Актуальные задачи и результаты в области компьютерного зрения и машинного обучения

Визильтер Юрий Валентинович, д.ф.-м.н., проф. РАН,
директор по направлению «Технологии искусственного интеллекта»,
начальник подразделения ФАУ «ГосНИИАС», *viz@gosniias.ru*

ITNT - 2022

VIII Международная конференция
«Информационные технологии и нанотехнологии»
(ИТНТ-2022)

Самара, 23.05.2022

Актуальные задачи и результаты в CV и ML (2020-21)
(обзор по материалам OpenTalks.AI/2022 и не только…)

The subjective overview of most interesting approaches, results and trends in CV and ML 2020-2021

- ***CNN → Transformers***
  - ***Transformers in NLP,…***
  - ***Transformers for Vision***
- ***Reinforcement Learning***
  - ***Decision Transformer***
  - ***Open-Ended Learning***
- ***Data Generation: Diffusion Models***
- ***Аппаратные средства для ИИ. Сравнение производительности***

2011-2016, 2017-2020+: две волны технологической революции в ИИ

Ранний период развития ИИ: созданы основные методы и подходы…

«Кривая хайпа» для ИИ

Завышенные ожидания

2000
«Зима ИИ»
Прогноз: 2040+

Разочарование

Рост интереса

Алгоритмы ИИ сильно проигрывают людям

2016-17
- **Глубокие соревнующиеся сети**
- **ГО с подкреплением**
- **ГО с использованием моделей, баз знаний и логического вывода**
- **Автоматическое обучение ГНС**

Вторая волна революции ГНС

Все задачи ИИ (НИР)

Первая волна революции ГНС

Распознавание (НИР->ОКР)

Функциональный («слабый») ИИ

Прогноз: 2020+

- **Появление глубоких нейронных сетей (ГНС)**
- **Решение задач компьютерного зрения, обработки сигналов и анализа больших данных на уровне человека и выше (superhuman)**

2011-12

Close future trends?

?

Сегодня доклад о том, как и куда все это развивается после 2020 г.

…2011+: в области ИИ началась технологическая революция ГНС

Глубокие нейронные сети

(краткое напоминание)

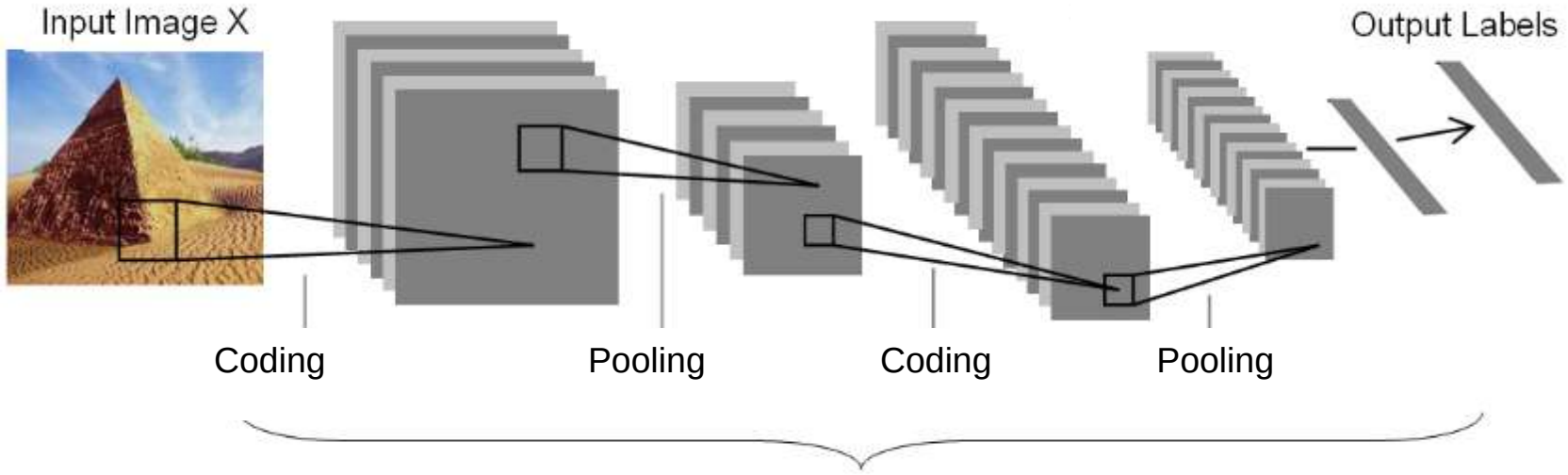
Глубокие конволюционные нейронные сети – новое поколение алгоритмов обнаружения и распознавания объектов на изображениях

2011: First Superhuman Visual Pattern Recognition

http://people.idsia.ch/~juergen/deeplearning.html

Глубокие сверточные сети: Почему это так здорово сработало?

Сверточные сети, глубокое обучение – это современная реинкарнация персептронов, объединяющая идеи *компьютерного зрения* и *обучения нейросетей*



Архитектура “Coding + Pooling” = “Кодирование + Объединение ”

- Иерархическое выделение признаков при обучении и/или самообучении на сверхбольших выборках изображений
- Много десятков уровней (кодирование + объединение данных)
- Инвариантность к сдвигу изображения (линейные фильтры)
- На верхних уровнях – собственно классификатор (полносвязный)

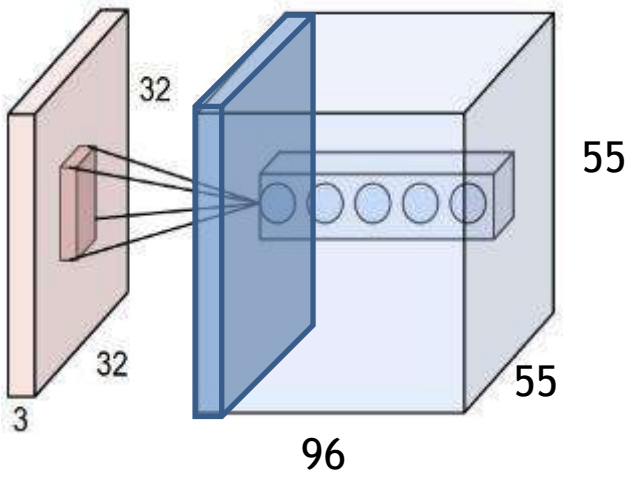
Сверточный нейрон это обучаемый линейный фильтр (веса = элементы маски фильтра) + нелинейность

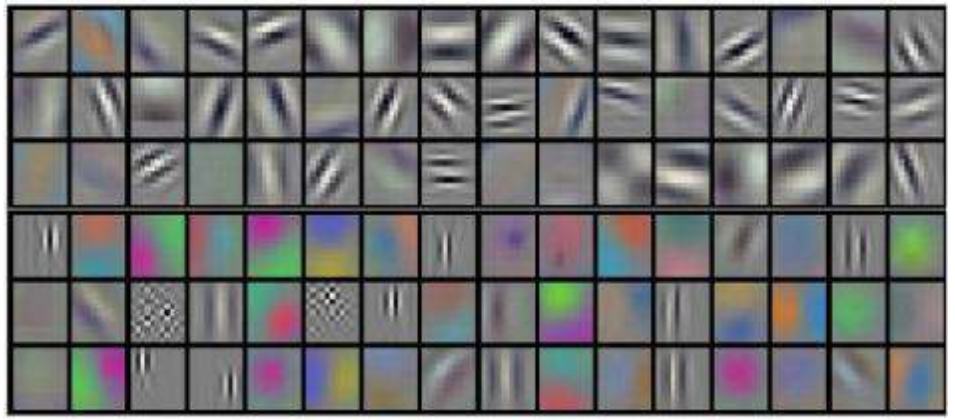
Идеи Deep Learning

(Kai Yu, 2012)

Математическая модель: CNN как эволюция расслоения базового многообразия

Качественная модель: адаптивный сверточный фильтр



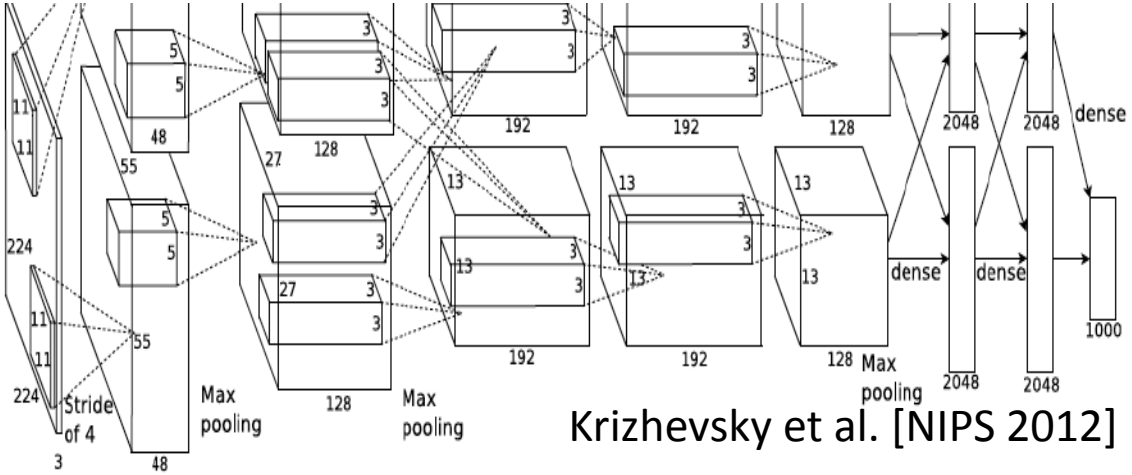


Всего 55*55*96 нейронов т.е. 290400 нейронов.

Если каждый нейрон связан с окрестностью 11x11 то всего 105,705,600 связей.

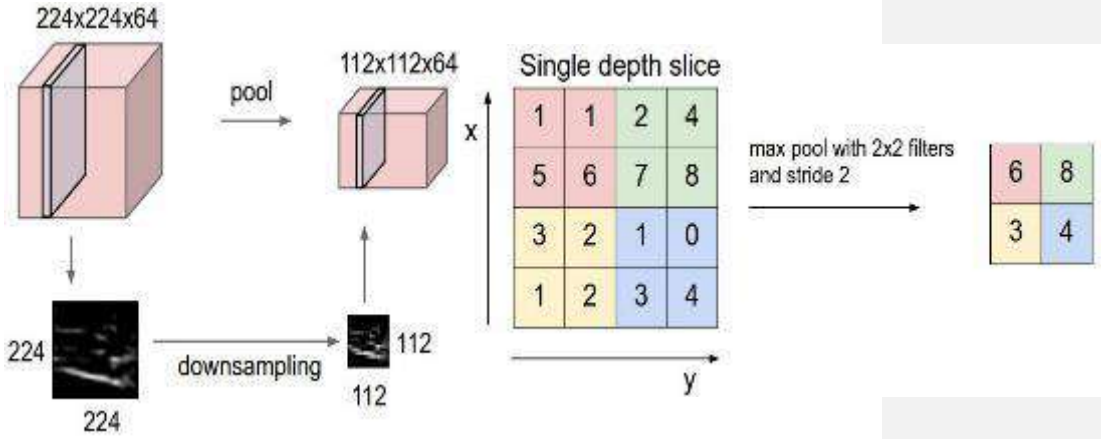
Все нейроны в пределах одного слоя глубины имеют одинаковые параметры (веса синапсов и пороги)

В данном случае есть только 96 уникальных наборов параметров
Каждый такой набор называется *фильтром* или *ядром*



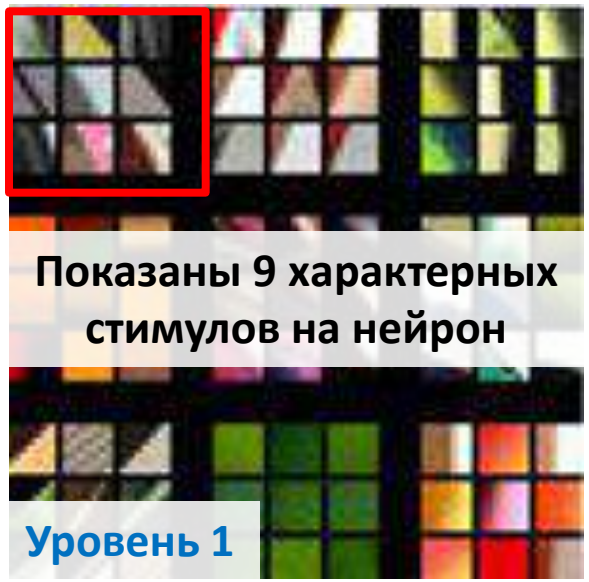
Krizhevsky et al. [NIPS 2012]

7 hidden layers, 650,000 neurons, 60,000,000 parameters



Слой пространственного объединения нужен для снижения размерности и уменьшения точности пространственной локализации признаков

Качественная модель: абстрагирование по пирамиде

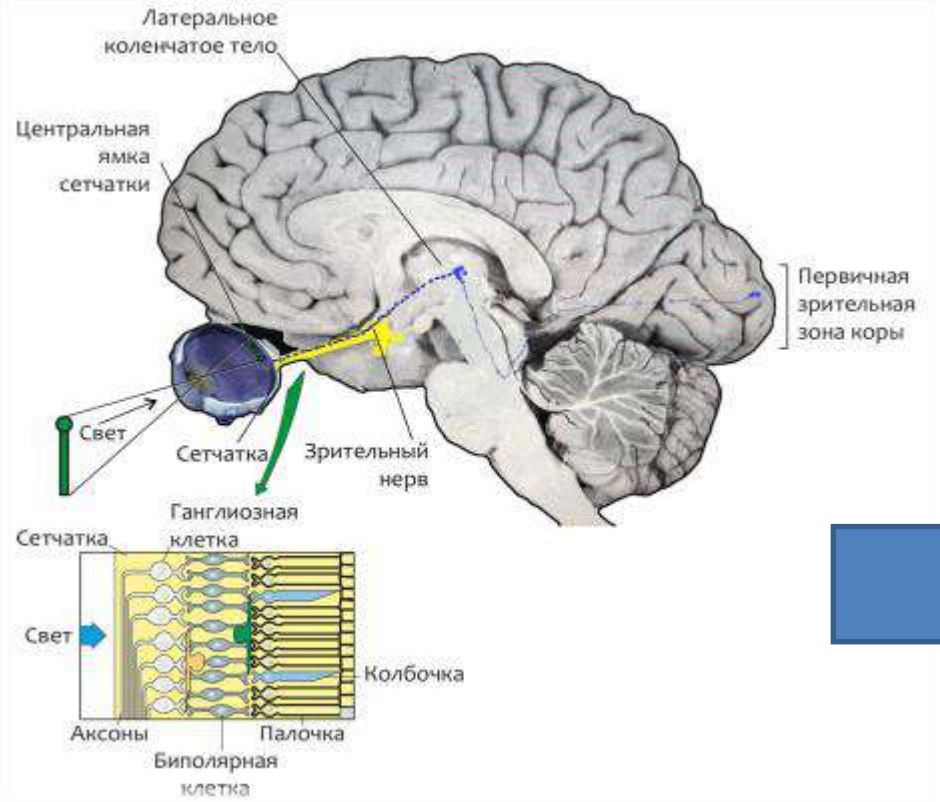


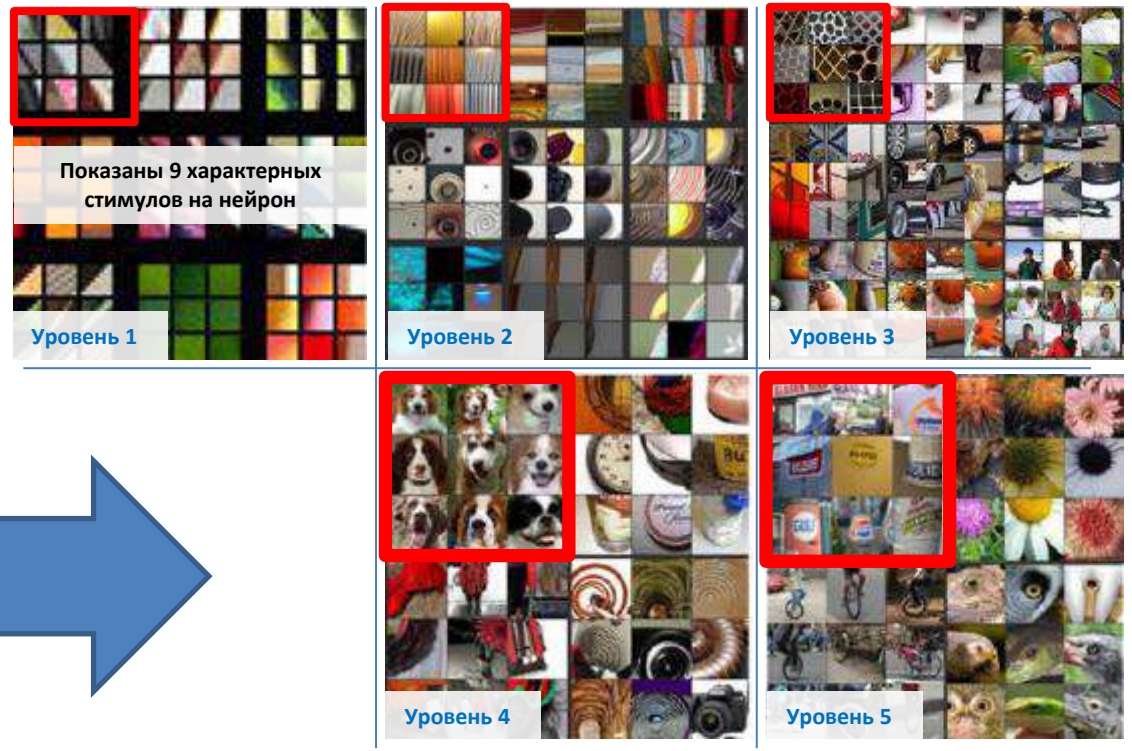
Какие элементы изображения распознают нейроны разных уровней:
чем выше слой сети, тем выше уровень абстракции

Visualizing and Understanding Convolutional Networks, Matthew Zeiler and Rob Fergus, ECCV'14

CV+NN = CNN

В качественном плане имеется значительное сходство между иерархическим пространством признаков ГНС и работой зрительной коры человека…





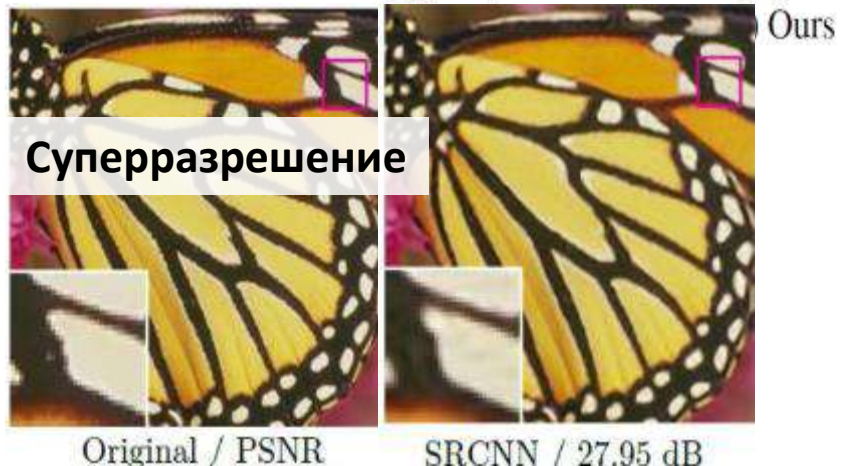
«**Зрительная кора** расположена в затылке. Первичную зрительную информацию (от сетчатки через реле таламуса) получают колонки самой задней ее части – эта область называется соответственно первичной зрительной корой. Нейроны ее колонок реагируют на самые элементарные, самые конкретные зрительные стимулы... С продвижением от затылка в сторону лба колонки начинают реагировать на все более и более абстрактные свойства зрительного объекта. За первичной зрительной корой следует вторичная зрительная кора, за ней третичная и так далее. Нейроны, реагирующие на точки и линии, сменяются нейронами, реагирующими на углы разной ориентации. Им на смену приходят нейроны, реагирующие на формы и предметы, и наконец, на категории предметов: дом, лицо, инструмент. Менее абстрактные репрезентации плавно перетекают в более абстрактные».

Н. Кукушкин «Хлопок одной ладонью»

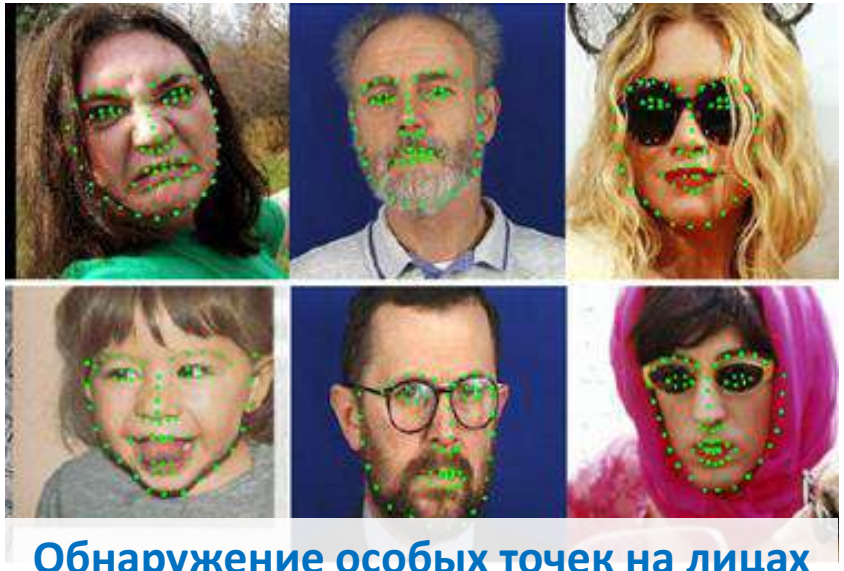
2015-16: CNN решают все задачи компьютерного зрения и царят до 2020…

Фильтрация шумов

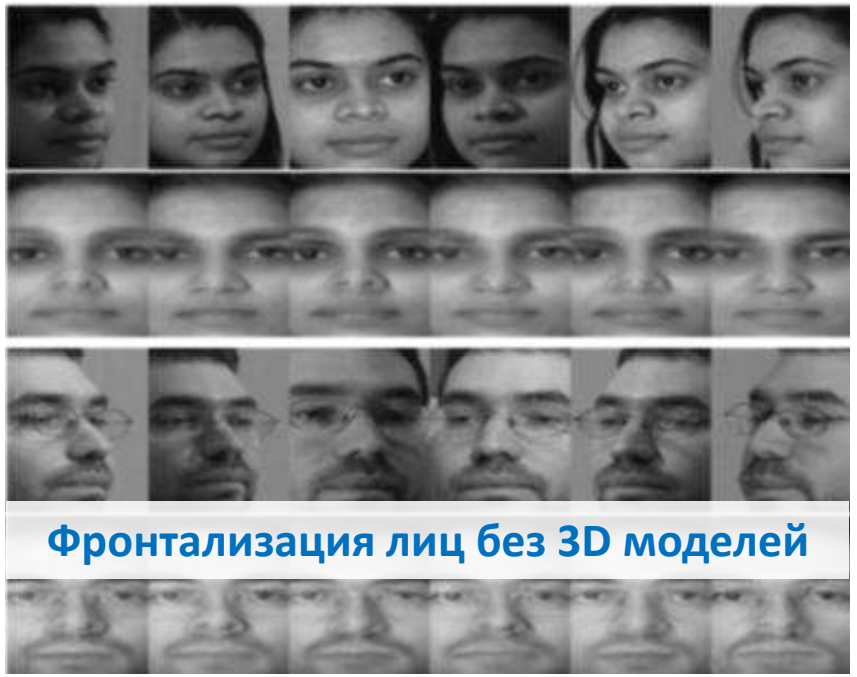
Удаление смаза



Суперразрешение



Обнаружение особых точек на лицах

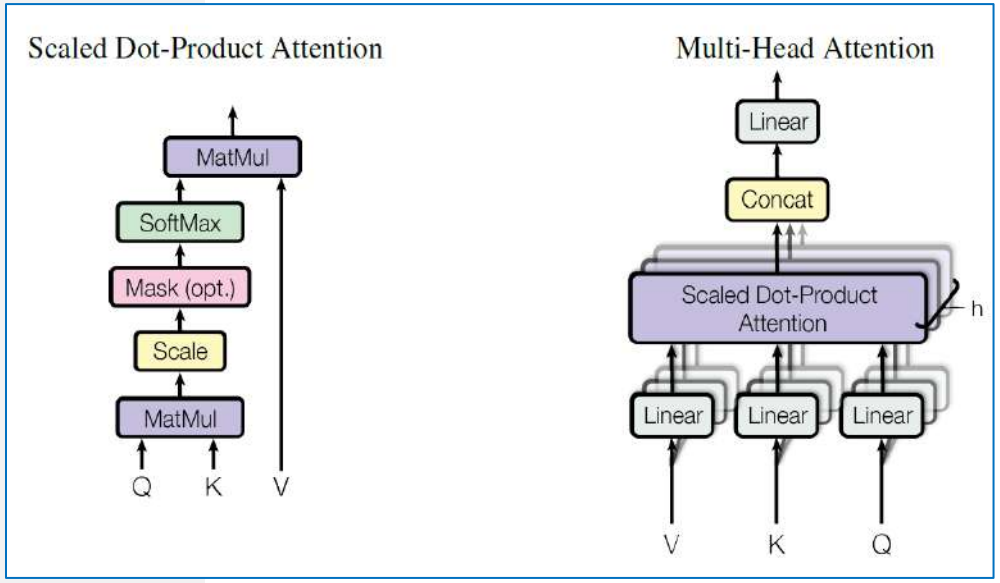


Фронтализация лиц без 3D моделей

Революция трансформеров

(появление трансформеров в области NLP, 2017-2019)

Attention Is All You Need (2017)



$$Q = XW^Q, \quad K = YW^K, \quad V = YW^V$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

Q - Query. Множество векторов, для которых мы хотим посчитать attention.
K - Key. Множество векторов, с которыми мы ходим посчитать attention.
V - Value. Обычно совпадает с K.
Вычисляется скалярное произведение Q со всеми K, от результата берется softmax и с полученными весами перевзвешивается V.

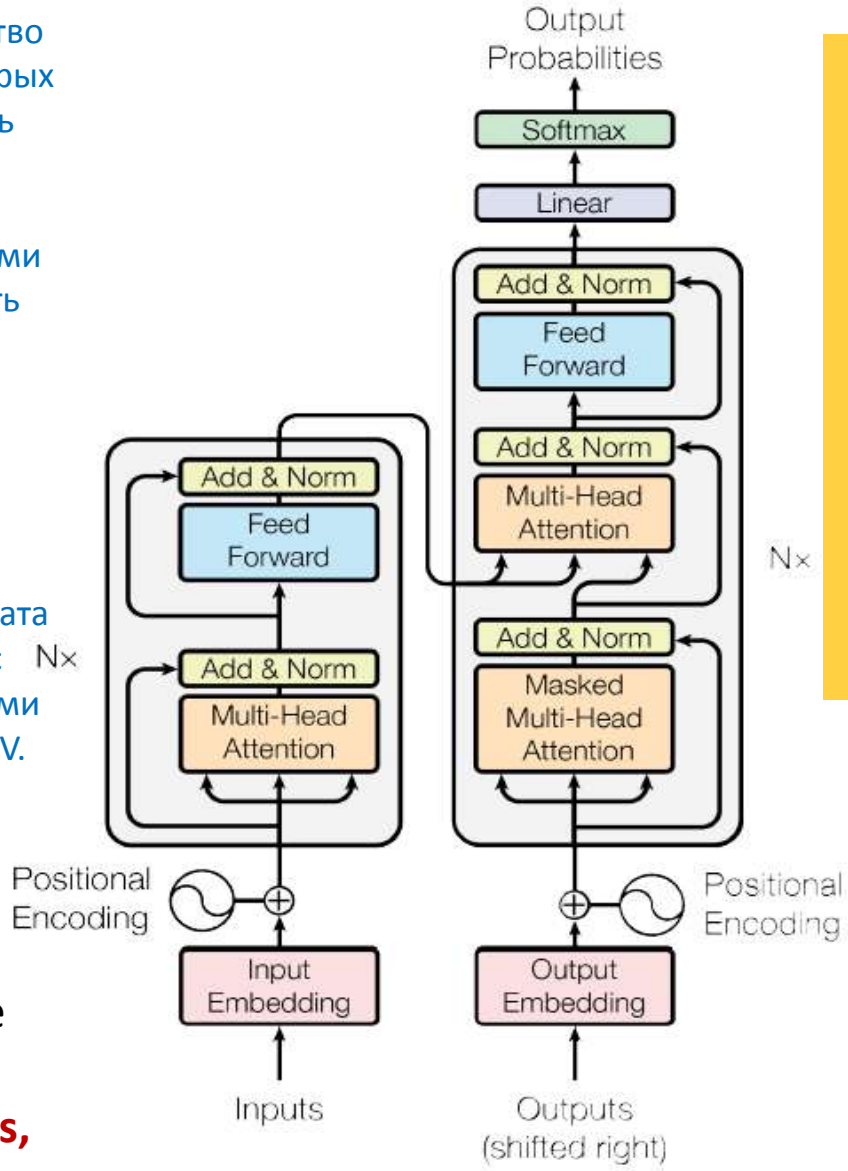


Figure 1: The Transformer - model architecture.

Управление вниманием (адаптивные парные веса) *оказались ключом к решению самых сложных задач!*

An attention function can be described as **mapping a Query and a set of Key-Value pairs to an output**, where the query, keys, values, and output are all vectors.

In the area of Natural Language Processing where **Transformer** style models have become state of the art on many tasks**. Motivation: Learning long-range dependencies is a key challenge for conv layers, which is important for symbolic sequences.**

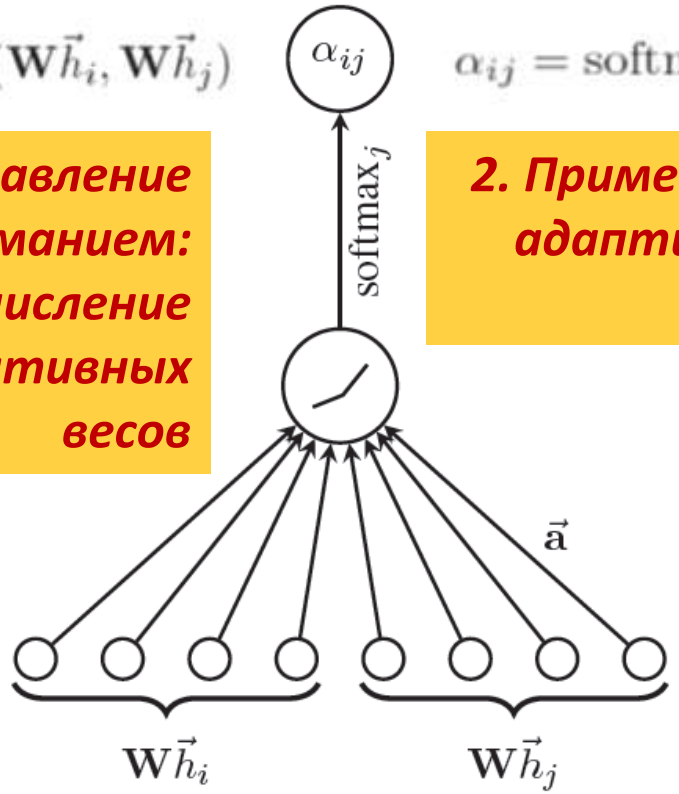
Attention Is All You Need (Ashish Vaswani et.al., 2017)

Attention как нейрон с адаптивными весами

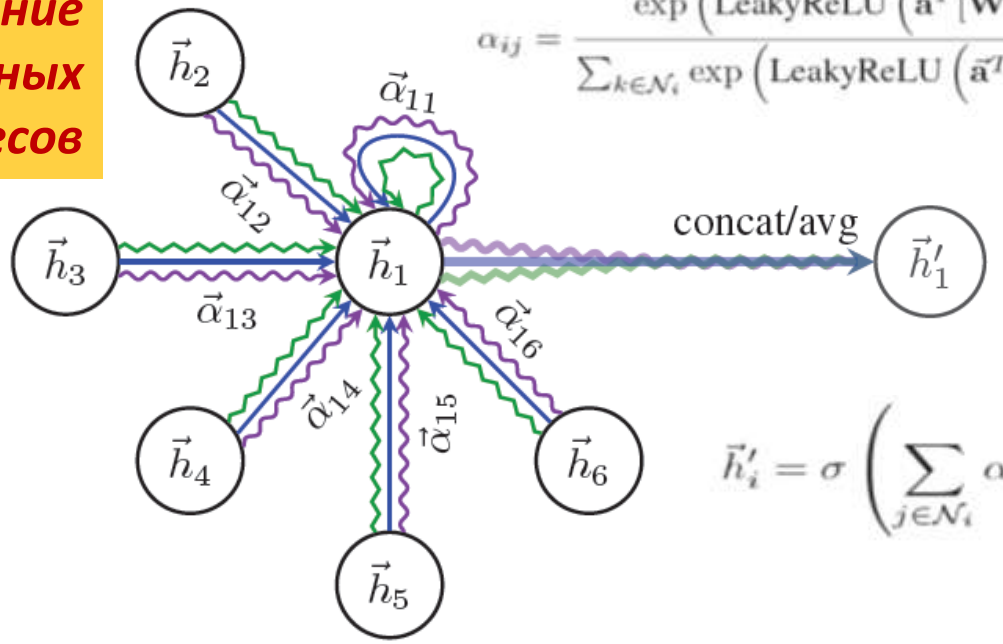
$$e_{ij} = a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j) \qquad \alpha_{ij} = \mathrm{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}_i}\exp(e_{ik})}.$$

Управление вниманием:
1. Вычисление адаптивных весов

2. Применение адаптивных весов



$$\alpha_{ij} = \frac{\exp\left(\mathrm{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_j]\right)\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\mathrm{LeakyReLU}\left(\vec{\mathbf{a}}^T[\mathbf{W}\vec{h}_i \| \mathbf{W}\vec{h}_k]\right)\right)}$$



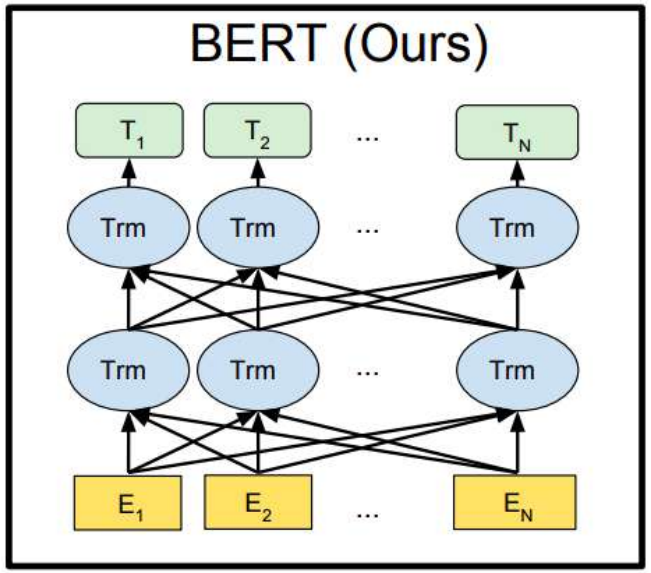
$$\vec{h}'_i = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\mathbf{W}\vec{h}_j\right)$$

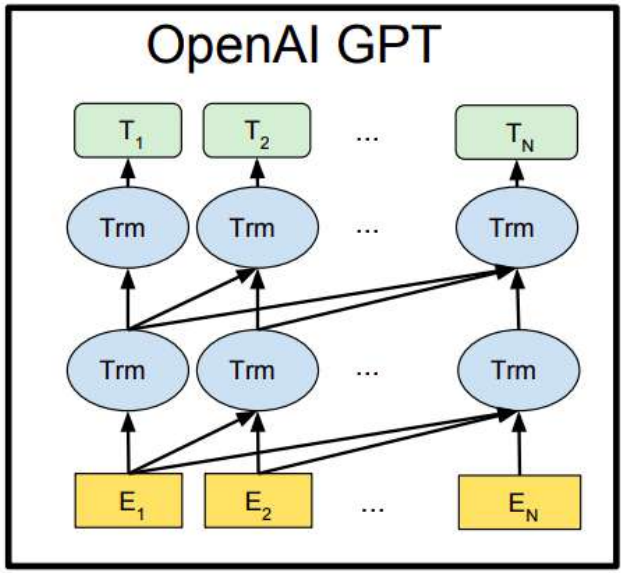
Figure 1: **Left:** The attention mechanism $a(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j)$ employed by our model, parametrized by a weight vector $\vec{\mathbf{a}} \in \mathbb{R}^{2F'}$, applying a LeakyReLU activation. **Right:** An illustration of multi-head attention (with $K = 3$ heads) by node 1 on its neighborhood. Different arrow styles and colors denote independent attention computations. The aggregated features from each head are concatenated or averaged to obtain $\vec{h}'_1$.

