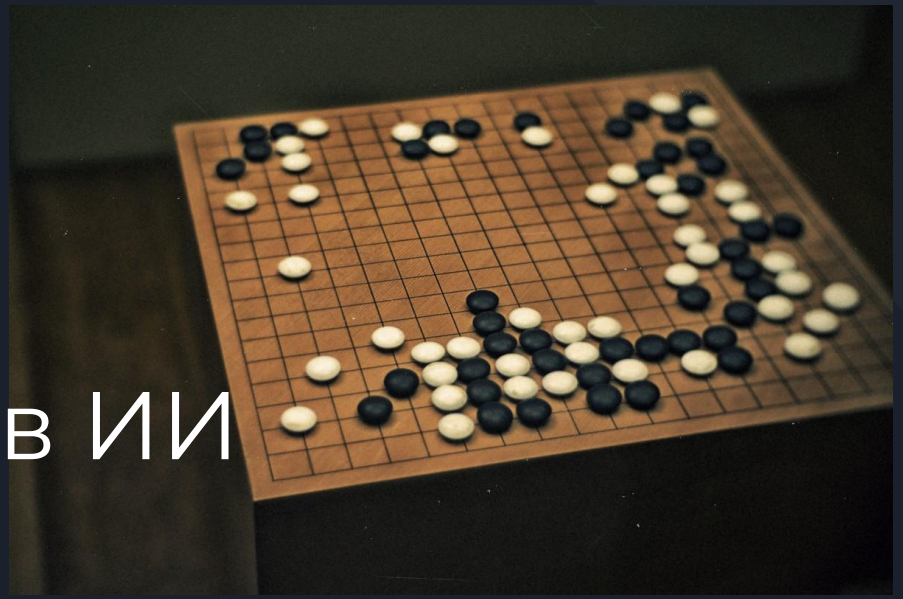
Введение в ИИ

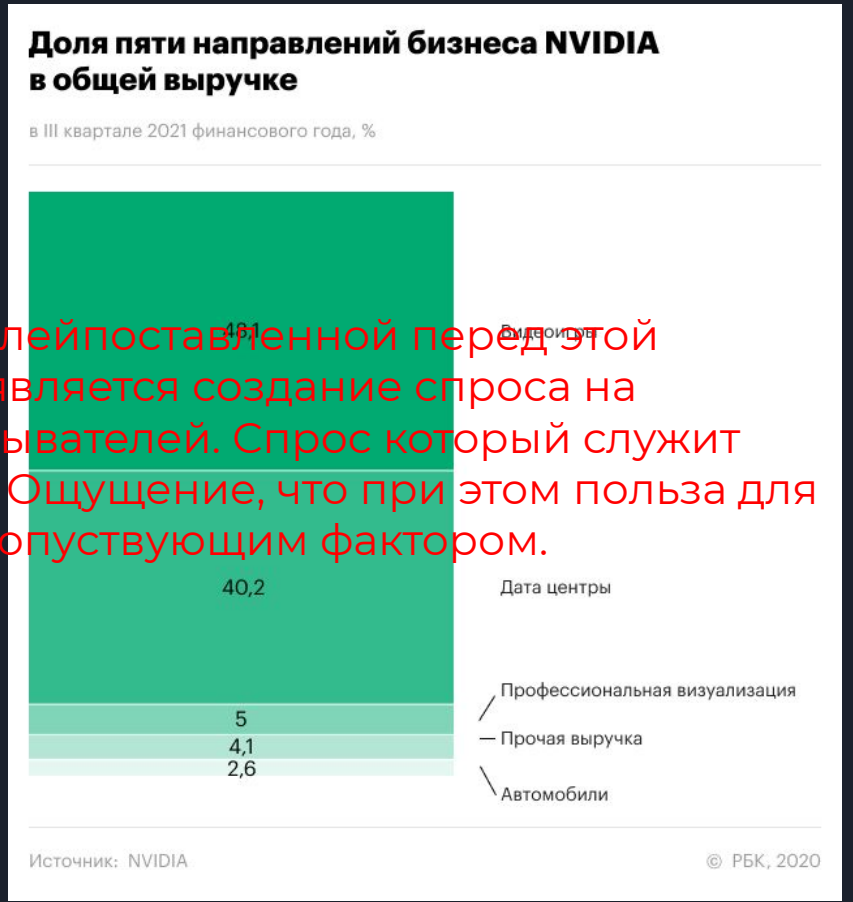


- Что из себя представляет ИИ?
- Области его применения
- Современные тренды, Deep Learning
- Примеры конкретных моделей
- Перспективы ближайшего будущего

