

ТРАНСФОРМЕРЫ В КОМПЬЮТЕРНОМ ЗРЕНИИ

Липкин С.М., 08.07.2022

ПРЕДСТАВЛЕНИЕ ТЕКСТОВЫХ ДАННЫХ

Словарь эмбеддингов

Слово	1.1	0.25	-0.91	0.77
кот	1.1	0.25	-0.91	0.77
кошка	1.0	0.27	0.78	0.88
автобус	-0.55	0.98	-0.42	0.21
межзвездный	0.05	-1.5	-0.52	1.2
ехать	0.25	0.72	0.01	-0.42

Одушевленность

Пол

Эмбеддинги близких по смыслу слов имеют схожие значения

ПРЕДСТАВЛЕНИЕ ТЕКСТОВЫХ ДАННЫХ

Кот и кошка едут на межзвездном автобусе

1.1	1.0	-0.55
0.25	0.27	0.98
-0.91	0.78	-0.42
0.77	0.88	0.21

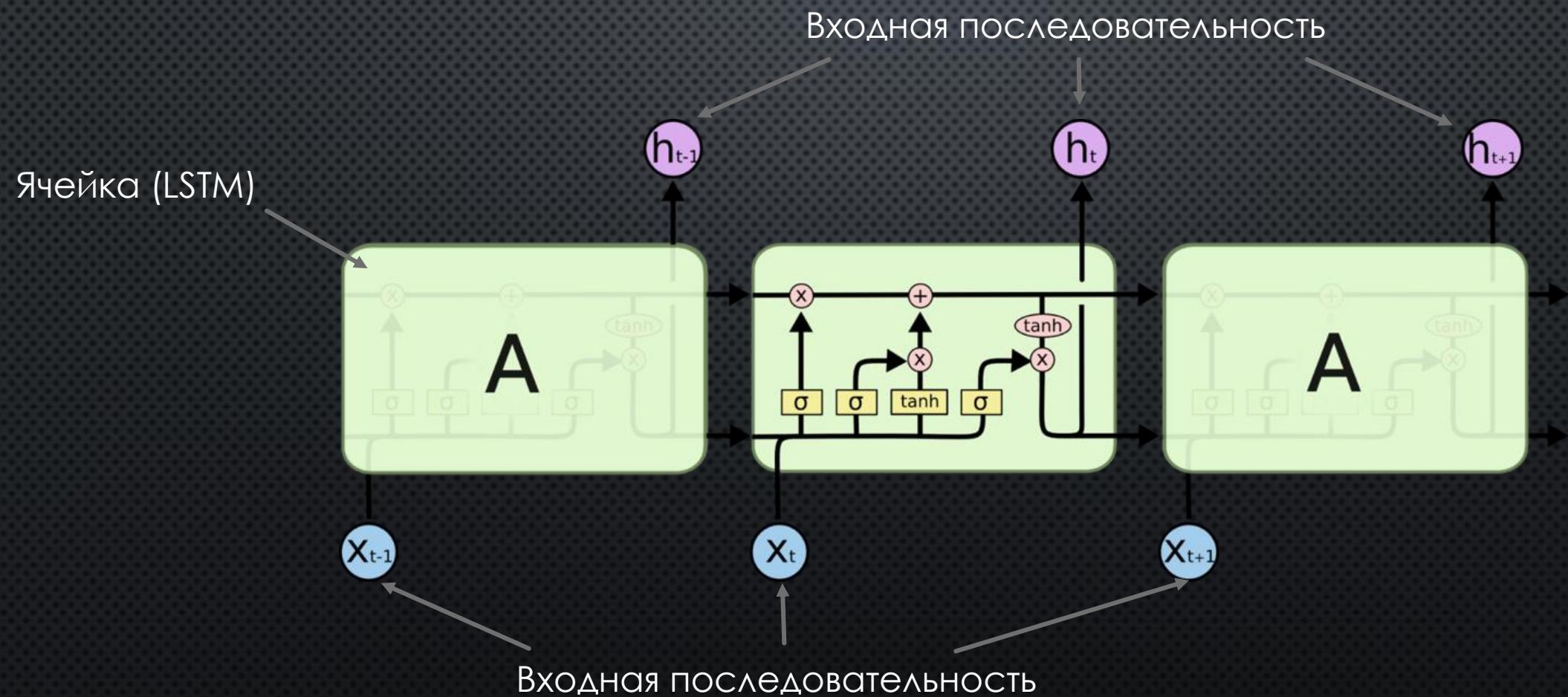
0.05
-1.5
-0.52
1.2

0.25
0.72
-0.01
-0.42



Эмбеддинг
предложения: [1, 5, 4]

РЕКУРРЕНТНАЯ НЕЙРОННАЯ СЕТЬ



РЕКУРРЕНТНАЯ НЕЙРОННАЯ СЕТЬ. НЕДОСТАТКИ

Медленное обучение

Исчезающие/взрывающиеся градиенты

Отсутствие параллелизации

Плохо справляется с длинными последовательностями

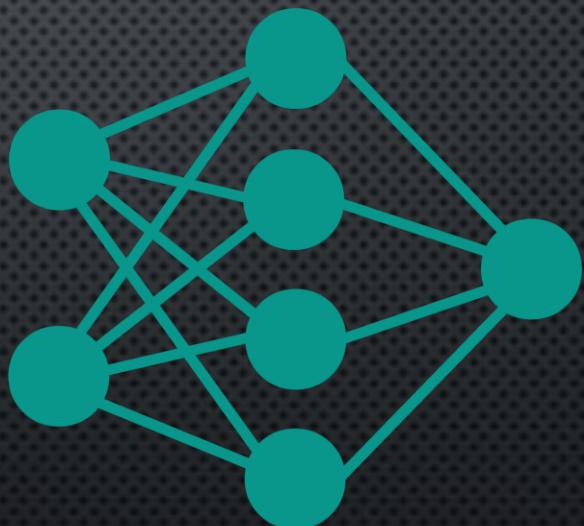
ЯЗЫКОВЫЕ МОДЕЛИ

Кот и кошка едут на межзвездном автобусе ...