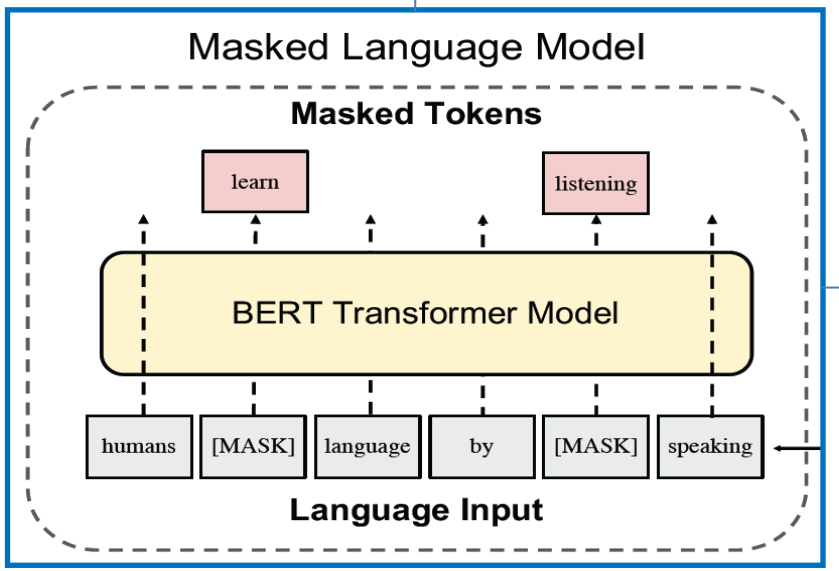
Graph Attention Networks (Petar Velickovic et.al., 2018)

Трансформеры и BERT (2018-19)

Обучение в стиле BERT!







Bidirectional Encoder Representations from Transformers (BERT)

BERT — это такой «трансформер», у которого увеличили число и размеры слоев, убрали декодирующую часть и научили **смотреть на контекст в обе стороны**. Внимание простого трансформера всегда направлено на токены слева от данного (слова справа заменяются особым словом [MASK], что обнуляет вес внимания). BERT же «маскирует» только то, что нужно предсказать, значит, внимание направлено на все токены входной последовательности — и слева, и справа.

https://sysblok.ru/knowhow/kak-ustroena-nejroset-bert-ot-google/

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, 2019)

BERT предобучается на «маскированной языковой модели». Нужно предсказать слово не в конце предложения, а где-то посередине: не «человек пошел в магазин за ???», а «человек пошел в ??? за молоком». **Модель называется маскированной, поскольку искомый токен заменяется токеном [MASK].**

OpenAI GPT-3: “may be the biggest thing since bitcoin”

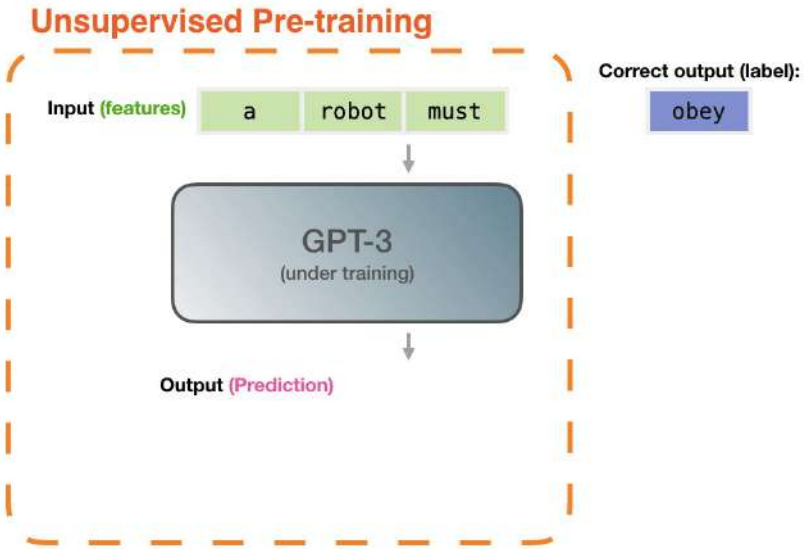
Параметры сети:

Размер сети 175 млрд. параметров(350 Гб fp16)
Контекстное окно: 2048 токена
96 трасформеров
96 self attention head
Размер батча: 3.2 млн

База данных:

The CommonCrawl data - 45 Тб
Прочие БД – 40%

Результаты сравнимые со state-of-the-art в 42 задачах NLP



Ключевая идея: обучение на огромной выборке на продолжение фразы (без учителя)

Суперкомпьютер:

10000 GPU Tesla V100
400 Гбит/сек
285 тыс. CPU ядер
Pytorch

Требует 355 лет и 5 млн. $ при аренде GPU Cloud

Архитектуры:

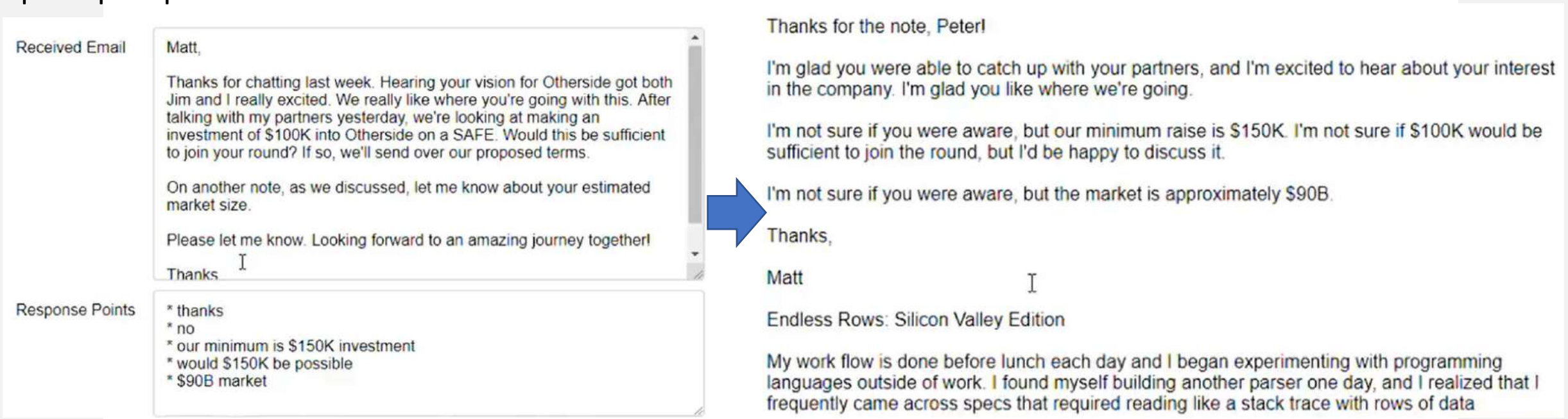
Модификация GPT-2
BERT
Transformer

- GPT-3 может отвечать на вопросы на естественном языке
- GPT-3 умеет переводить с любого языка, отвечать на вопросы, писать статьи и т.д. вплоть до написания исходного кода программ

ХИТ ИИ 2020!

OpenAI GPT-3: “may be the biggest thing since bitcoin”

Примеры приложений:



Автоматическое написание писем по нескольким тезисам

- Общение с ИИ на естественном языке
- Универсальная модель
- Команды для РТК/ИИ
- Построение моделей по уставу/правилам

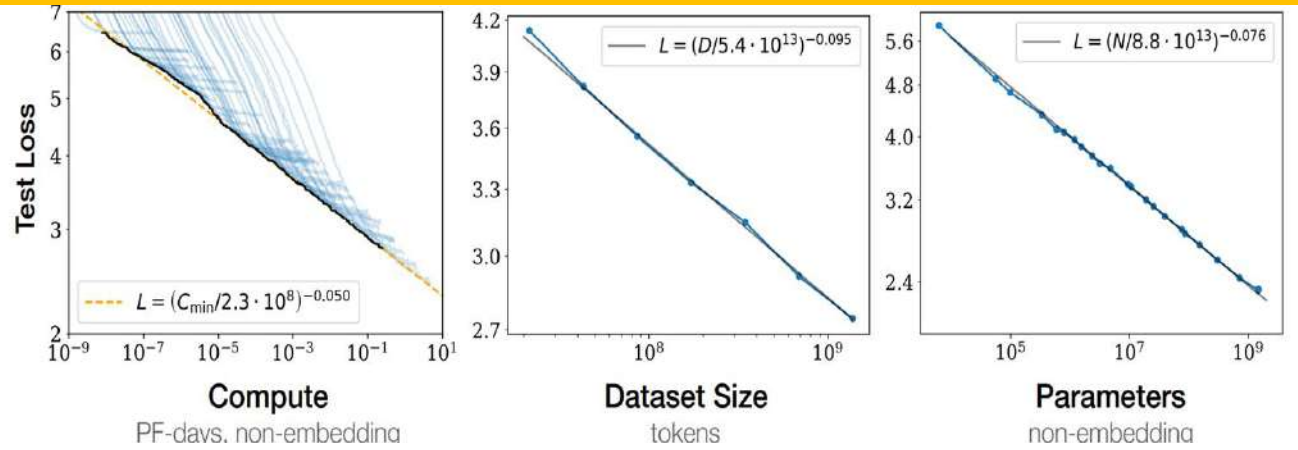
Революция трансформеров

(развитие трансформеров в области NLP, 2020-2021)

Large models in 2021/22

Один из главных трендов: модели становятся все больше

- (English) **GPT-Neo** (2.7B), **GPT-J** (6B), **Jurassic-1** (7.5B/178B), **Gopher** (280B), **GPT-NeoX** (20B), **MT NLG** (530B), …
- (Russian) **ruGPT-3** (13B)
- (Chinese) **CPM-2** (11B/198B* - MoE), **M6** (10B/100B), **Wu Dao 2.0** (1.75T*), **PangGu-α** (2.6B/13B/207B), **ERNIE 3.0** (10B) **Yuan 1.0** (245B), **ERNIE 3.0 Titan** (260B)
- (Korean) **HyperCLOVA** (204B)
- (Code) **OpenAI Codex** (12B), **Google's** (up to 137B)
- **ByT5** (up to 12.9B)
- **XLM-R XL/XXL** (3.5B/10.7B), **T-ULRv5 XL** (2.2B)
- **DeBERTa v1/v2** (1.5B), **v3** (300M)
- **Switch-C** (1.6T*), **GLaM** (1.2T*), **MoE LM** by Meta (1.1T*)
- **DALL·E** (12B)



"Scaling Laws for Neural Language Models"
https://arxiv.org/abs/2001.08361

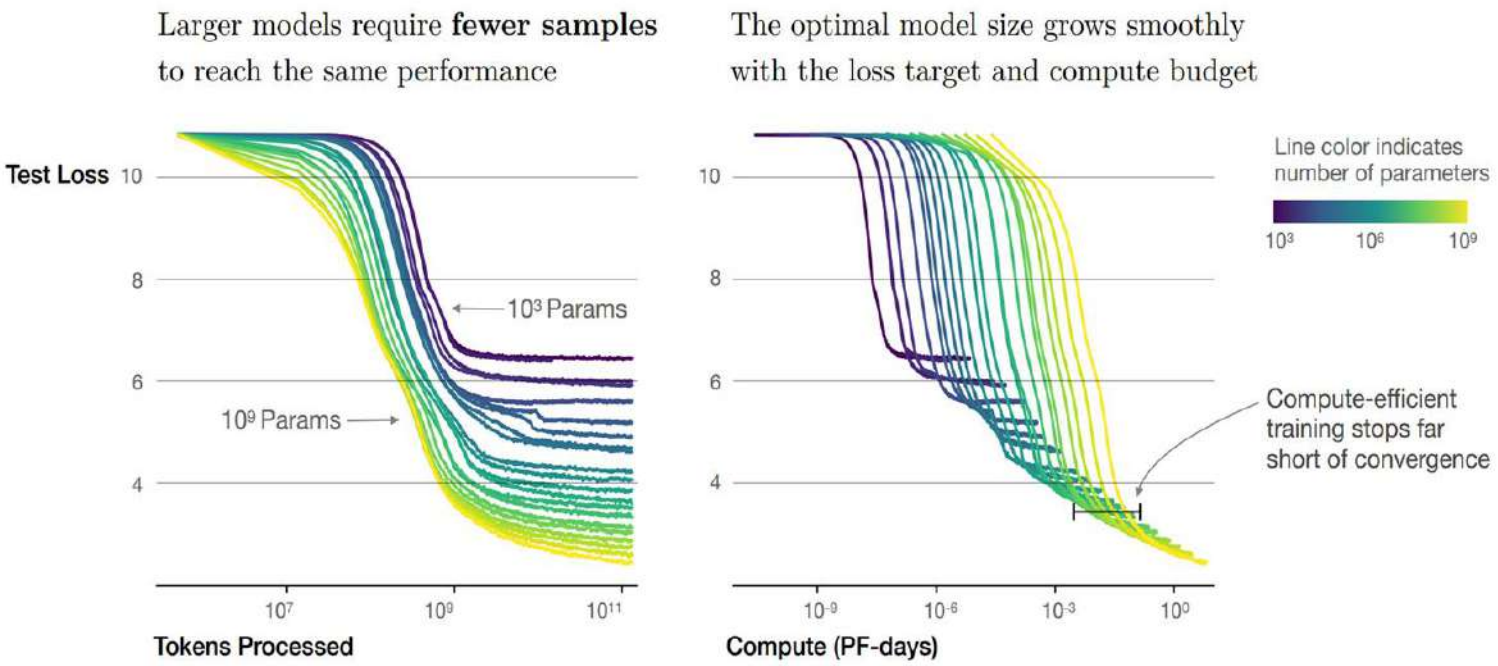
- «Мы исследовали **эмпирические законы масштабирования для производительности языковой модели** при кросс-энтропийной функции потерь.
- **Потери масштабируются по степенному закону в зависимости от размера модели, размера набора данных и объема вычислений при обучении**. При этом некоторые зависимости охватывают более семи порядков.
- Другие **архитектурные детали, такие как ширина или глубина сети, оказывают минимальное влияние в широком диапазоне масштабов (!)**.

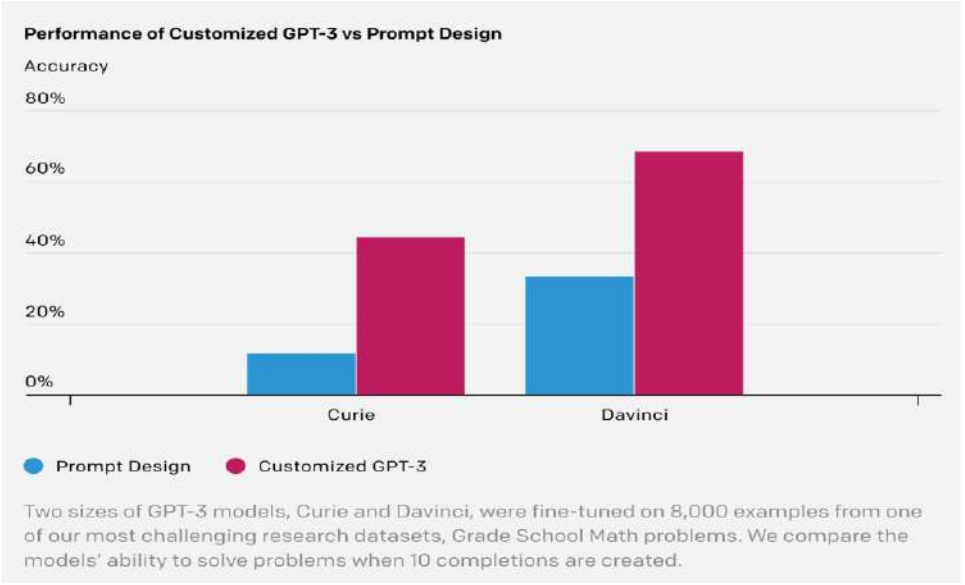
Идея слайда: Grigory Sapunov, NLP in 2021, OpenTalks.AI / 2022

Large models in 2021/22

Один из главных трендов: модели становятся все больше

Впрочем, для их обучения далеко не всегда нужны гигантские базы..





GPT fine-tuning on 8000 samples

"Scaling Laws for Neural Language Models"
https://arxiv.org/abs/2001.08361

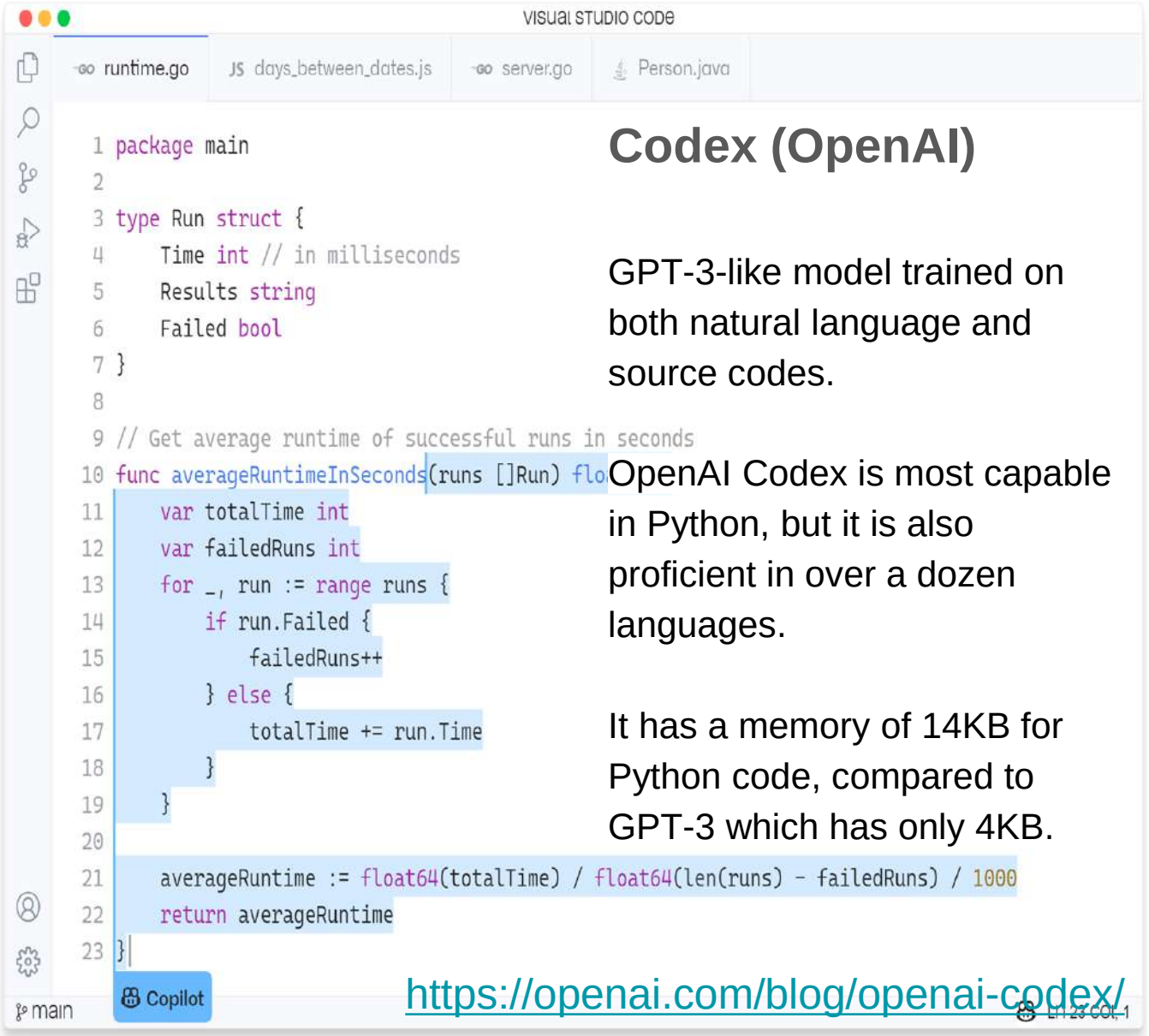
.

https://openai.com/blog/customized-gpt-3/

Простые уравнения управляют зависимостью переобучения от размера модели/набора данных и зависимостью скорости обучения от размера модели. **Более крупные модели значительно более эффективны**, поэтому обучение с оптимальной вычислительной эффективностью включает **обучение очень больших моделей на относительно скромном количестве данных с остановкой значительно раньше, чем будет достигнуто насыщение**».

Идея слайда: Grigory Sapunov, NLP in 2021, OpenTalks.AI / 2022

Приложения: Автоматическое программирование



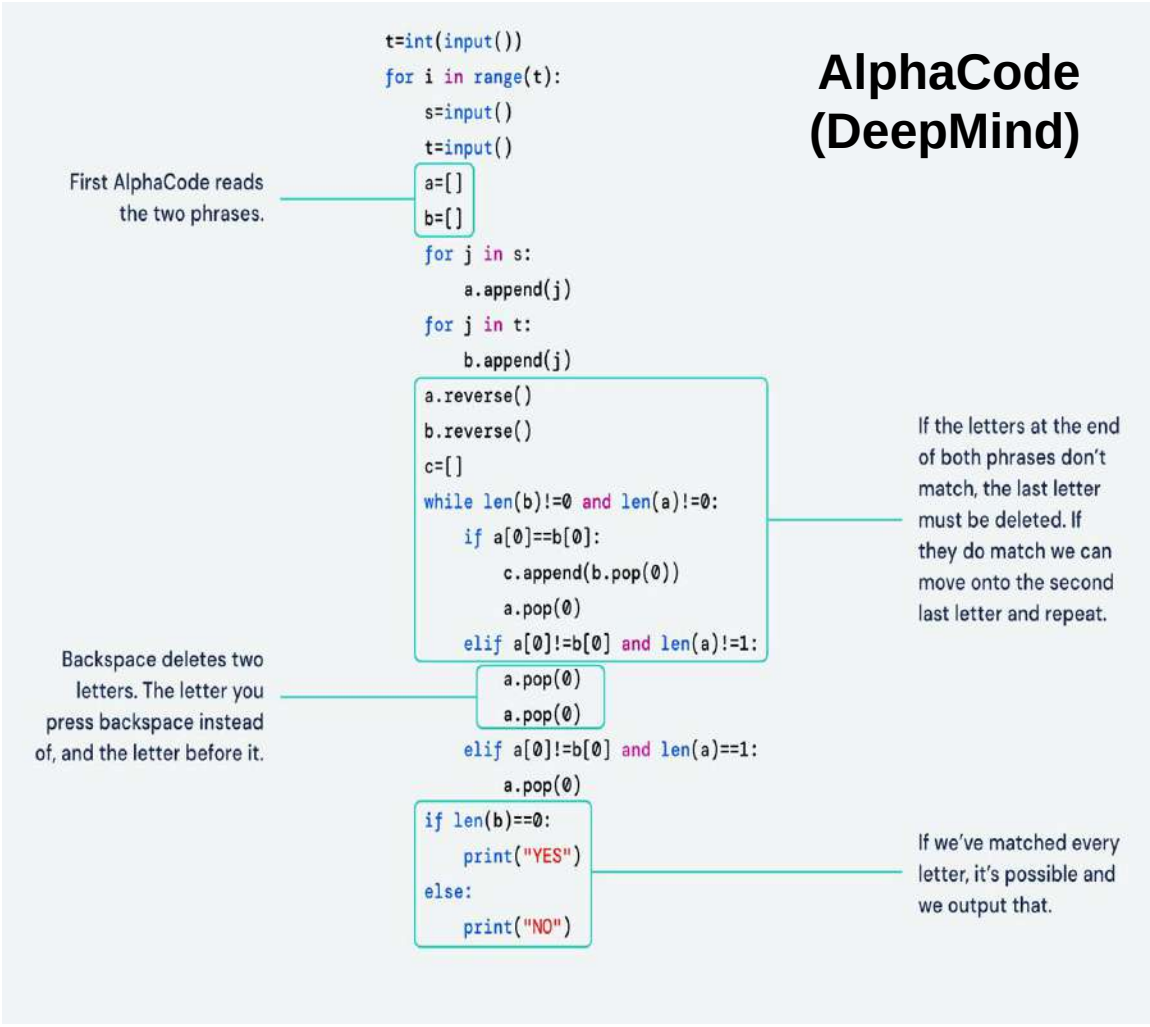
Codex (OpenAI)

GPT-3-like model trained on both natural language and source codes.

OpenAI Codex is most capable in Python, but it is also proficient in over a dozen languages.

It has a memory of 14KB for Python code, compared to GPT-3 which has only 4KB.

https://openai.com/blog/openai-codex/

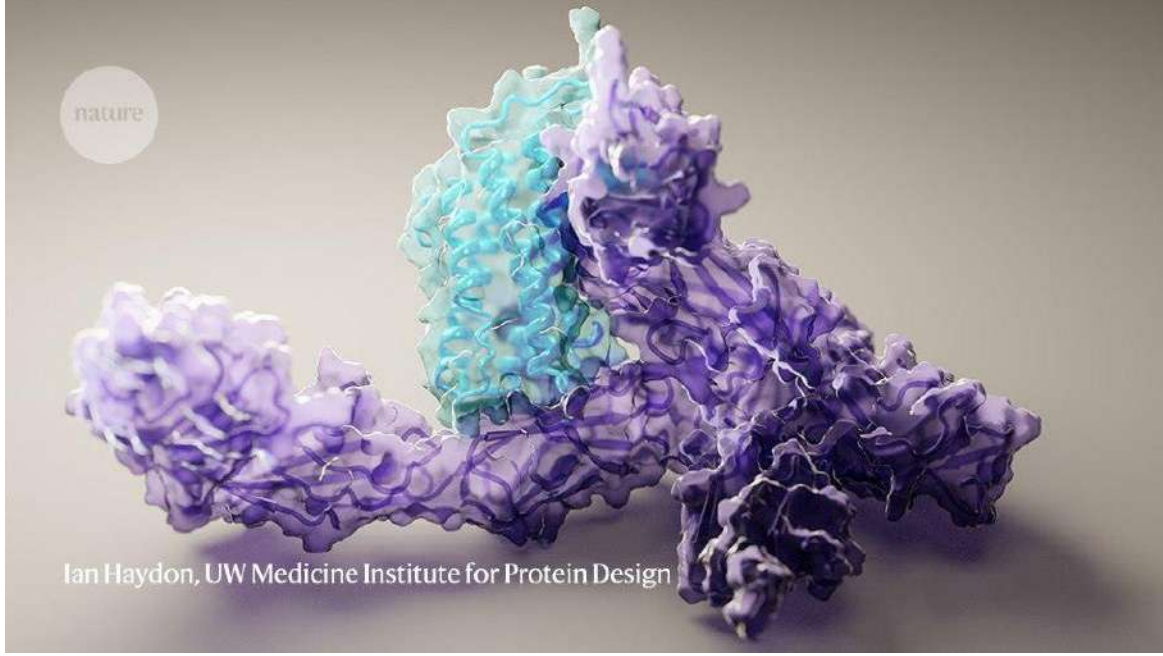


https://deepmind.com/blog/article/Competitive-programming-with-AlphaCode

Идея слайда: Grigory Sapunov, NLP in 2021, OpenTalks.AI / 2022

Приложения: Transformers in Biology

Finally transformers came into biology!



- **ESM-1b** protein language model (https://www.pnas.org/content/118/15/e2016239118)
- **MSA Transformer** for multiple sequence alignment (https://www.biorxiv.org/content/10.1101/2021.02.12.430858v1)
- **RoseTTAFold** for predicting protein structures (includes graph transformers) (https://www.science.org/doi/abs/10.1126/science.abj8754)
- **AlphaFold2** for predicting protein structures (https://www.nature.com/articles/s41586-021-03819-2)

Foundation Models

Идея: обучить одну модель на большом объеме данных, а потом сразу применять ее для множества приложений

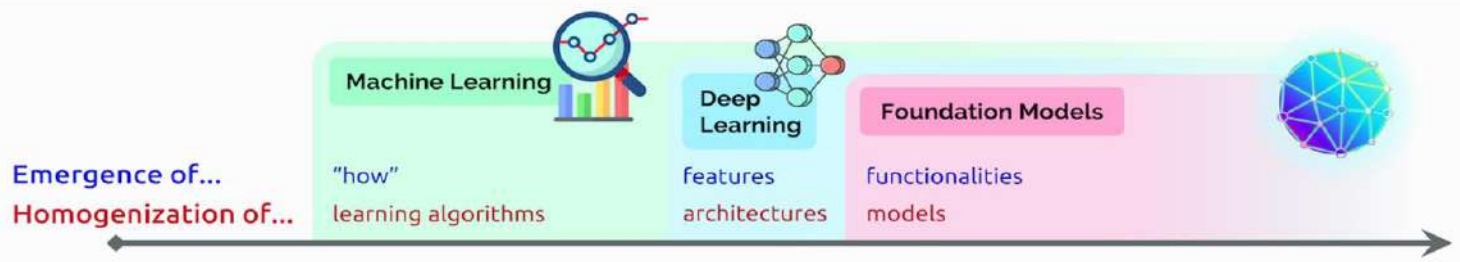


Fig. 2. A foundation model can centralize the information from all the data from various modalities. This one model can then be adapted to a wide range of downstream tasks.

Center for Research on Foundation Models

HAI Stanford University Human-Centered Artificial Intelligence

In recent years, a new successful paradigm for building AI systems has emerged: **Train one model on a huge amount of data and adapt it to many applications. We call such a model a foundation model**.

Foundation models (e.g., GPT-3) have demonstrated impressive behavior, but can fail unexpectedly, harbor biases, and are poorly understood. Nonetheless, they are being deployed at scale.

https://arxiv.org/abs/2108.07258

https://blog.inten.to/foundation-models-b89e7610057

Идея слайда: Grigory Sapunov, NLP in 2021, OpenTalks.AI / 2022

Multi-task: 107 supervised NLP tasks

Идея: можно сразу учить на большое число задач

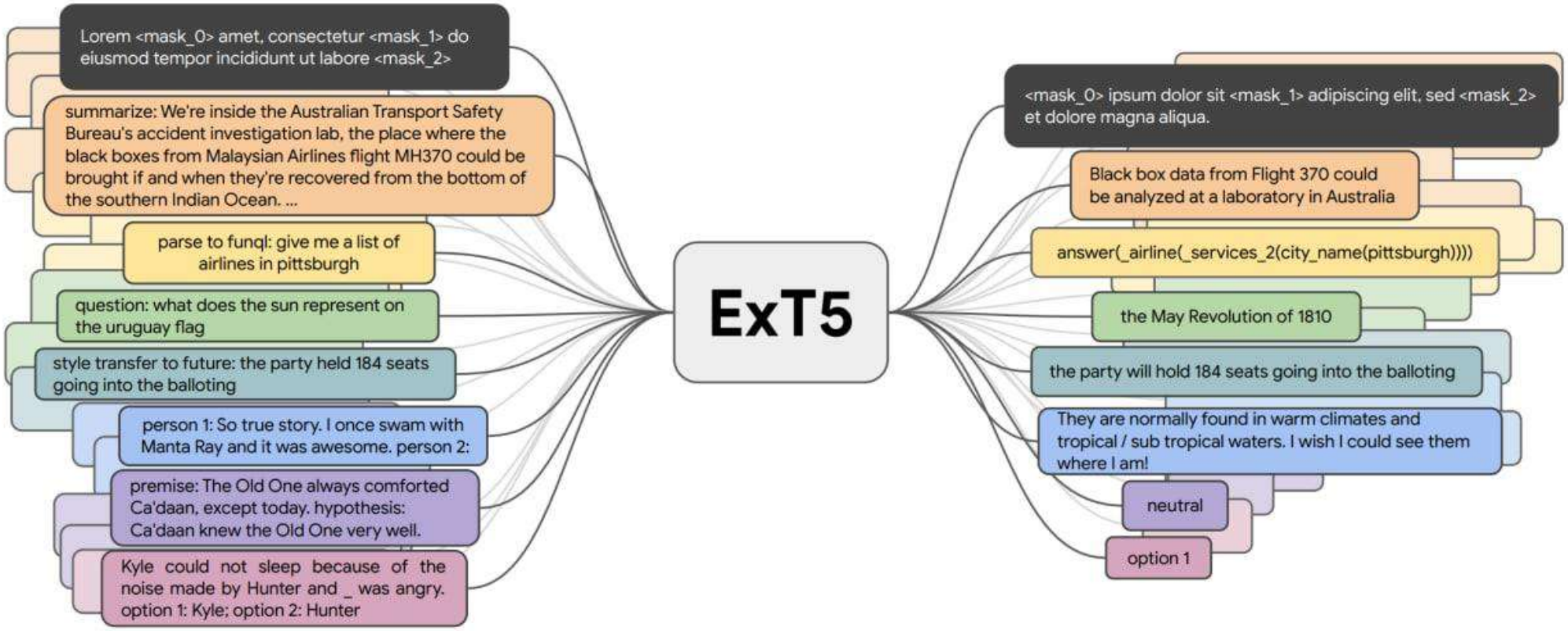


Figure 1: ExT5 pre-trains on self-supervised C4 and supervised ExMix: a massive multi-task mixture.

"ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning"
https://arxiv.org/abs/2111.10952
https://openai.com/blog/webgpt/

Слайд: Grigory Sapunov, NLP in 2021, OpenTalks.AI / 2022

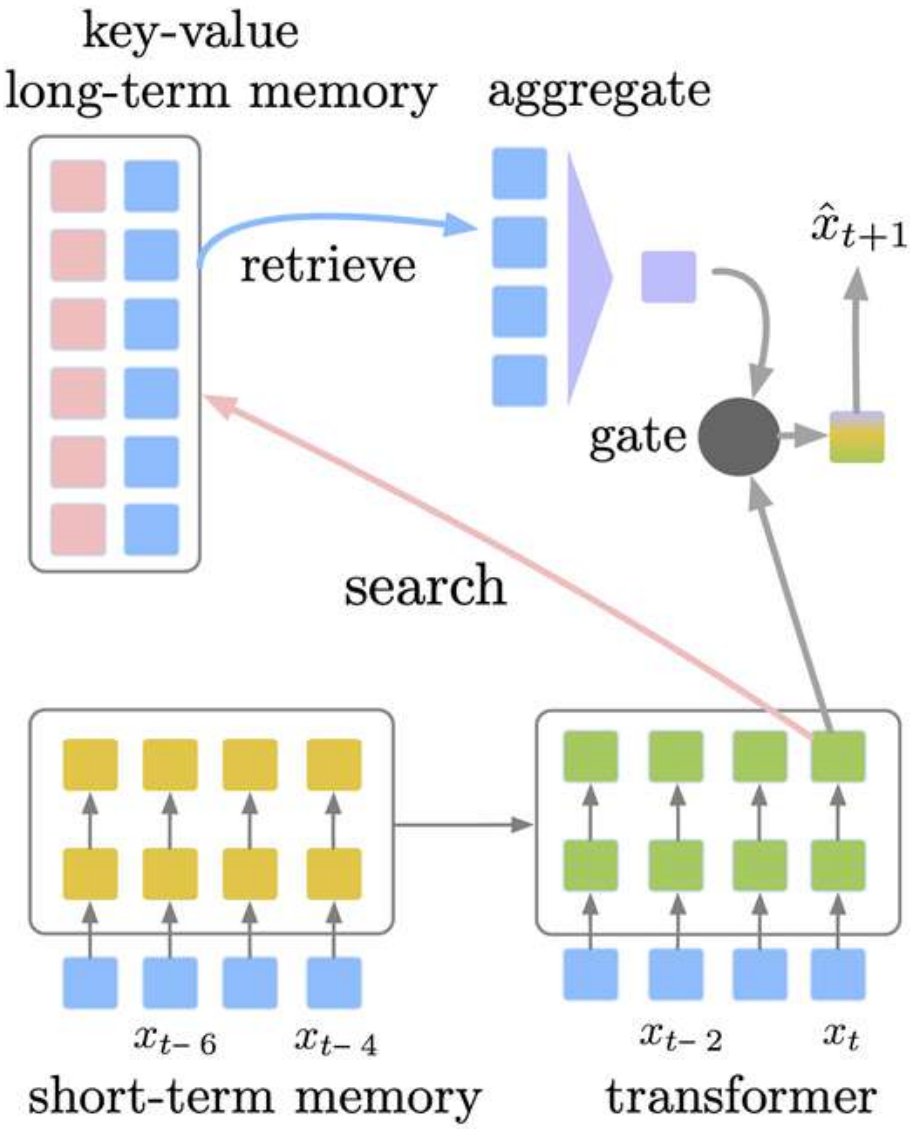
Retrieval-based models 2021

Поиск в коллекциях вместо/для супер-сетей (открытые базы)

- **EMDR²** ("End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering", https://arxiv.org/abs/2106.05346)
- **SPALM** ("Adaptive Semiparametric Language Models", https://arxiv.org/abs/2102.02557)
- **BlenderBot 2.0** ("Internet-Augmented Dialogue Generation", https://arxiv.org/abs/2107.07566)
- **FiD** ("Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering", https://aclanthology.org/2021.eacl-main.74/)
- **DeepMind RETRO** ("Improving language models by retrieving from trillions of tokens", https://arxiv.org/abs/2112.04426)
- **OpenAI WebGPT** ("WebGPT: Browser-assisted question-answering with human feedback", https://arxiv.org/abs/2112.09332)

With a 2 trillion token database, RETRO (Retrieval-Enhanced Transformer) obtains **comparable performance to GPT-3** and Jurassic-1 on the Pile, despite **using 25× fewer parameters**.

With RETRO the model is not limited to the data seen during training– it has access to the entire training dataset through the **retrieval mechanism**.