О ИИ евангелистах

Быть техническим евангелистом трудно — нужно знать матчасть, поэтому основная масса IT-евангелистов — примызывающиеся к IT люди, которые имеют к нему лишь очень косвенное отношение. Особенно много их расплодилось в среде всяких скрам-аджайл-тусовок, где писать код не нужно, зато можно с умным видом молоть языком о процессах, эстимейтах и стори-пойнтах и тут же получать внимание публики. К слову, одна из причин, почему тунеядцы ломятся в спикеры и IT-евангелисты — непомерная жажда внимания и славы при минимуме усилий. В то время как действительно профессионалы молча делают свое дело и не светятся на публике без лишней надобности. (itpravda.ru)

Роль евангелистов в ИИ



Если соотносится с ПФУ, то одной из целейпоставленной перед этой публикой, объемлющим управлением является создание спроса на микросхемы и софт в среде рядовых обывателей. Спрос который служит для развития IT индустрии как таковой. Ощущение, что при этом польза для народного хозяйства является скорее сопуствующим фактором.

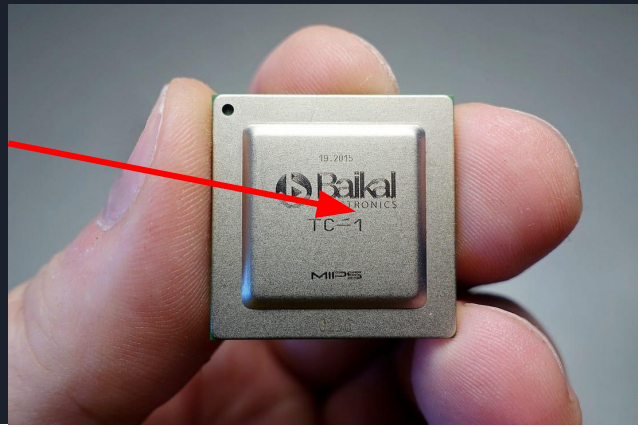
Основы ИТ в bleeding edge физике твердого тела

- На чем базируется успех современных ИТ компаний и ИИ как такового?
- Что значит свой процессор, если ли у России таковой?
- Градация корпораций ИТ индустрии:
  - *Социумы и культуры где все нижеперечисленное возможно (в основном это Запад, либо примкнувшие к нему страны)
  - Исследовательские институты и лаборатории как кузница кадров и методов создания кристаллов. (лаборатории США и Западной Европы)
  - Компании производящие степперы-литографы. (ASML, Nikkon, IBM?)
  - Компании строящие производственные линии. (Intel, Samsung)
  - Компании проектирующие кристаллы в соответствие с техническими возможностями производственных линий. (Intel, Nvidia, ARM, Samsung)
  - Компании производящие потребительский софт. (Microsoft, Google, Amazon, Facebook)

Обратная связь

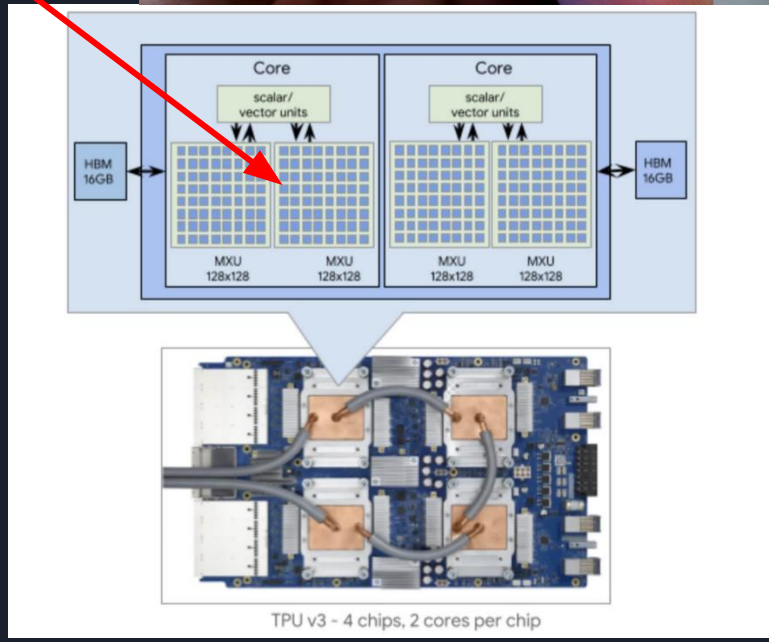
Российский процессор?