1.1	1.0	-0.55
0.25	0.27	0.98
-0.91	0.78	-0.42
0.77	0.88	0.21

0.05
-1.5
-0.52
1.2

0.25
0.72
-0.01
-0.42



... на дачу

ТРАНСФОРМЕР. ВНИМАНИЕ

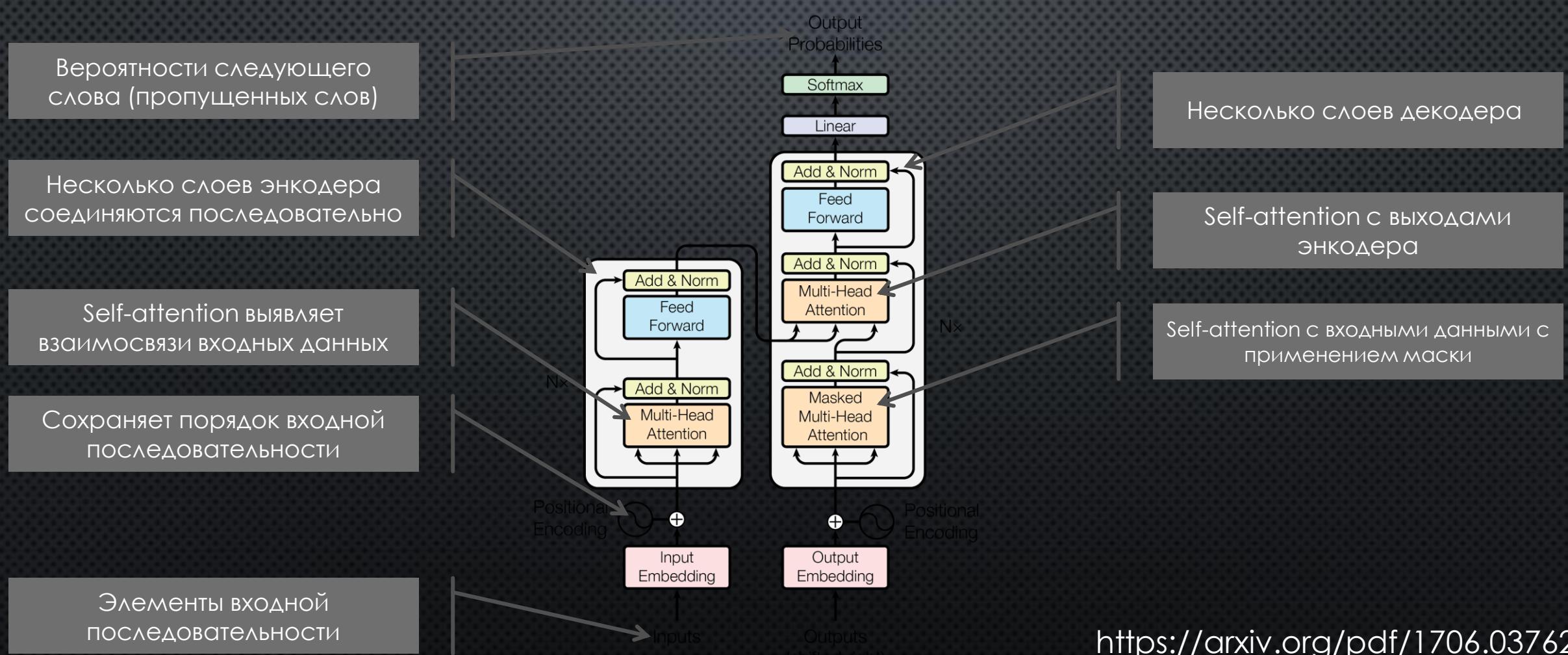


База зананий

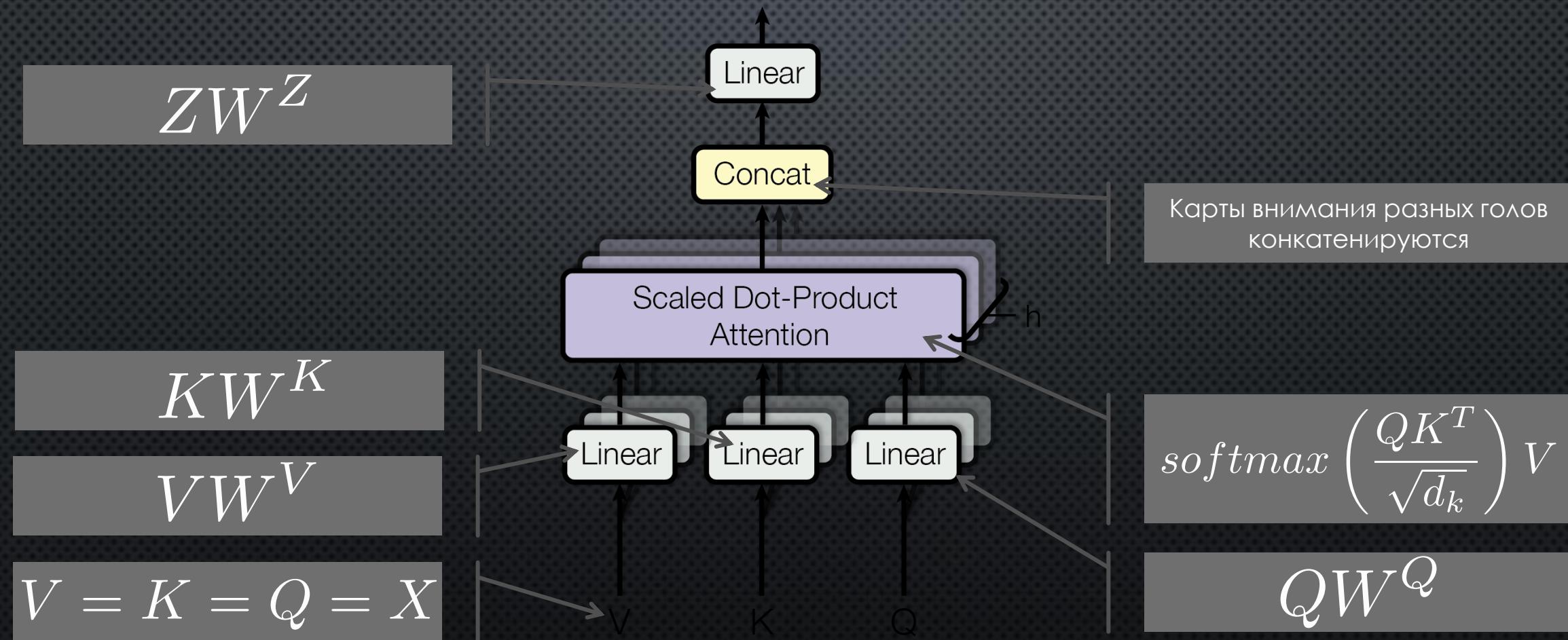
Ключи	Значения
k_1	v_1
k_2	v_2
k_3	v_3
k_4	v_4