Идея слайда: Grigory Sapunov, NLP in 2021, OpenTalks.AI / 2022

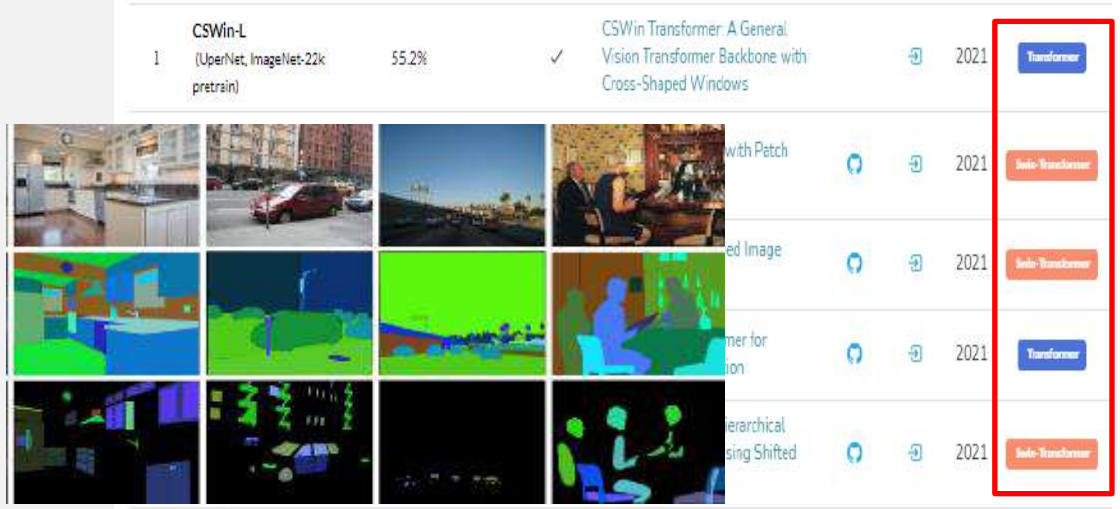
Трансформеры
в задачах зрения
(развитие трансформеров
в области CV, 2020-2021)

Трансформеры идут в CV! (2020+)

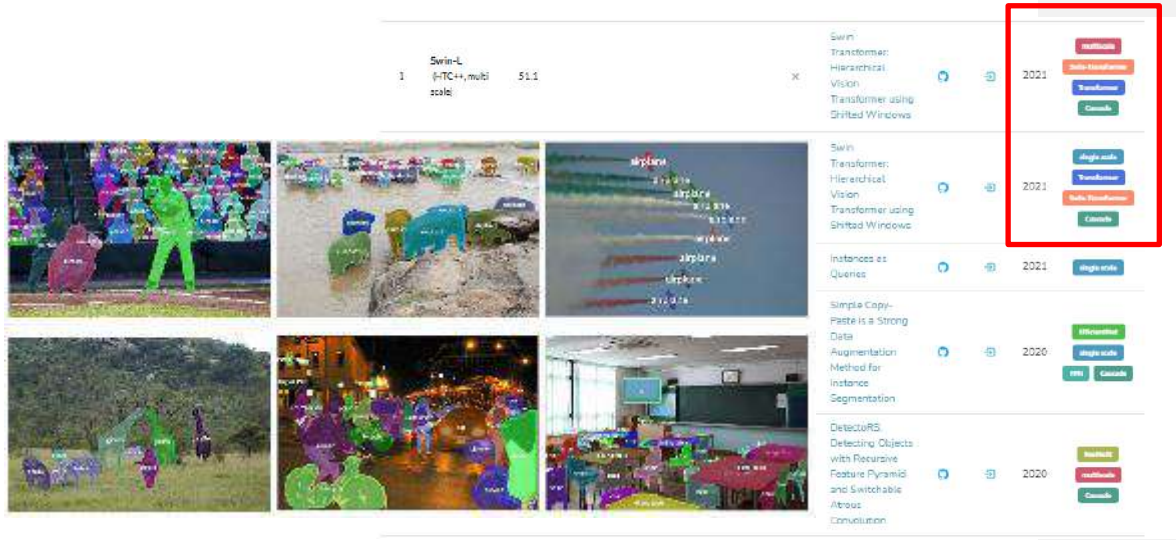
Классификация ImageNet

Обнаружение MS COCO

Сегментация ADE20K



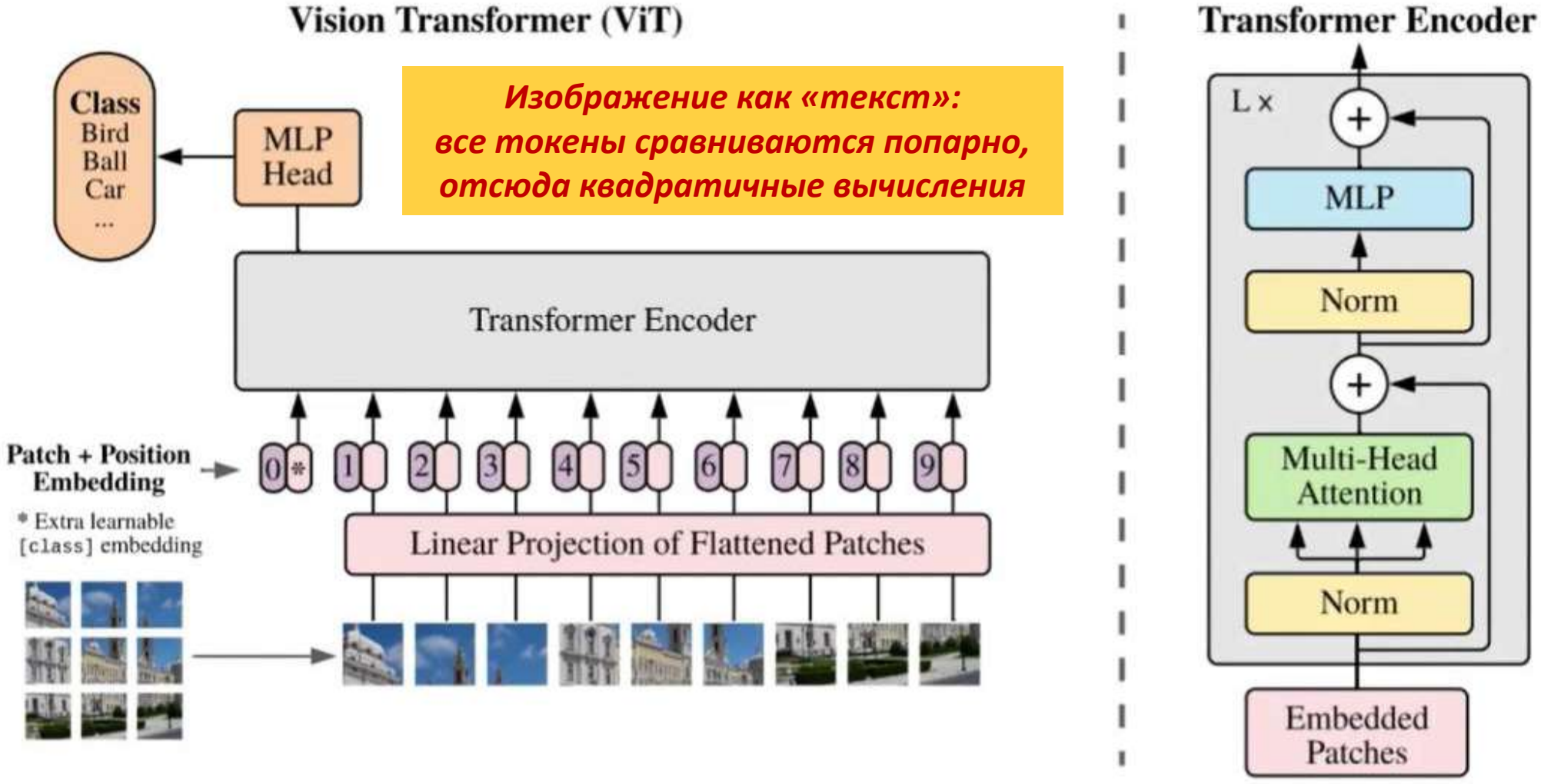
Сегментация(instance) MS COCO



По данным https://paperswithcode.com/

Vision Transformer architecture

Базовая идея применения трансформеров к задачам зрения: вместо текстовых токенов - фрагменты изображений



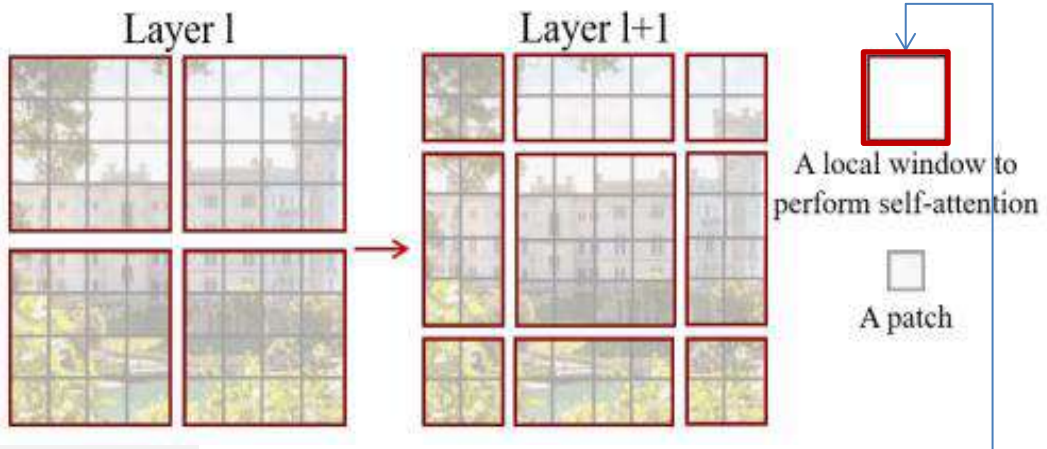
Изображение как «текст»: все токены сравниваются попарно, отсюда квадратичные вычисления

(C) Слайд: Alexey Dosovitskiy, Transformers for Vision, OpenTalks.AI/2022

Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 202

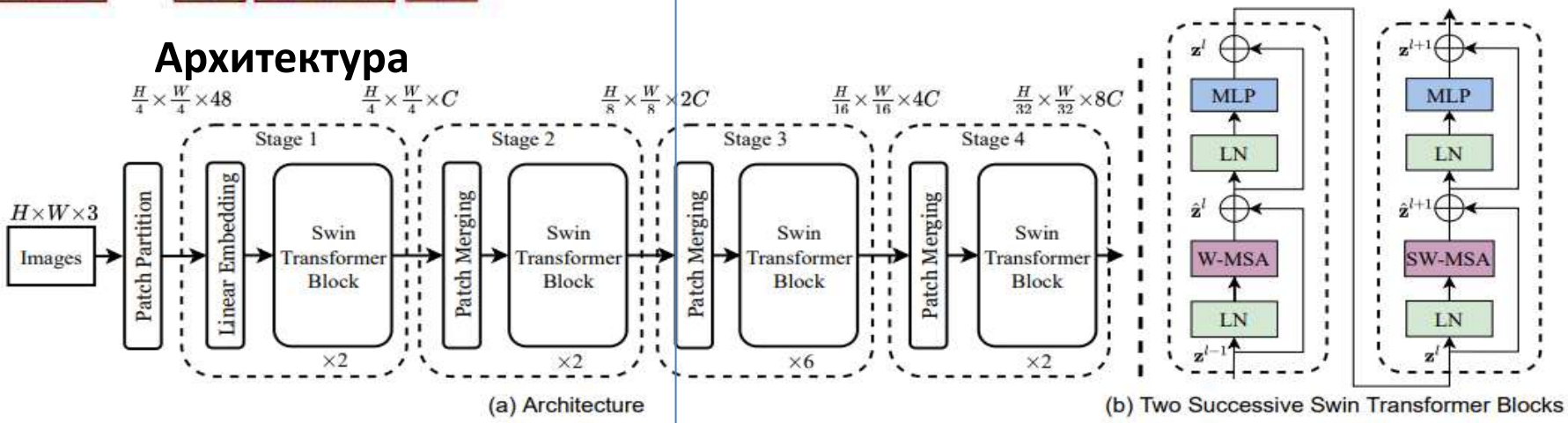
SWIN Transformer (2021)

Shifted Windows



Не сверточная сеть!
Используются трансформеры!
Лучшее качество в многих задачах!

Архитектура



Трансформеры горзадо медленнее сверточных сетей из-за квадратичных вычислений, но если их блоки применять локально, то скорость существенно улучшается!

https://arxiv.org/pdf/2105.01601.pdf

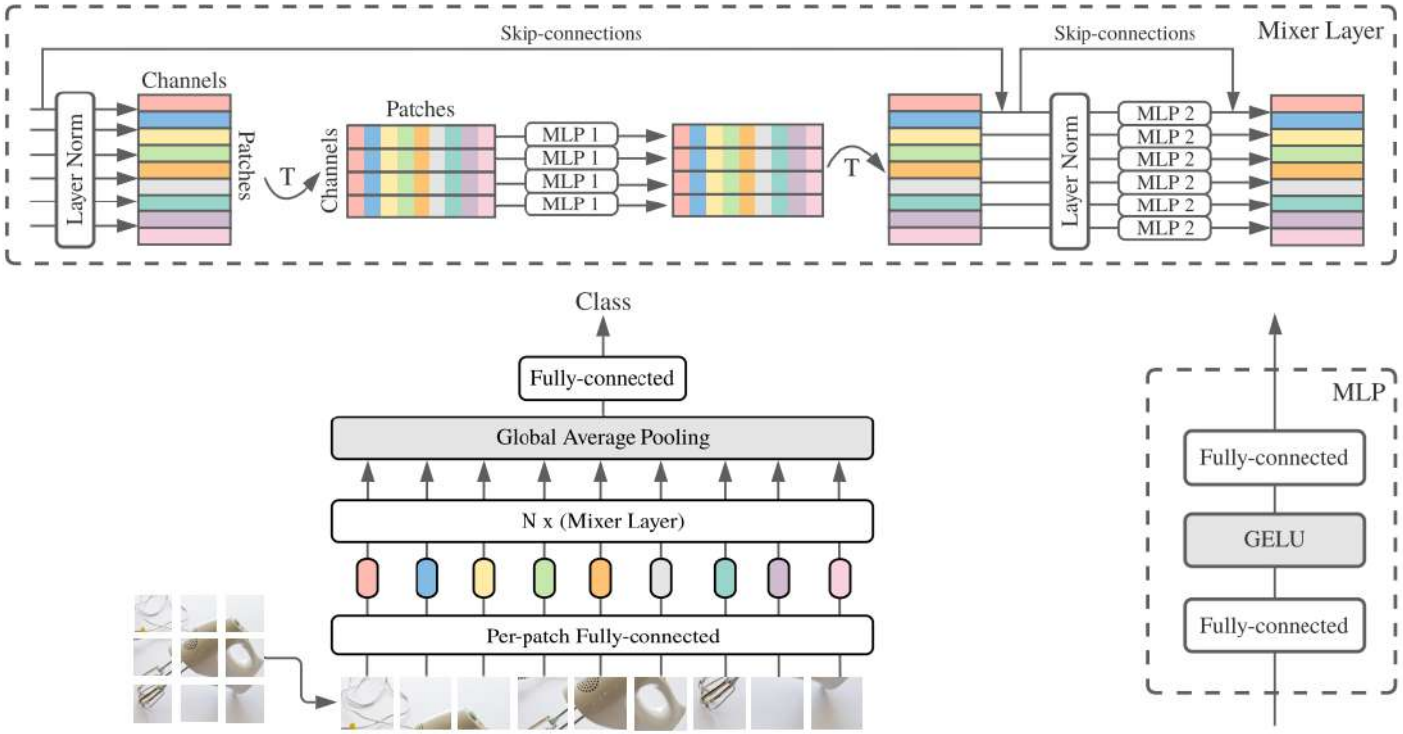
MLP-Mixer: An all-MLP Architecture for Vision (2021)

MLP = Multi-Layer Perceptron

От сверточных сетей
и трансформеров
обратно к персептронам?

Не сверточная сеть! Не трансформер!
Иерархия блоков полносвязных сетей!
Лучшее качество в многих задачах!

Архитектура



MLP-Mixer consists of per-patch linear embeddings, Mixer layers, and a classifier head.
Mixer layers contain one token-mixing MLP and one channel-mixing MLP, each consisting of two fully-connected layers and a GELU nonlinearity.
Other components include: skip-connections, dropout, and layer norm on the channels.

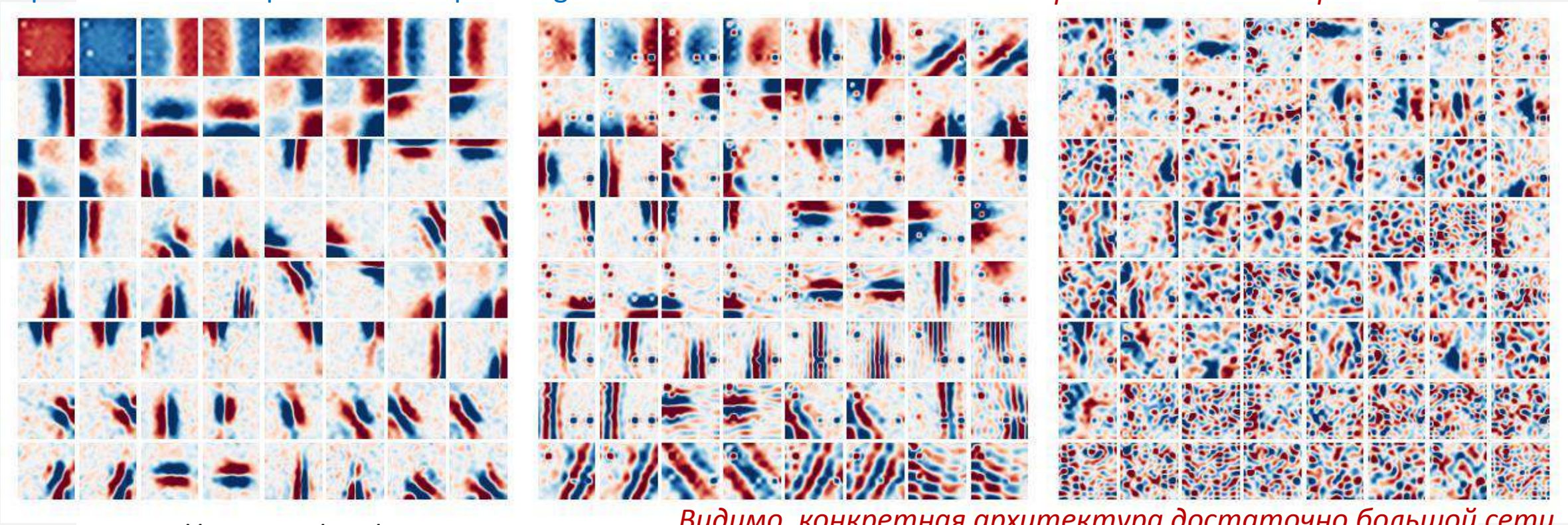
https://arxiv.org/pdf/2103.14030.pdf

MLP-Mixer: An all-MLP Architecture for Vision (2021)

Hidden units in the first (left), second (center), and third (right) token-mixing MLPs of a Mixer-B/16 model. Each unit has 196 weights, one for each of the 14 × 14 incoming patches. Pairs are sorted by filter frequency.
In contrast to the kernels of convolutional filters, where each weight corresponds to one pixel in the input image, one weight in any plot from the left column corresponds to a particular 16 × 16 patch of the input image.

Не сверточная сеть! Не трансформер!
Иерархия блоков полносвязных сетей!
Лучшее качество в многих задачах!

Аналогия со зрительной корой сохраняется и в несверточных сетях…



https://arxiv.org/pdf/2103.14030.pdf

Видимо, конкретная архитектура достаточно большой сети не так важна как целевая функция и способ обучения…

Трансформеры: Классификация изображений (2021)

Классификация ImageNet

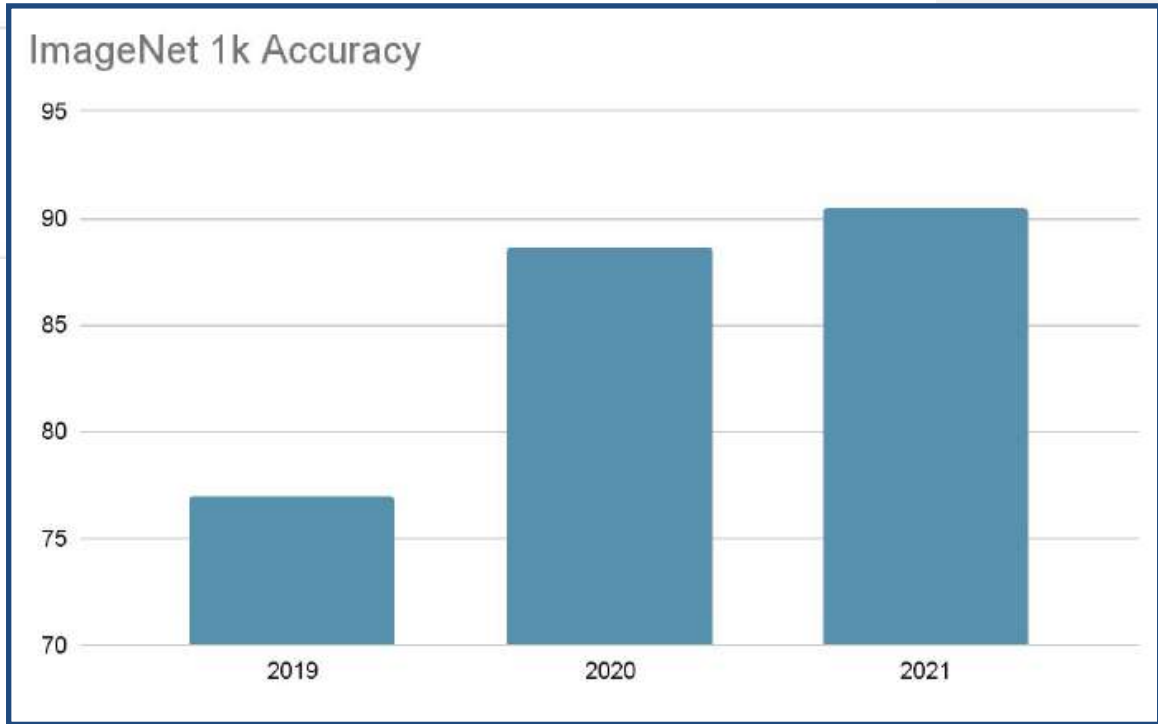
1	ViT-G/14	90.45%		1843M	✓	Scaling Vision Transformers		2021	Transformer
2	ViT-MoE-15B (Every-2)	90.35%		14700M	✓	Scaling Vision with Sparse Mixture of Experts		2021	Transformer
3	Meta Pseudo Labels (EfficientNet-L2)	90.2%	98.8%	480M	✓	Meta Pseudo Labels		2021	EfficientNet
4	Meta Pseudo Labels (EfficientNet-B6-Wide)	90%	98.7%	390M	✓	Meta Pseudo Labels		2021	EfficientNet
5	NFNet-F4+	89.2%		527M	✓	High-Performance Large-Scale Image Recognition Without Normalization		2021	

По данным https://paperswithcode.com/

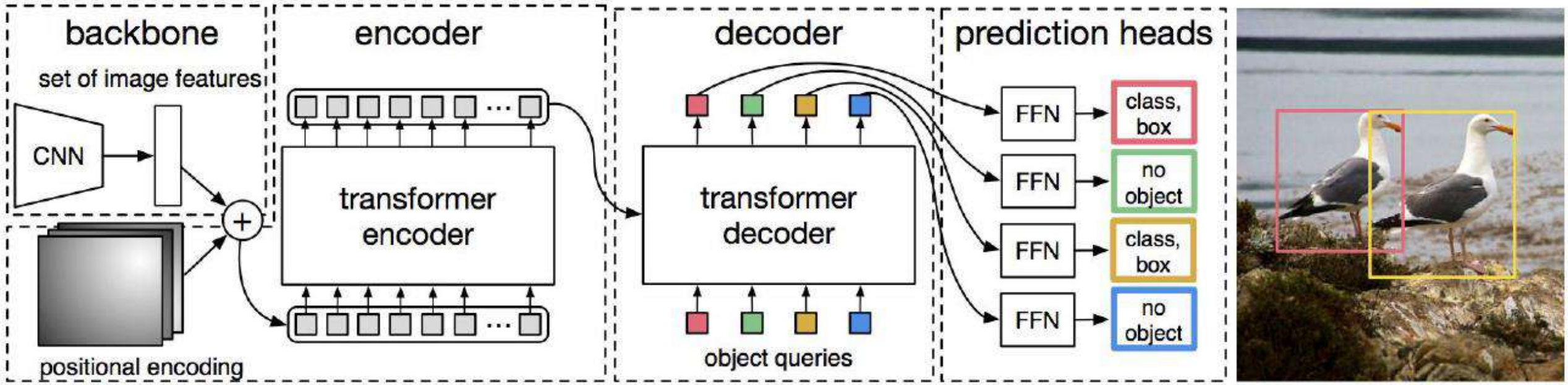
Self-Attention for Vision

Ashish Vaswani[1], Prajit Ramachandran[1], and Aravind Srinivas[2]

[1] Google Research, [2] UC Berkeley



DETR: End-to-End Object Detection with Transformers



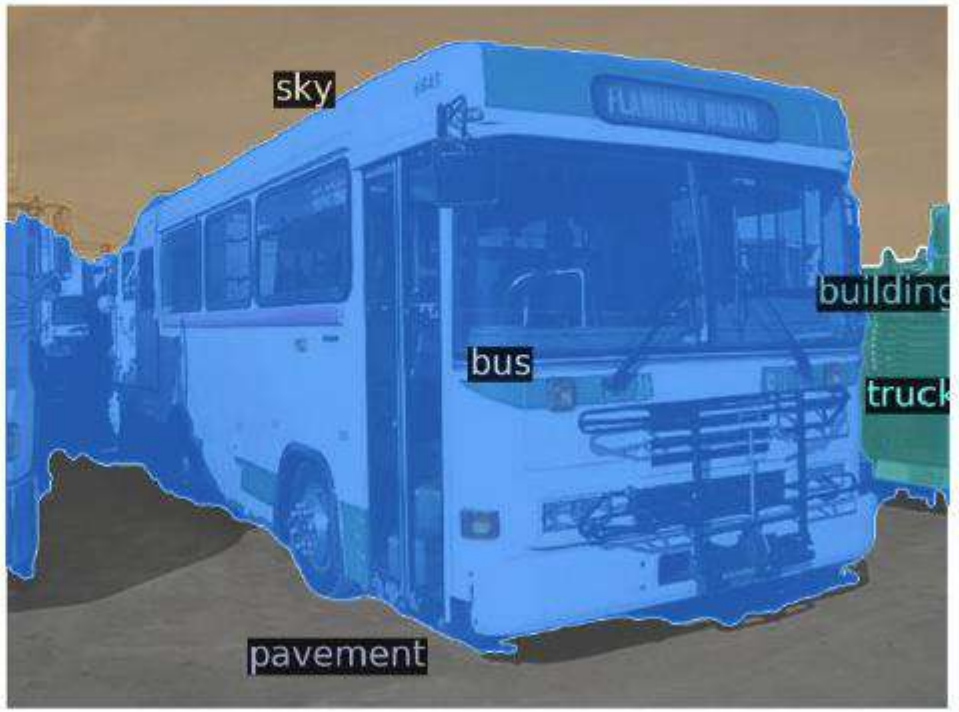
Self-Attention for Vision

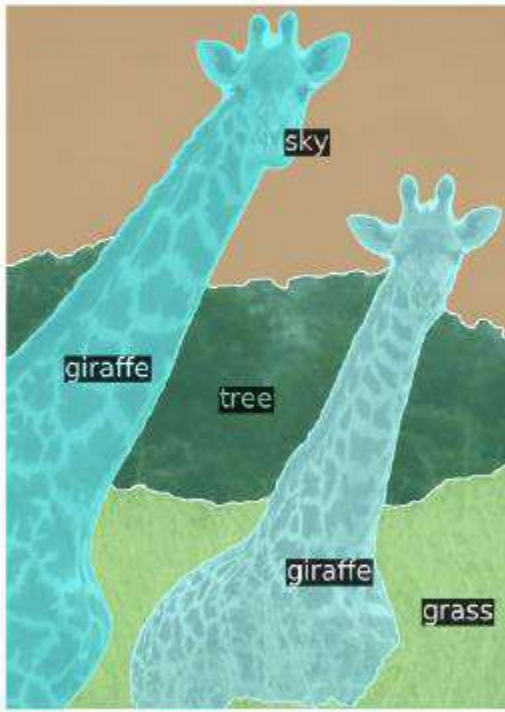
Ashish Vaswani[1], Prajit Ramachandran[1], and Aravind Srinivas[2]

[1] Google Research, [2] UC Berkeley

Carion et al 2020

DETR can be modified to perform panoptic segmentation





Self-Attention for Vision

Ashish Vaswani[1], Prajit Ramachandran[1], and Aravind Srinivas[2]

[1] Google Research, [2] UC Berkeley

Carion et al 2020

ViViT: Video ViT

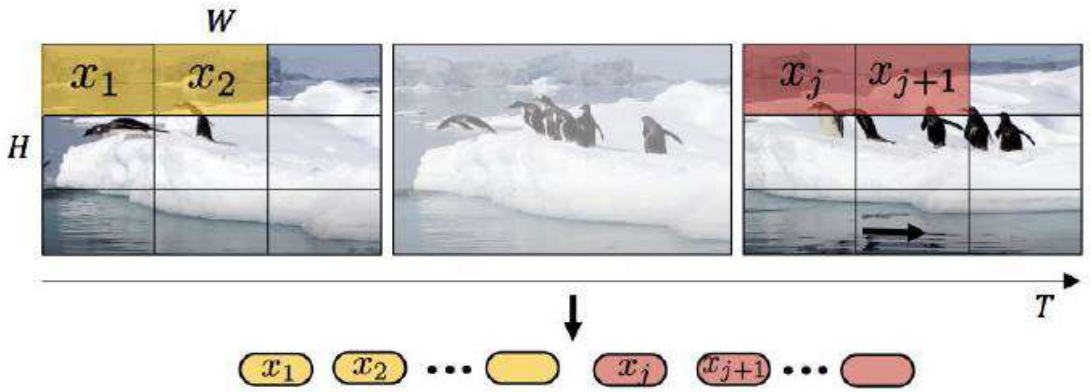
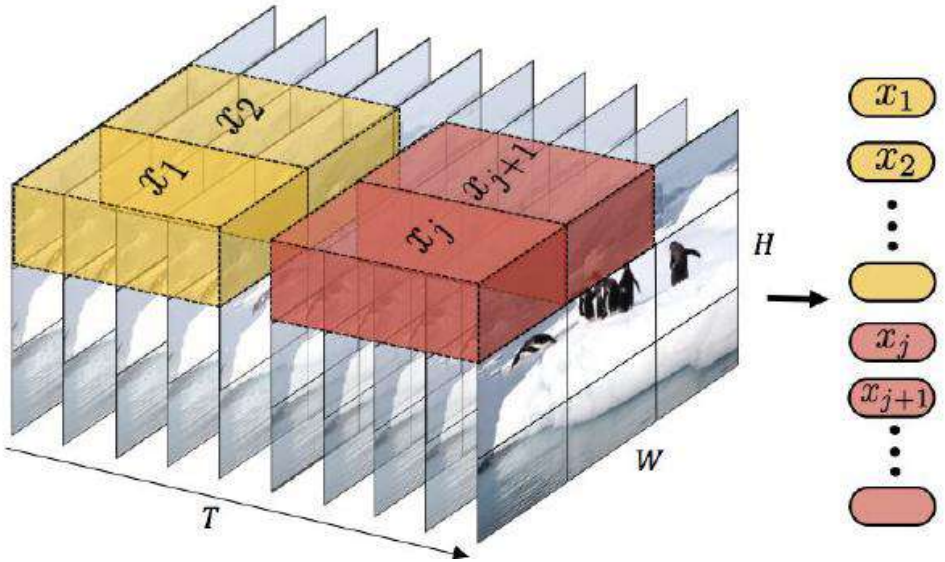


Figure 2: Uniform frame sampling: We simply sample n_t frames, and embed each 2D frame independently following ViT [15].



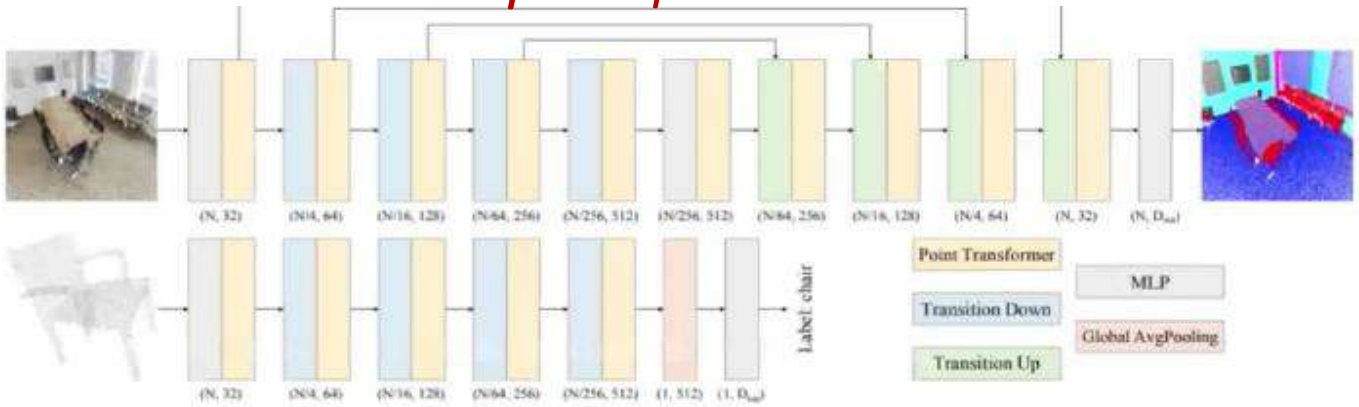
(a) Kinetics 400

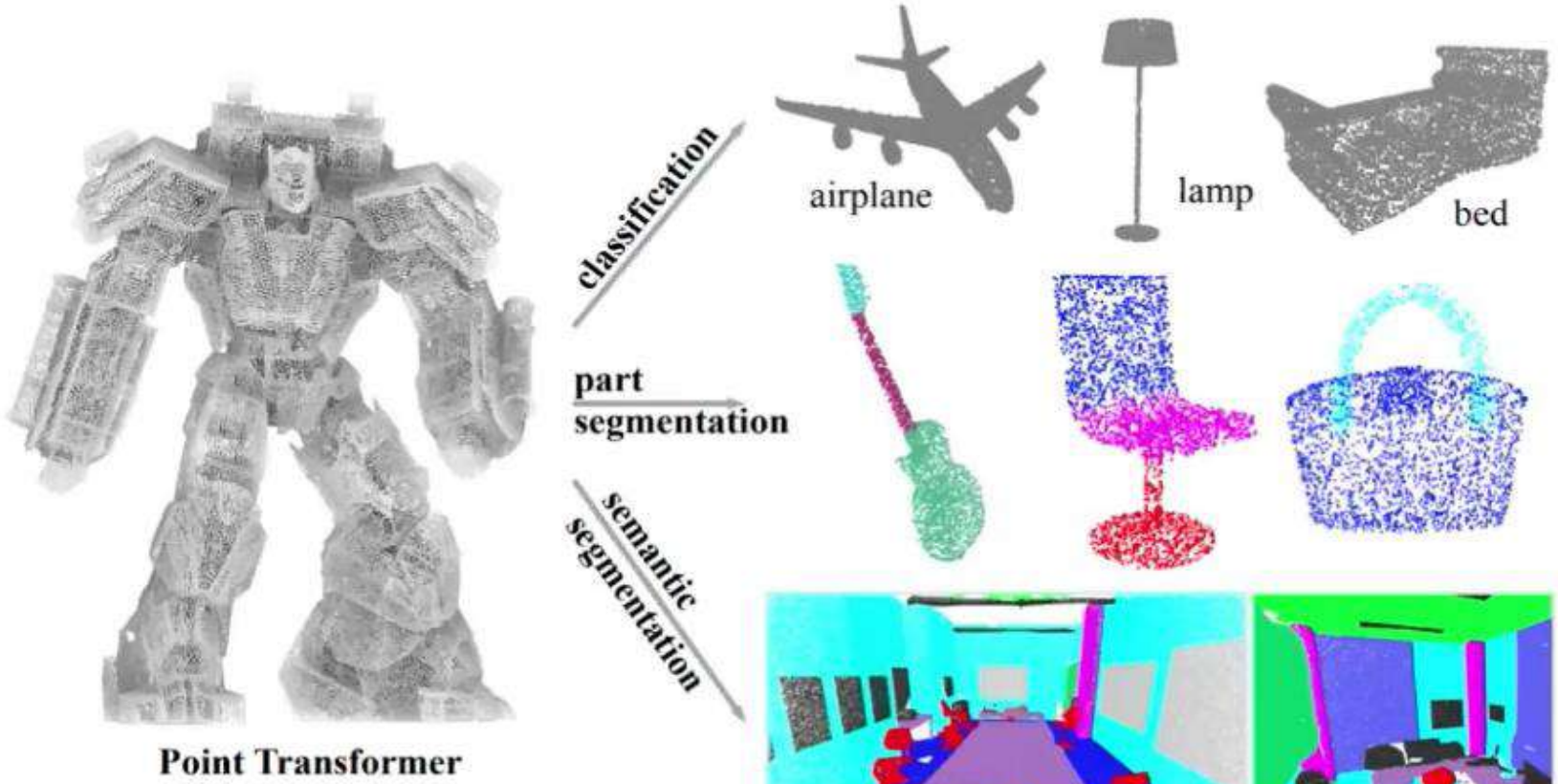
Method	Top 1	Top 5	Views
blVNet [16]	73.5	91.2	–
STM [30]	73.7	91.6	–
TEA [39]	76.1	92.5	10 × 3
TSM-ResNeXt-101 [40]	76.3	–	–
I3D NL [72]	77.7	93.3	10 × 3
CorrNet-101 [67]	79.2	–	10 × 3
ip-CSN-152 [63]	79.2	93.8	10 × 3
LGD-3D R101 [48]	79.4	94.4	–
SlowFast R101-NL [18]	79.8	93.9	10 × 3
X3D-XXL [17]	80.4	94.6	10 × 3
TimeSformer-L [2]	80.7	94.7	1 × 3
ViViT-L/16x2	80.6	94.7	4 × 3
ViViT-L/16x2 320	**81.3**	**94.7**	4 × 3
Methods with large-scale pretraining			
ip-CSN-152 [63] (IG [41])	82.5	95.3	10 × 3
ViViT-L/16x2 (JFT)	82.8	95.5	4 × 3
ViViT-L/16x2 320 (JFT)	83.5	95.5	4 × 3
ViViT-H/16x2 (JFT)	**84.8**	**95.8**	4 × 3