MIPS (сокращение от названия соответствующего проекта Стэнфордского университета англ. *Microprocessor without Interlocked Pipeline Stages*[1])



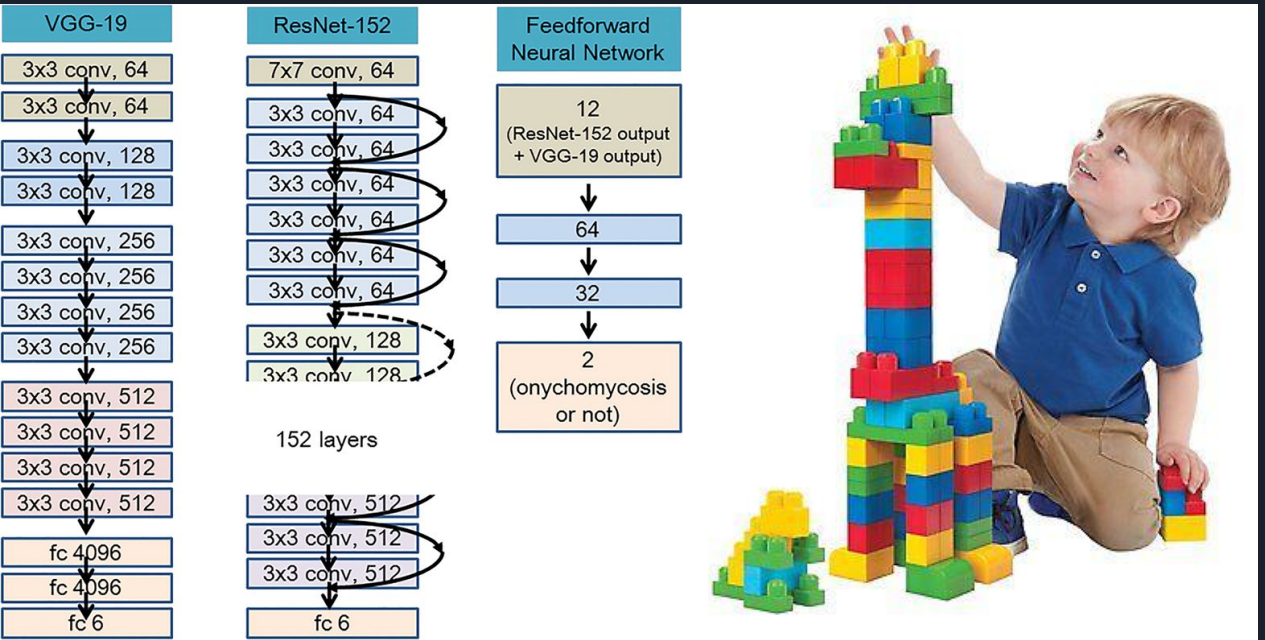
Универсальные
АЛУ

С точки зрения выше озвученной системной градации компании, проектирующие процессоры в РФ (Байкал, МЦСТ Эльбрус и.т.д), это детский сад - штаны на лямках. Современные вычислители, это прежде всего вычислители для ИИ. Ониимеют высоко регулярную, “простую архитектуру” по принципу конструктора, где в кристалле сосуществуют уже давно спроектированные, едва ли не open source процессора (как правило ARM) и однотипные структуры data шин и универсальных вычислителей чему пример Google TPU который был запроектирован едва ли не за полгода.



