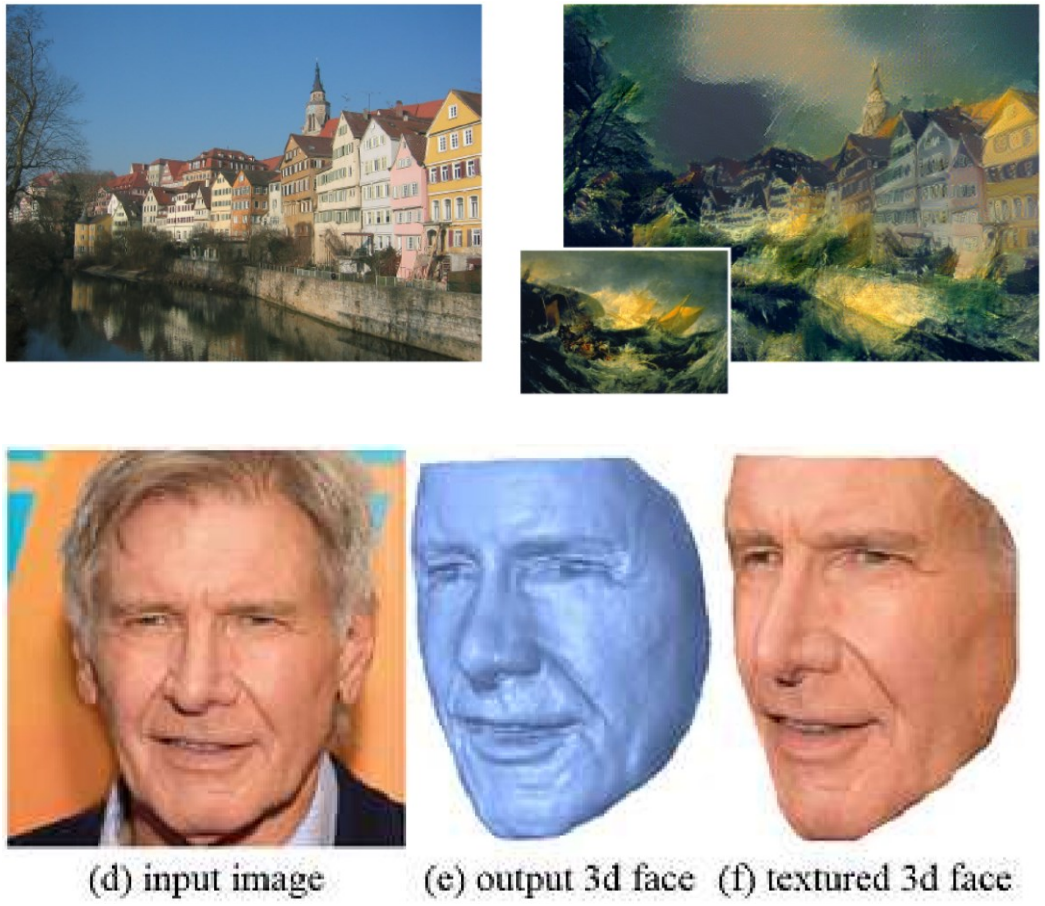
Antonia Creswell et al. Generative Adversarial Networks: an overview. 2017.

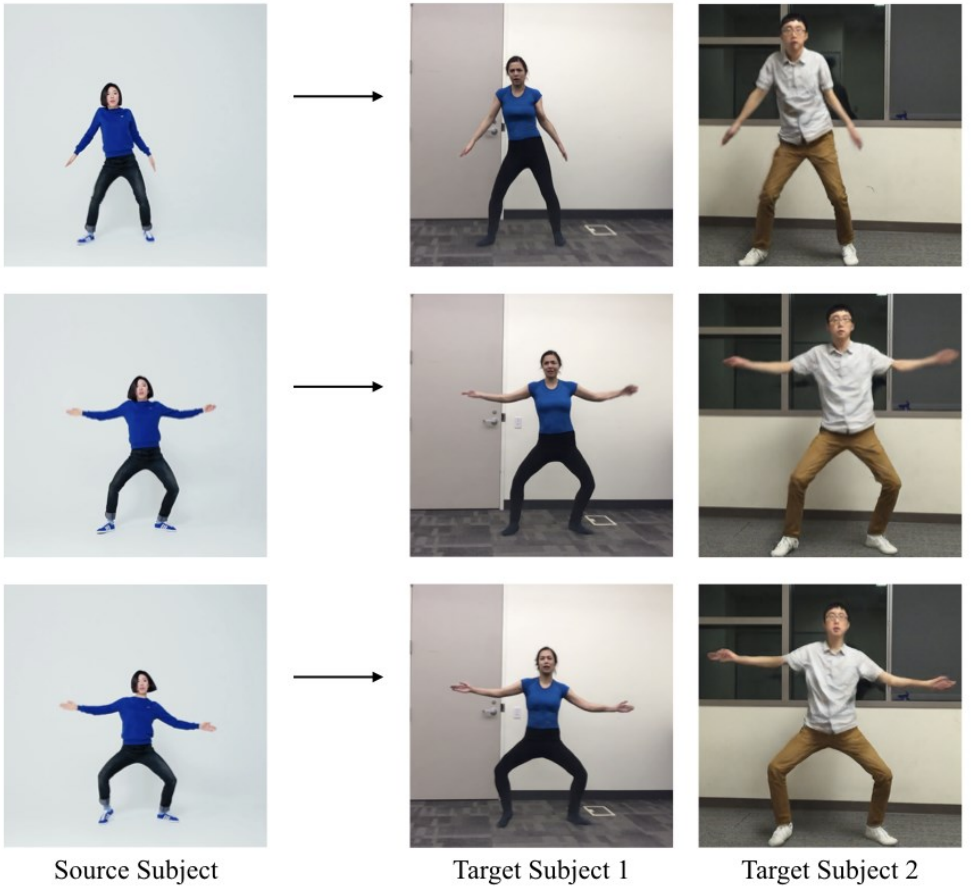
Zhengwei Wang et al. Generative Adversarial Networks: a survey and taxonomy. 2019.