$[n, 1]$ $[n, 1]$

ТРАНСФОРМЕР. ОБЩАЯ АРХИТЕКТУРА



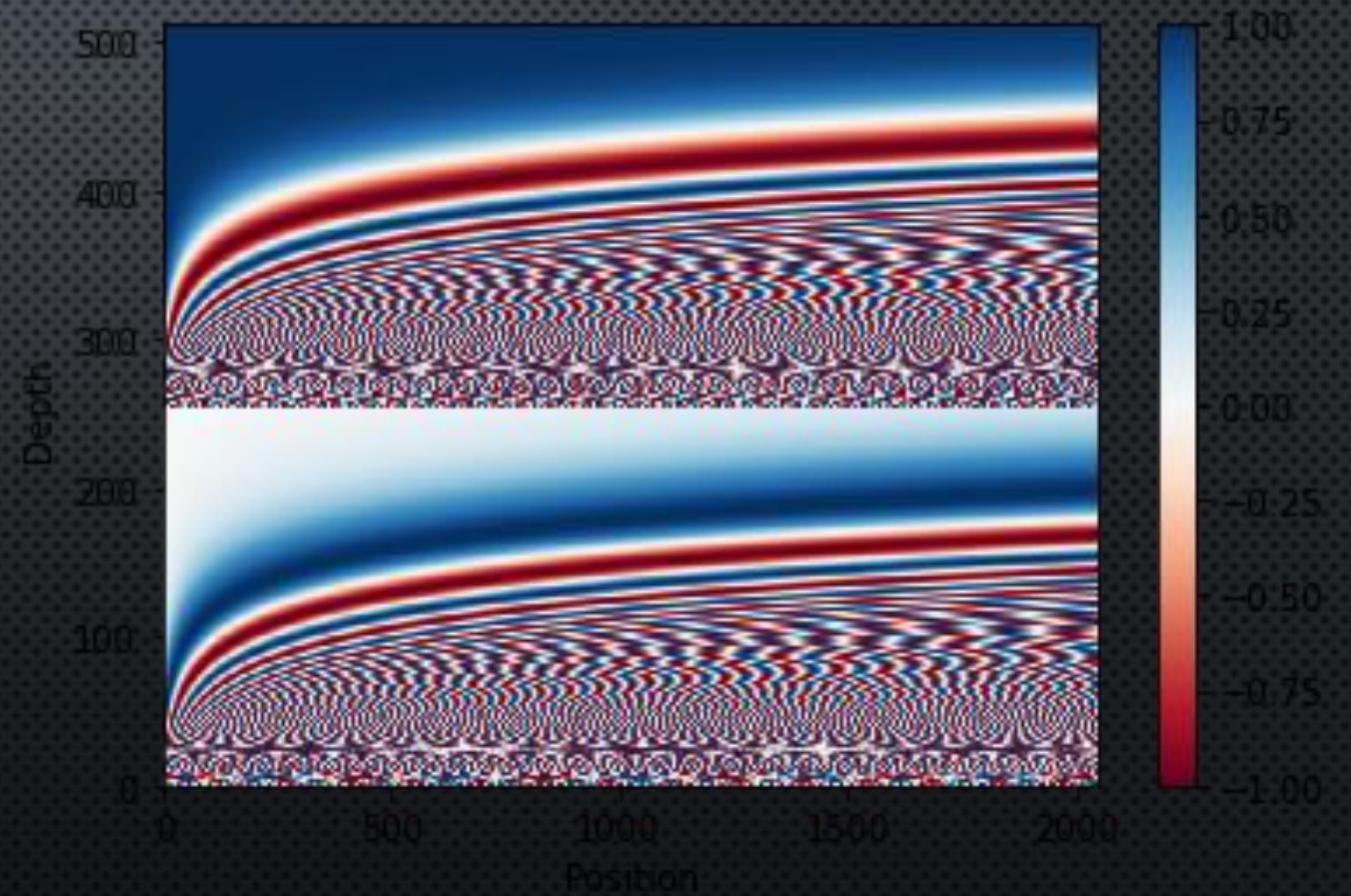
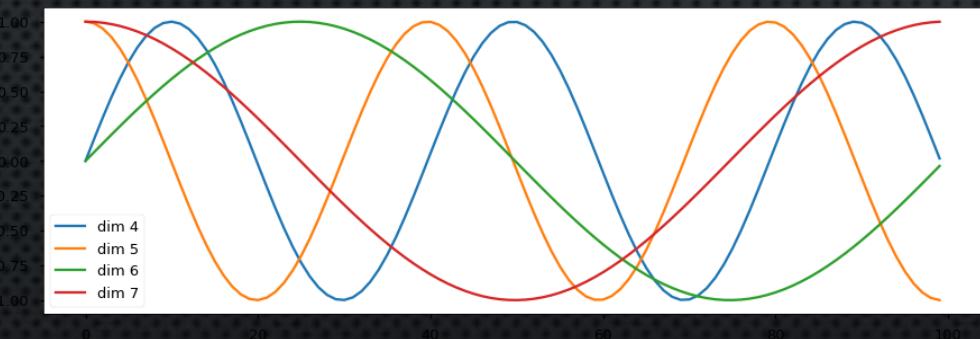
ТРАНСФОРМЕР. ВНИМАНИЕ