Deep Learning как лего конструктор

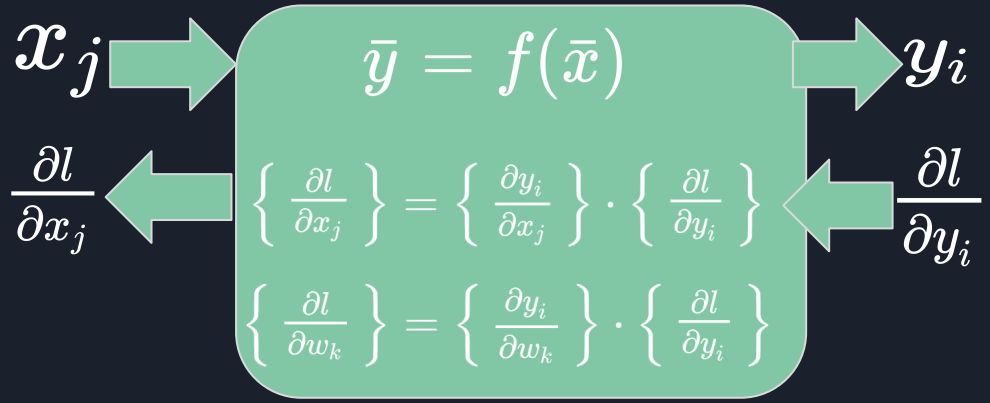
С учетом современных тенденций, современные ИИ системы это DL. А это оптимизированные фреймоорвки для обучения и скоринга (Pytorch, TensorFlow, Mxnet). Создание моделей в них подобно lego конструктору, где кубик за кубик собираешь нужную тебе архитектуру. Средства автоматизации и эту деятельность превратили в рутинную. Таким образом для получения посредственного, но работающего результата необходим минимум знаний, но для получения выдающегося требуется много труда.

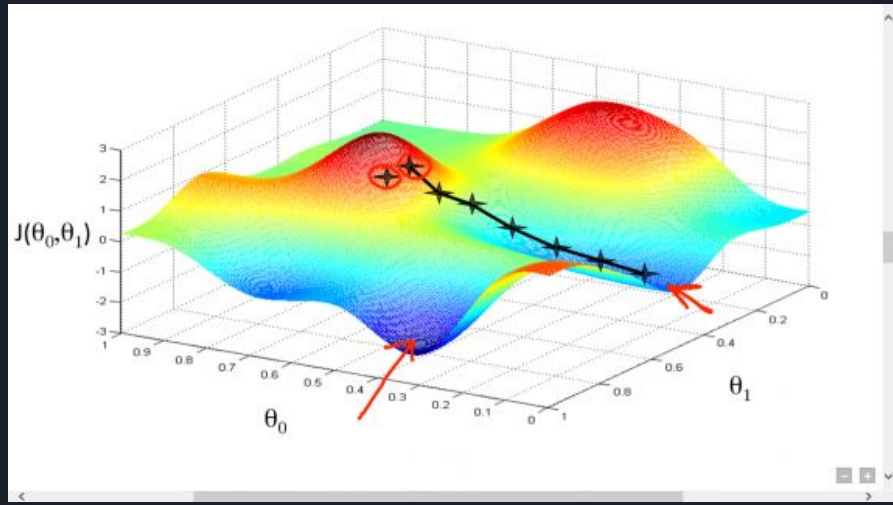


Поговорим о технике (backprop)

Рабочей лошадкой современных ИИ систем является backprop алгоритм, являющийся по сути цепным правилом.

$$J_x = \left\{ \frac{\partial y_i}{\partial x_j} \right\} \qquad J_w = \left\{ \frac{\partial y_i}{\partial w_j} \right\}$$