Chris Nicholson. A Beginner's Guide to Generative Adversarial Networks. 2019.



Синтез изображений и видео





Caroline Chan, Shiry Ginosar, Tinghui Zhou, Alexei A. Efros. Everybody Dance Now. ICCV-2019.

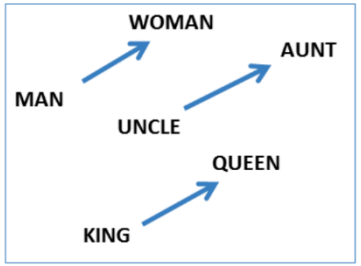
Эволюция подходов в обработке текстов

Декомпозиция задач по уровням «пирамиды NLP»

- морфологический анализ, лемматизация, опечатки, …
- синтаксический анализ, выделение терминов, NER, …
- семантический анализ, выделение фактов, тем, …

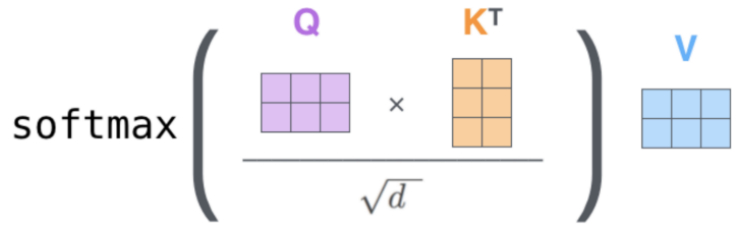
Модели векторизации слов (эмбедингов)

- модели дистрибутивной семантики:
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016], …
- тематические модели LDA [Blei, 2003], ARTM [2014], …