Arnab et al 2021

(C) Слайд: Alexey Dosovitskiy, Transformers for Vision, OpenTalks.AI/2022

Point Transformer



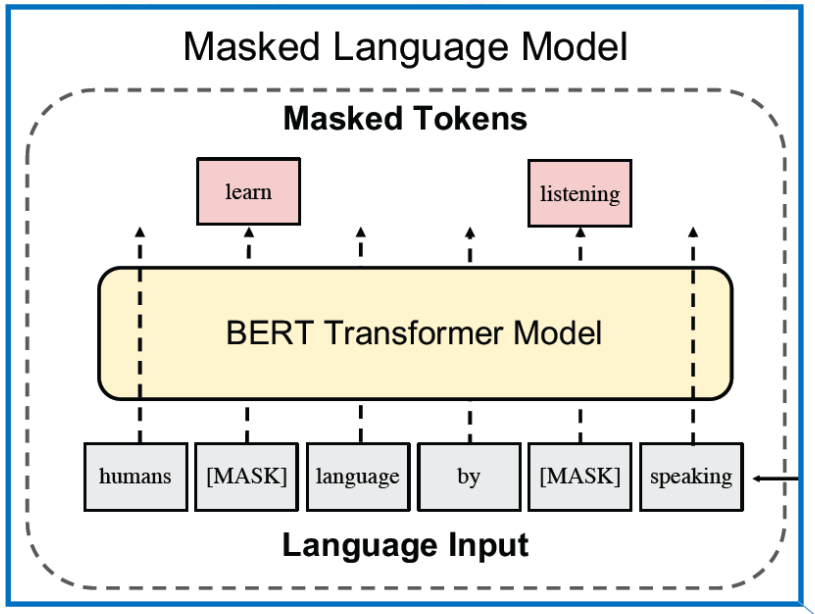


(C) Слайд: Alexey Dosovitskiy, Transformers for Vision, OpenTalks.AI/2022

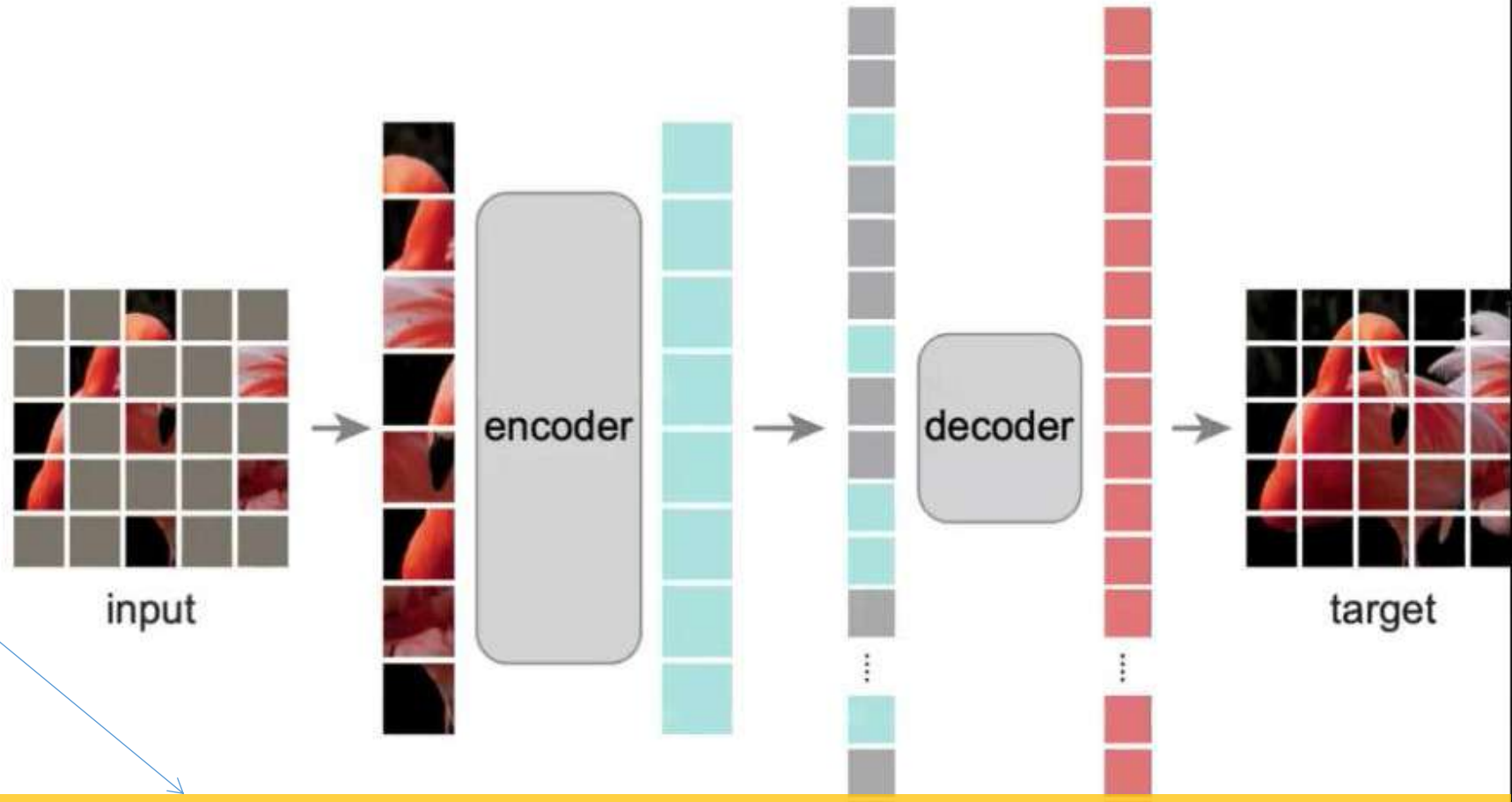
Zhao et al., Point Transformer, ICCV 2021

MAE: Masked Autoencoders Are Scalable Vision Learners

Better BERT-style training



Simple autoencoding of RGB pixel values works extremely well!



Сети типа BERT обучают на восстановление «замаскированных» фрагментов текста. Оказалось для изображений это тоже работает!

(C) Слайд: Alexey Dosovitskiy, Transformers for Vision, OpenTalks.AI/2022

Kaiming He et al., Masked Autoencoders Are Scalable Vision Learners, arXiv 2021

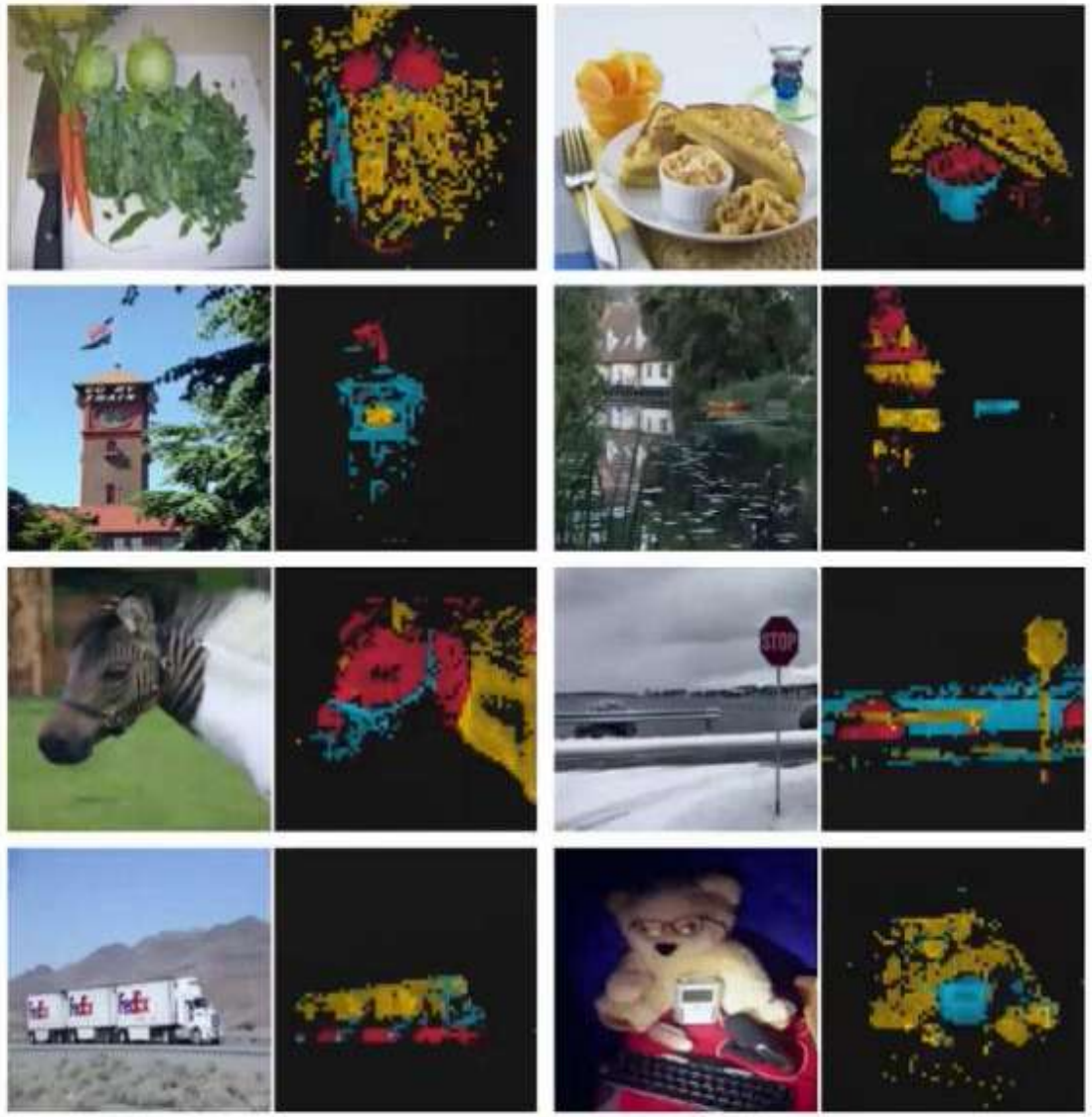
DINO: Emerging Properties in Self-Supervised Vision Transformers

"Standard" vision self-supervised learning + transformers = ❤️

Good results on classification

Interesting emerging segmentations

Good results on DAVIS (video segmentation)

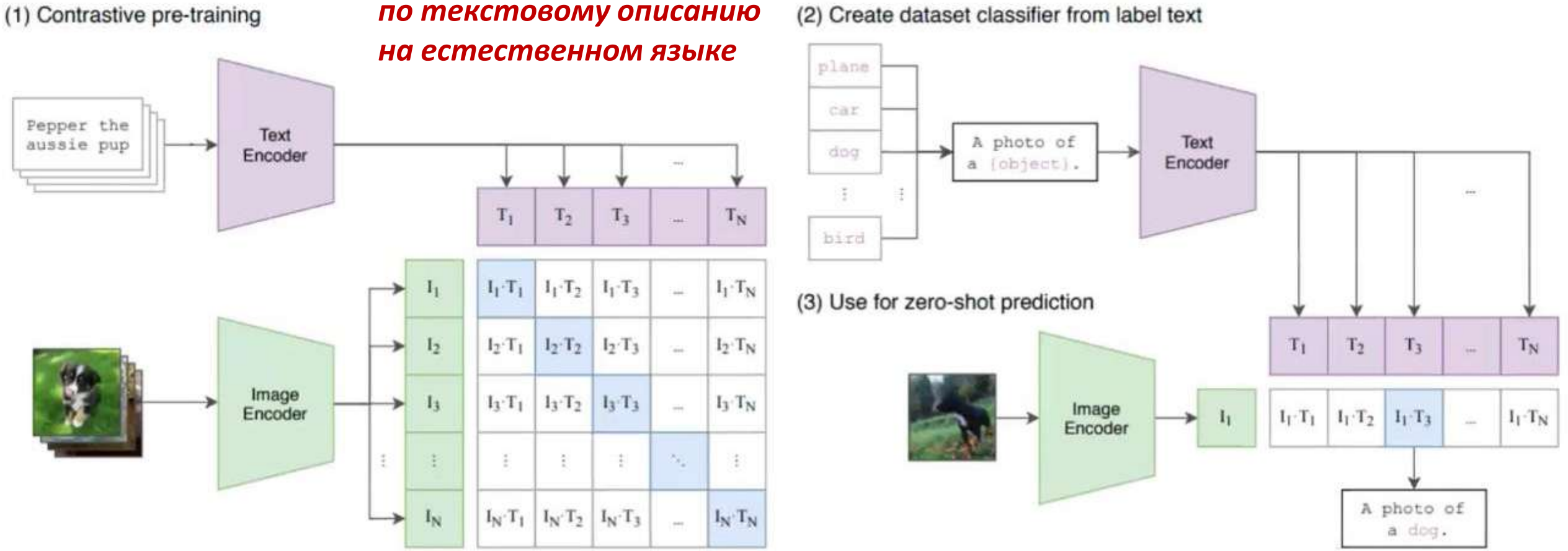


(C) Слайд: Alexey Dosovitskiy, Transformers for Vision, OpenTalks.AI/2022

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, ICCV 2021

CLIP: Visual representations from image-and-text data

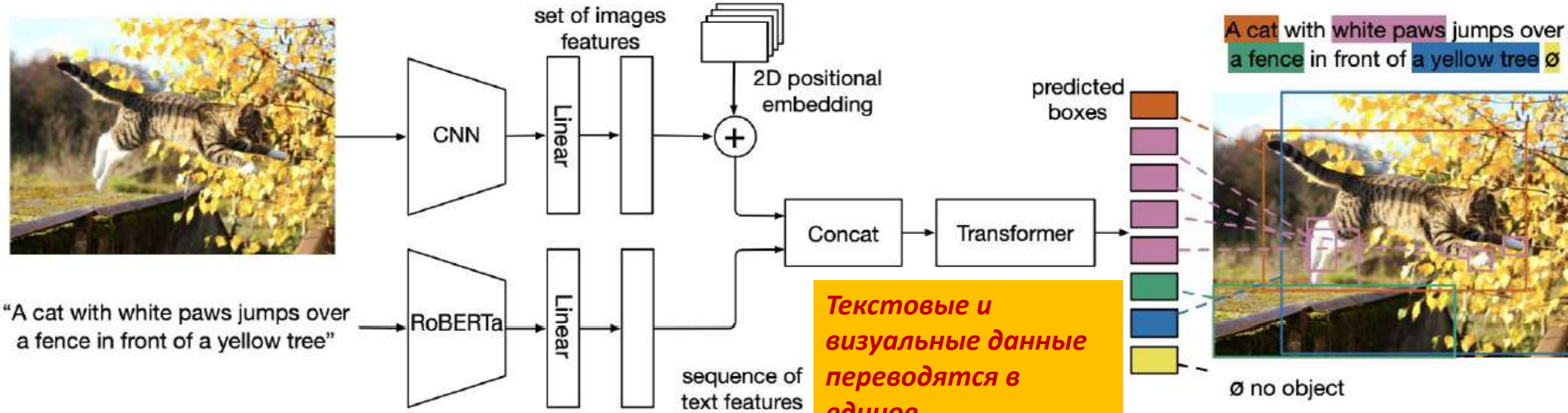
Генерация визуальных данных по текстовому описанию на естественном языке



(C) Слайд: Alexey Dosovitskiy, Transformers for Vision, OpenTalks.AI/2022

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

Multimodal DETR (M-DETR)



Текстовые и визуальные данные переводятся в единое представление и потом обрабатываются совместно!

Kamath et al 2021

Self-Attention for Vision

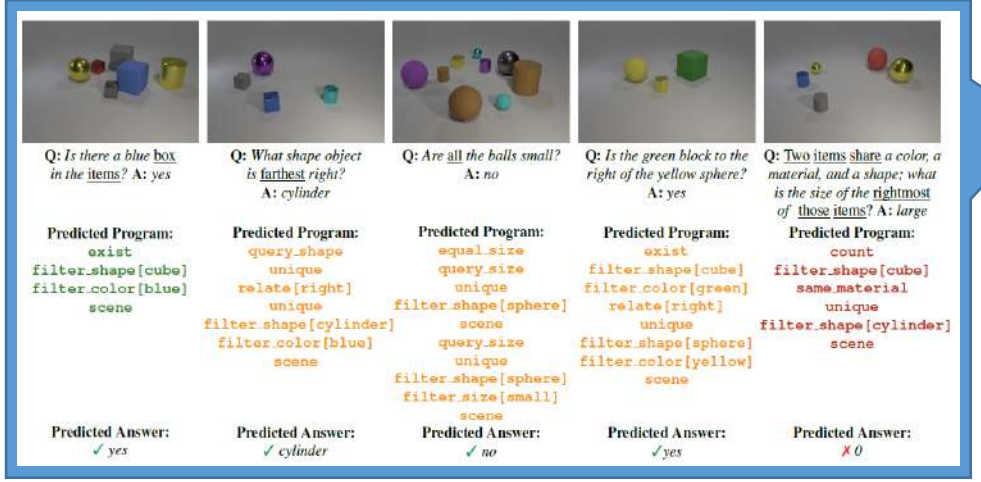
Ashish Vaswani[1], Prajit Ramachandran[1], and Aravind Srinivas[2]

[1] Google Research, [2] UC Berkeley

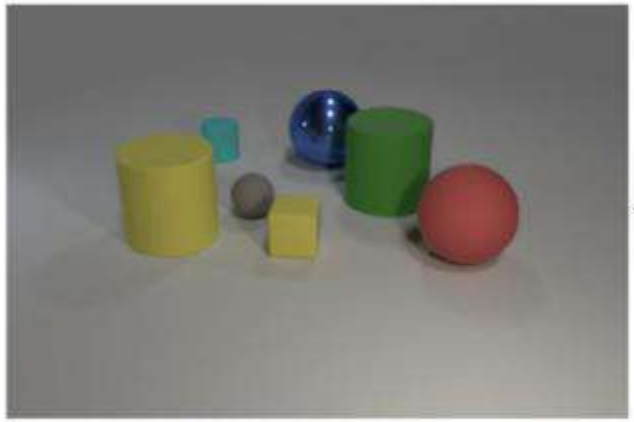
Multi-modal Learning

M-DETR

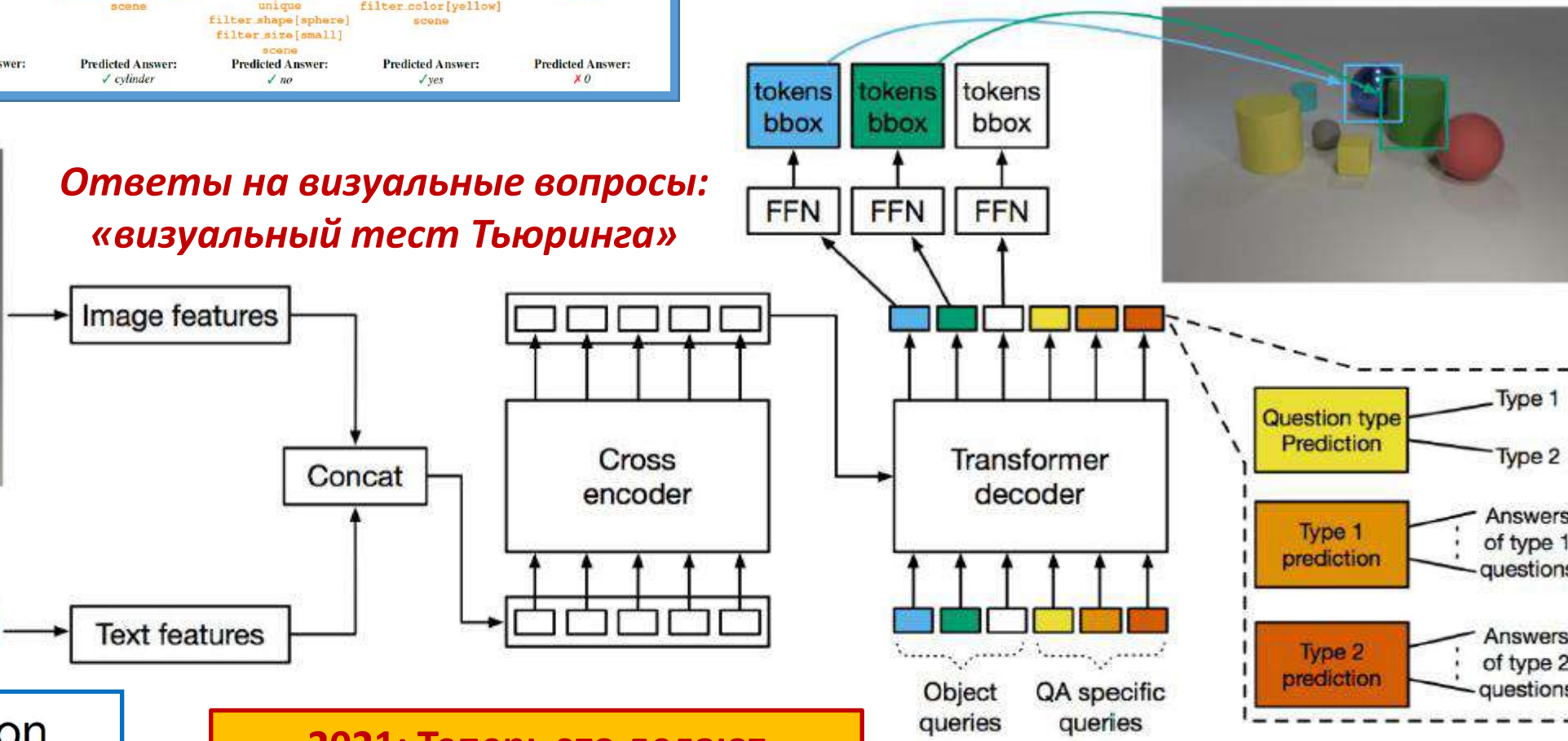
В условном мире…



2017: ИИ-1 встречается с ИИ-2!!!



Ответы на визуальные вопросы: «визуальный тест Тьюринга»



2021: Теперь это делают трансформеры

Self-Attention for Vision

Ashish Vaswani[1], Prajit Ramachandran[1], and Aravind Srinivas[2]

[1] Google Research, [2] UC Berkeley

Kamath et al 2021

M-DETR

В реальном мире

Self-Attention for Vision

Ashish Vaswani[1], Prajit Ramachandran[1], and Aravind Srinivas[2]

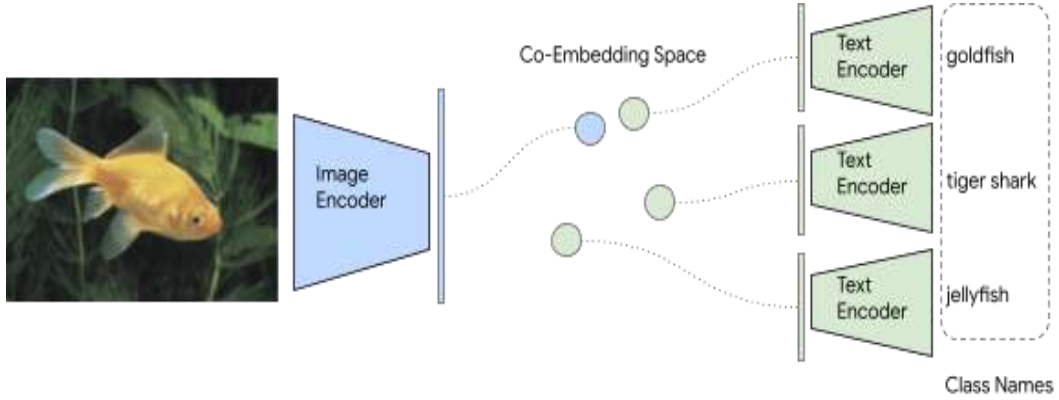
[1] Google Research, [2] UC Berkeley

Kamath et al 2021

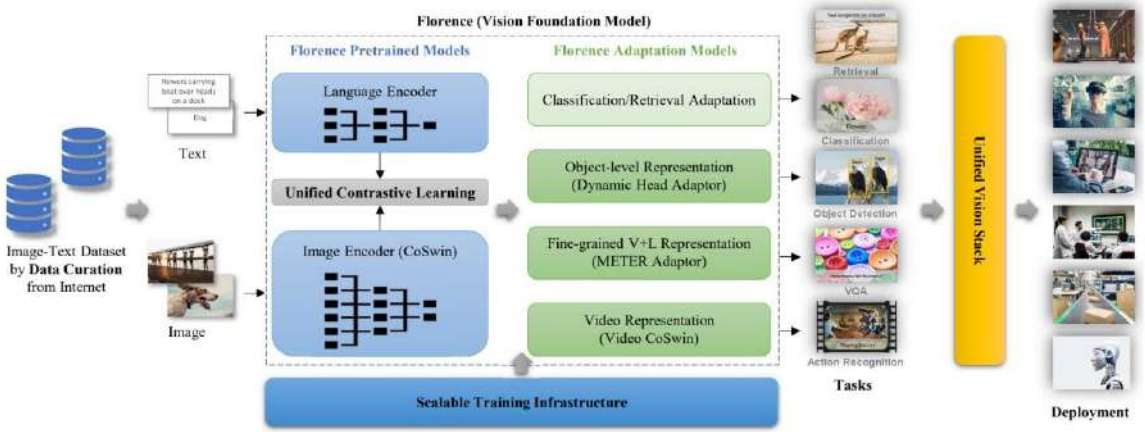
Multi-modal Learning

DALL·E (OpenAI) “Zero-Shot Text-to-Image Generation” https://arxiv.org/abs/2102.12092

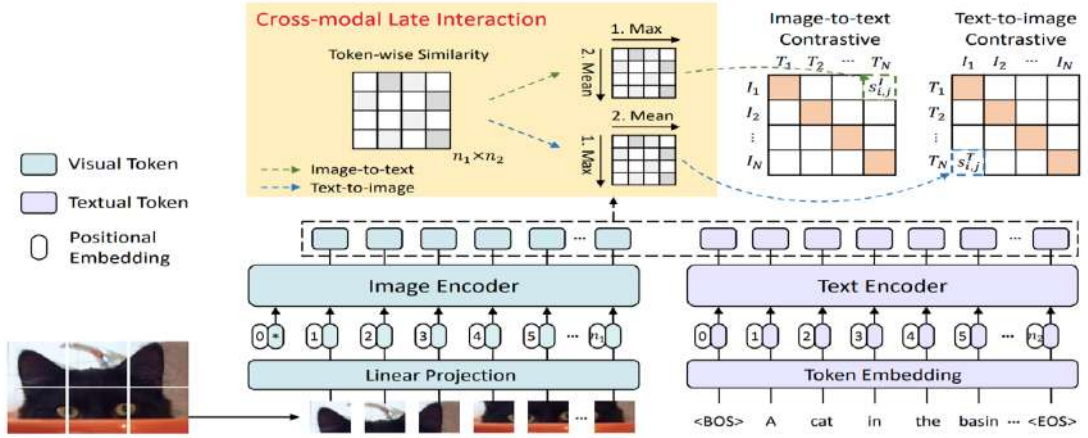
ALIGN (Google) “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision” https://arxiv.org/abs/2102.05918



Florence (Microsoft) “Florence: A New Foundation Model for Computer Vision” https://arxiv.org/abs/2111.11432

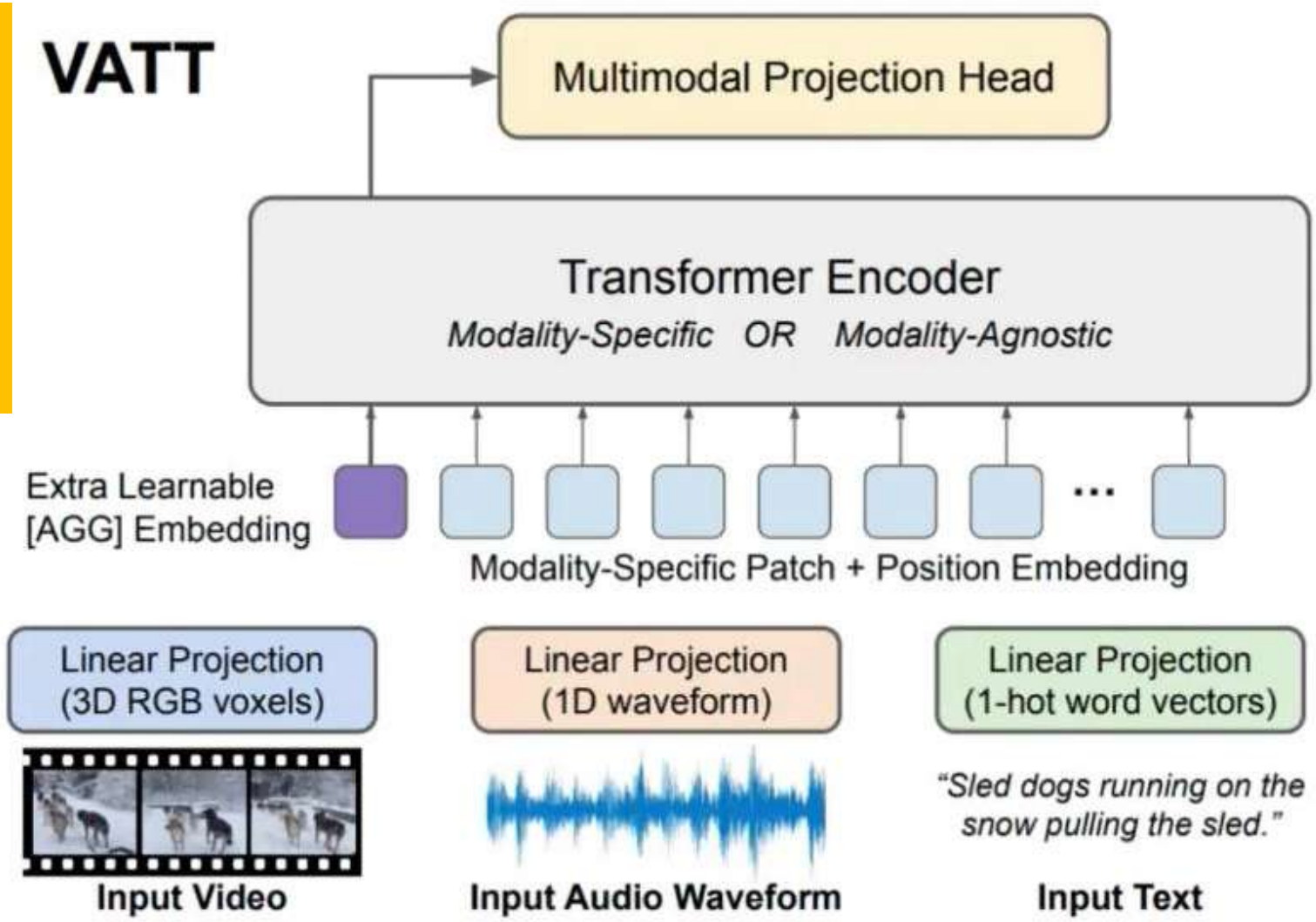


FILIP (Huawei) “FILIP: Fine-grained Interactive Language-Image Pre-Training” https://arxiv.org/abs/2111.07783



VATT: Learning from video, audio, and text

Чем больше модальностей мы объединяем, тем более впечатляющим может быть результат!



Данные любых модальностей переводятся в единое универсальное представление и потом обрабатываются совместно

(C) Слайд: Alexey Dosovitskiy, Transformers for Vision, OpenTalks.AI/2022

Akbari et al., VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text, NeurIPS 2021

Transformers vs. CNN

В бортовых приложениях CNN пока все еще предпочтительнее…

Плюсы трансформеров:

1. Большой receptive field начиная уже с первого слоя.
2. Высокие показатели метрик для изображений высокого разрешения.

Минусы трансформеров:

1. Ресурсоемкость из за большого receptive field.
2. Отсутствие пространственной инвариантности.
3. Скорость inference.
4. Потеря низкоуровневых признаков.

CenterNet2 + BiFPN 1280x1280 (Comparison if all else equal)						
Backbone	AP	AP50	AP75	AP small	AP medium	AP large
ViT-B16	49.9	67.4	54.4	31.0	55.0	67.0
Res2Net (3x times faster)	49.4	66.4	53.9	31.0	54.3	64.2
Relative diff, %	1	1	1	0	1	4

	AP	AP50	AP75	AP small	AP medium	AP large
DyHead SWIN-L 2000x1200 (TTA) ~ 1 FPS	58.7	77.1	64.5	41.7	62.0	72.8
YOLOR 1280x1280 = 34 FPS	57.3	75.0	62.7	40.4	61.2	69.2
Relative diff, %	2	3	3	3	1	5

Главное преимущество трансформеров: большие данные любых модальностей, включая визуальную, переводятся в единое универсальное представление и обрабатываются единообразно

Из материалов семинаров ГосНИИАС: Swin Transformer, А. Зимин

Reinforcement Learning
(обучение с подкреплением)

Advances in RL-2022:

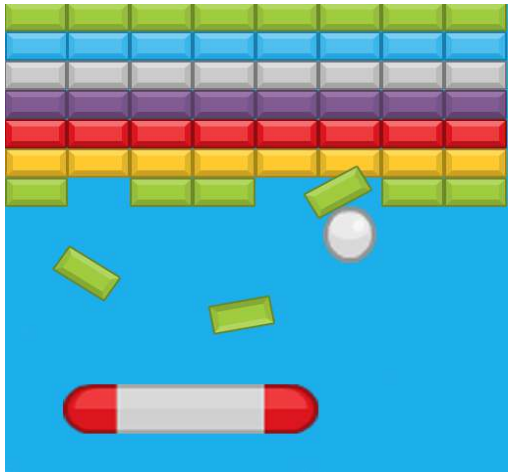
- SUNRISE
- **Decision Transformer**
- Implicit Behavior Cloning
- Navigation Turing Test
- **Open-Ended Learning**

Мой выбор

Наводки: Valentin Malykh, Advances in Reinforcement Learning in 2022, OpenTalksAI/2022

Краткое напоминание

Deep Reinforcement Learning: обучение с подкреплением



- **Reward $R(t)$:** score you earned at current step

  Вознаграждение (выигрыш после хода)

- **State S**: current screen

  Состояние (что видим на экране)

- **Action i:** move your board left / right

  Действие (что делаем)

- **Action value function $\hat{Q}(S, i)$:** ← **Функция Беллмана!**

  your predicted future total rewards

  Стоимость (выигрыш в будущем)

- **Policy $\pi(s)$**: How to choose your action

  Стратегия (какой ход предпочтительнее)

Главный ключ к решению задач глубокой оптимизации: ***глубокое обучение с подкреплением для выучивания эвристик!***

Как научить сеть играть в компьютерную игру

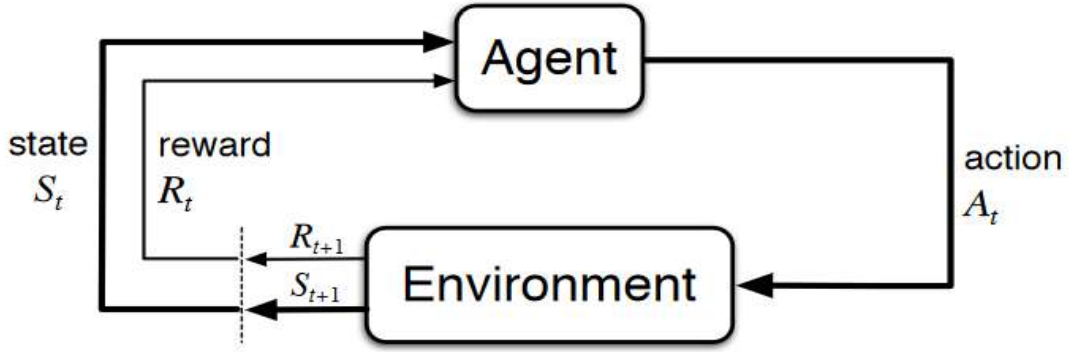


Схема знакома всем специалистам по теории управления

Human-level control through deep reinforcement learning (Minh, et al. Nature, 2015)

Технологии глубокого обучения в исследовании операций и управлении

AI Beats a Fighter Pilot in a Virtual Dogfight (2016)

AI ALPHA, built by Psibernetix, Inc. with Air Force Research Laboratory

Оперативно-тактическое и групповое управление

Показатель зрелости: игровые задачи

StarCraft II - полноценный тактический военный симулятор с упрощённой моделью ведения боя.

Прогноз 2018: *ГНС смогут выиграть у людей* ***лет через 5***

Автономное управление и оптимизация облика изделий

Оптимизация бизнес-процессов и производства

Технологии глубокого обучения в исследовании операций и управлении

AI Beats a Fighter Pilot in a Virtual Dogfight (2016)

AI ALPHA, built by Psibernetix, Inc. with Air Force Research Laboratory

Оперативно-тактическое и групповое управление

Показатель зрелости: игровые задачи

Сочетание технологий ГНС, RL и генетического отбора

2019+: Дорога открыта!