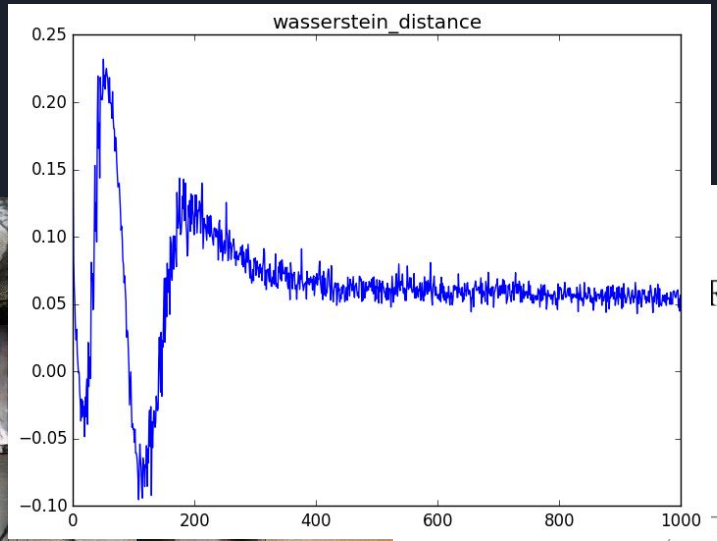


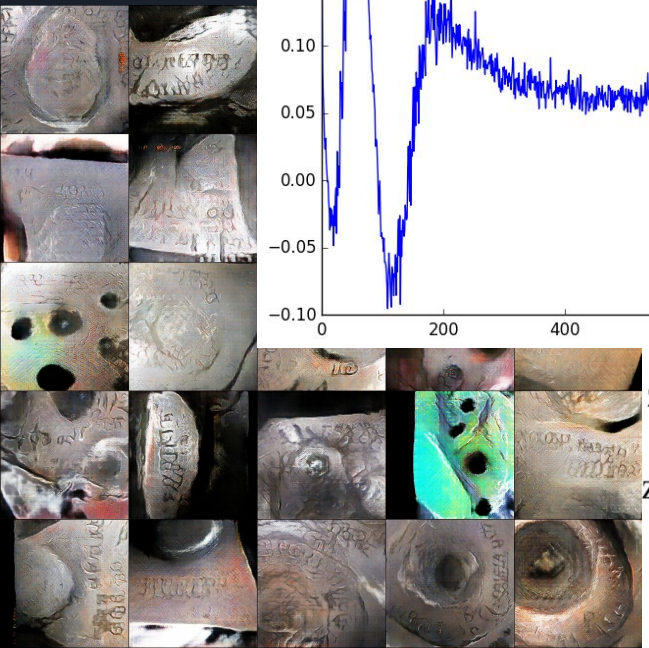
$$w_k = w_k - lr * \left\{ \frac{\partial l}{\partial w_k} \right\}$$

ИИ есть оптимизация критериев заданных извне (1)

$$L_D^{wgan} = -E_{x\sim p_d}[D(x)] + E_{x\sim p_g}[D(\hat{x})]$$

$$L_G^{wgan} = -E_{x\sim p_g}[D(\hat{x})]$$





$$\nabla_w[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))]$$

Real image x

z~N(0,1)
or
z~U(-1,1)

Generator

Critic

f

Loss

$$-\nabla_\theta \frac{1}{m}\sum_{i=1}^{\tilde{m}} f_w(g_\theta(z^{(i)}))$$

ИИ есть оптимизация критериев заданных извне (2)

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} (1 - \hat{Y}_{xyc})\alpha log(\hat{Y}_{xyc}) \; if \; Y_{xyc} = 1 \\ (1 - Y_{xyc})^{\beta} (\hat{Y}_{xyc})^{\alpha} log(1 - \hat{Y}_{xyc}) \end{cases}$$

$$L_off = \frac{1}{N} \sum_{p \subset P} \left| \hat{O}_{\tilde{p}} - \left(\frac{p}{R} - \tilde{p} \right) \right|$$

$$L_{size} = \frac{1}{N} \sum_{p_k \subset P} \left| S_{p_k} - \boldsymbol{s}_k \right|$$

$$L_{det} = L_k + \lambda_{size} L_{size} + \lambda_{off} L_{off}$$

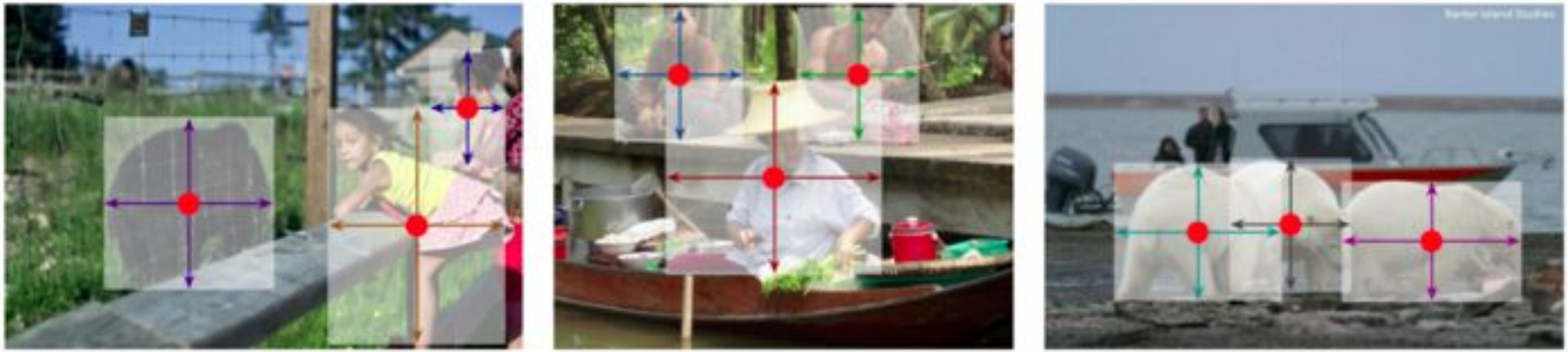


Figure 2: We model an object as the center point of its bounding box. The bounding box size and other object properties are inferred from the keypoint feature at the center. Best viewed in color.