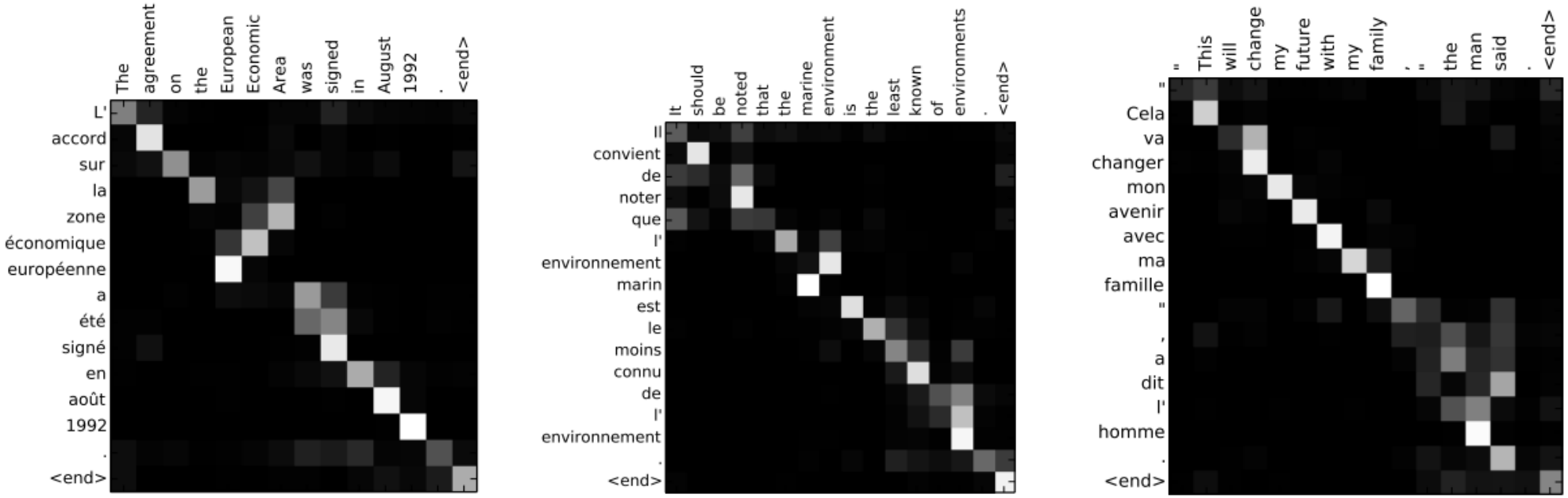


Нейросетевые модели контекстной векторизации

- рекуррентные нейронные сети: LSTM, GRU, …
- «end-to-end» модели внимания и трансформеры:
  машинный перевод [2017], BERT [2018], GPT-4 [2023], …



Модели внимания: машинный перевод



Интерпретация моделей внимания: *матрица семантического сходства* A[t,i] показывает, на какие слова x[i] входного текста модель обращает внимание, когда генерирует слово перевода y[t]

Bahdanau et al. Neural machine translation by jointly learning to align and translate. 2015

Модели внимания: аннотирование изображений

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

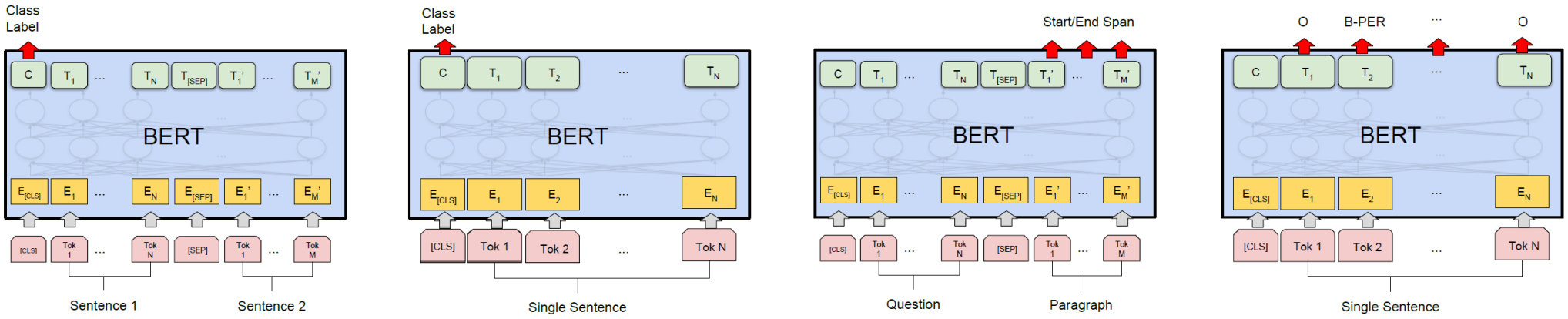
A giraffe standing in a forest with trees in the background.

Интерпретация: на какие области модель обращает внимание, генерируя подчёркнутое слово в описании изображения

Kelvin Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2016

Трансформеры: нейросетевые модели языка

- Обучаются векторизовать и предсказывать слова по контексту
- Обучаются по терабайтам текстов, «они видели в языке всё»
- Мультиязычны: обучаются на десятках языков
- Мультизадачны: для каждой новой задачи NLP/NLU достаточно предобученной модели или дообучения на небольшой выборке



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

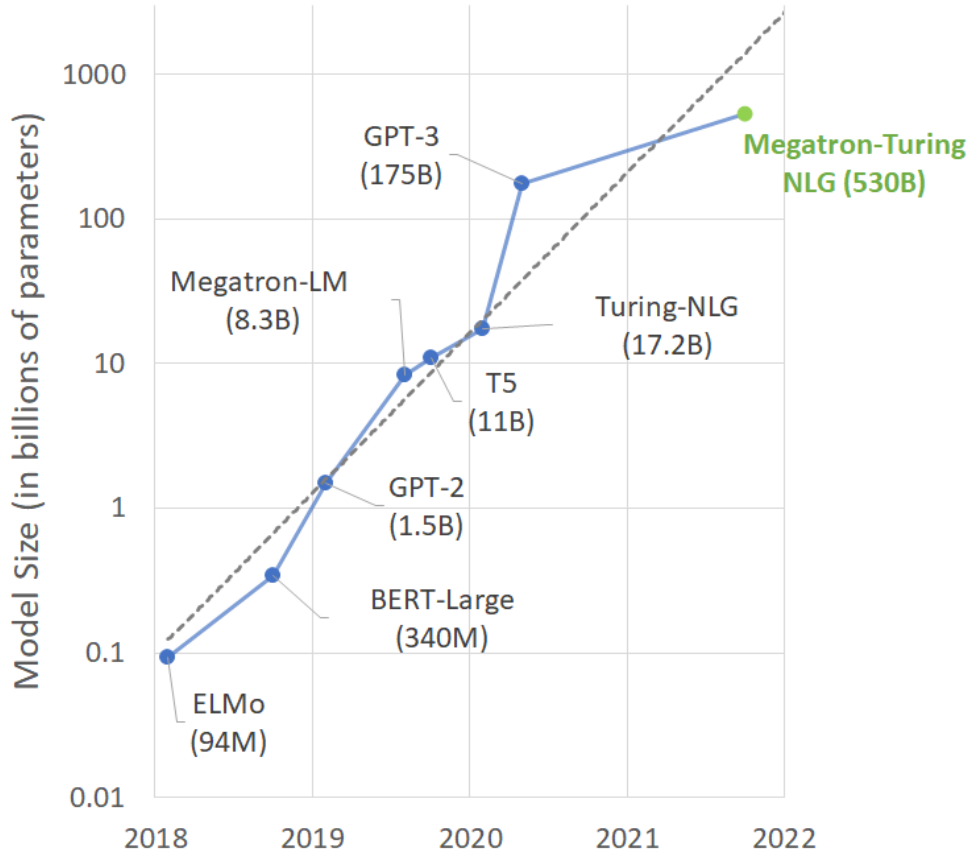
(b) Single Sentence Classification Tasks: SST-2, CoLA