ТРАНСФОРМЕР. ПОЗИЦИОННАЯ КОДИРОВКА

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{model}})$$



ТРАНСФОРМЕРЫ В CV. НАПРАВЛЕНИЯ

Извлечение признаков

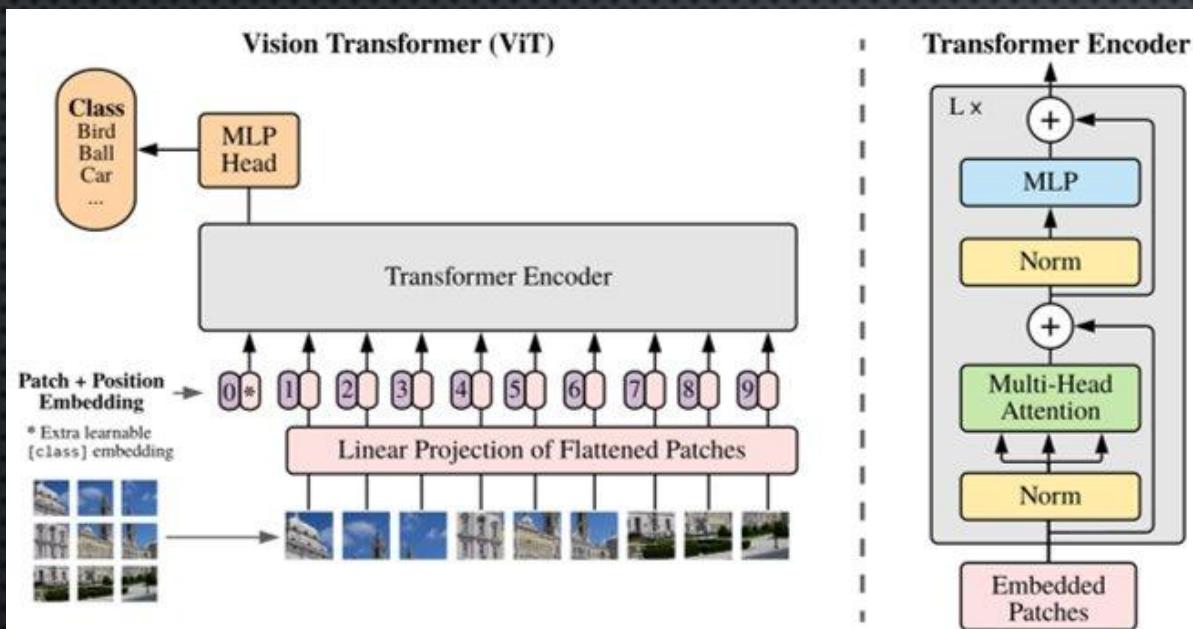
Высокоуровневые задачи

Низкоуровневые задачи

Обработка видео

Мультимодальные задачи

ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ. VISION TRANSFORMER (ViT)

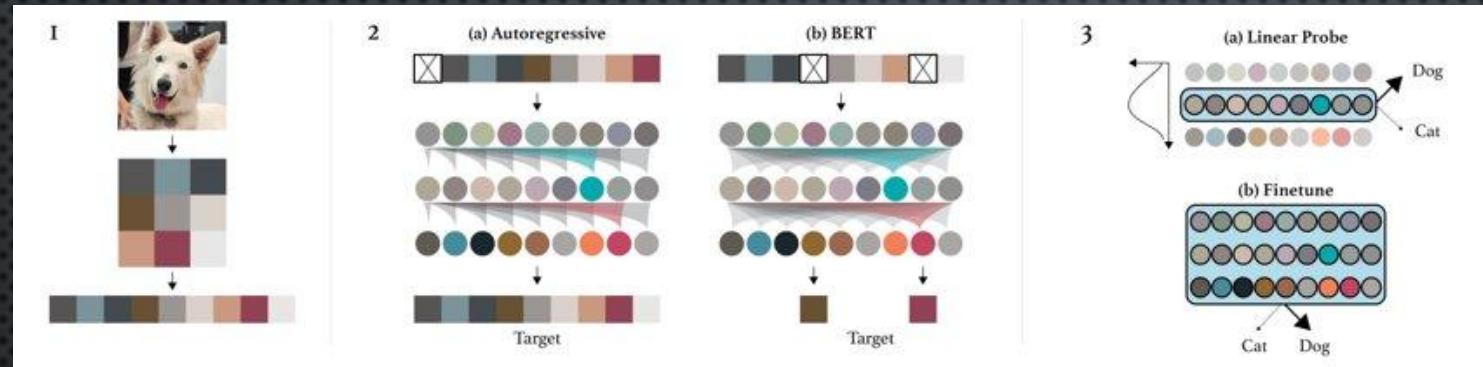


	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21K (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.55 ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	90.72 ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	99.50 ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	—
CIFAR-100	94.55 ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	—
Oxford-IIIT Pets	97.56 ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	—
Oxford Flowers-102	99.68 ± 0.02	99.74 ± 0.00	99.61 ± 0.02	99.63 ± 0.03	—
VTAB (19 tasks)	77.63 ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	—
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

SOTA

Огромное
число
параметров

ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ. iGPT



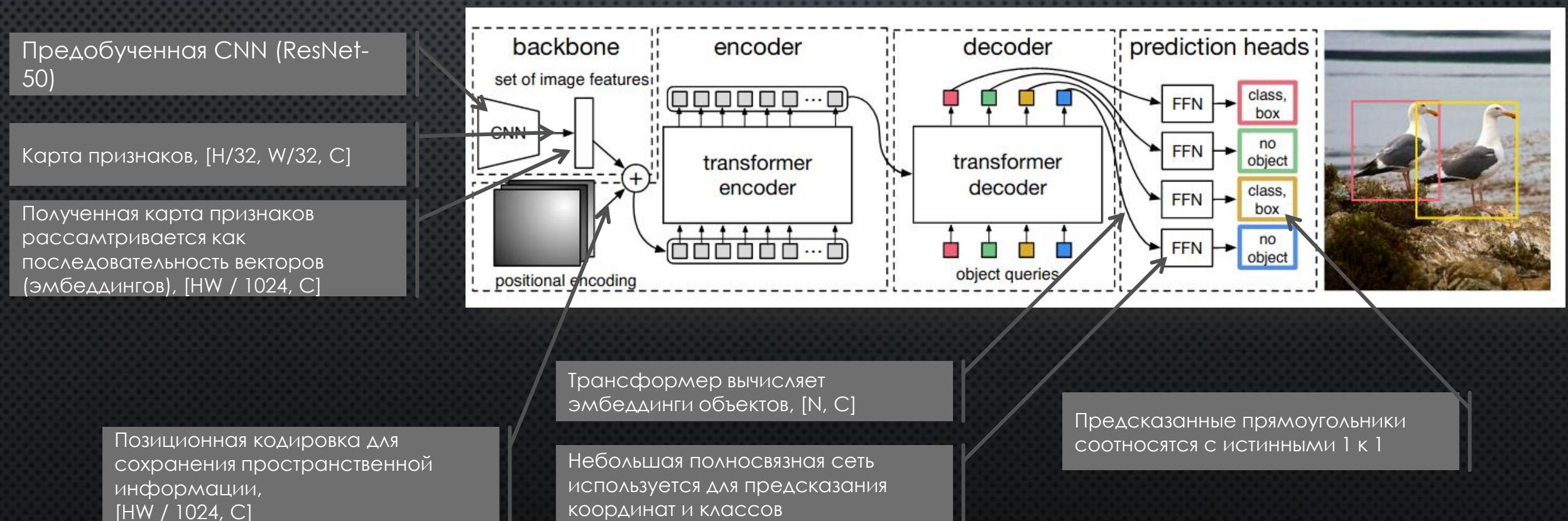
Evaluation	Dataset	Our Result	Best non-iGPT Result
Logistic regression on learned features (linear probe)	CIFAR-10	96.3 iGPT-L 32x32 w/ 1536 features	95.3 SimCLR w/ 8192 features
	CIFAR-100	82.8 iGPT-L 32x32 w/ 1536 features	80.2 SimCLR w/ 8192 features
	STL-10	95.5 iGPT-L 32x32 w/ 1536 features	94.2 AMDIM w/ 8192 features
	ImageNet	72.0 iGPT-XL 64x64 w/ 15360 features	76.5 SimCLR w/ 8192 features
Full fine-tune	CIFAR-10	99.0 iGPT-L 32x32, trained on ImageNet	99.0 GPipe, trained on ImageNet
	ImageNet 32x32	66.3 iGPT-L 32x32	70.2 Isometric Nets