Ещё пример модель языка (GPT-3)

https://theaisummer.com/self-attention/

OpenAI

GPT-3, an autoregressive language model with 175 billion parameters

$$P(\alpha_i|\alpha_{i-1}\dots\alpha_{i-n}) = f(\theta_{trans}, \alpha_{i-1}\dots\alpha_{i-n})$$

$$\overbrace{O_{m\times d_k}}^{\text{Output values}} = \overbrace{softmax\left(\frac{Q_{m\times d_k}K^T_{d_k\times n}}{\sqrt{d_k}}\right)}^{\text{Attention matrix } m\times n}\overbrace{V_{n\times d_k}}^{\text{Values}}$$

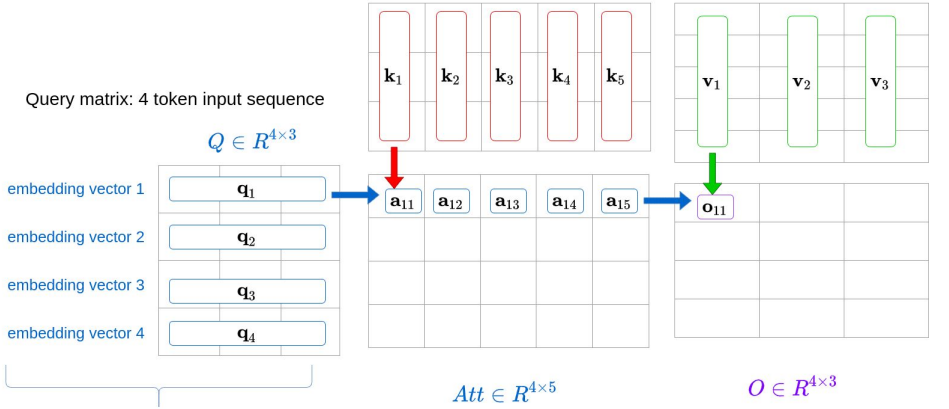
Keys, Values: 5 token sequence to associate with the input queries

$K^T \in R^{3\times 5}$

$V \in R^{5\times 3}$

Query matrix: 4 token input sequence

$Q \in R^{4\times 3}$



$Att \in R^{4\times 5}$

$O \in R^{4\times 3}$

Embeddings are processed in parallel with a simple matrix multiplication

Attention matrix
each dot product show the similarity between a query and key vector (softmax is applied row-wise)