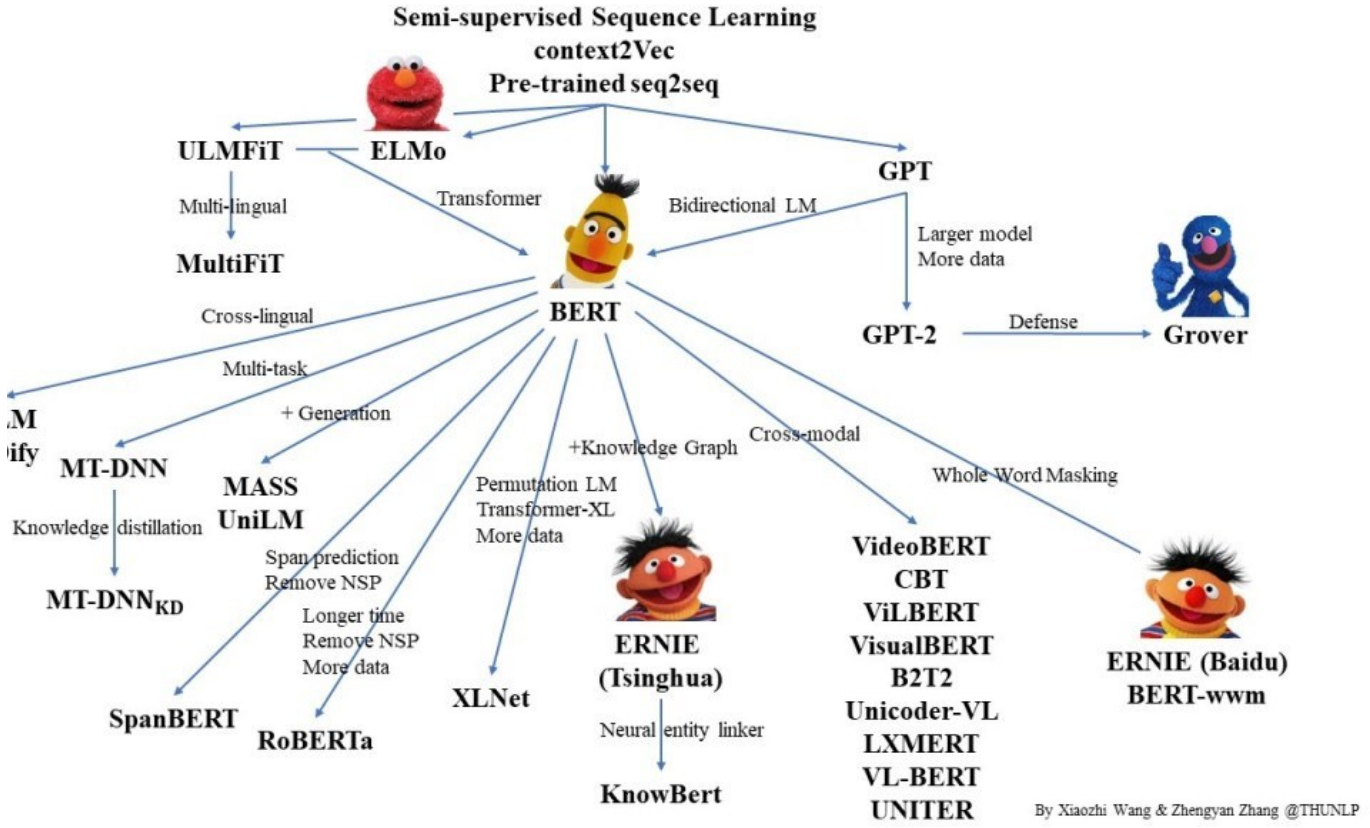
(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Трансформеры: нейросетевые модели языка