SOTA

Предобучение
на
неразмечены
ханных

Огромное
число
параметров

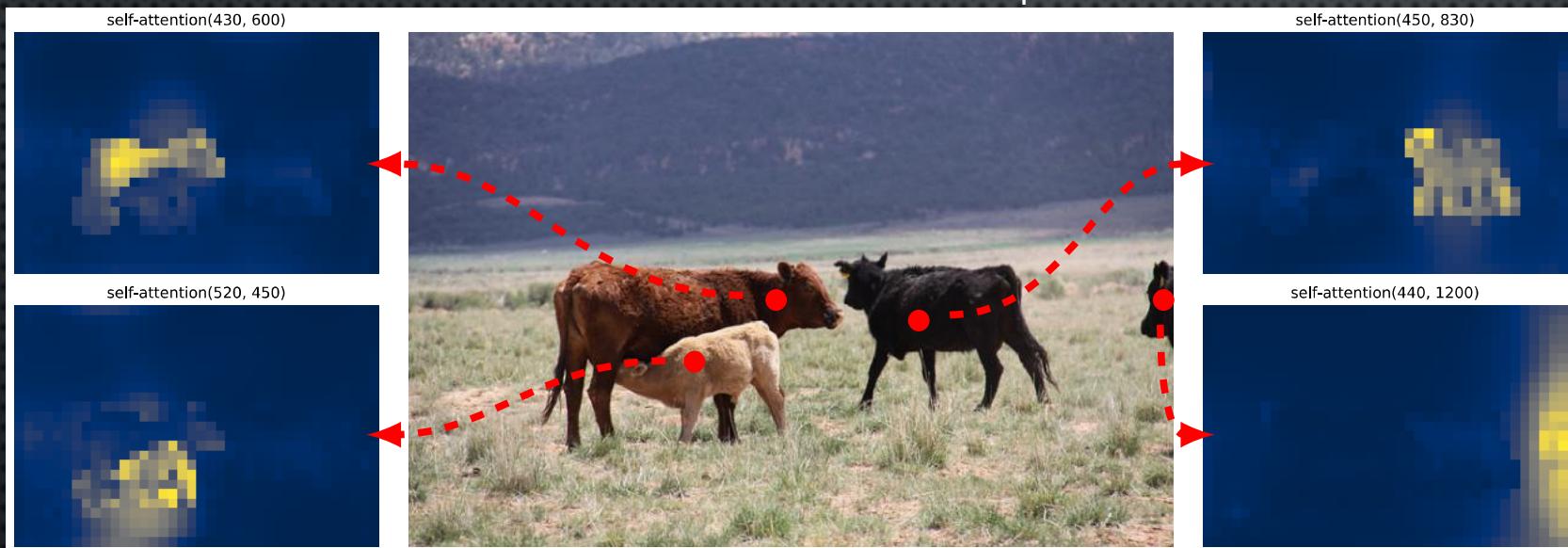
ДЕТЕКТИРОВАНИЕ ОБЪЕКТОВ DETECTION TRANSFORMER (DETR)



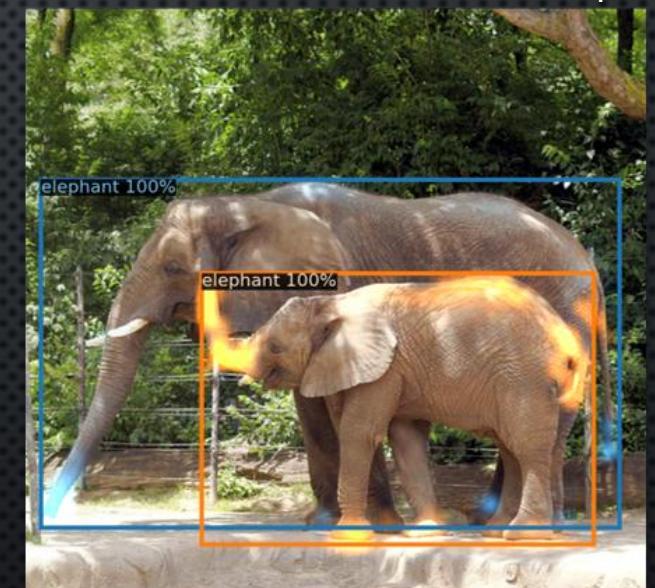
ДЕТЕКТИРОВАНИЕ ОБЪЕКТОВ

DETECTION TRANSFORMER (DETR)