Автономное управление и оптимизация облика изделий

StarCraft II - полноценный тактический военный симулятор с упрощённой моделью ведения боя.

~~***Прогноз 2018: ГНС смогут выиграть у людей через 5 лет***~~

25.01.2019: DeepMind AlphaStar со счетом 11:1 победила ведущих профессиональных игроков в StarCraft II

TOP AI 2019!

Оптимизация бизнес-процессов и производства

Технологии глубокого обучения в исследовании операций и управлении

25 января 2019 г. компания DeepMind опубликовала релиз: ГНС AlphaStar со счетом 11:1 победила ведущих профессиональных игроков в StarCraft II !!!

Глубокая оптимизация в задачах теории игр

2019: ГНС строят модели тактических ситуаций и проводят анализ путей решения стратегических задач

AlphaStar: Mastering the Real-Time Strategy Game StarCraft II

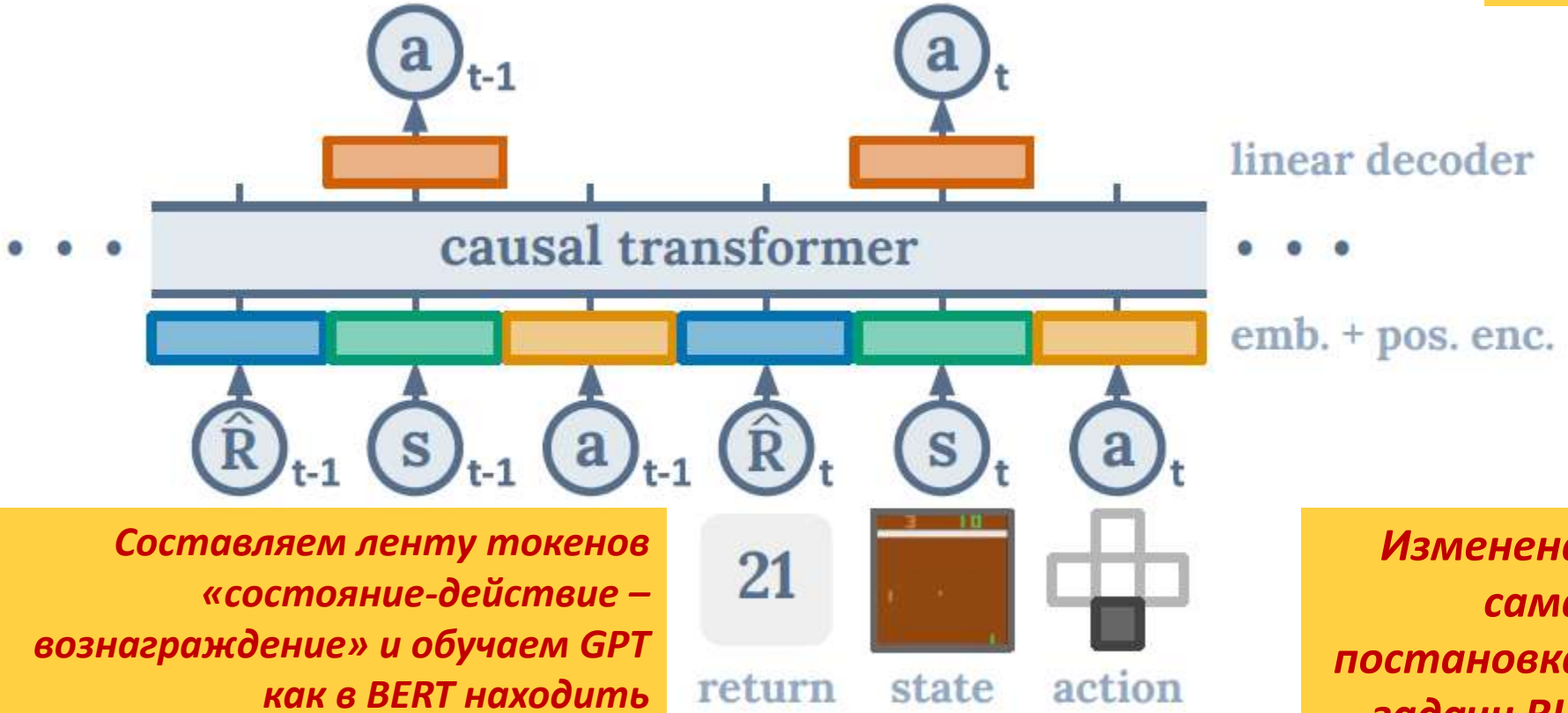
TOP AI 2019!

Как это сделано: Сочетание глубоких сетей, обучения с учителем, обучения с подкреплением и генетического отбора

https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

Decision Transformer

2021: Трансформер для обучения с подкреплением



Составляем ленту токенов «состояние-действие – вознаграждение» и обучаем GPT как в BERT находить «замаскированное» от нас наилучшее будущее решение

Изменена сама постановка задачи RL!

"We study, if **generative trajectory modeling** – i.e. *modeling the joint distribution of the* ***sequence of states, actions, and rewards*** – can serve as a **replacement for conventional RL algorithms**."

Decision Transformer architecture. States, actions, and returns are fed into modality-specific linear embeddings and a positional episodic timestep encoding is added. Tokens are fed into a GPT architecture which predicts actions autoregressively using a causal self-attention mask.

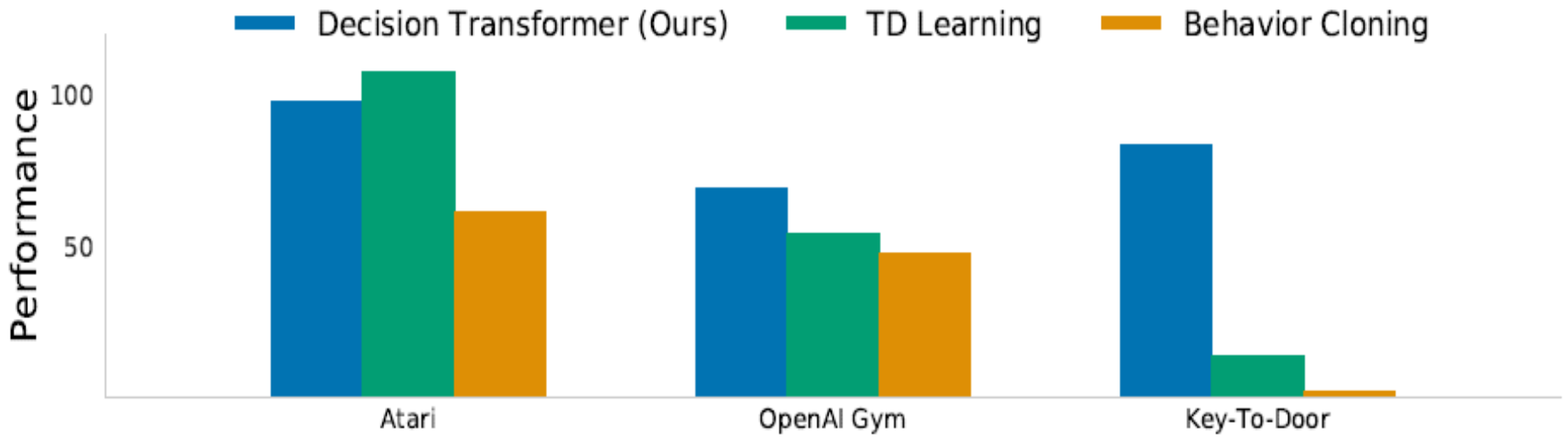
Decision Transformer: Reinforcement Learning via Sequence Modeling, *Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, Igor Mordatch*, 2021.

Decision Transformer

Трансформер для обучения с подкреплением

"We introduce a framework that **abstracts Reinforcement Learning (RL) as a sequence modeling problem**. Unlike prior approaches to RL that fit value functions or compute policy gradients, **Decision Transformer simply outputs the optimal actions by leveraging a causally masked Transformer.** By conditioning an autoregressive model on the **desired return (reward), past states, and actions**, our Decision Transformer model can **generate future actions** that achieve the desired return".

Очень может быть, что и в RL трансформеры также будут вскоре играть ключевую роль!



Decision Transformer matches or exceeds the performance of state-of-the-art model-free offline RL baselines on Atari, OpenAI Gym, and Key-to-Door tasks.

Decision Transformer: Reinforcement Learning via Sequence Modeling, *Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, Igor Mordatch*, 2021.

Open-Ended Learning (Открытое обучение)

2021: Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?

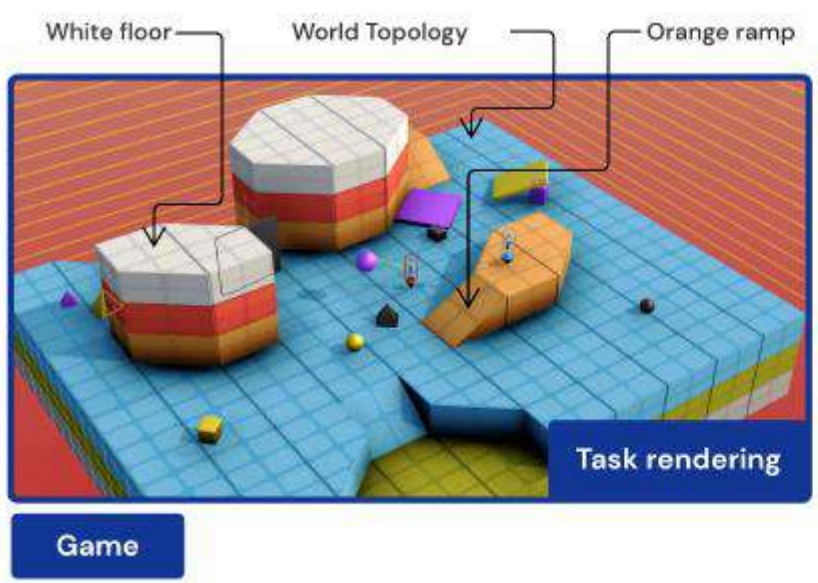
«В этой работе мы создаем **агентов, которые могут работать далеко за пределами одной, индивидуальной задачи.**

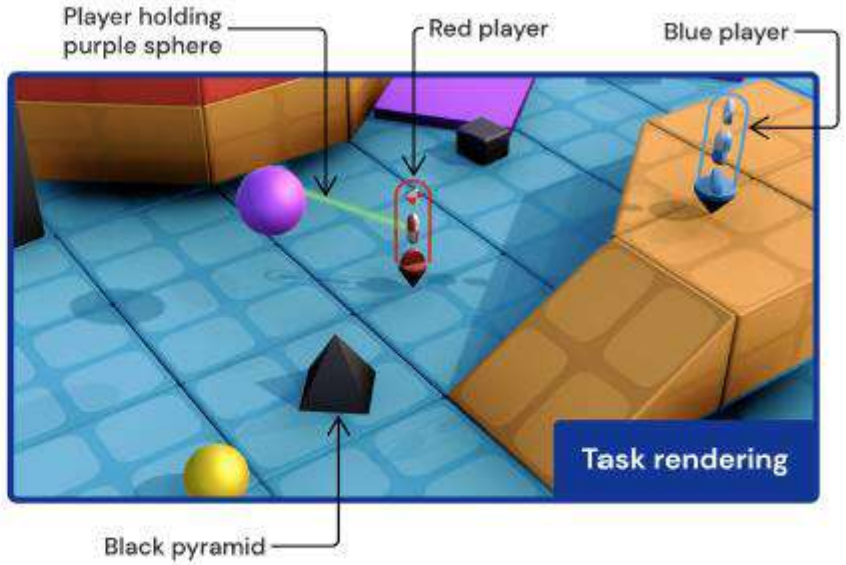
Для этого мы формируем целую **вселенную задач** и демонстрируем возможность **обучать агентов, которые смогут работать во всей этой огромной вселенной и за ее пределами.**»

Open-Ended Learning Leads to Generally Capable Agents, *Open-Ended Learning Team*, Adam Stooke, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, Nat McAleese, Nathalie Bradley-Schmieg, Nathaniel Wong, Nicolas Porcel, Roberta Raileanu, Steph Hughes-Fitt, Valentin Dalibard and Wojciech Marian Czarnecki,* **DeepMind, 2021.**

Open-Ended Learning

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?





Задача = Игра + Мир + Игроки

An instance of a **task** within the XLand environment space, composed of the **world** – the layout of the topology, initial object and player positions, and player gadgets – as well as the **game** – the specification of rewarding states for each player in this task.

«**Среда изначально является многоагентной**, охватывающей континуум конкурентных, кооперативных и независимых игр, которые расположены в искусственно сгенерированных мирах, моделирующих физику упрощенного 3D мира.

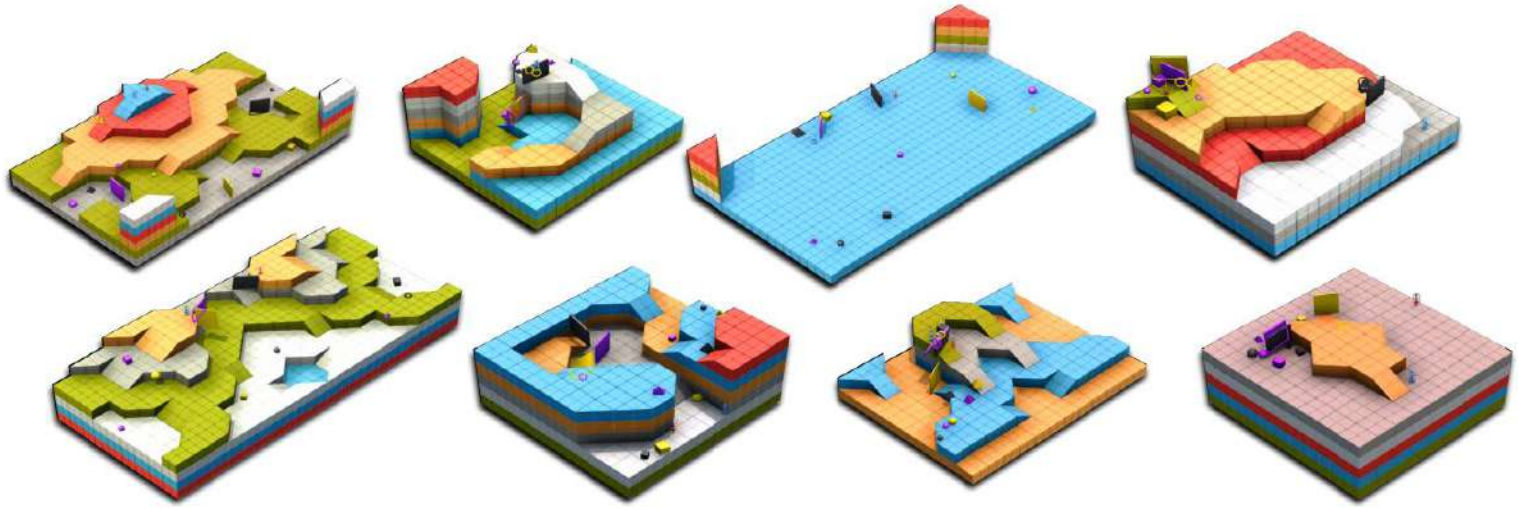
Пространство игр исключительно разнообразно с точки зрения задач, стоящих перед агентами, и **даже само измерение прогресса обучения агента является открытой исследовательской проблемой**.»

Open-Ended Learning Leads to Generally Capable Agents, *Open-Ended Learning Team*,* **2021.**

Open-Ended Learning

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?

Five examples of some of the 42 tasks in the **hand-authored evaluation task set**.



An example array of **worlds from the XLand** environment space.

«**Среда изначально является многоагентной**, охватывающей континуум конкурентных, кооперативных и независимых игр, которые расположены в искусственно сгенерированных мирах, моделирующих физику упрощенного 3D мира.

Пространство игр исключительно разнообразно с точки зрения задач, стоящих перед агентами, и **даже само измерение прогресса обучения агента является открытой исследовательской проблемой**.»

Open-Ended Learning Leads to Generally Capable Agents, *Open-Ended Learning Team*,* **2021.**

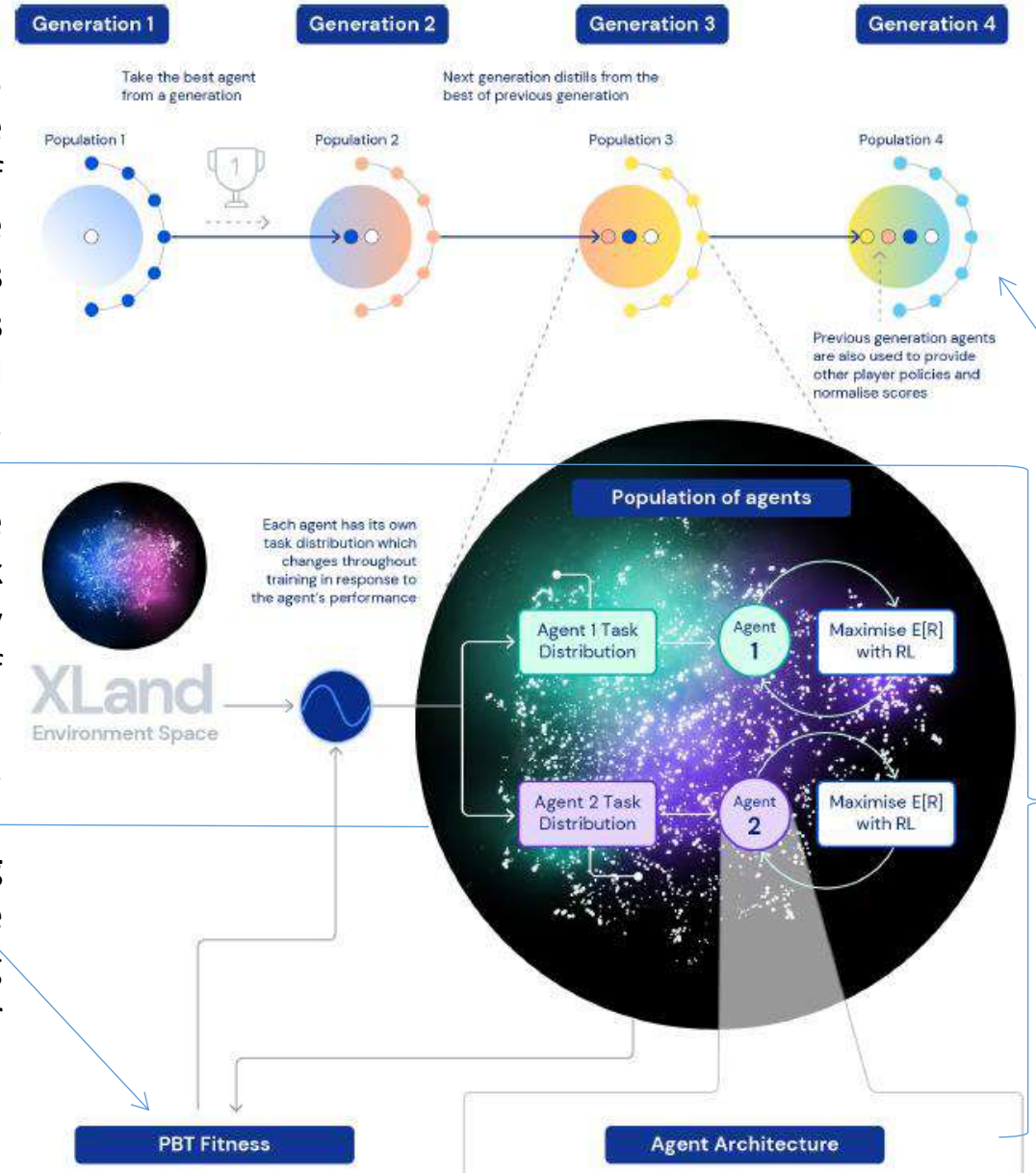
Open-Ended Learning

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?

Combined learning process. Generations of agents are trained, composed of populations of agents where the **best performing agents become distillation teachers of the next generation** as well as **co-players to train against**.

In each population, agents are trained with **dynamic task generation** that continuously adapts the distribution of training tasks Pπk (ℵ) for each agent πk.

Population based training (PBT) modulates the generation process by trying to Pareto dominate other agents with normalised percentiles metric.



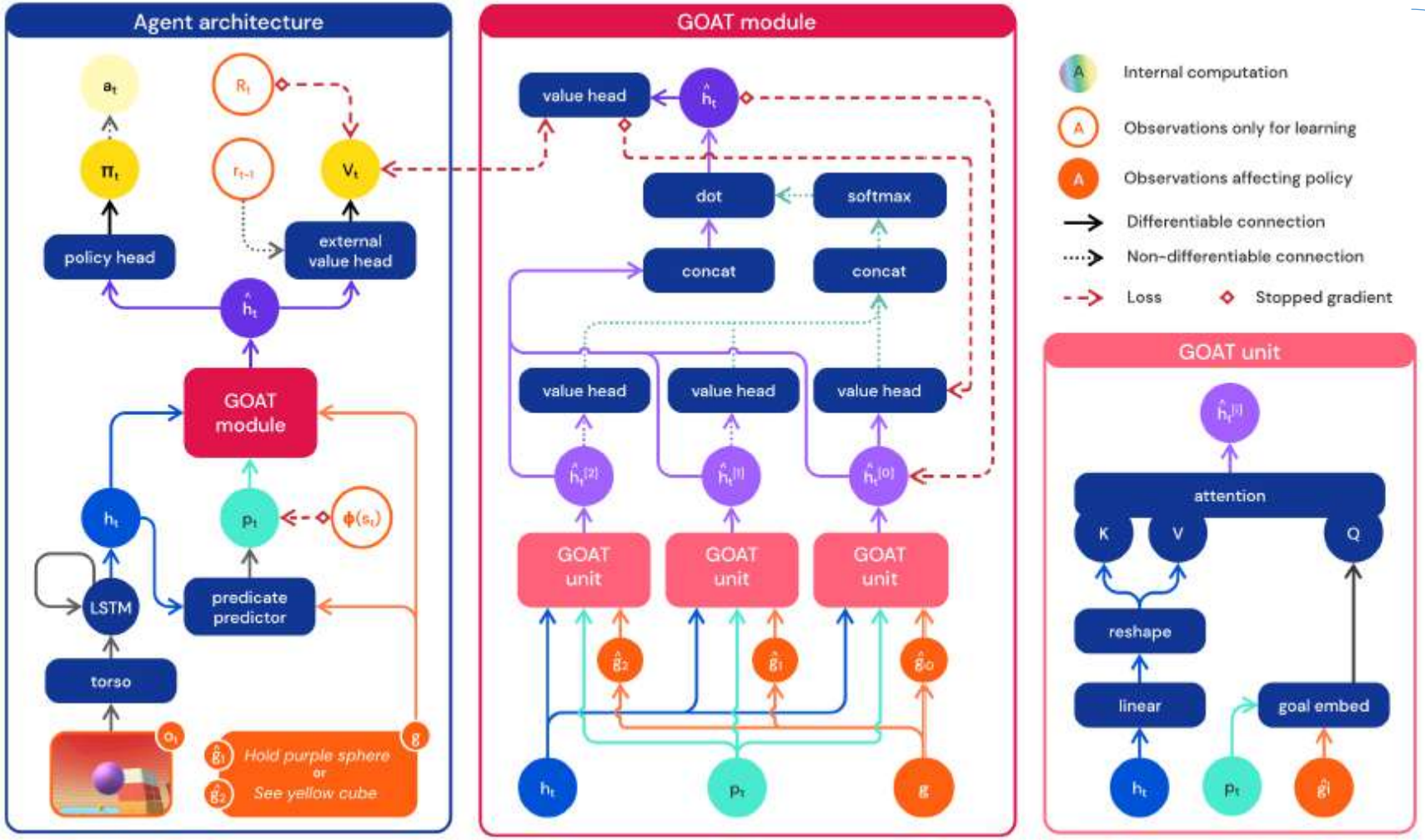
«Мы используем **итеративное улучшение между поколениями агентов**, а **не стремление к максимизации единственной цели**, позволяя **количественно оценить прогресс, несмотря на то, что задачи несравнимы с точки зрения достижимых вознаграждений**.

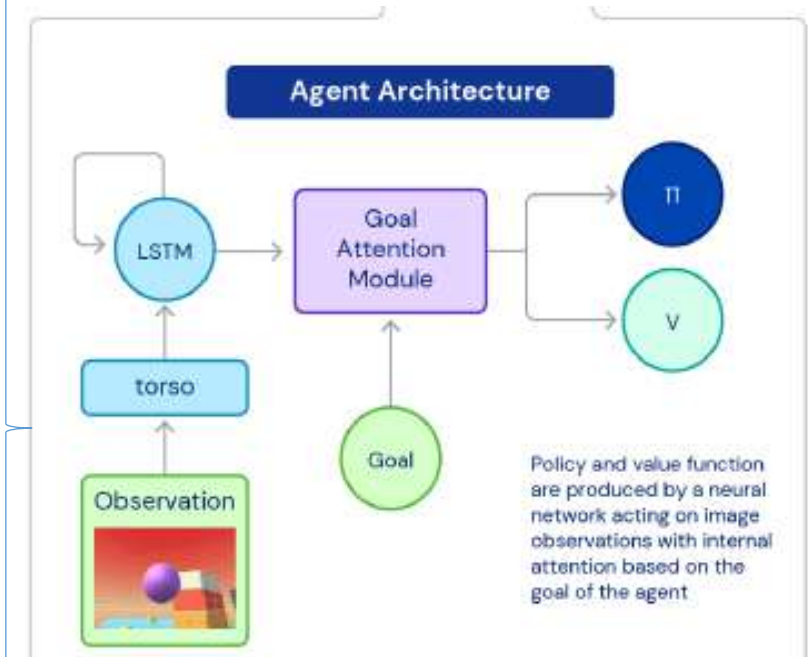
При огромном пространстве задач **чистое обучение с подкреплением на фиксированном распределении учебных задач не работает**. Необходим **открытый процесс обучения**, который динамически изменяет распределение задач обучения и цели обучения так, что **агент никогда не прекращает обучение** и **учится учиться решению неизвестных задач в процессе их решения**.»

Open-Ended Learning Leads to Generally Capable Agents, *Open-Ended Learning Team*,* **2021.**

Open-Ended Learning

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?





Как это сделано: LSTM + goal-ATTN

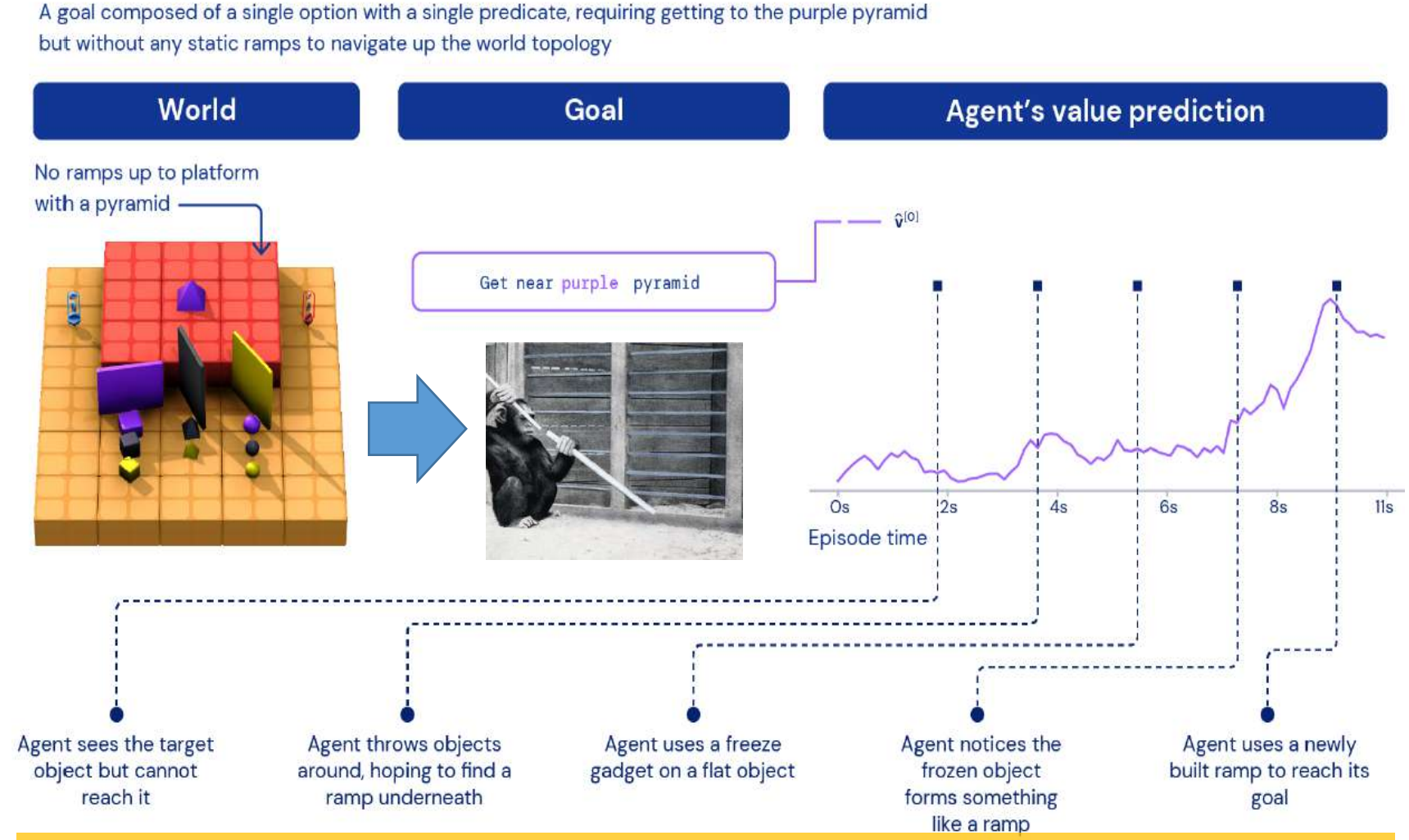
Each agent trains with deep reinforcement learning and consists of a **neural network producing the policy π and value function v.** The input observations consist of images, proprioception and the goal.

The agent processes the observations through the **torso** and a **recurrent core** to produce h_t , which is used for the predicate predictor, producing π_t . **GOAT module** performs logical analysis of the goal using value consistency.

Open-Ended Learning Leads to Generally Capable Agents, *Open-Ended Learning Team*,* **2021.**

Open-Ended Learning

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?



Пример: для достижения новой цели в новом окружении агент начинает использовать предметы окружения как орудия

«Полученные в результате агенты обладают успешным **поведением, обобщающимся на многие точки во вселенной задач**. Примеры такого zero-shot обобщения демонстрируют хорошую производительность в играх Hide and Seek, Capture the Flag и Tag.

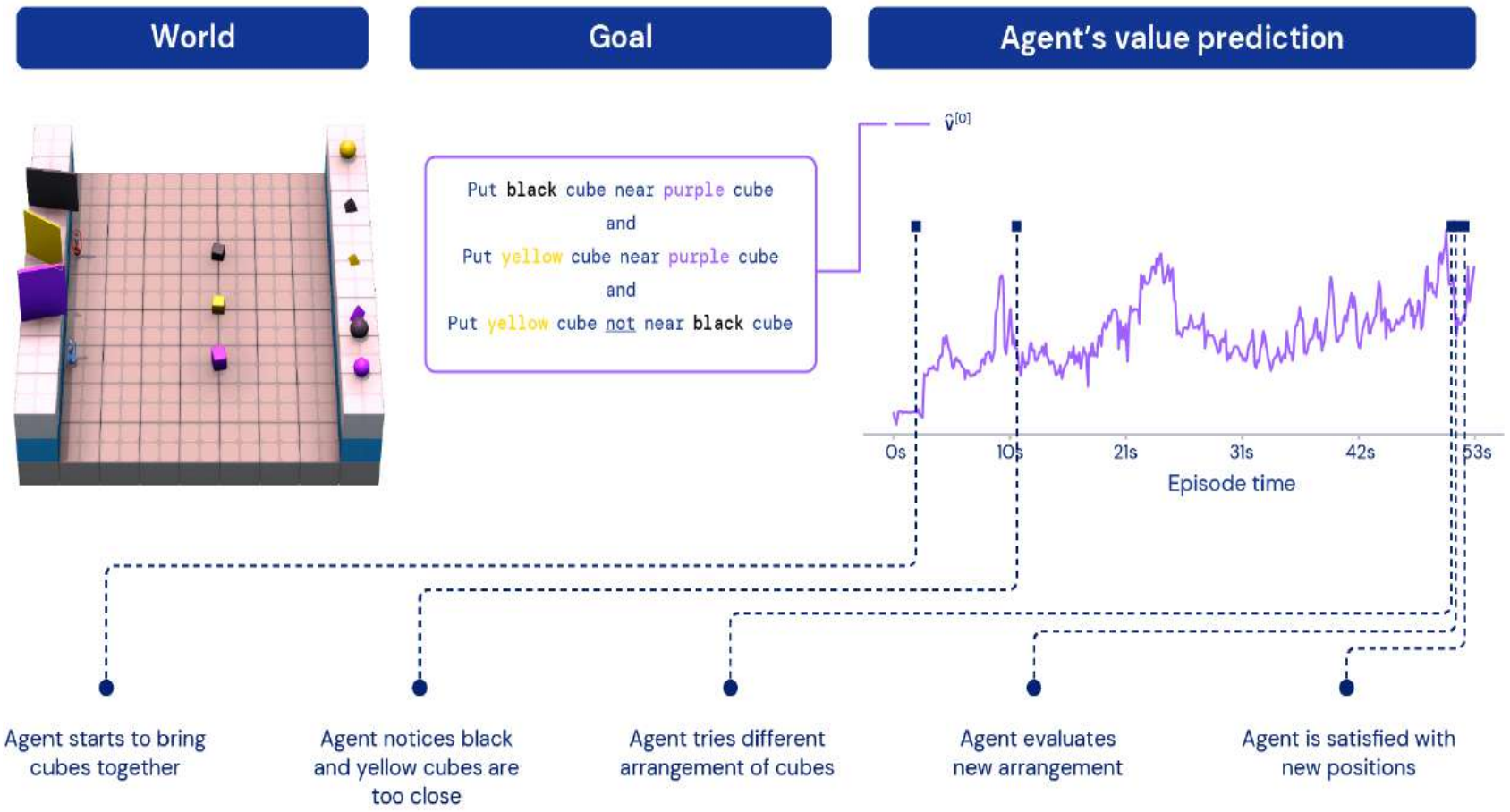
С помощью анализа кривых обучения и созданных специально тестовых задач мы исследовали поведение наших агентов и обнаружили интересные новые эвристические **модели поведения**, такие как **экспериментирование методом проб и ошибок, создание и использование инструментов, переключение стратегий и сотрудничество**.»

Open-Ended Learning Leads to Generally Capable Agents, *Open-Ended Learning Team*,* **2021.**

Open-Ended Learning

Открытое обучение учит не задачам, а когнитивному поведению!

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?



Вывод: Такие модели поведения (экспериментирование, использование инструментов, сотрудничество**) характерны для когнитивного поведения людей и животных и необходимы для самообучающихся роботов.**

Пример: при решении логической головоломки агент использует эксперименты (метод проб и ошибок). Этому его не учили!

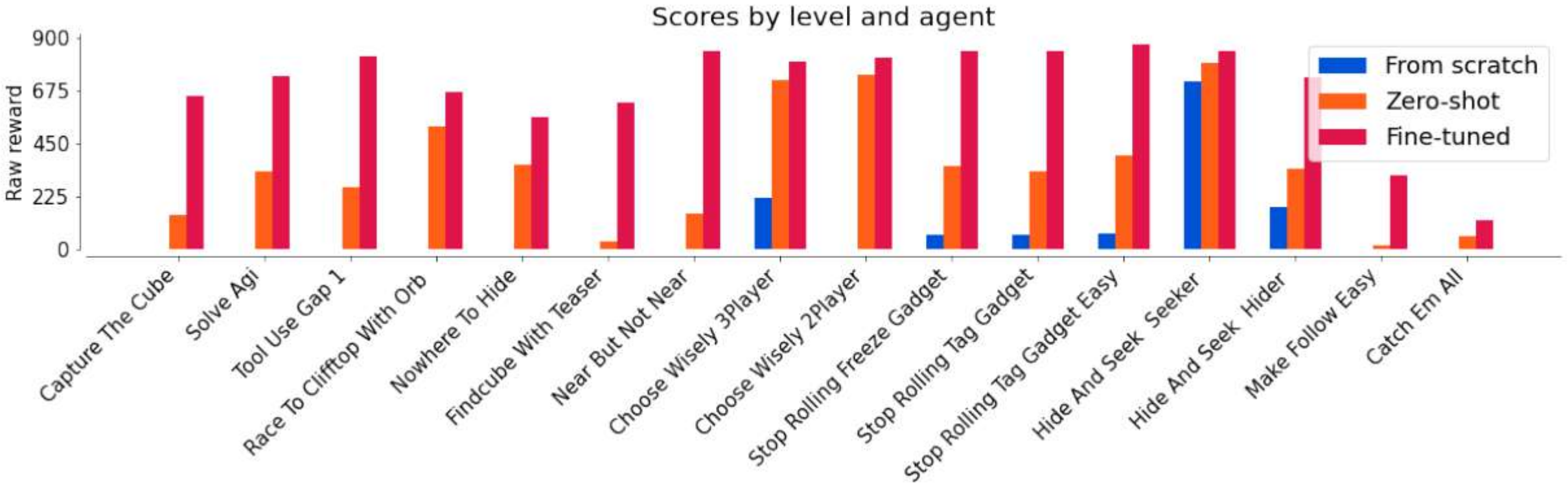
Open-Ended Learning Leads to Generally Capable Agents, *Open-Ended Learning Team*,* **2021.**

Open-Ended Learning

Открытое обучение учит не задачам, а когнитивному поведению!