Рост числа параметров в больших языковых моделях (LLM)





Проблески общего искусственного интеллекта

Sparks of Artificial General Intelligence:
Early experiments with GPT-4

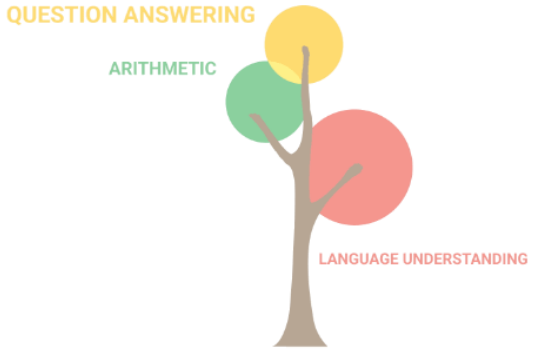
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research *(27 March 2023)*

Эмерджентность — *новые навыки модели, не закладывавшиеся при обучении:*

- объяснять свои ответы, перефразировать, переводить на другие языки
- реферировать, генерировать планы, сценарии, шаблоны
- строить аналогии, менять тональность, стиль, глубину изложения
- генерировать программный код на различных языках
- решать некоторые логические и математические задачи
- искать и исправлять собственные ошибки по подсказке

Эмерджентность: новые способности модели

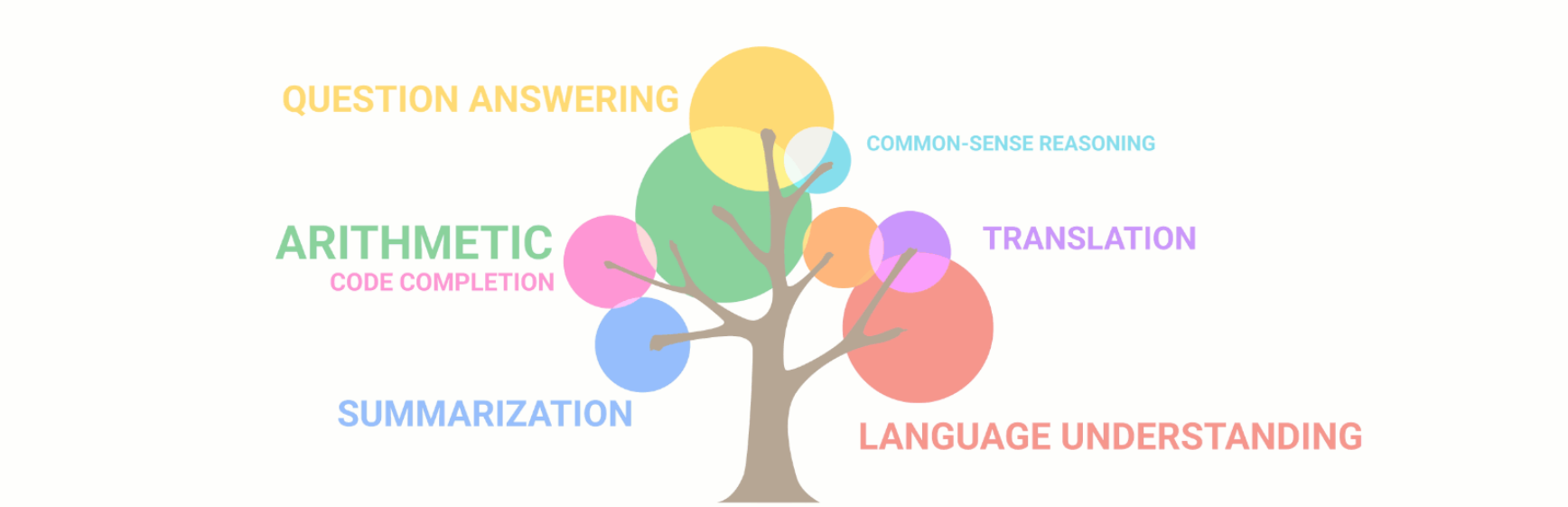


GPT-2: 14-Feb-2019

1,5 млрд. параметров, корпус 10 млрд. токенов (40Gb), контекст 768 слов (1,5 стр.)