Внимание энкодера

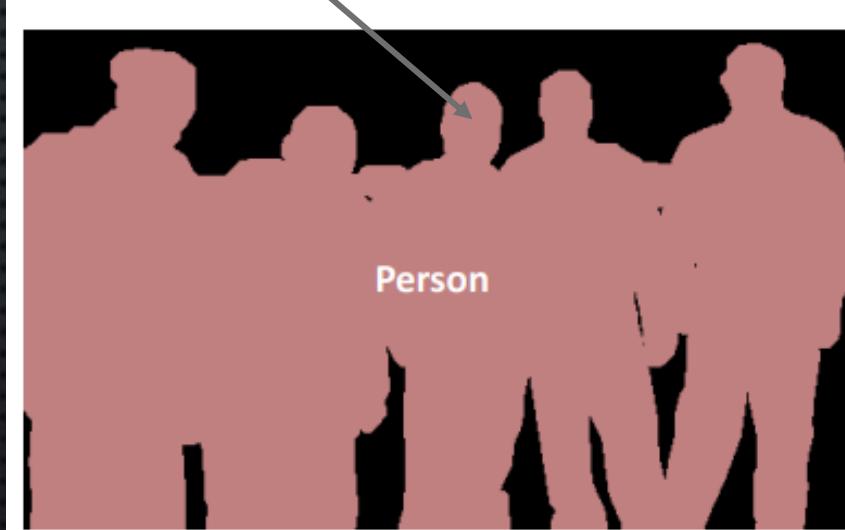


Внимание декодера



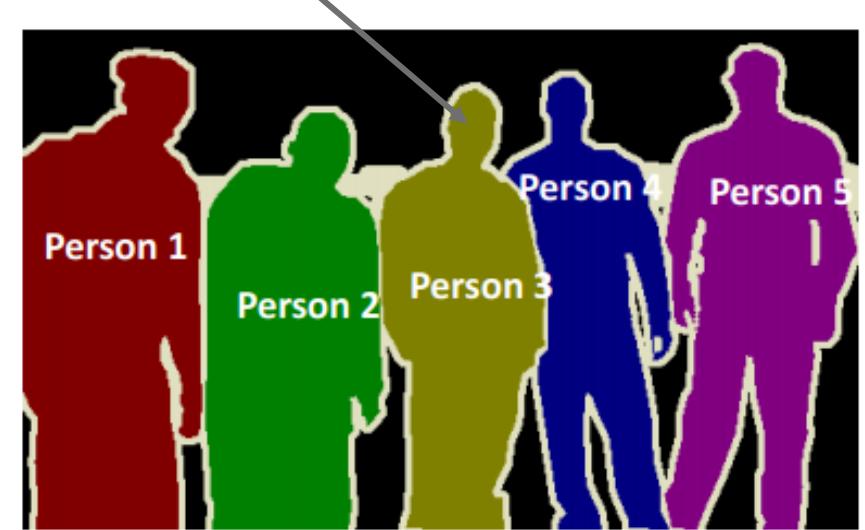
СЕГМЕНТАЦИЯ

Маска вычисляется для изображения в целом



Семантическая сегментация

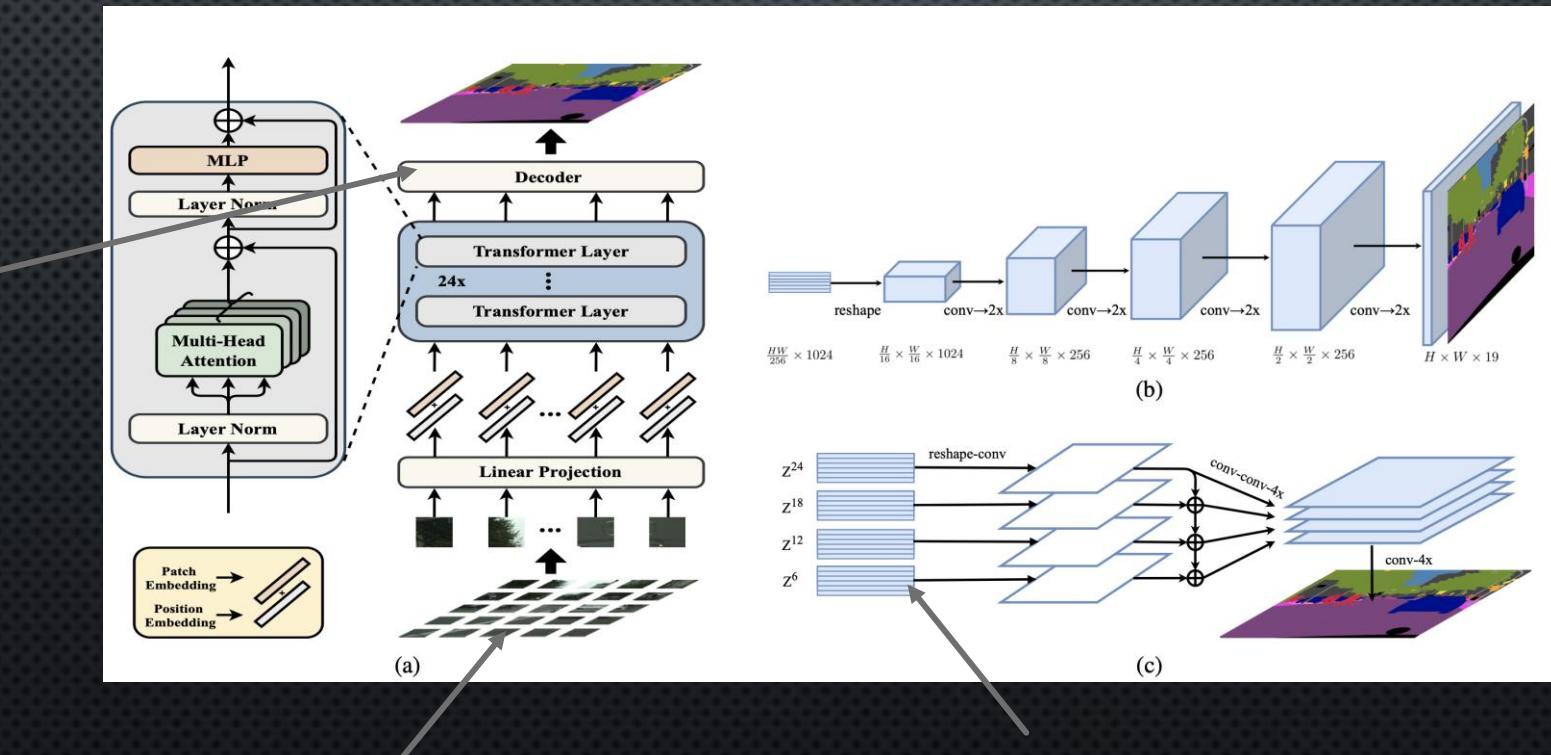
Маска вычисляется для каждого объекта



Объектная сегментация

СЕМАНТИЧЕСКАЯ СЕГМЕНТАЦИЯ. SEGMENTATION TRANSFORMER (SETR)