The output weighted values: information is aggregated/routed

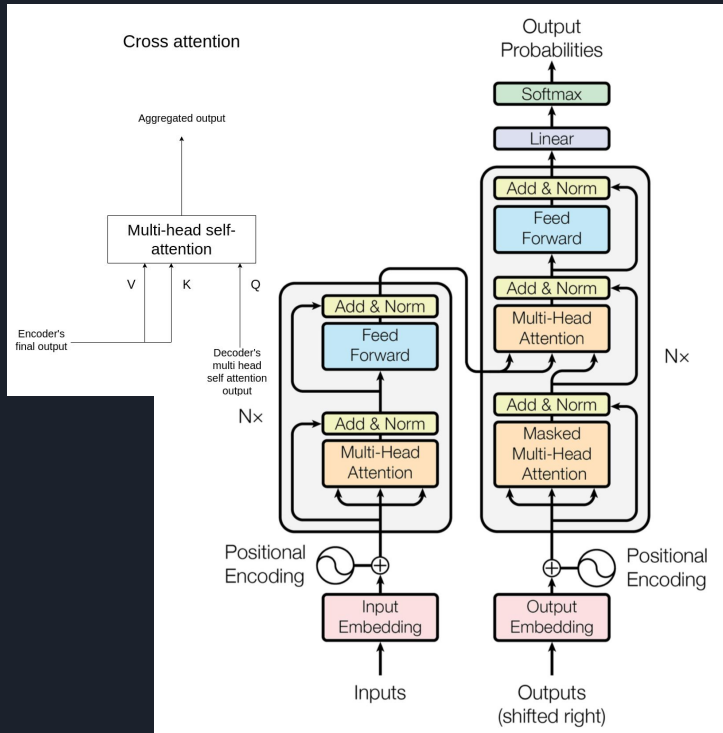
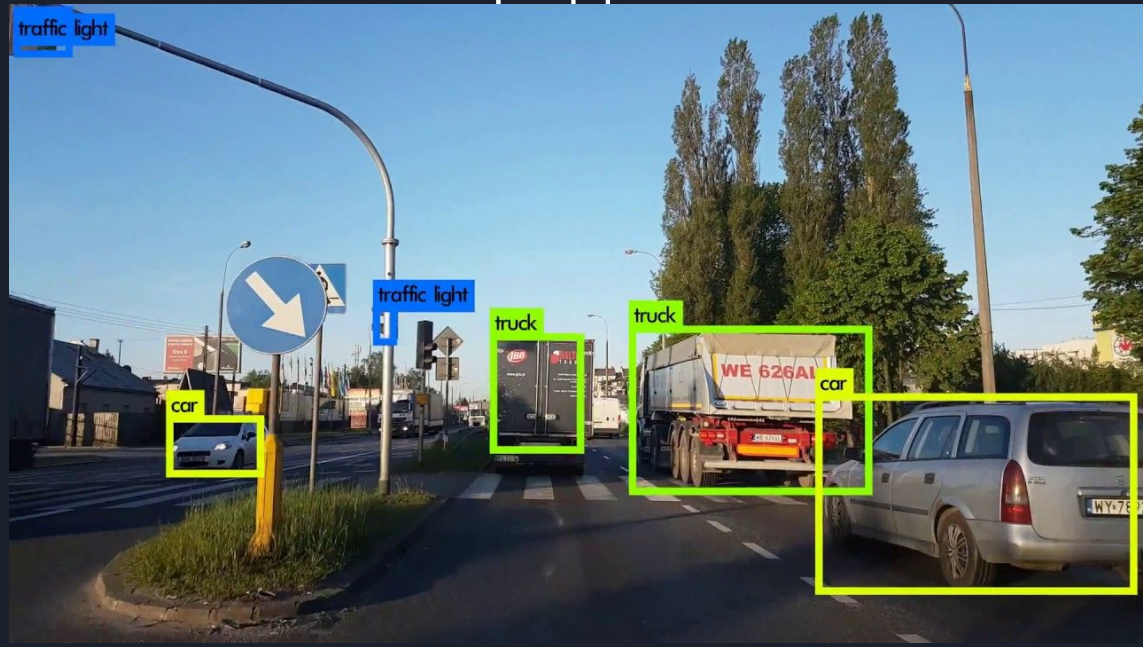


Figure 1: The Transformer - model architecture.

Сдвиг алгоритмической парадигмы

Подобный подход стал возможным, так как фундаментально было доказано и проверено практике, что любая функция представима в виде серии линейных матричных преобразований + нелинейность. Что позволило перестроится алгоритмистам для эффективного использования подобных TPU архитектур, которые по своему еще проще чем то, что делает Nvidia. Вследствие чего был смещен фундаментальный акцент с попытки ускорить существующие алгоритмы к перестройке самих алгоритмов, что в итоге предопределило отставание, а в будущем и вероятный крах Intel и традиционного подхода к вычислениям как такового. Мир имеет аналоговую природу, и если вы не в состоянии считать его в точности (а вы не в состоянии) то требование дальнейшей абсолютной строгости в обработке информации просто лишено смысла.



По сути современные ИИ вычислители моделируют аналоговую природу мира в его дискретизированном-цифровом виде, что имеет как свои удобства для отладки так и издержки, но в целом можно ожидать выхода на сцену вычислителей вообще аналоговой природы, как уже было в истории развития вычислительной техники.

Так что же есть ИИ?

ИИ - это набор подходов и методов оптимизации статистики, либо натуральной, либо сгенерированной в процессе игры.
Подход к ИИ как к некой логической схеме уже умер и только затуманивает суть дела.