- способность написать эссе, которое конкурсное жюри не смогло отличить от написанного человеком

Эмерджентность: новые способности модели

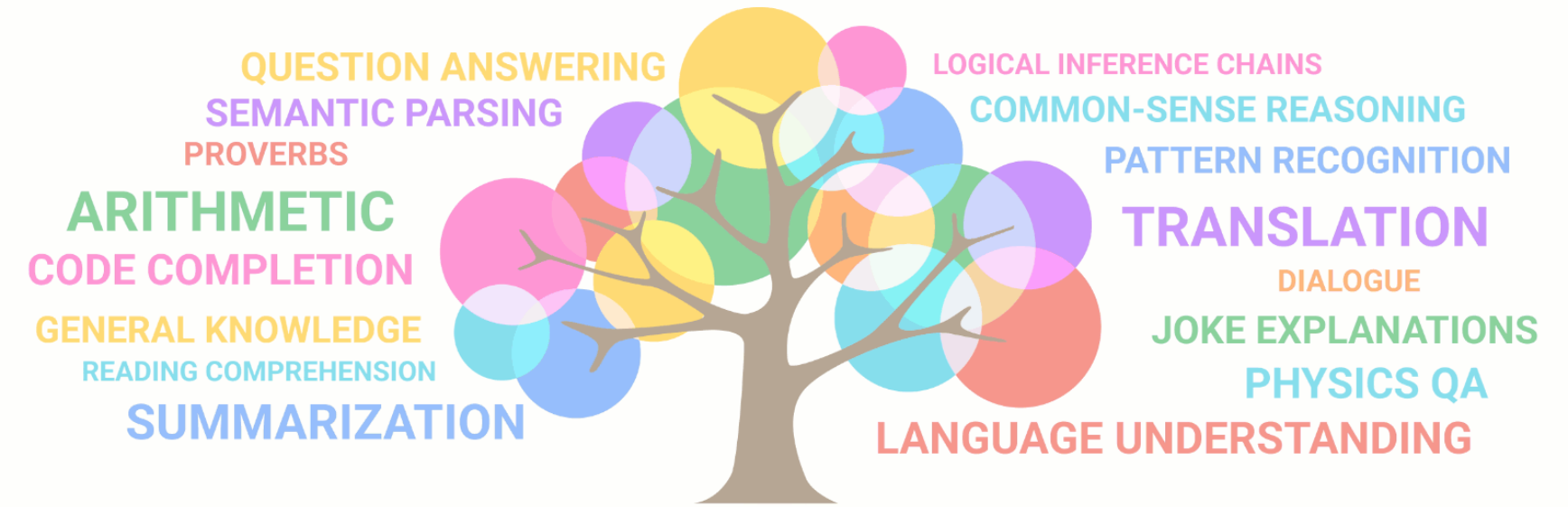


GPT-3: 11-Jun-2020

175 млрд. параметров, корпус 500 млрд. токенов, контекст 1536 слов (3 стр.)

- способность делать перевод на другие языки
- способность решать логические и простейшие математические задачи
- способность генерировать программный код по текстовому описанию

Эмерджентность: новые способности модели



GPT-4: 14-Mar-2023

>1 трл. параметров, корпус >1Tb, контекст 24 000 слов (48 страниц)

- способность описывать и анализировать изображения
- способность реагировать на подсказки вроде «Let's think step by step»
- способность решать качественные физические задачи по картинке

Возможности и угрозы

Чаты GPT уже способны помогать с рутинно-творческой работой:

- служить языковым интерфейсом к знаниям человечества
- делать обзоры, рефераты, сводки на разных языках
- сообщать новости, поддерживать разговор по теме
- генерировать документы или сайты по описанию
- в том числе юридические документы по шаблонам
- генерировать программный код по описанию
- уточнять и дополнять контент по просьбе, в диалоге
- разговаривать с детьми с учётом возрастных особенностей
- выполнять функции воспитателя, учителя, наставника
- оказывать психологическую помощь

Возможности и угрозы

Чаты GPT способны (даже не обладая автономностью):

- «галлюцинируя», давать неверные сведения, касающиеся здоровья человека, других людей, событий, технологий, норм, правил, законов
- вызывать необоснованное доверие и манипулировать
- побуждать человека к действиям, не выгодным ему
- побуждать изменить точку зрения, замалчивая информацию
- поддерживать предрассудки и лженаучные представления
- поддерживать пропагандистские медиа-кампании
- влиять на формирование мировоззрения детей и подростков
- оказывать депрессивное воздействие на психику

Содержание

1. Фундаментальные основы искусственного интеллекта
   - От экспертных систем к машинному обучению
   - Задачи оптимизации в машинном обучении
   - Обучаемая векторизация и глубокие нейронные сети
2. Проблески общего искусственного интеллекта
   - Модели внимания, трансформеры и генеративные модели
   - Большие языковые модели (LLM)
   - Свойство эмерджентности
3. Фундаментальные проблемы технологий ИИ
   - Математические проблемы
   - Технологические проблемы
   - Социо-гуманитарные и организационные проблемы