Свертки
применяются в
декодере

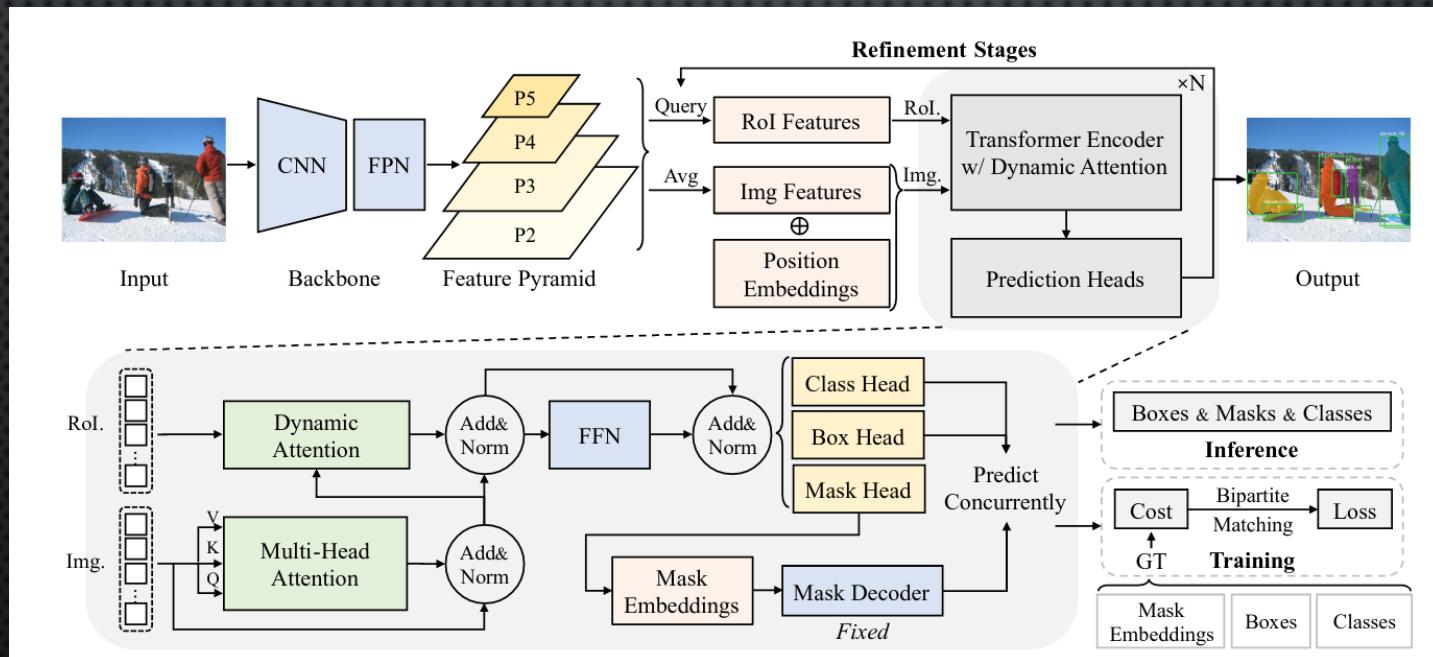


Изображение
обрабатывается напрямую

Применяется аналог FPN

SOTA

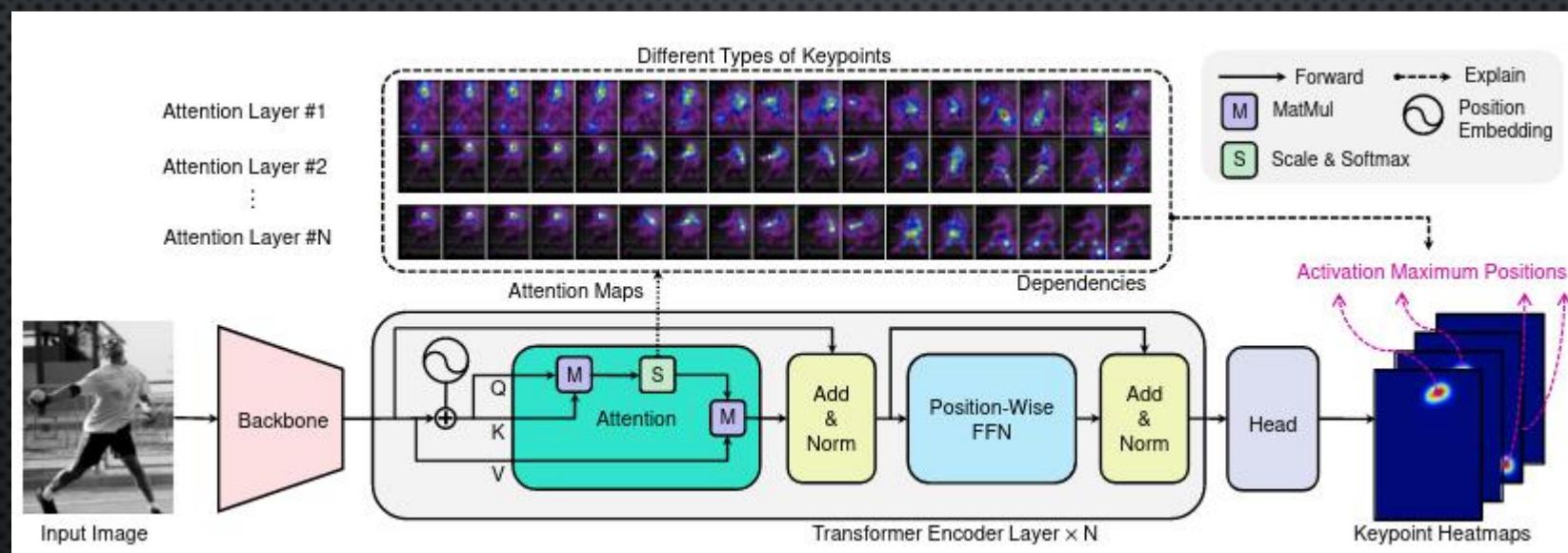
ОБЪЕКТНАЯ СЕГМЕНТАЦИЯ. INSTANCE SEGMENTATION TRANSFORMER



ДЕТЕКТИРОВАНИЕ КЛЮЧЕВЫХ ТОЧЕК



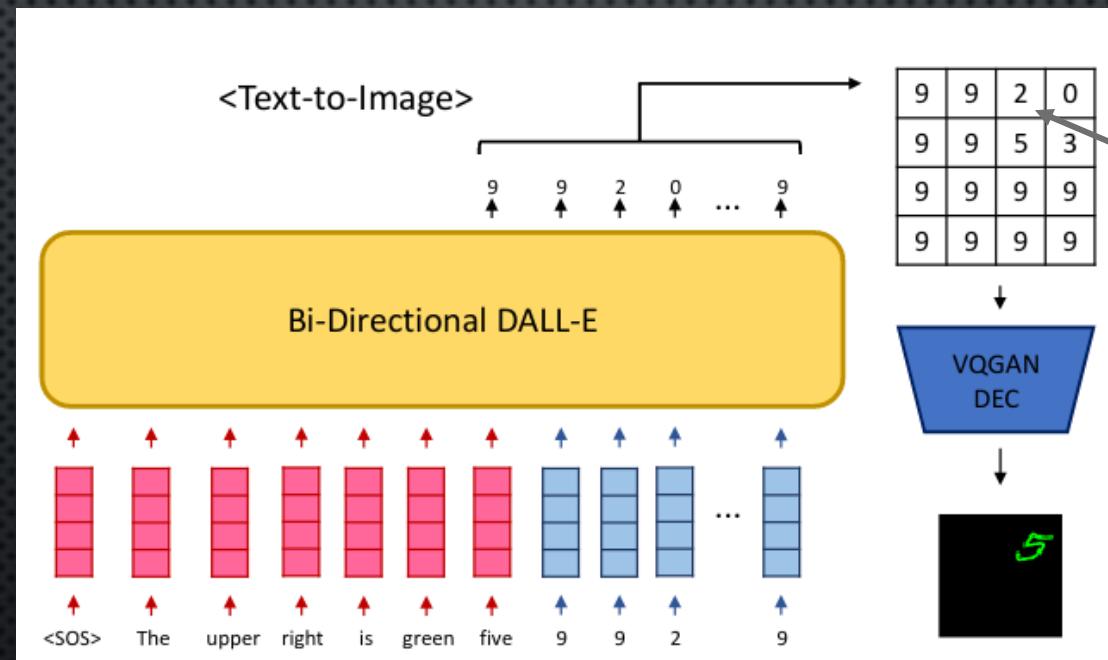
ДЕТЕКТИРОВАНИЕ КЛЮЧЕВЫХ ТОЧЕК. TRANSPOSE