Но учиться (дообучаться) на конкретную задачу все равно нужно!

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?



Comparison of three agents from different training regimes on a range of hand-authored levels. **Scratch:** An agent trained from scratch for 100 million steps. **Zero-shot:** the agent trained using our methodology and evaluated on these held out levels zero-shot. **Fine-tuned:** the same agent but trained for an additional 100 million steps on the level. 100 million steps is equivalent to 30 minutes of wall-clock time in our setup. **This rapid finetuning improves the agent score significantly compared to zero-shot, and in the majority of cases training from scratch does not achieve any reward**.

Open-Ended Learning Leads to Generally Capable Agents, *Open-Ended Learning Team*,* **2021.**

Close-Ended Learning vs. Open-Ended Learning

Произойдет ли смена парадигмы в обучении? Если да, то когда?

Можно ли научить ИИ решать заранее неизвестные задачи в совершенно новой обстановке?

Close-Ended RL

Цель обучения: обучение конкретным навыкам

Процесс обучения

Обучение с учителем

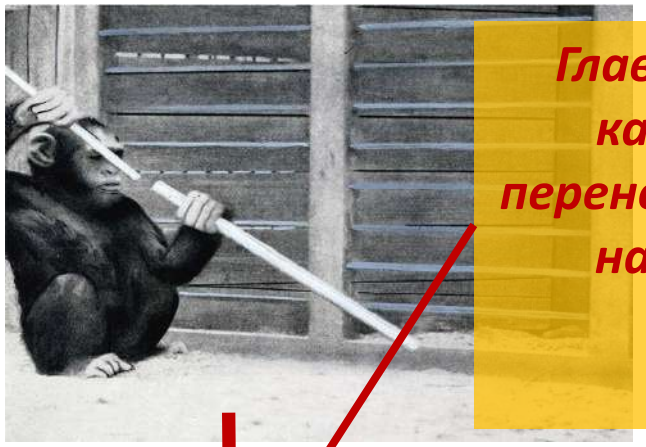
Обучение с подкреплением и генетический отбор

Исходные данные и форма их представления

Задача представлена обучающей выборкой. Требуется набрать большой объем примеров действий людей.

Open-Ended RL

Цель обучения: обучение когнитивному поведению



Главная проблема – как осуществить перенос когнитивных навыков из «мира кубиков» в мир реальных задач

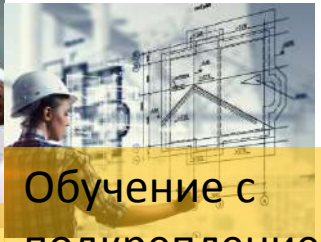
Процесс обучения

(I) Обучение когнитивному поведению

Обучение с подкреплением и генетический отбор

Обучение с учителем

(II) Изучение задачи



Обучение с подкреплением и генетический отбор

(III) Когнитивный поиск решения

Исходные данные и форма их представления

Задача представлена текстовым (формальным) описанием (**zero-shot**) и небольшой обучающей выборкой примеров (**дообучение**). Нужна база знаний предметной области (**ИИ1 + ИИ2!**).

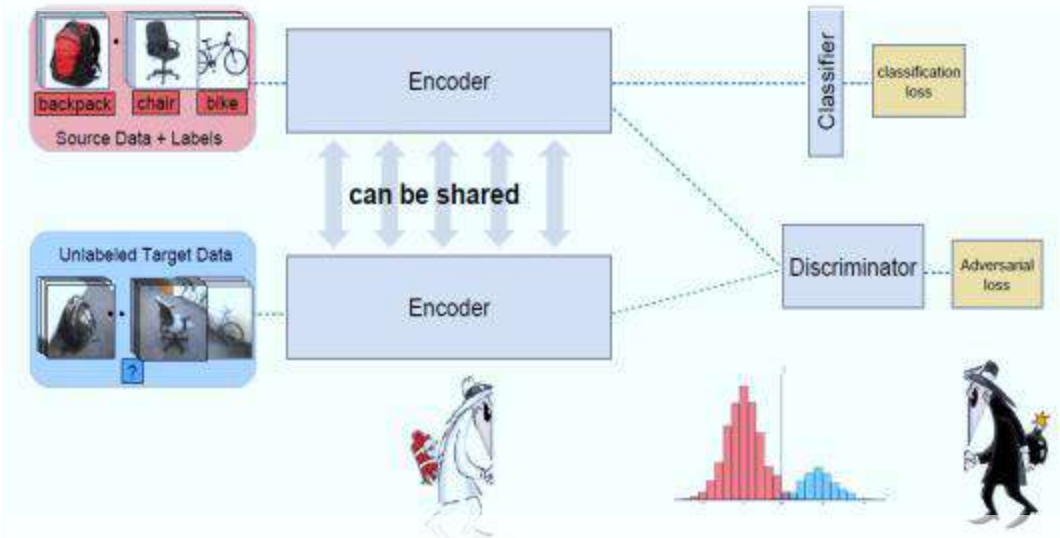
Data Generation:
GAN → Diffusion models

Наводка: Dmitry Vetrov, Head of BayesGroup Professor at HSE Leading researcher at AIRI, OpenTalks.AI, 17.02.2022

Generative models

- GAN
  - Realistic samples
  - Suffer from mode-collapse
- Normalizing flow
  - Explicit model
  - Tend to overfit training data
  - Not flexible enough
- VAE
  - Covers training sample
  - Unrealistic samples (e.g. blurry images)

Генератор создает визуальные образы, стараясь обмануть Дискриминатор…



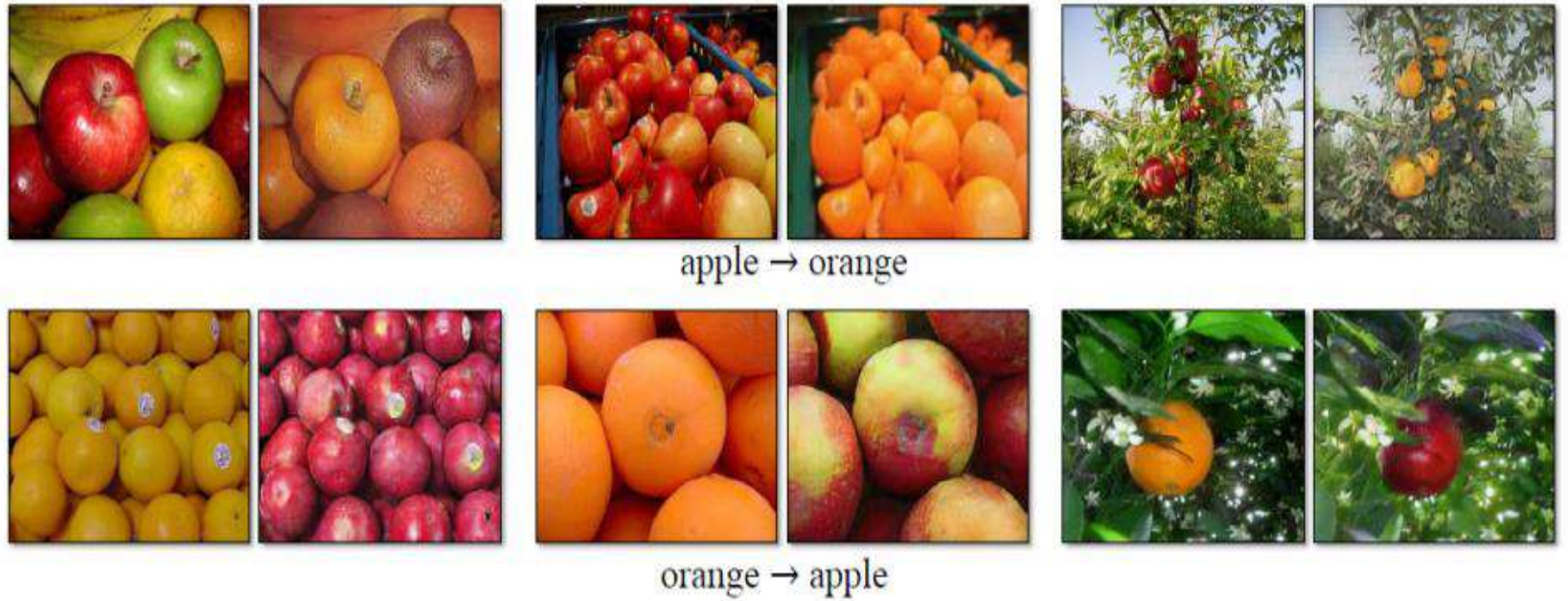
….Дискриминатор старается отличить фантазии Генератора от реальности

Dmitry Vetrov, OpenTalks.AI/2022

Adaptive Deep Models across Visual Domains, Kate Saenko, ICCV, 2017

Generative models

- GAN
  - Realistic samples
  - Suffer from mode-collapse
- Normalizing flow
  - Explicit model
  - Tend to overfit training data
  - Not flexible enough
- VAE
  - Covers training sample
  - Unrealistic samples (e.g. blurry images)

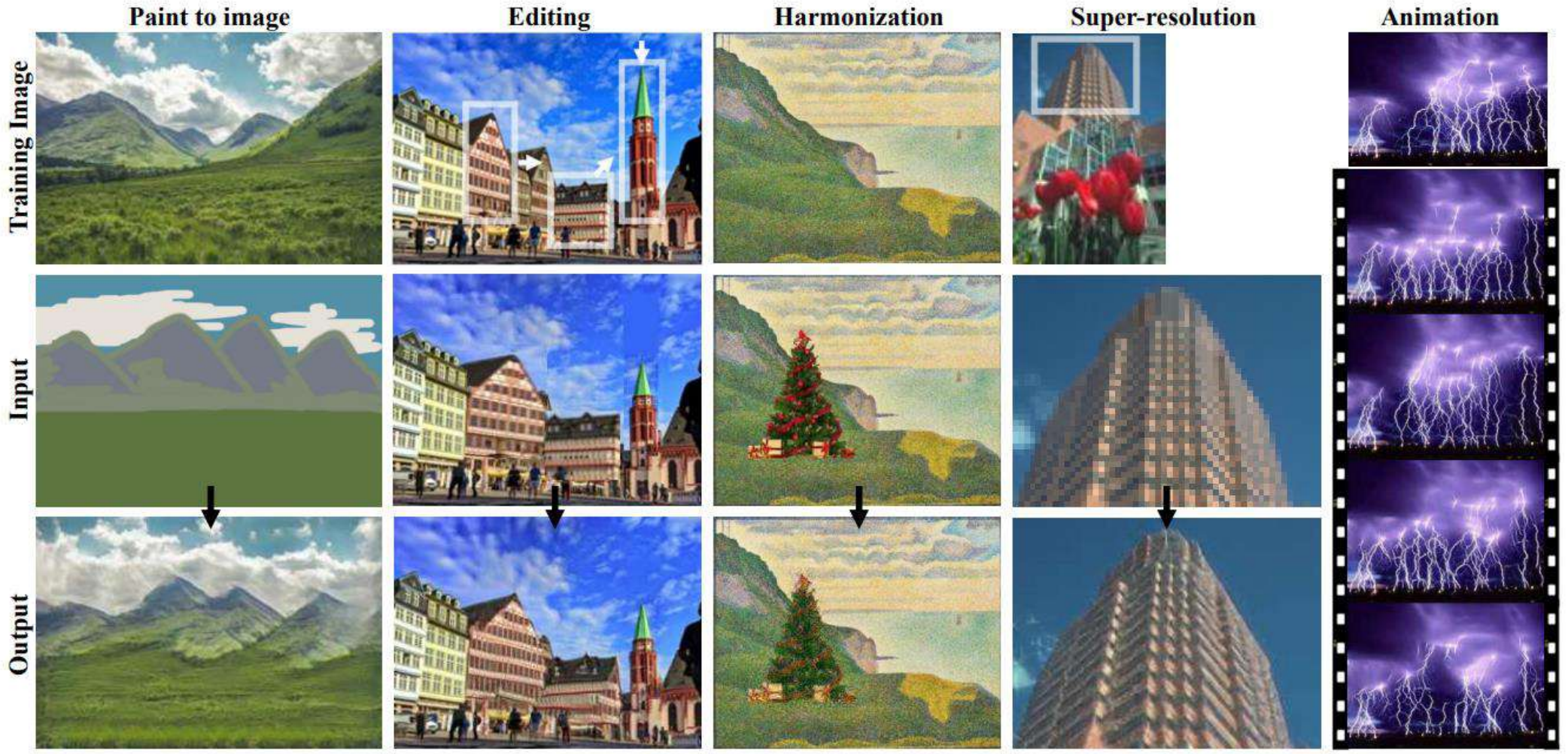


Неоспоримое лидерство GAN до 2020…

Dmitry Vetrov, OpenTalks.AI/2022

Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, Jun-Yan Zhu et al., ICCV, 2017

SinGAN: Learning a Generative Model from a Single Natural Image

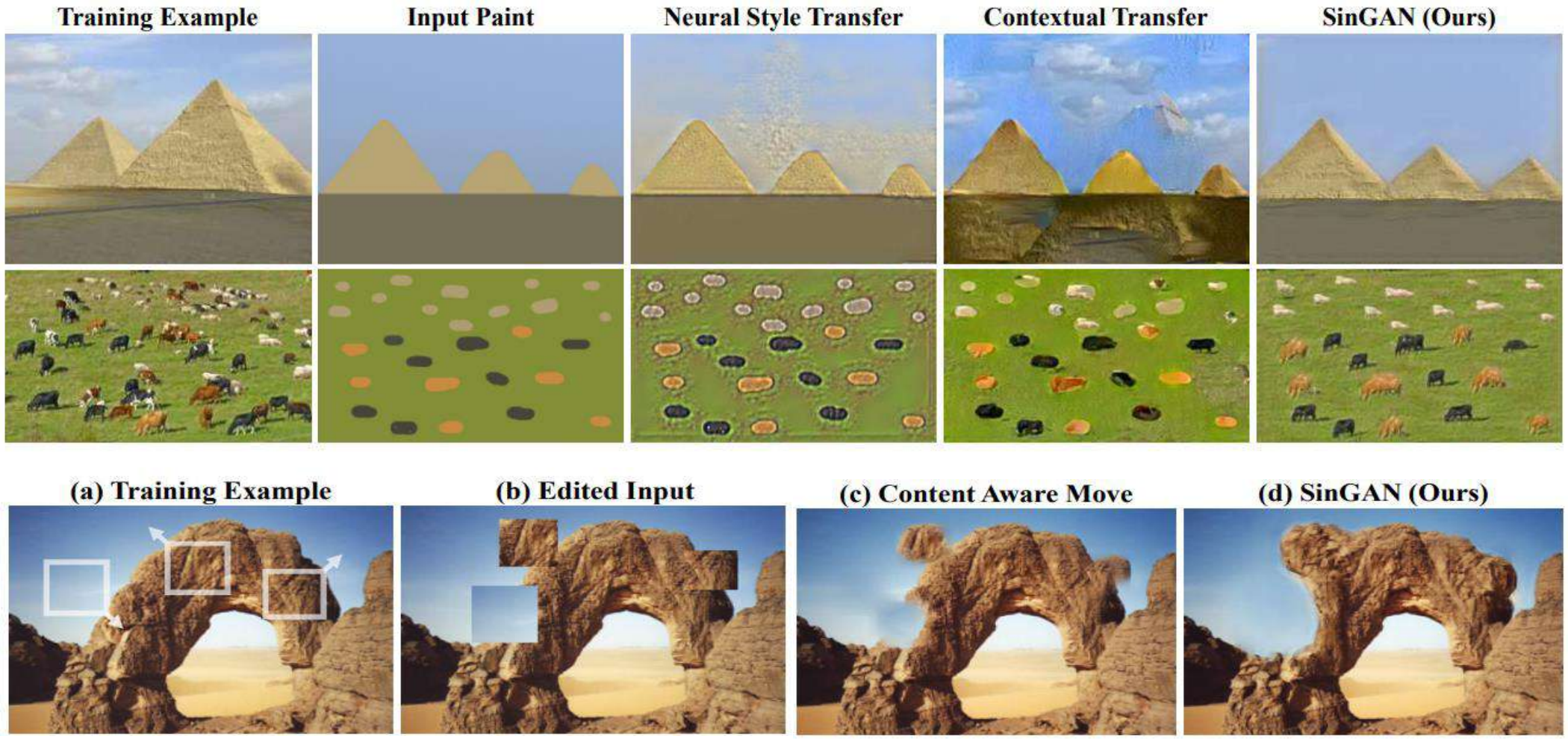


Для обучения реалистичного генератора изображений больше не нужны большие базы примеров!

SinGAN: Learning a Generative Model From a Single Natural Image, ICCV-2019,
Tamar Rott Shaham, Tali Dekel, Tomer Michaeli

ICCV 2019
Seoul, Korea

SinGAN: Learning a Generative Model from a Single Natural Image



Практически неограниченные возможности реалистичного манипулирования данными…

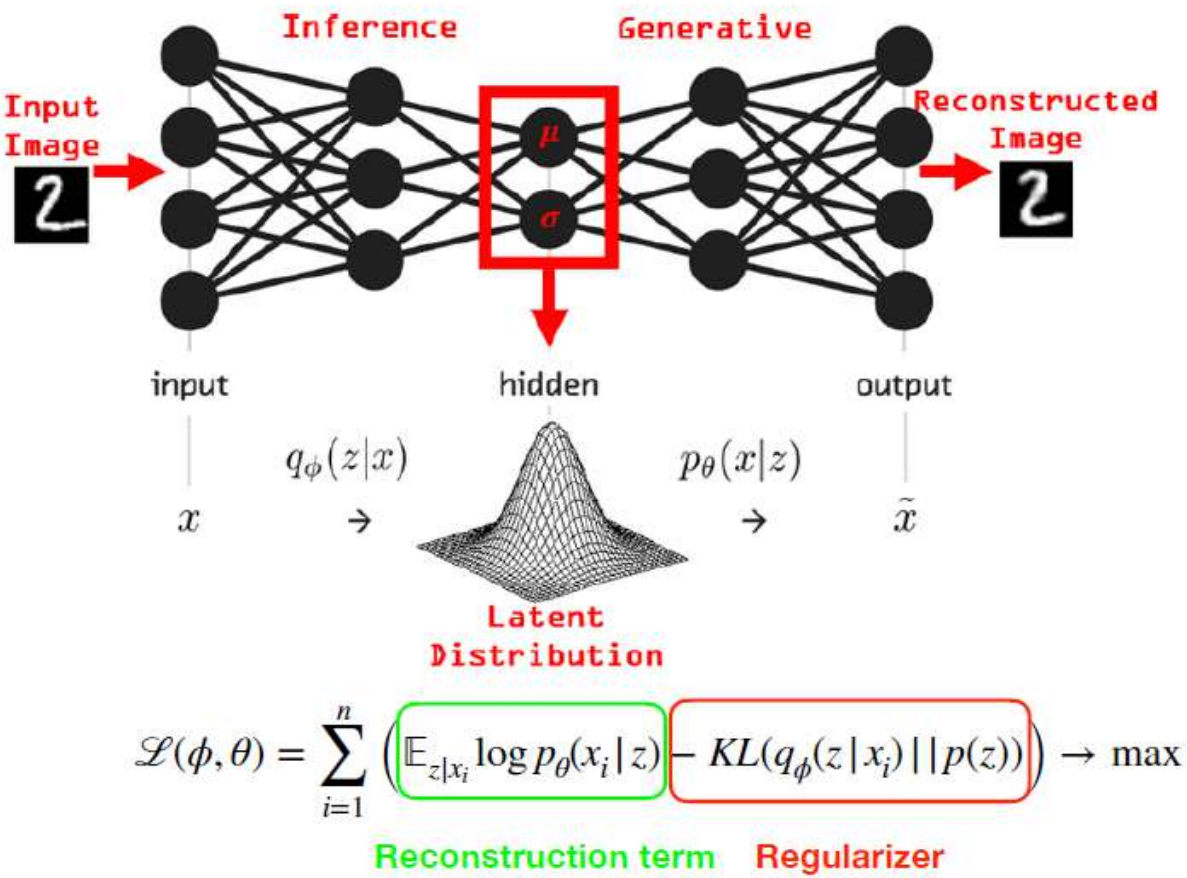
SinGAN: Learning a Generative Model From a Single Natural Image, ICCV-2019,
Tamar Rott Shaham, Tali Dekel, Tomer Michaeli

Неоспоримое лидерство GAN до 2020…

Generative models

- GAN
  - Realistic samples
  - Suffer from mode-collapse
- Normalizing flow
  - Explicit model
  - Tend to overfit training data
  - Not flexible enough
- VAE
  - Covers training sample
  - Unrealistic samples (e.g. blurry images)

VAE



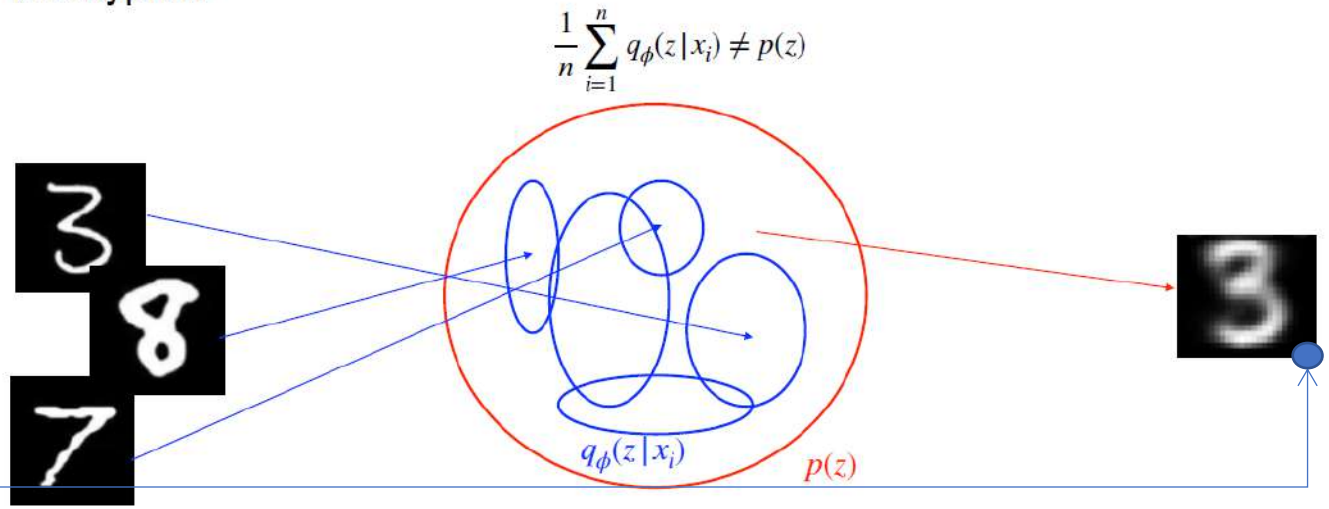
2020: VAE?

Generative models

- GAN
  - Realistic samples
  - Suffer from mode-collapse
- Normalizing flow
  - Explicit model
  - Tend to overfit training data
  - Not flexible enough
- VAE
  - Covers training sample
  - Unrealistic samples (e.g. blurry images)

VAE

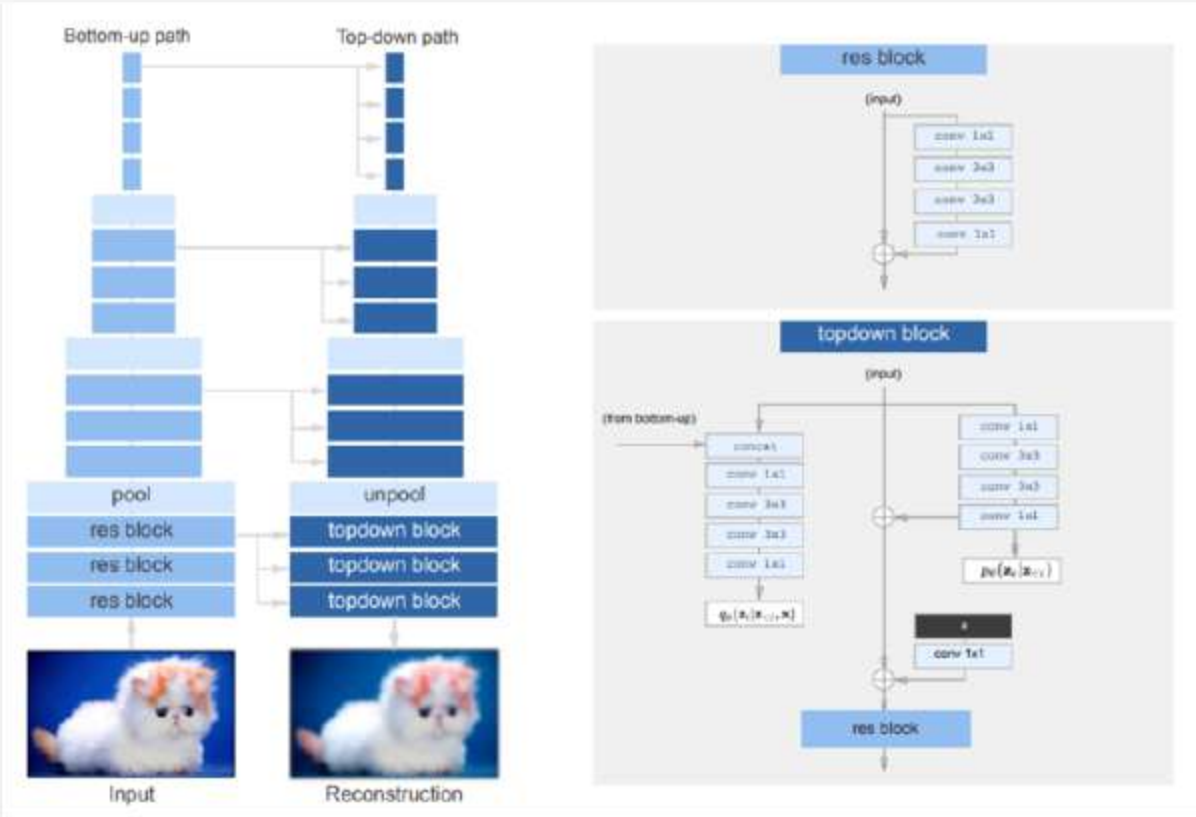
- In VAE we can ensure good reconstructions but samples from prior may be atypical



2020: VAE?

Very Deep VAE

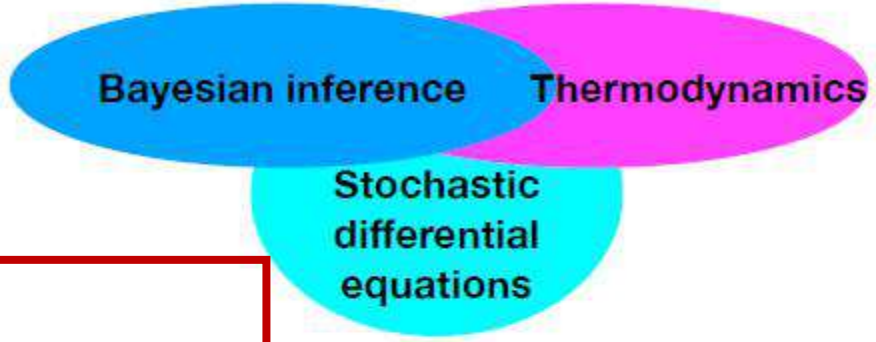
- Very deep VAE achieved SOTA on image synthesis (Child20)
- They beat auto-regressive models
- The dimension of latents is much larger than the dimension of observed variables
- VD VAE has hierarchical structure

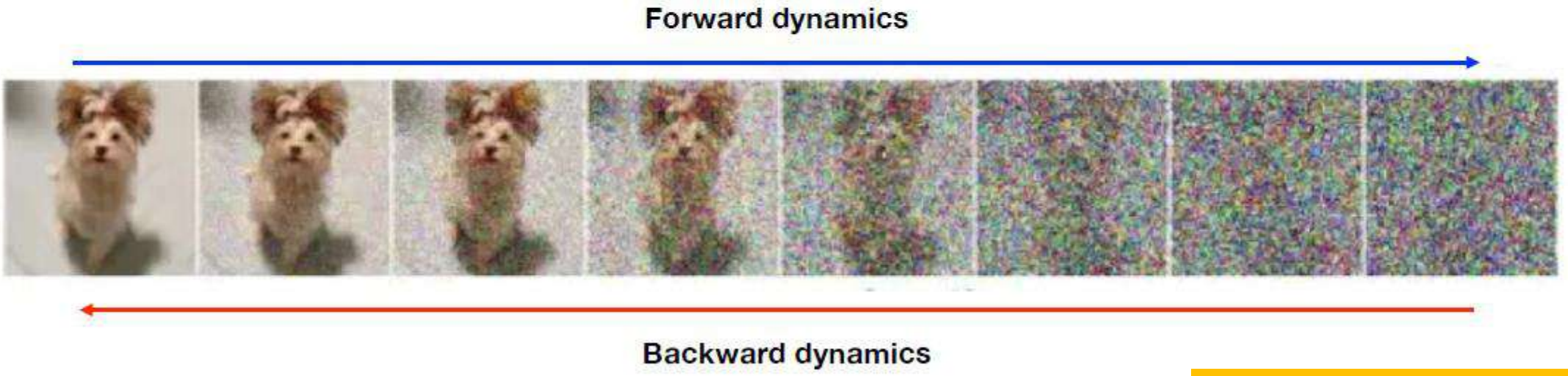


2020: VAE?

Diffusion models

- A breakthrough in generative modeling (Ho20)
- The object is gradually noisified until white noise is obtained
- Reverse process allows to generate objects from noise
- Each object from training set is mapped to **exactly the same distribution**





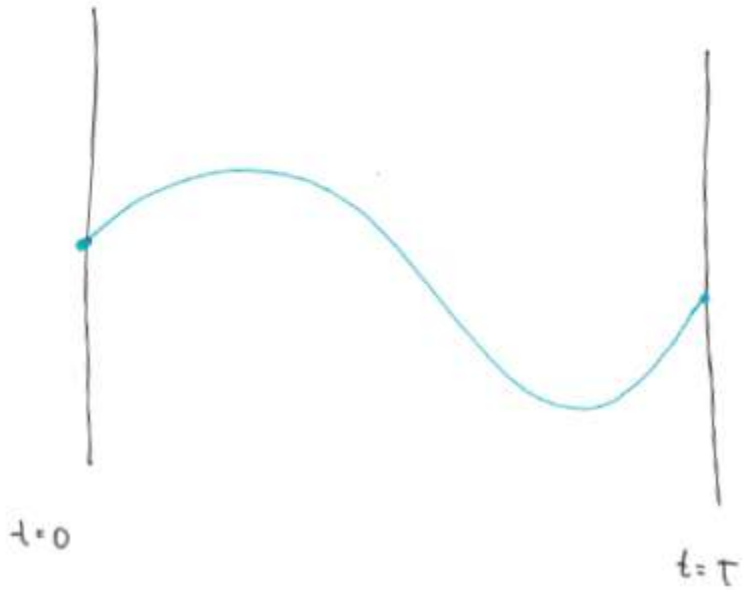
2020: Diffusion Models?

Diffusion models

Stochastic differential equations

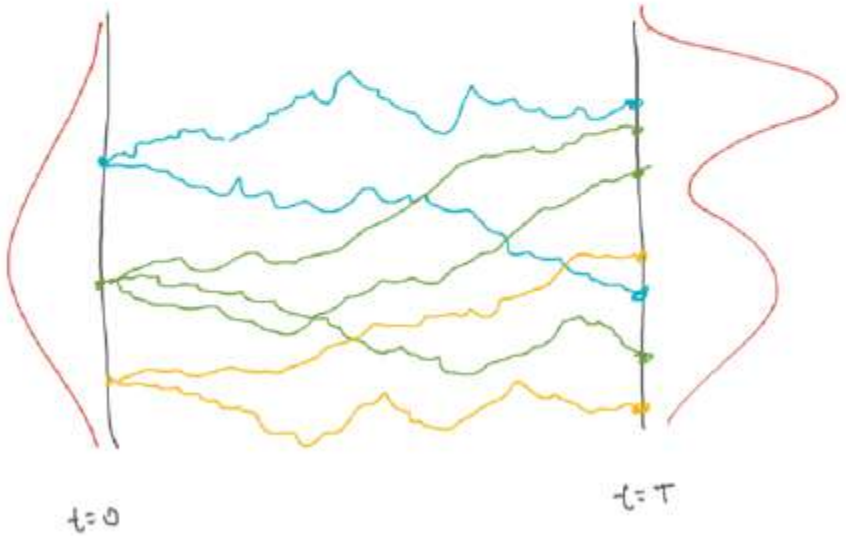
- Ordinary differential equation defines the evolution of a point

$$dx = f(x,t)dt$$



- Stochastic differential equation defines the evolution of a distribution

$$dx = f(x,t)dt + g(x,t)dW$$



2020: Diffusion Models?

Diffusion models

Backward dynamics

- Consider SDE that maps some complex distribution to white noise

$$dx = \alpha x dt + dW, \quad \alpha < 0, \quad x_0 \sim p_0(x)$$

- Let $p_t(x)$ be a distribution at time t

- Is it possible to reconstruct initial distribution from noise?
- Yes! By using backward dynamics SDE (Song20)

$$dx = [\alpha x - s(x,t)]dt + dW, \quad \alpha < 0, \quad x_T \sim p_T(x)$$

- Note that backward SDE essentially uses **the score**

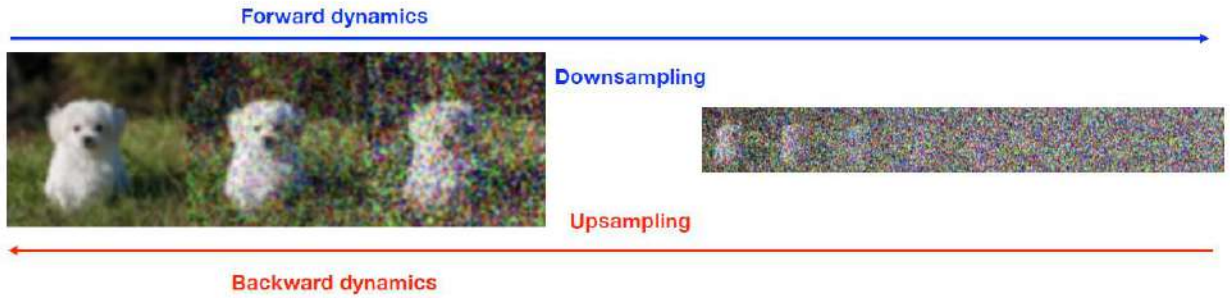
$$s(x,t) = \frac{\partial}{\partial x} \log p_t(x)$$

2020: Diffusion Models?

Diffusion models

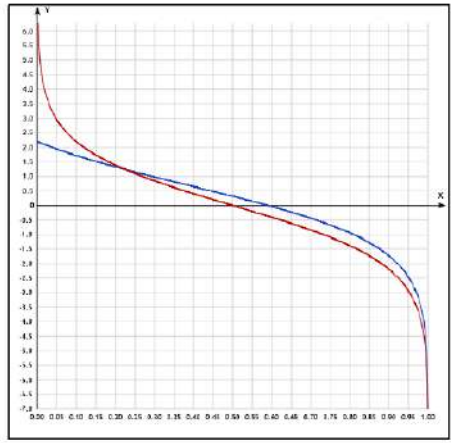
Reduction in size

- High resolution in very noisy image is useless
- Reduction in size significantly accelerates both dynamics
- But when one should downsample?



Downsampling

- The information is gradually lost during diffusion process
- Can measure the loss in terms of SNR
- Downsampling
  1. loses some signal because of dimension reduction
  2. reduces the noise due to averaging of patches
- Can make make multiple downsampling for further acceleration



Conclusion

- **Diffusion models seems to be a new SOTA generative framework**
- The bottleneck is inference time
- Dimension reduction is one (of many) way to speed up generation

Paradigm	Quality	Diversity	Speed
VAE	✗	✓	✓
GAN	✓	✗	✓
Diffusion	✓	✓	✗

2020+: Diffusion Models!

Diffusion models

References

• D. Kingma, M. Welling. Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114, 2013.

• R. Child. Very deep vaes generalize autoregressive models and can outperform them on images.arXiv preprint arXiv:2011.10650, 2020.

• J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239, June 2020.

• Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020.