Фундаментальные математические проблемы

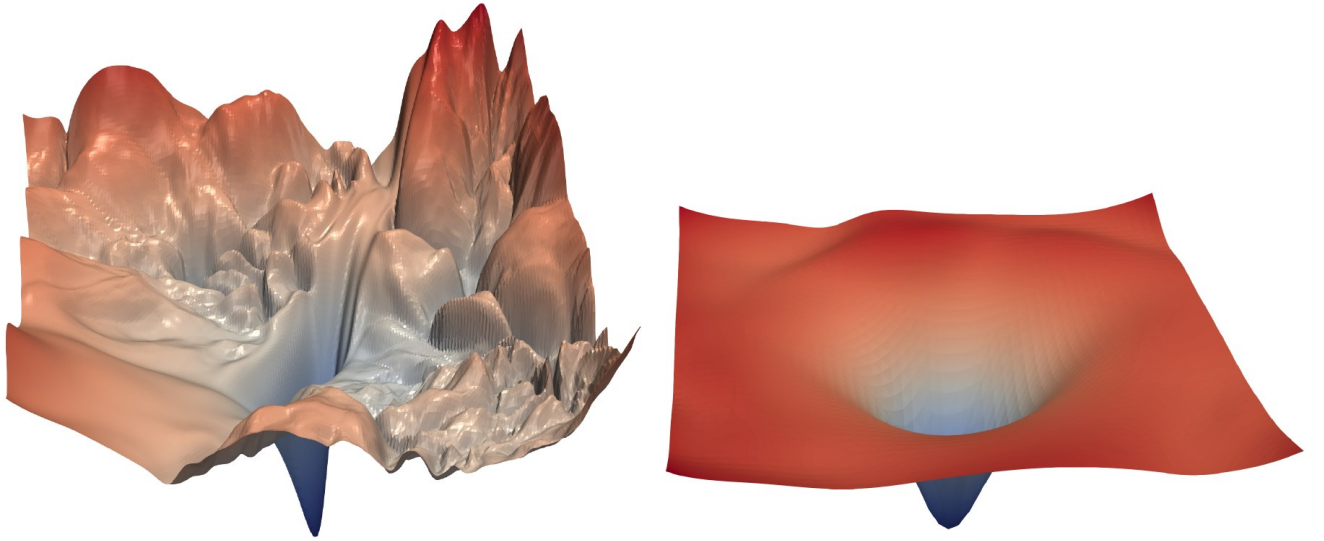
1. Поиск более компактных архитектур ИНС (проблема $100M)
   - разреживание полносвязных слоёв, квантизация, дистилляция
   - использование низкоранговых автокодировщиков (матричных разложений)
2. Исследование ландшафта оптимизируемого критерия
   - методы поиска устойчивых (широких, непереобученных) экстремумов
   - влияние архитектуры, регуляризации, функции потерь на экстремумы
3. Алгоритмы распределённого и федеративного обучения
4. Распознавание синтетических текстов и синтез «водяных знаков»
5. Полная или хотя бы частичная интерпретируемость моделей

Peter Belcak, Roger Wattenhofer. Exponentially Faster Language Modeling. ArXiv, 21 Nov 2023
Eduard Tulchinskii et al. Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts. 2023

Сквозные связи (skip connection, ResNet)

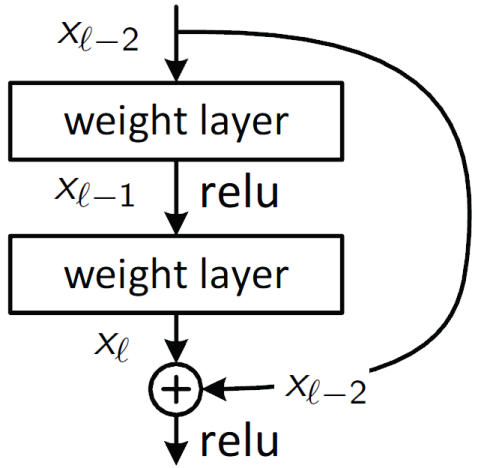
Сквозная связь слоя l с предшествующим слоем $l - d$

Упрощается ландшафт оптимизируемого критерия, устраняются локальные экстремумы и седловые точки:



without skip connections

with skip connections



Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. 2015

Проблемы интерпретируемости моделей

Тематические модели языка интерпретируемы благодаря тому, что векторы параметров w — дискретные вероятностные распределения,

$$w_i \geq 0, \quad \sum_i w_i = 1.$$

Градиентный шаг для численного решения задачи $Q(w) \to \max$:

$$w := w + h\nabla Q(w)$$

Метод простых итераций для оптимизации$Q(w)$ на 1-симплексе:

$$w := \mathrm{norm}(w \otimes \nabla Q(w))$$

Ирхин И. А., Воронцов К. В. Сходимость алгоритма аддитивной регуляризации тематических моделей. 2020.

Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. URSS, 2023. ISBN 978-5-9519-4345-3.

Проблемы представления знаний

Устранение «галлюцинаций» LLM как результата неполного, неточного, недостоверного представления фактов и знаний.