ИИ - состоит из двух главных компонент лежащих в его основе:

1. ***Интерпретация*** придает некоторый смысл статистическим данным, множеству картинок, видеофайлов, звуковых дорожек, каких то статистик. Интерпретация это искусство.
2. ***Оптимизация***, как правило градиентный спуск с использованием цепного правила, если модель дифференцируема (99% случаев).

Суть ИИ

Задание критериев (того что получило название inductive human defined biases) поиска в двух аспектах

1. Объема параметрического (фазового как сказали бы физики) пространства
2. Меры “хорошести” найденного.

Но в целом, как уже было сказано суть машины неизменная это вычислительная мощь поиска того что нужно пилоту - сознанию им управляющим.

Прорыв Deep Learning раздвинул границы того что теперь как мы понимаем могут искать машины. В частности теперь ясно что машина может и музыку сочинять и картины писать.

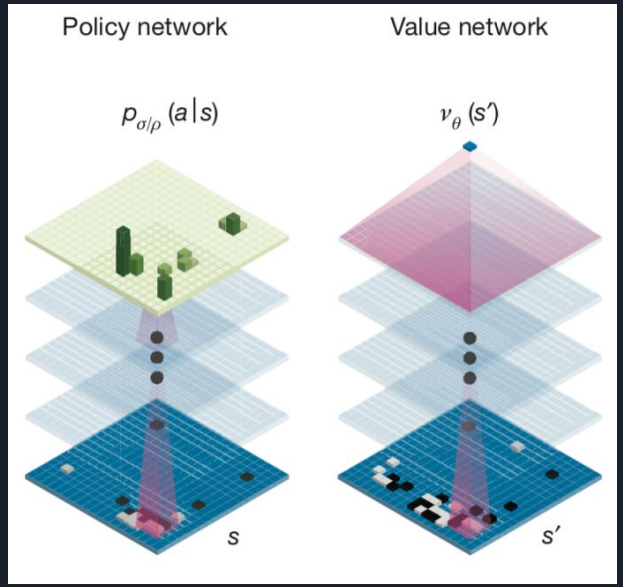
Поиск = оптимизация

Что может ИИ?

Все то же что может человек только ***гораздо лучше***:

- Писать картины
- Сочинять музыку
- Подделывать голос
- Создавать образы объектов
- Играть в ГО/шахматы/карты/стрелялки
- Быть собеседником
- Оптимизировать программы, писать вряд ли (проблема остановки)

Но ИИ не может создать новый критерий по которому будет создан другой ИИ



Филосовский аспект, можно ли сравнивать И с ИИ?

На заре развития ИИ, бытовало мнение что компьютер способен заменить левое, вычислительно полушарие человека, а потому лучше играет в шахматы, карты но не способен к творчеству. На текущий момент ясно что это не совсем так. Глубокие сети способны впитать всю визуальную культуру человечества, и в ряде случаев со значительным запасом превосходят возможности человеческой психики (его бессознательного) в интерпретации образов (данных в виде видео/аудио/и.т.д). Современные нейросети с успехом бьют человека в шахматы и даже Го, до сих пор считавшийся трудной задачей, с успехом справляются с проектированием конструкций с заданными параметрами (это более оптимизаторы даже, а не ИИ). По сему видимо стоит сделать вывод что ИИ подход в целом описывает работу бессознательных уровней человека. Однако сознания человека, будучи “пилотом” все же остается тайной. А ведь это именно тот уровень на котором и происходит интерпретация данных, уровень который наделяет их тем или иным смыслом. Хотя машина, которую создали сознания-пилоты программистов и в состоянии превзойти своих создателей вычислительной мощи, все же сам “пилот” остается за сценой.

Машина стала способный очень быстро перебирать пространство вариантов согласно заданному критерию и в этом смысле есть смычка с эзотерическими учениями утверждающими что наше бессознательное - это биокомпьютер

Машина способна творить и неплохо, что критерии красоты-хорошести всегда дается извне и в этом состоит фундаментальная разница между И и ИИ.

Перспективы ИИ

1. Фотоника проще, быстрее, дешевле.
2. Длина стадии конвейера равна длине волны
3. Низкое тепловыделение
4. Возможен квантовый параллелизм
5. Если удастся подключить биомашины к созданию микро и нано ИИ структур, то цена будет просто копеечной.
6. Вычислительная техника уже некогда была аналоговой, на новом витке спирали она станет гибридной.