Denoising diffusion probabilistic models

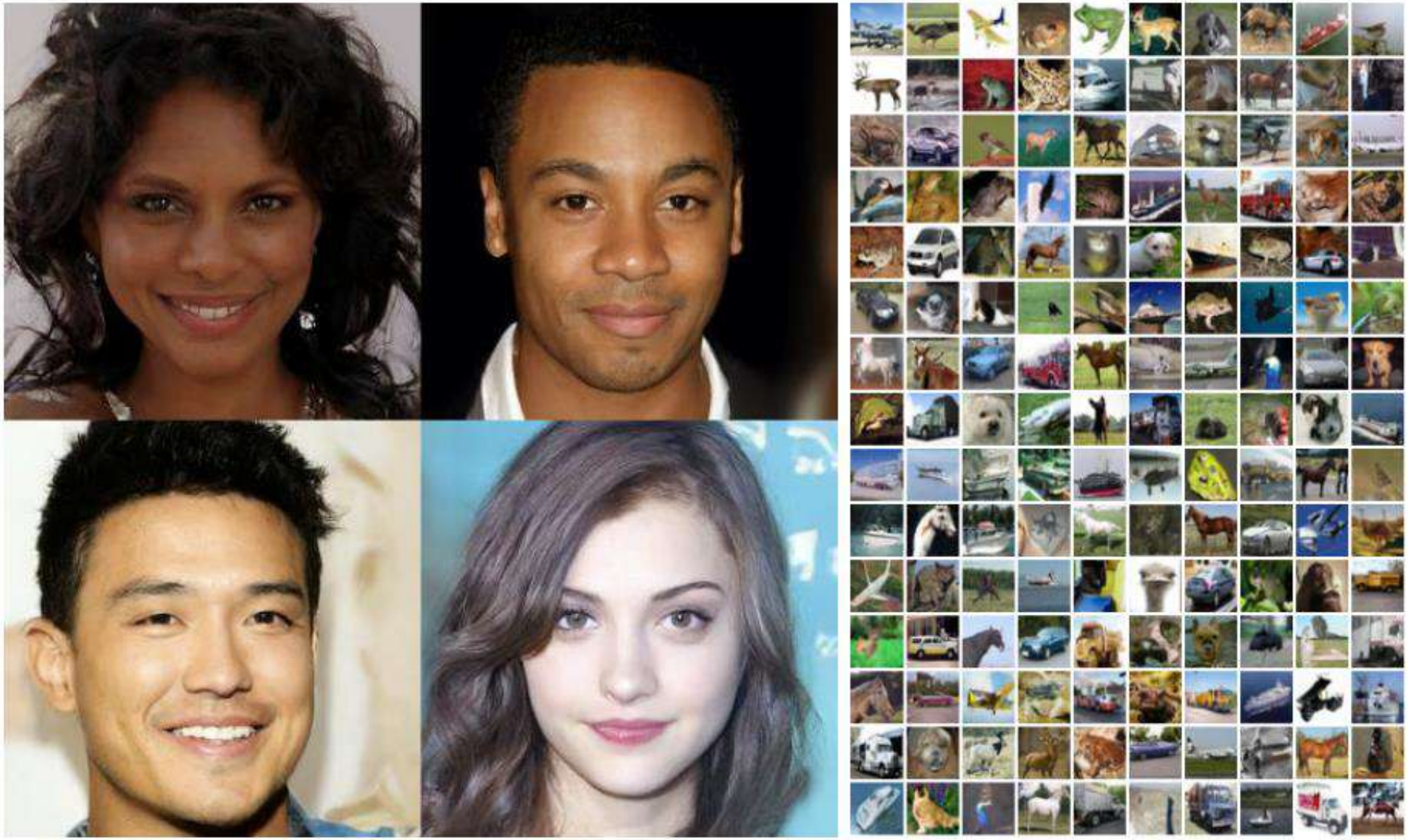
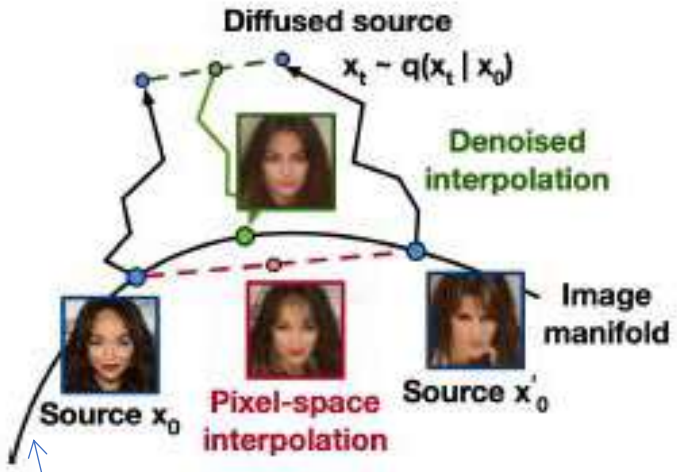


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020

Denoising diffusion probabilistic models

Интерполяция между двумя лицами



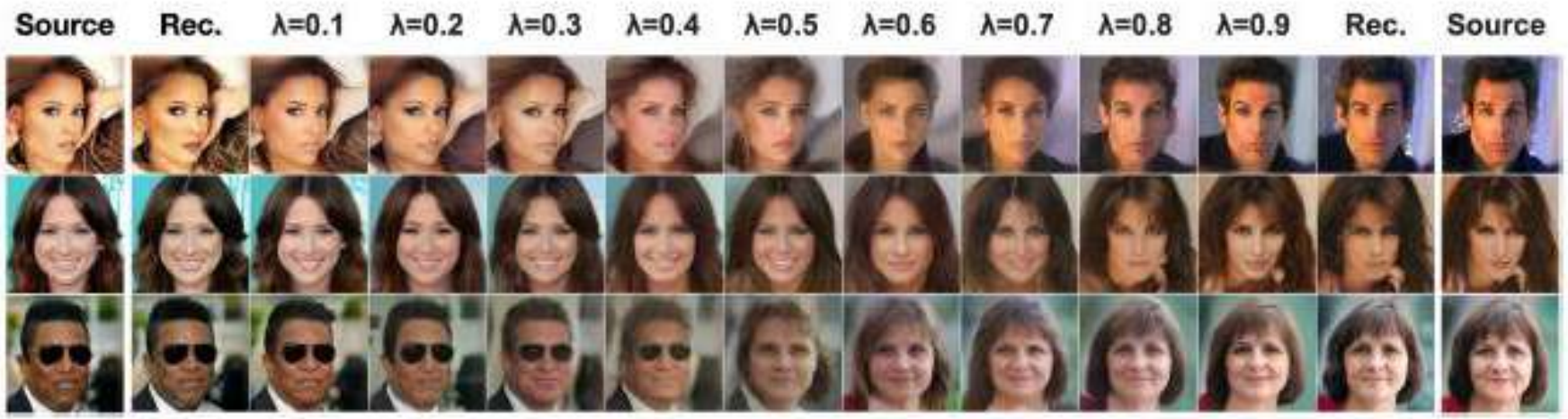


Figure 8: Interpolations of CelebA-HQ 256x256 images with 500 timesteps of diffusion.

Непрерывное пространство изображений (manifold) (распределение)

Coarse-to-fine interpolation Figure 9 shows interpolations between a pair of source CelebA 256×256 images as we vary the number of diffusion steps prior to latent space interpolation. Increasing the number of diffusion steps destroys more structure in the source images, which the model completes during the reverse process. This allows us to interpolate at both fine granularities and coarse granularities. In the limiting case of 0 diffusion steps, the interpolation mixes source images in pixel space. On the other hand, after 1000 diffusion steps, source information is lost and interpolations are novel samples.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020

Denoising diffusion probabilistic models

Интерполяция между двумя лицами

Непрерывное пространство изображений (manifold) (распределение)

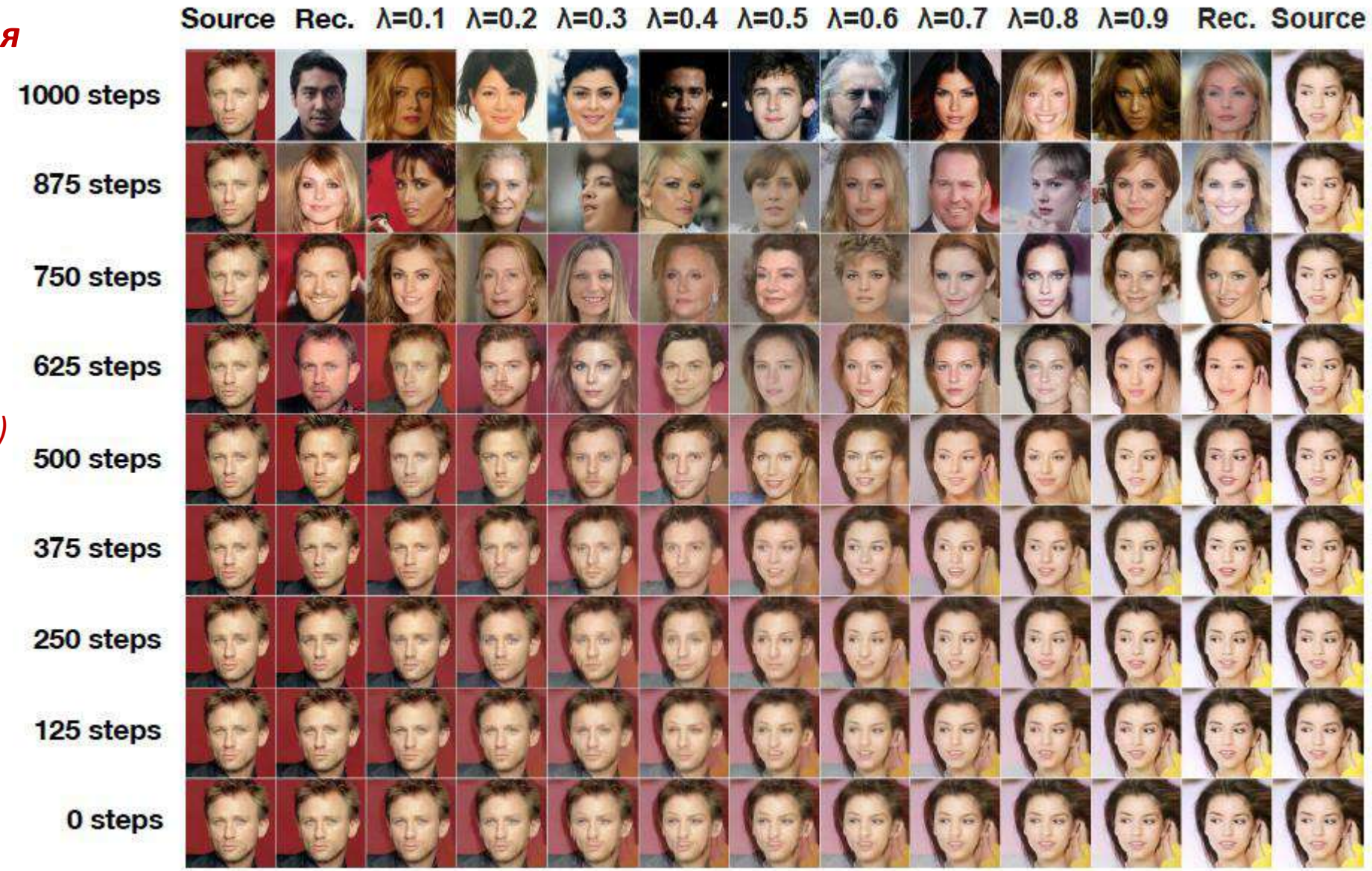


Figure 9: Coarse-to-fine interpolations that vary the number of diffusion steps prior to latent mixing.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020

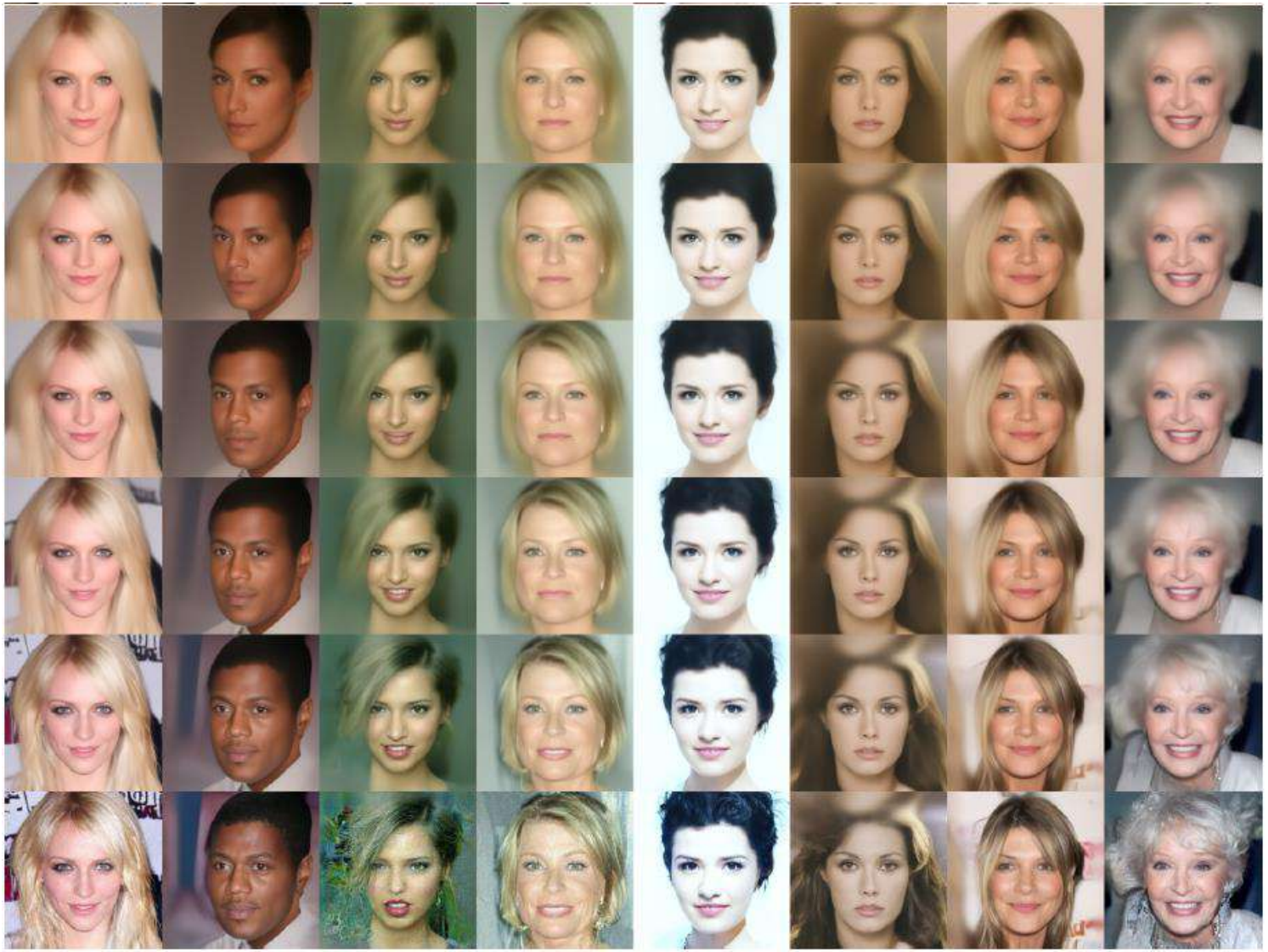
Score-based generative modeling through stochastic differential equations

Samples on 1024 × 1024 CelebA-HQ

Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations, ICLR 2021

Score-based generative modeling through stochastic differential equations

Изменение «температуры» изображения как нормы эмбеддинга

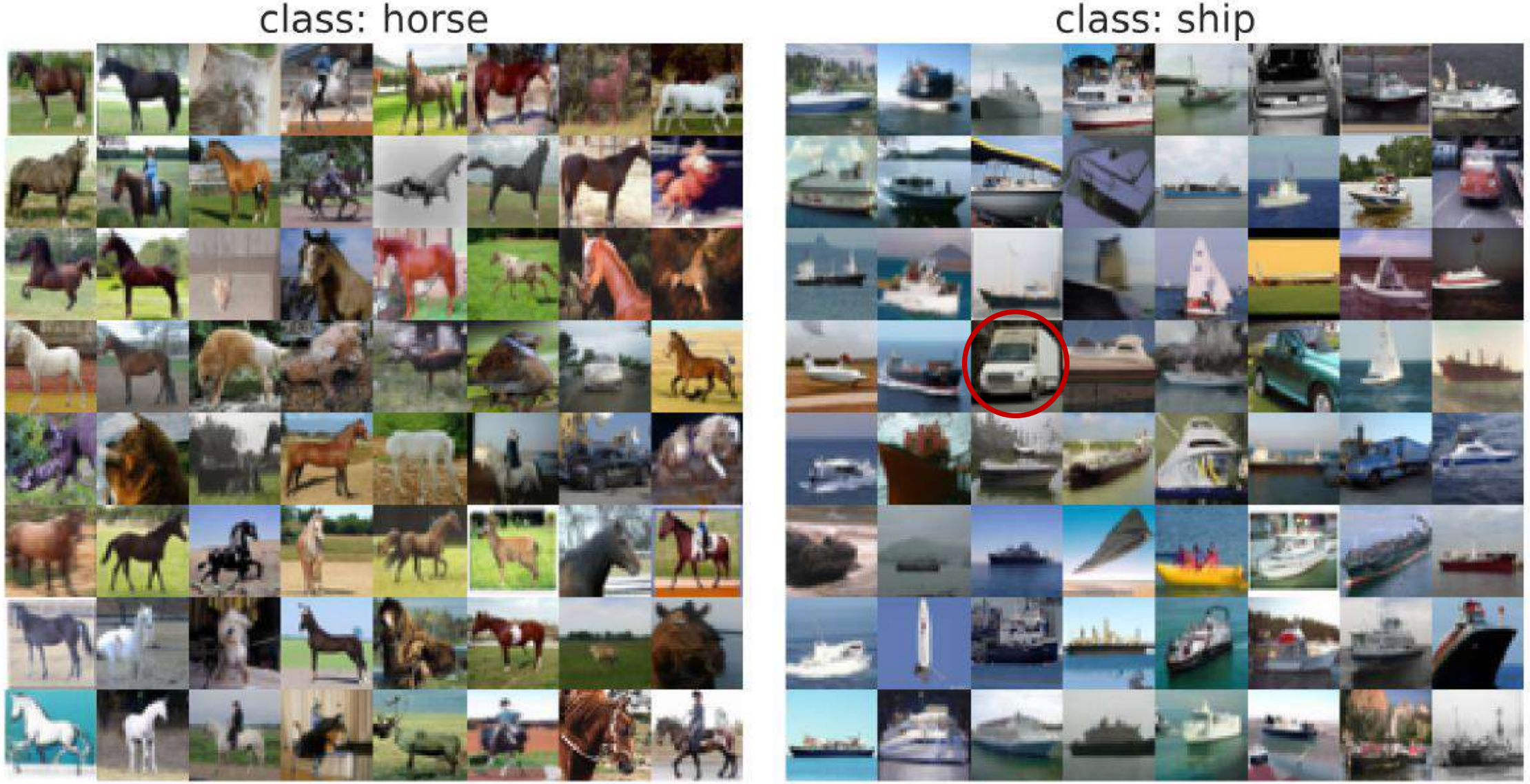


Непрерывное пространство изображений (manifold) (распределение)

temperature rescaling (reducing norm of embedding).

Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations, ICLR 2021

Score-based generative modeling through stochastic differential equations

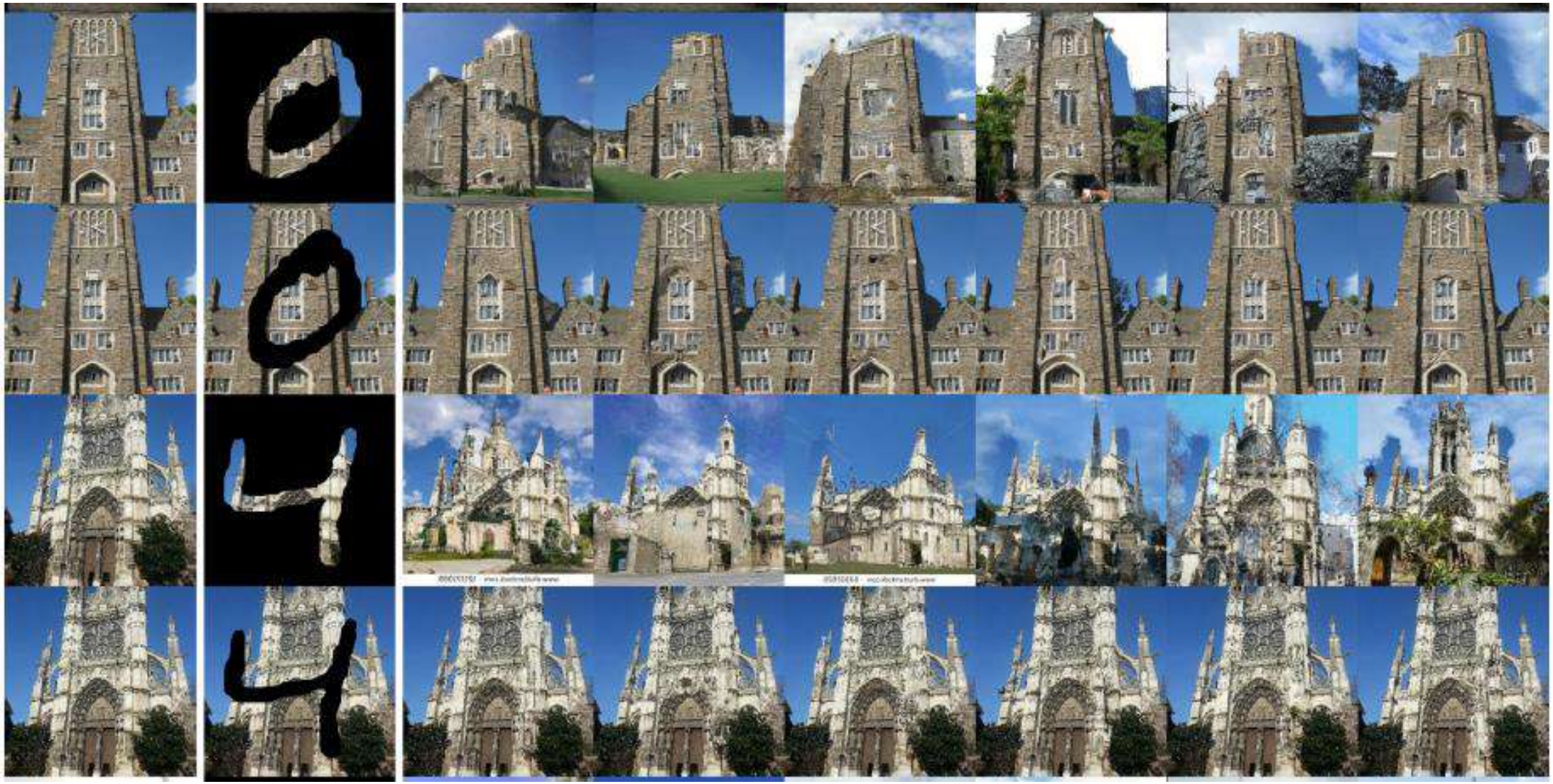


Генерация для указанного класса

Class-conditional image generation

Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations, ICLR 2021

Score-based generative modeling through stochastic differential equations

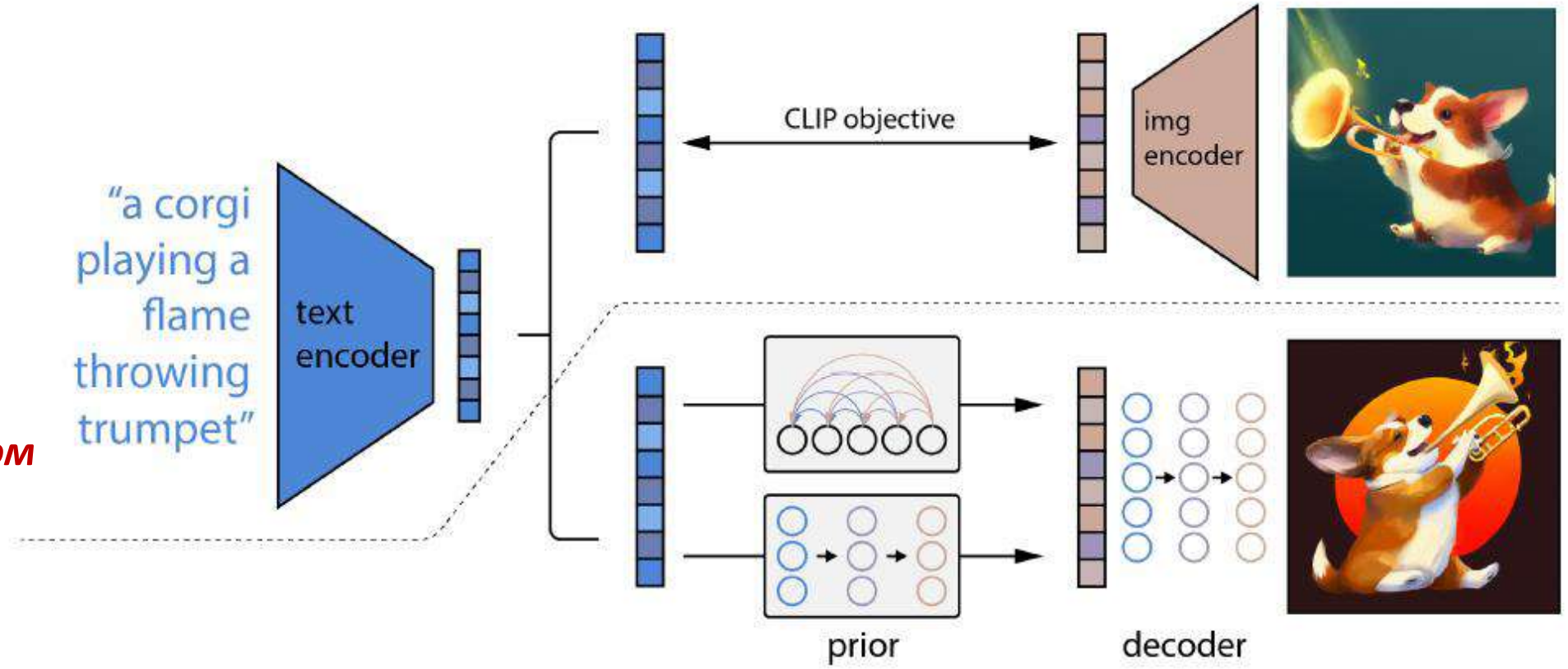


Реконструкция скрытого inpainting results for 256×256 church images.

Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations, ICLR 2021

DALL·E 2: images and art from natural language description

Генерация изображений и рисунков по текстовому описанию на естественном языке



A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or **diffusion prior** to produce an image embedding, and then this embedding is used to condition a **diffusion decoder** which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

Hierarchical Text-Conditional Image Generation with CLIP Latents
Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

We develop methods for training **diffusion priors in latent space**, and show that they achieve comparable performance to autoregressive priors, while being more compute-efficient.

DALL·E 2: images and art from natural language description

Непрерывное пространство изображений (manifold) (распределение)

a photo of a cat → an anime drawing of a super saiyan cat, artstation

a photo of a victorian house → a photo of a modern house

a photo of an adult lion → a photo of lion cub

a photo of a landscape in winter → a photo of a landscape in fall

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

diffusion priors in latent space

DALL·E 2: images and art from natural language description

Our decoder model provides a unique opportunity to explore CLIP latent space by allowing us to directly visualize what the CLIP image encoder is seeing. Visualization of *reconstructions of CLIP latents from progressively more PCA dimensions (20, 30, 40, 80, 120, 160, 200, 320 dimensions), with the original source image on the far right.* **The lower dimensions preserve coarse-grained semantic information**, whereas the **higher dimensions encode finer-grained details** about the exact form of the objects in the scene.

Непрерывное пространство изображений (manifold) (распределение)

Hierarchical Text-Conditional Image Generation with CLIP Latents
Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

diffusion priors in latent space

DALL·E 2: images and art from natural language description

Diffusion models really seems to be a new SOTA generative framework!

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

a dolphin in an astronaut suit on saturn, artstation

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

https://openai.com/dall-e-2/#demos

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

diffusion priors in latent space

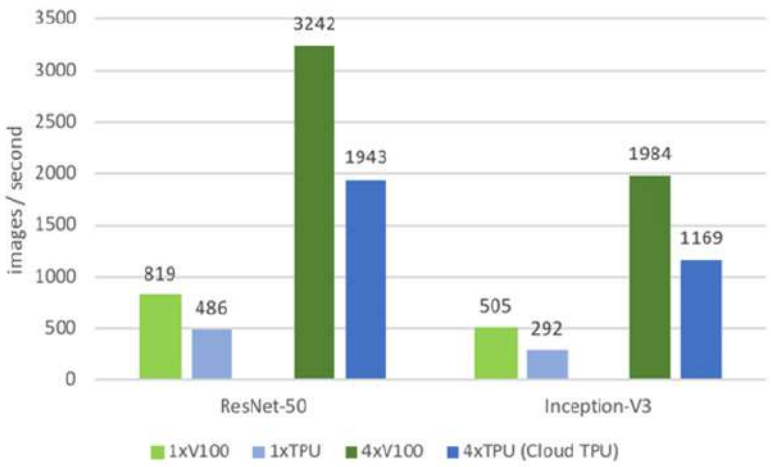
Сравнение производительности аппаратных средств для ИИ

Результаты первого сравнительного тестирования отечественных аппаратных средств в задачах ИИ (компьютерного зрения)

Наводки: Доклады Д.Ватолина и Б.В. Вишнякова на OpenTalks.AI/22

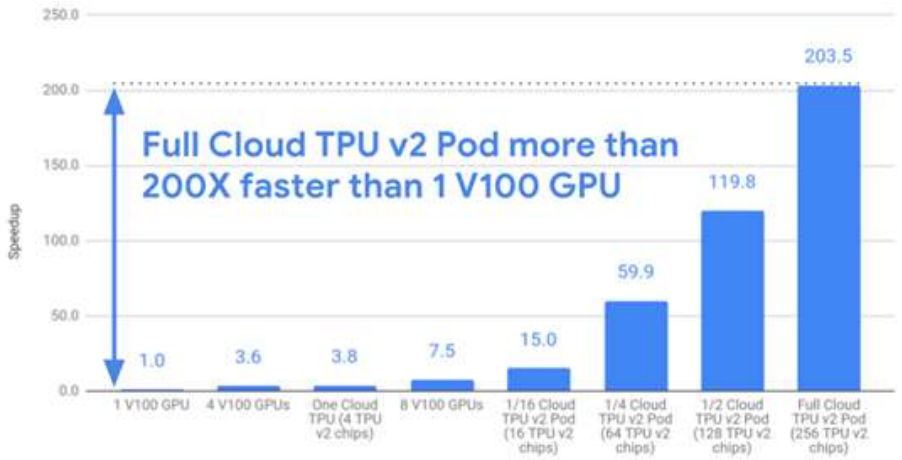
Сравнение производительности: Каждый считает себя лучшим

NVIDIA в 1,6 раз быстрее TPU на ResNET-50



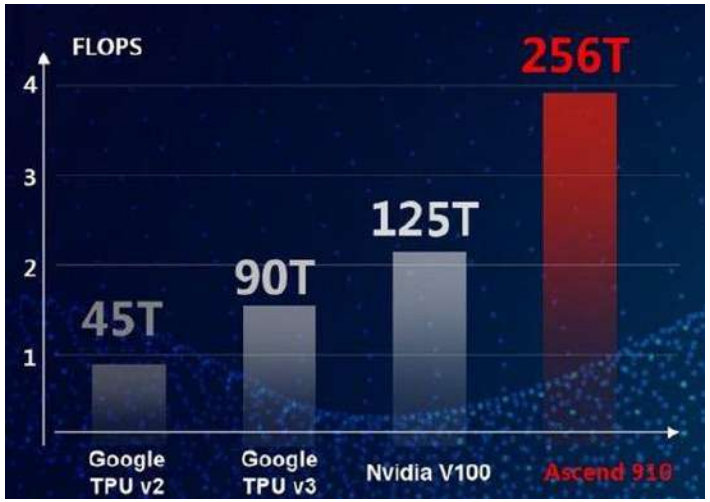
NVIDIA

TPU в 203,5 раз быстрее NVIDIA на ResNET-50



Google

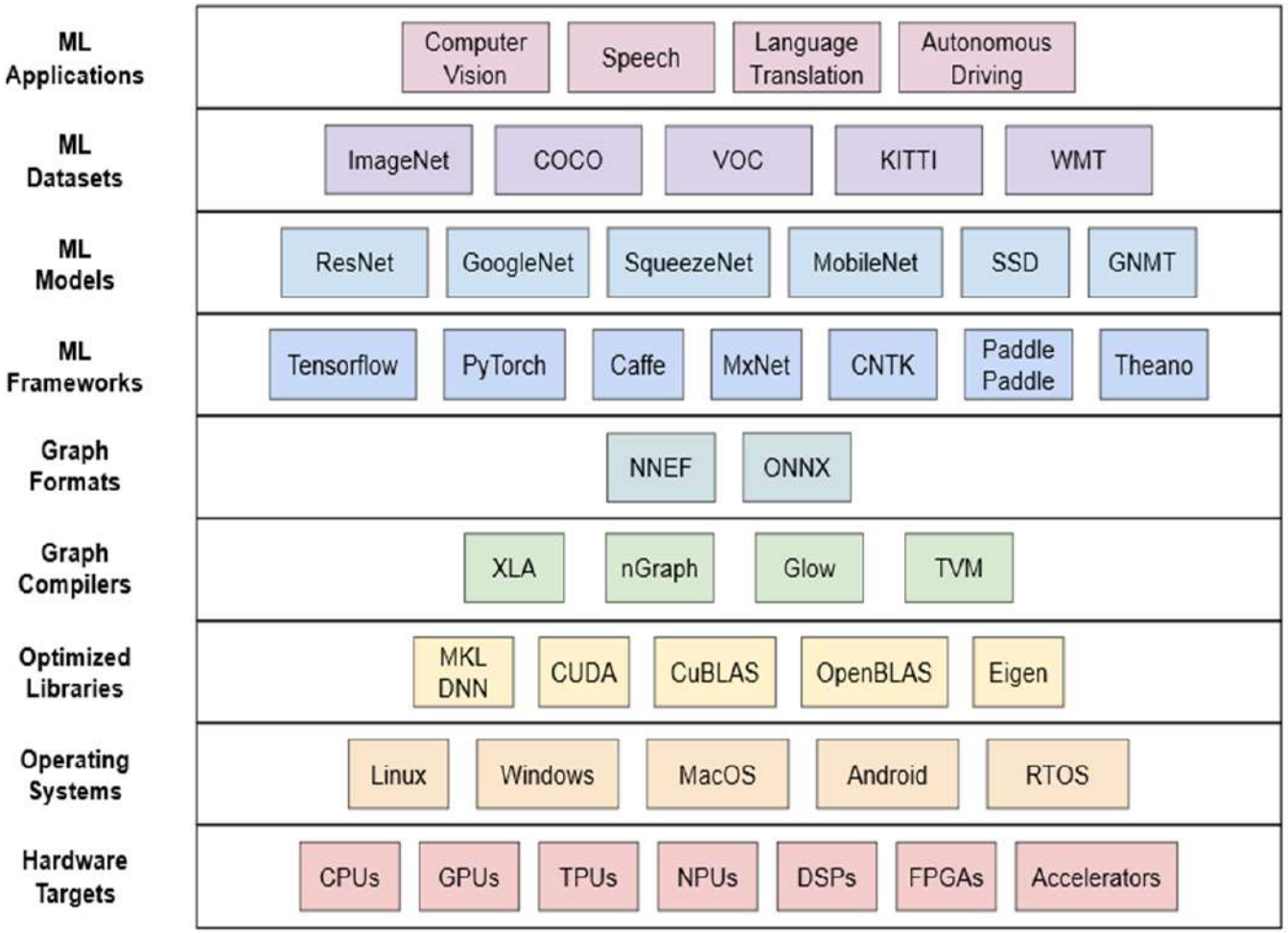
Huawei обгоняет TPU v3 в 2,8 раз и NVIDIA в 2 раза



Huawei

Слайд: Дмитрий Ватолин, Hardware для ИИ — главное в 2021, OpenTalksAI/2022

Как сравнивать запуск разных сетей в разных условиях?



https://edge.seas.harvard.edu/files/edge/files/mlperf_inference.pdf

Решение для всех — MLCommons (ранее MLPerf)

ML Commons

Результаты на данный момент:

- 30+ организаций-участников
- 230+ прогонов

Треки:

- **Training** (обучение)
- **Inference** (выполнение)
- **Closed division** — сравнение hardware
- **Opened division** — сравнение hardware, но можно менять метод предобработки и параметры обучения

Слайд на основе: Дмитрий Ватолин, Hardware для ИИ — главное в 2021, OpenTalksAI/2022

MLCommons: Категории участников

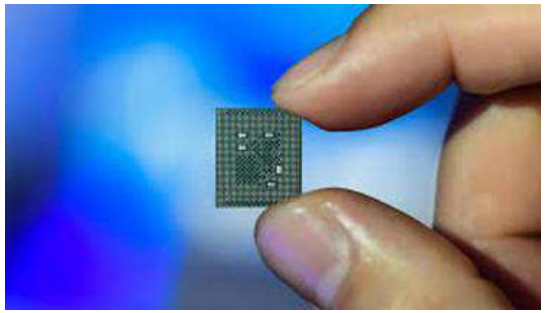
Решение для всех — MLCommons (ранее MLPerf)

Training

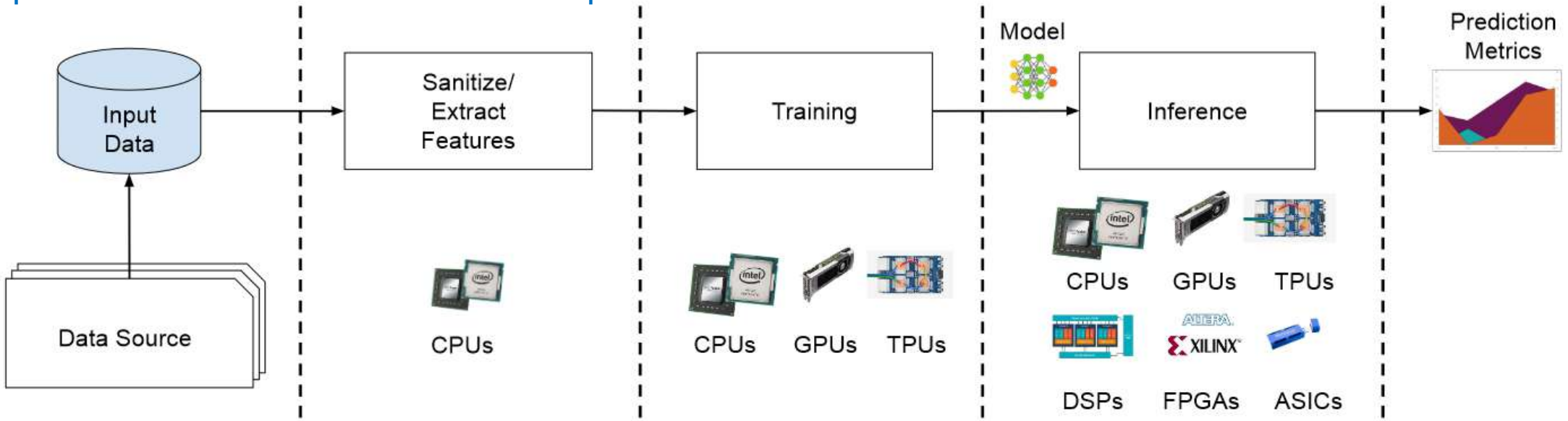
- HPC — суперкомпьютеры
- Regular — обычные компьютеры и облачные решения

Inference

- Datacenters — промышленные сервера
- Mobile systems — телефоны и ноутбуки
- Edge systems — периферийные устройства (микрокомпьютеры)



«Зоопарк» железа на Inference больше в разы



Слайд на основе: Дмитрий Ватолин, Hardware для ИИ — главное в 2021, OpenTalksAI/2022

https://mlcommons.org/en/

Решение для российского ОПК: программа сравнительного тестирования от ГосНИИАС

- ✓ Тестирование вычислителей отечественных производителей проводится на базе Платформы-ГНС
- ✓ На первом этапе рассматриваются широко распространённые задачи: классификация, обнаружение и сегментация
- ✓ В тестировании учувствуют классические архитектуры ГКНС

Набор задач для тестирования (ver. 02.2022)

Задача классификации изображений

БД для обучения и тестирования	ILSVRC 2014
Используемые архитектуры	✓ ResNet 18 ✓ ResNet 34 ✓ ResNet 50 ✓ MobileNet v1
Используемые метрики качества	Accuracy, Top5
Форма входных данных	3x224x224

Задача обнаружения объектов

БД для обучения и тестирования	MS COCO
Используемые архитектуры	✓ SSD-ResNet 34 ✓ SSD-Mobilenet v2
Используемые метрики качества	mAP
Форма входных данных	3x512x512

Задача семантической сегментации

БД для обучения и тестирования	CityScapes
Используемые архитектуры	✓ UNetVGG16
Используемые метрики качества	mIOU
Форма входных данных	3x512x512

Платформа ГНС © 2022 ФГУП «ГосНИИАС»

Решение для российского ОПК: программа сравнительного тестирования от ГосНИИАС

- ✓ **«Датацентр»** – категория подразумевает отсутствие ограничений на потребляемую мощность.
- ✓ **«Бортовое исполнение»*** – категория подразумевает ограничение в энергопотреблении вычислителя до 100 Ватт
- ✓ **«Бортовое исполнение (20 Ватт)»*** – категория подразумевает ограничение в энергопотреблении вычислителя до 20 Ватт.