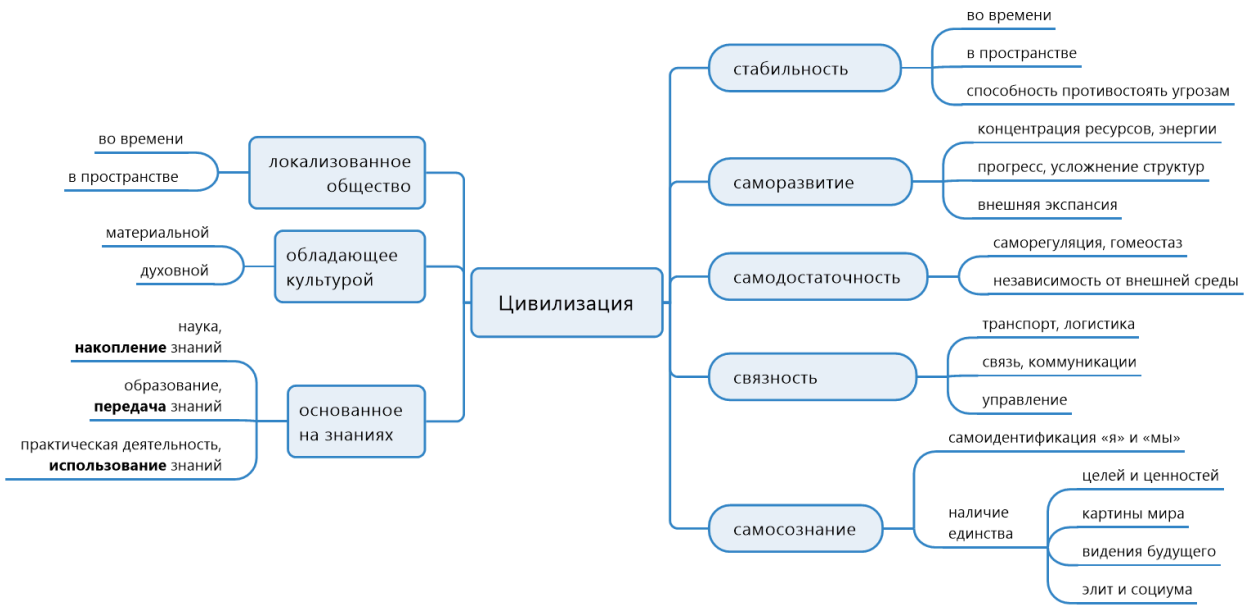
1. Обучить LLM навыкам корректного цитирования со ссылками.
2. Выделить в LLM в явном виде знания языков, знания о мире, правила рассуждений, правила коммуникации.
3. Обучить LLM человеческим приёмам работы со смыслом текста
   - разметка фрагментов текста (границы, теги, связи, комментарии)
4. Найти форму представления знаний, оптимальную (по критериям полноты, недвусмысленности, лаконичности, безопасности) для коммуникации между людьми и LLM
   - структурированный текст (например, в формате mind-map)

Задача синтеза структурированного текста

Автоматизировать синтез понятий из корпуса текстов, в виде некотрой структуры (аспекты, взаимосвязи, их важности), удобной как для компьютерной обработки, так и для восприятия человеком?

Пример. Структура mind-map для понятия «цивилизация»:

- легко прочитывается как линейный текст
- разбивает каждую идею на подыдеи
- отделяет важное от второстепенного
- допускает дальнейшее уточнение, детализацию



Фундаментальные технологические проблемы

1. Создание доверенных платформ для обучения моделей ИИ на отечественных программно-аппаратных средствах (Платформа-ГНС, PlatLib)
2. Создание бенчмарков для тестирования моделей ИИ на основе многокритериального подхода (Russian SuperGlue)
3. Интеграция LLM с отраслевыми решателями задач
4. Экология данных: проблема замусоривания Интернета фейковым и синтетическим контентом

Фундаментальные социо-гуманитарные, философские, организационные проблемы

1. Формирование LLM по данным на языках народов России
2. Создание технологий мониторинга информационного пространства

3. Каковы цели создания общего искусственного интеллекта (AGI)? Какова цель эволюции — развитие жизни или развитие разума? — ответы на эти вопросы формируют планы научных исследований и долгосрочные тренды развития

4. Создание открытых данных и бирж данных
5. Интенсификация (на порядок!) подготовки кадров в области ИИ

Выводы: задачи в области ИИ для обеспечения научно-технологического суверенитета РФ

- Развивать математические технологии доверенного, распределенного, федеративного обучения больших моделей по большим данным
- Развивать отечественные программно-аппаратные платформы
- Формировать конкурентоспособные LLM по большим доверенным данным на русском языке и языках народов России
- Создавать открытые данные и биржи данных
- Интенсифицировать (на порядок!) подготовку кадров в области ИИ