**размер пакета (batch) = 1*

Состав тестируемых аппаратных платформ (ver. 02.2022)

Датацентр

Xeon Silver 4112	Intel **(2017)**
Эльбрус 8 С	АО «МЦСТ» **(2014)**
1892ВМ248 «RoboDeus»	АО НПЦ «ЭЛВИС» **(2021)**

Бортовое исполнение

К1945ВМ028 IVA H TPU	IVA Technologies (ГК «ХайТэк») **(2021)**
Jetson AGX Xavier	Nvidia **(2020)**

Бортовое исполнение (20 Ватт)

1879ВМ8Я NM6408	АО НТЦ «Модуль» **(2017)**
1892ВМ248 «RoboDeus»	АО НПЦ «ЭЛВИС» **(2020)**
Jetson Nano	Nvidia **(2020)**
NUC, Pentium Silver J5005	Intel **(2019)**
NUC, UHD Graphics 605	Intel **(2019)**

Платформа ГНС © 2022 ФГУП «ГосНИИАС»

Тестирование отечественных вычислителей ver. 02.2022

Тестирование считается успешным, если пройден 5% порог качества по соответствующей метрике

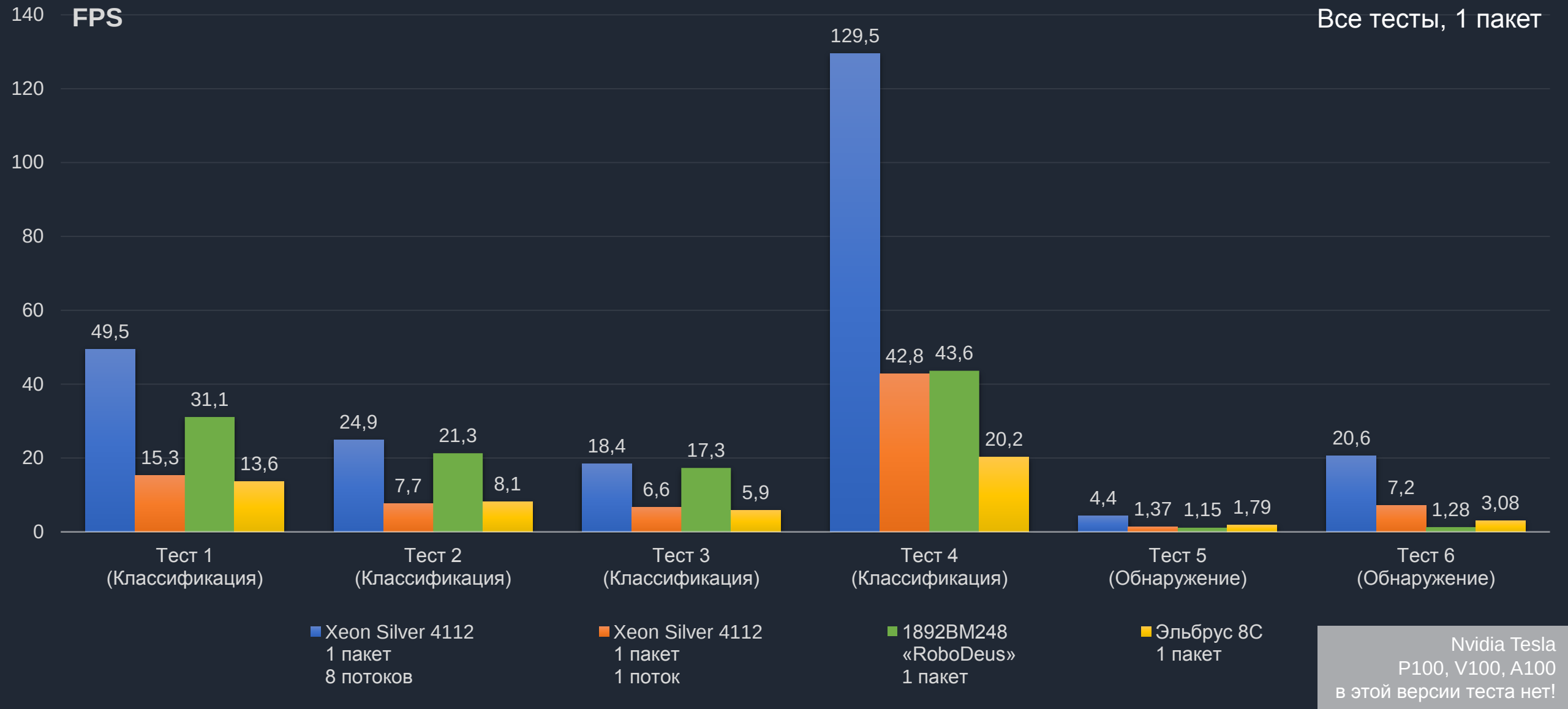
	CNN	Тест	NM6408	Эльбрус-8С	Robodeus	IVA H TPU	Xeon Silver 4112	NUC, Pentium Silver J5005	NUC, UHD Graphics 605	Nvidia Jetson AGX Xavier	Jetson Nano
Класси-фикация	**ResNet18**	**№1**	✓	✓	✓	✓	✓	✓	✓	✓	✓
	ResNet34	**№2**	✓	✓	✓	✓	✓	✓	✓	✓	✓
	ResNet50	**№3**	✓	✓	✓	✗	✓	✓	✓	✓	✓
	MobileNet-v1	**№4**	✓	✓	✓	✗	✓	✓	✓	✓	✓
Обнару-жение	**SSD-ResNet34**	**№5**	✓	✓	✓	?	✓	✓	✓	✓	✓
	SSD-MobileNet-v2	**№6**	✓	✓	✓	✓	✓	✓	✓	✓	✓
Сегмен-тация	**Unet-VGG16**	**№7**	✓	✗	?	?	✓	✓	✓	✓	✓

✓ - тест пройден

✗ - тест не пройден по качеству

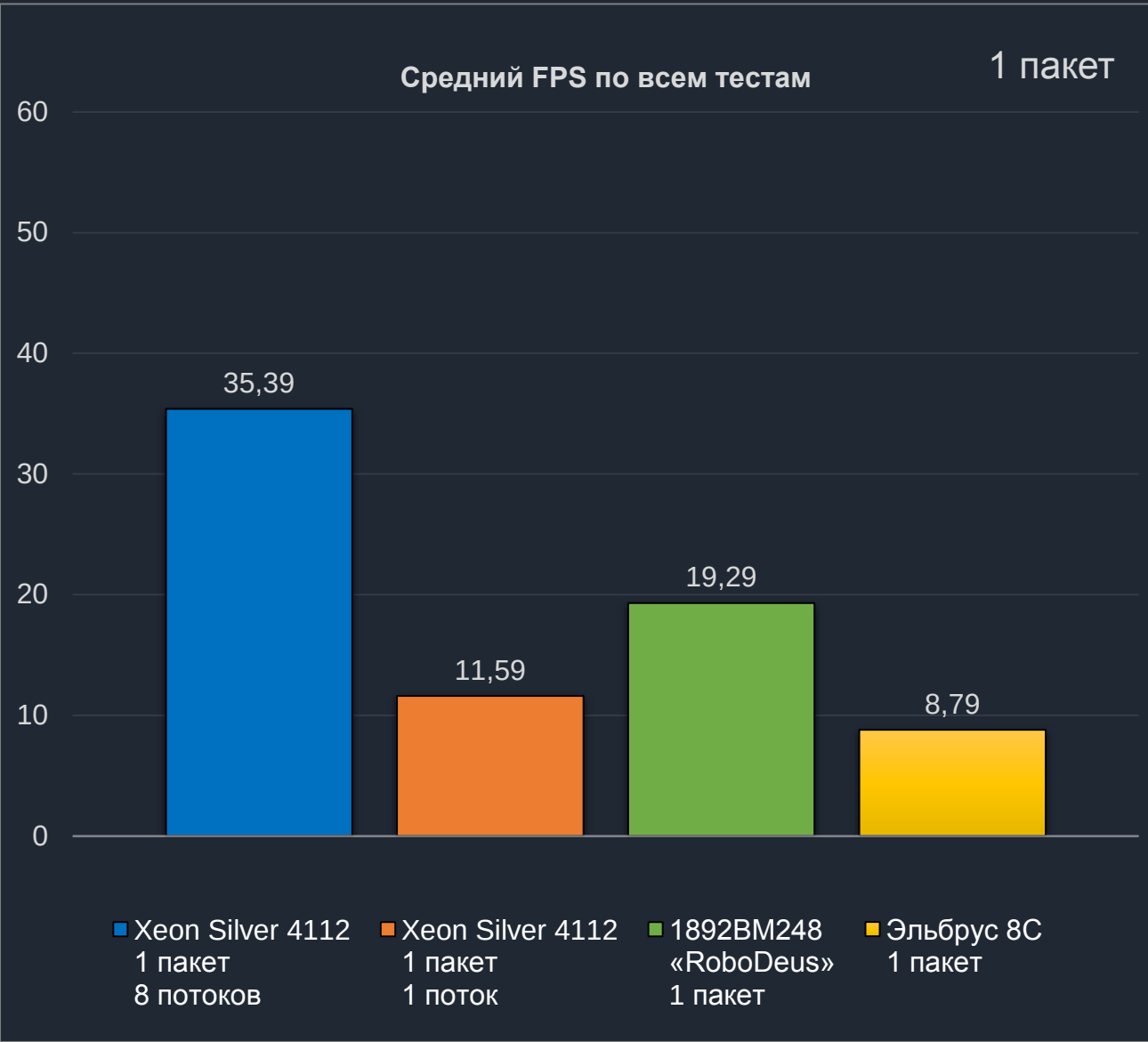
? - данные не предоставлены

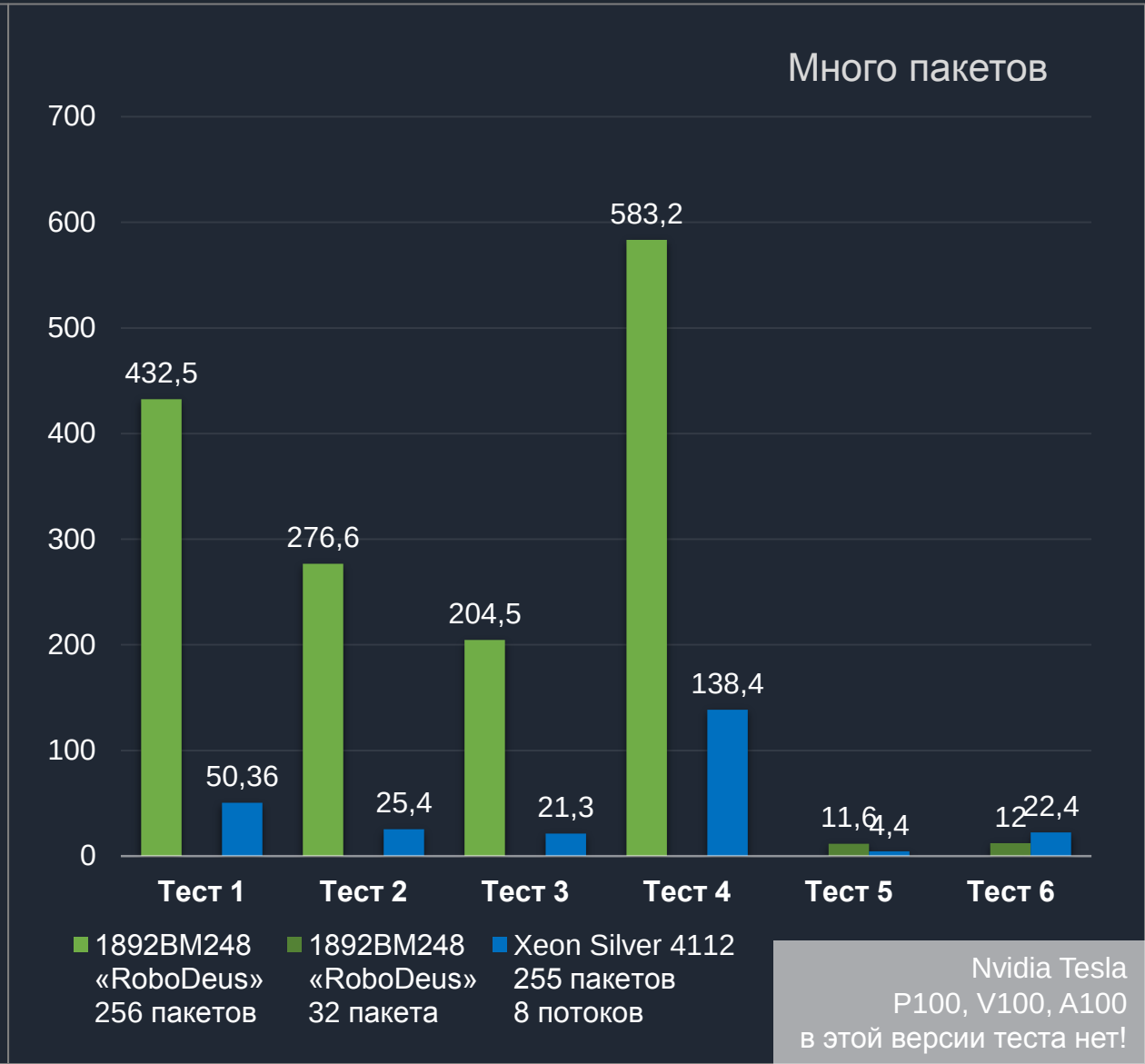
Результаты сравнительного тестирования (ver. 02.2022) **для категории «Датацентр»**



Nvidia Tesla P100, V100, A100 в этой версии теста нет!

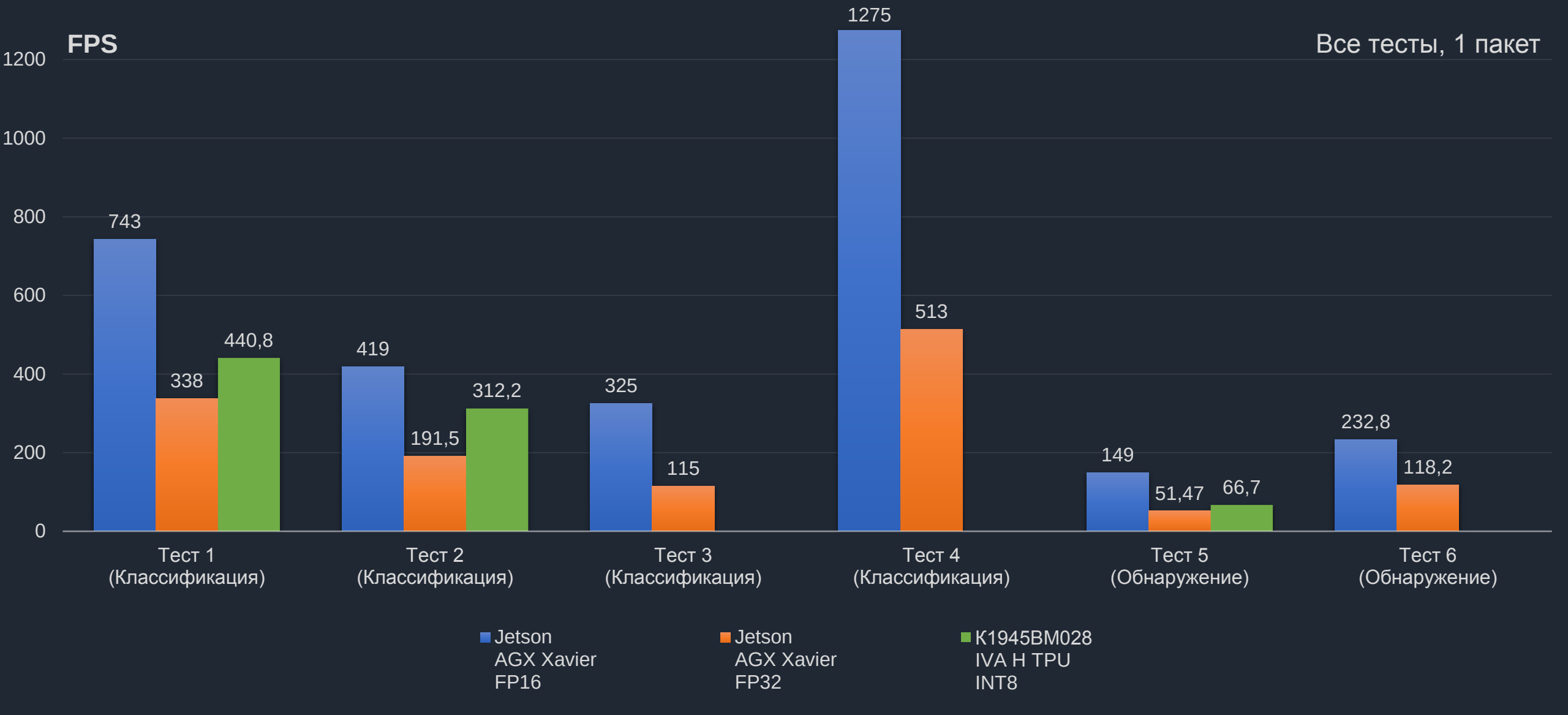
Результаты сравнительного тестирования (ver. 02.2022) **для категории «Датацентр»**



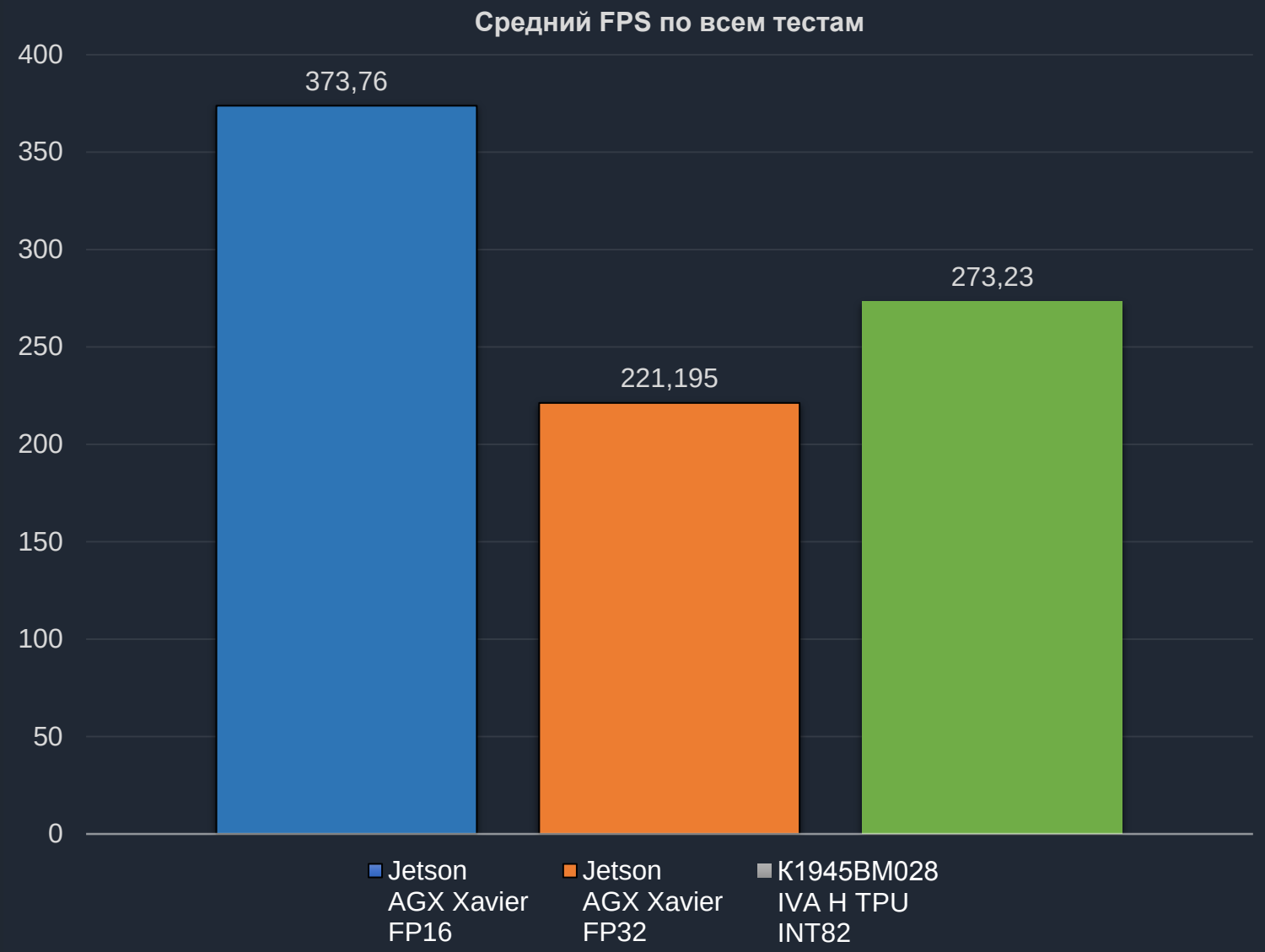


Nvidia Tesla
P100, V100, A100
в этой версии теста нет!

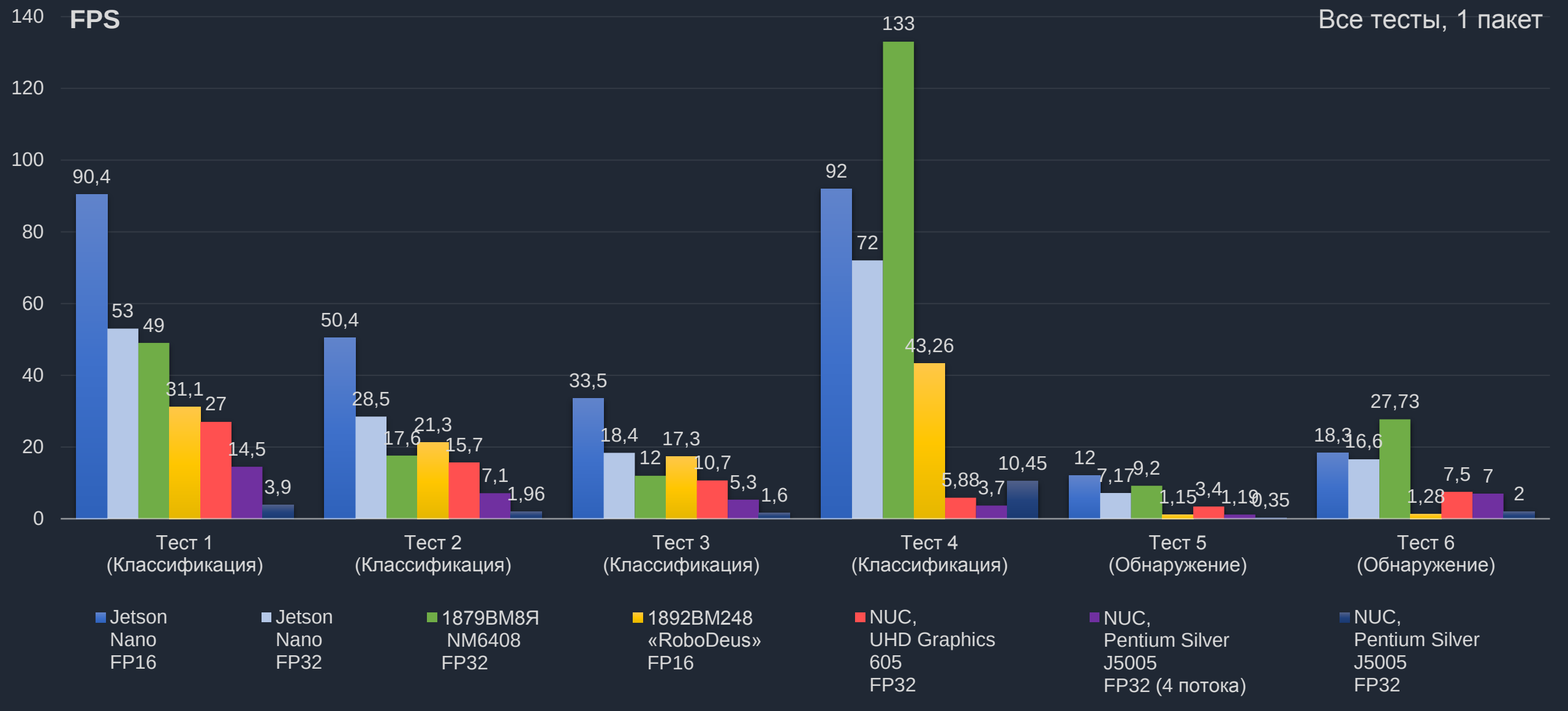
Результаты сравнительного тестирования (ver. 02.2022) **для категории «Бортовое исполнение»**



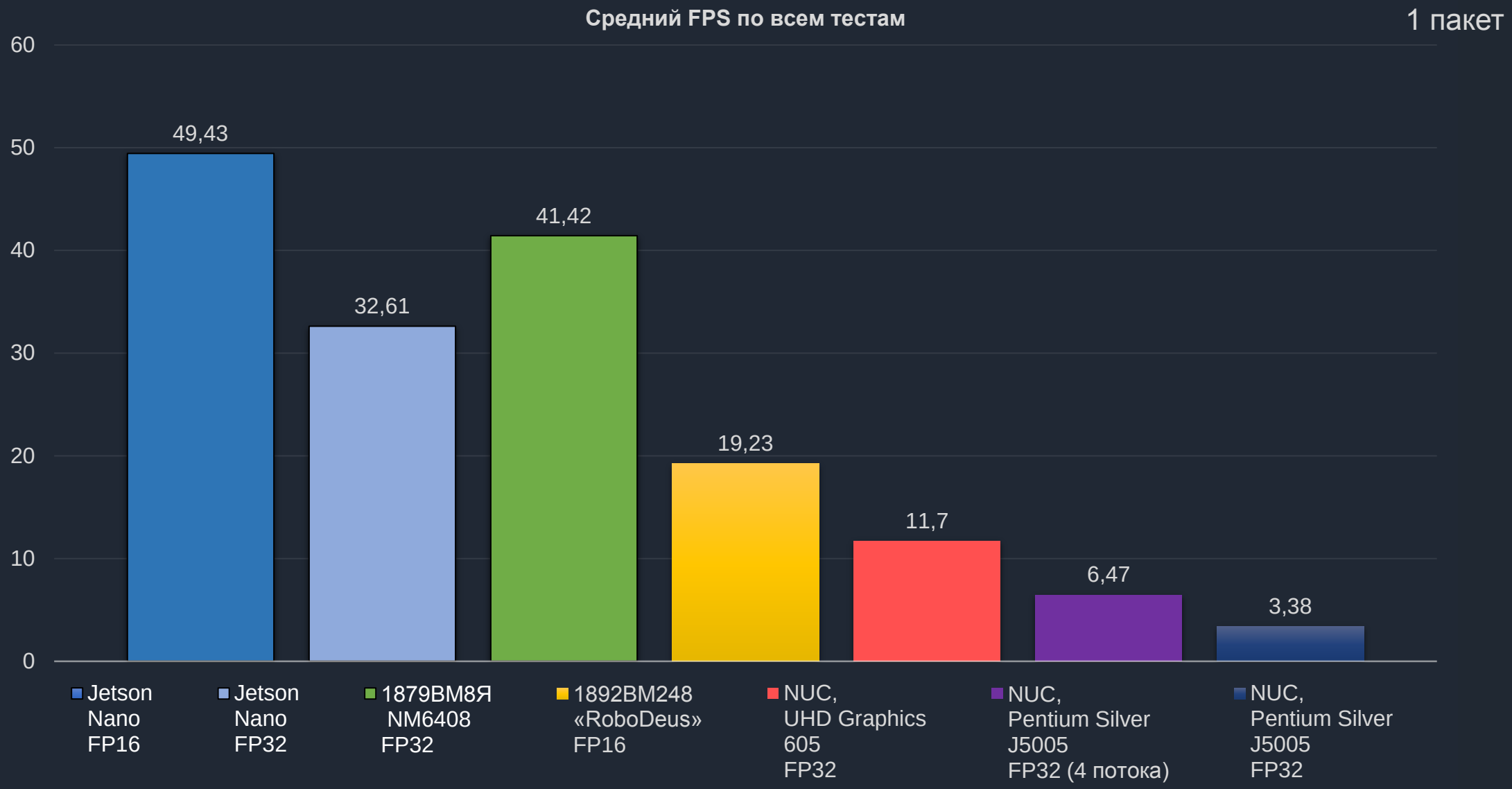
Результаты сравнительного тестирования (ver. 02.2022) **для категории «Бортовое исполнение»**



Результаты сравнительного тестирования (ver. 02.2022) **для категории «Бортовое исполнение (20 Ватт)»**



Результаты сравнительного тестирования (ver. 02.2022) **для категории «Бортовое исполнение (20 Ватт)»**

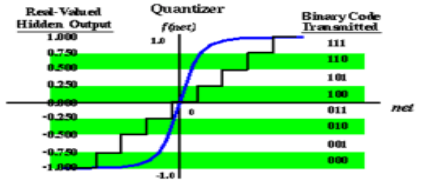


Инструмент: отечественная среда нейросетевой разработки «Платформа-ГНС» (PLAT)

Единая интегрированная среда

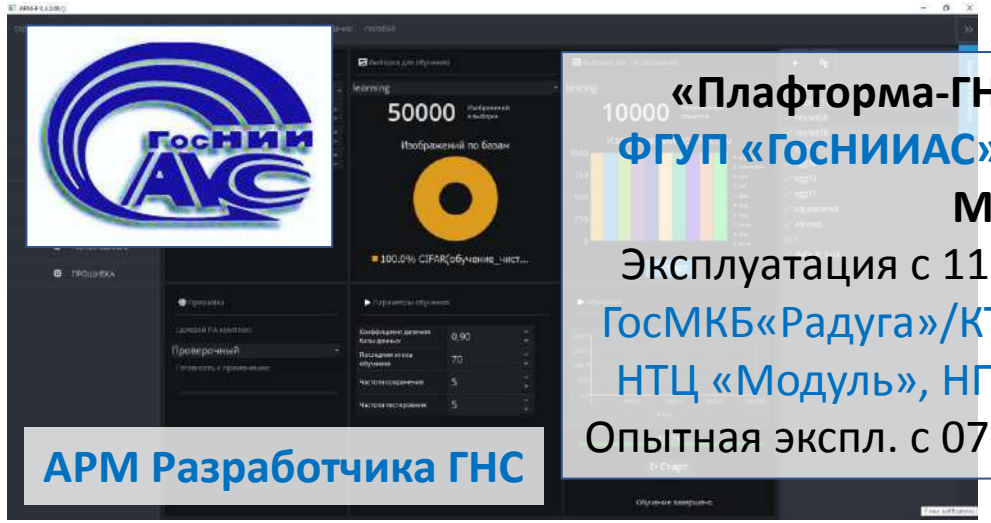
Обучение ГНС

Преобразование ГНС



Аппаратная реализация ГНС

OpenVINO
TensorRT
NeuroMatrix
reVISION Stack

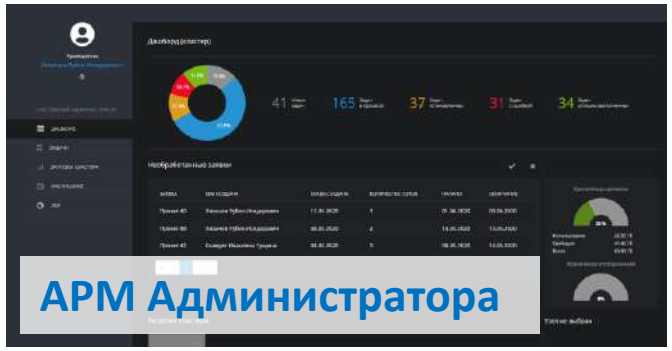


АРМ Разработчика ГНС

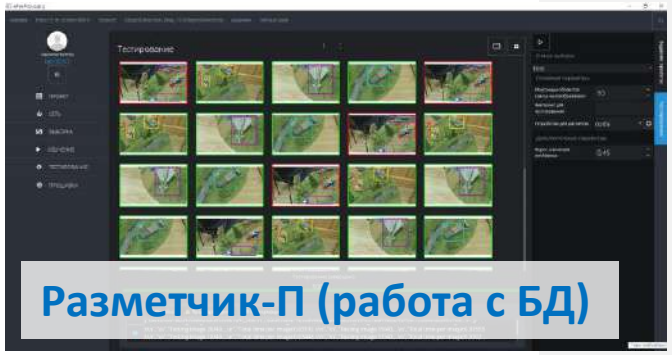
«Плафторма-ГНС»: разработана ФГУП «ГосНИИАС» при поддержке Минпромторга РФ

Эксплуатация с 11.2020: ГосНИИАС, ГосМКБ«Радуга»/КТРВ, ЗАО МНИТИ, НТЦ «Модуль», НПО МИС, АО РПКБ

Опытная экспл. с 07.2020: ВИТ «ЭРА»



АРМ Администратора



Разметчик-П (работа с БД)

Унифицированная платформа

- **Сертифицируемый на НДВ исходный код**
- Импорт/экспорт из основных фреймворков
- **Наличие типовых решений для основных задач обучения ГНС**
- **Поддержка отечественных аппаратных платформ и отечественных ОС**
- Контроль доступа к данным и проектам
- **Низкие требования к квалификации ИТР**

=> массовое внедрение ГНС в изделия ВВСТ

Поддерживаемые ОС: Windows, Linux, AstraLinux

Поддерживаемые процессоры: x86-64, Эльбрус, NVidia CUDA, Nvidia Tesla, RocM/HIP, FPGA Xillinx, ARM v12 и выше, Neuromatrix, Robodeus, AMD Radeon/Radeon Instinct

Импортные аналоги: NVIDIA DIGITS, BLVC Caffe, Pytorch Framework, TensorFlow

Модуль

МЦСТ ЭЛЬБРУС

ЭЛВИС

Актуальные задачи и результаты в области компьютерного зрения и машинного обучения

Визильтер Юрий Валентинович, д.ф.-м.н., проф. РАН,
директор по направлению «Технологии искусственного интеллекта»
ФАУ «ГосНИИАС», *viz@gosniias.ru*

Спасибо за внимание!

ITNT - 2022

VIII Международная конференция
«Информационные технологии и нанотехнологии»
(ИТНТ-2022)

Самара, 23.05.2022