SOTA

Интерпретуемые
предсказания

ГЕНЕРАЦИЯ ИЗОБРАЖЕНИЙ. DALL-E



Используется словарь
для элементов
изображения

НИЗКОУРОВНЕВЫЕ ЗАДАЧИ. ГЕНЕРАЦИЯ ИЗОБРАЖЕНИЙ

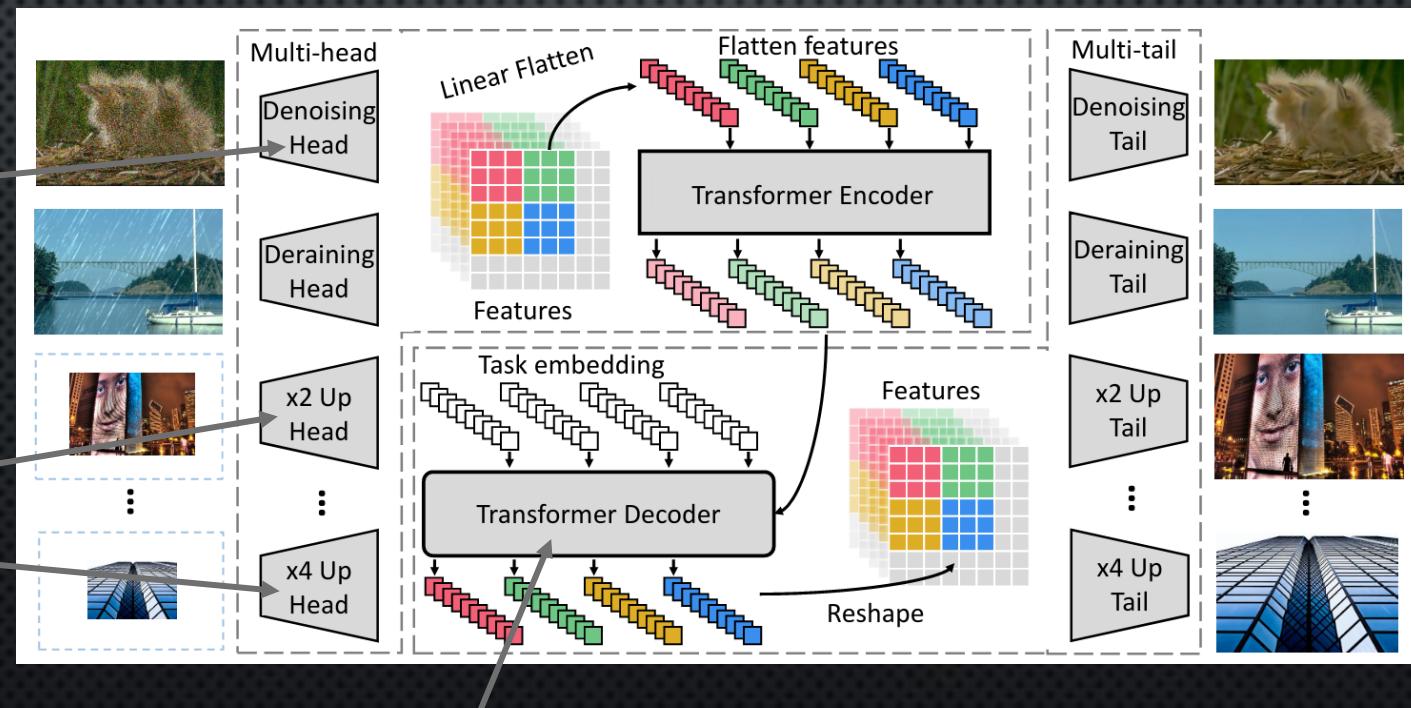
A pentagonal green clock. A green clock in the shape of pentagon



ОБРАБОТКА ИЗОБРАЖЕНИЙ. IMAGE PROCESSING TRANSFORMER

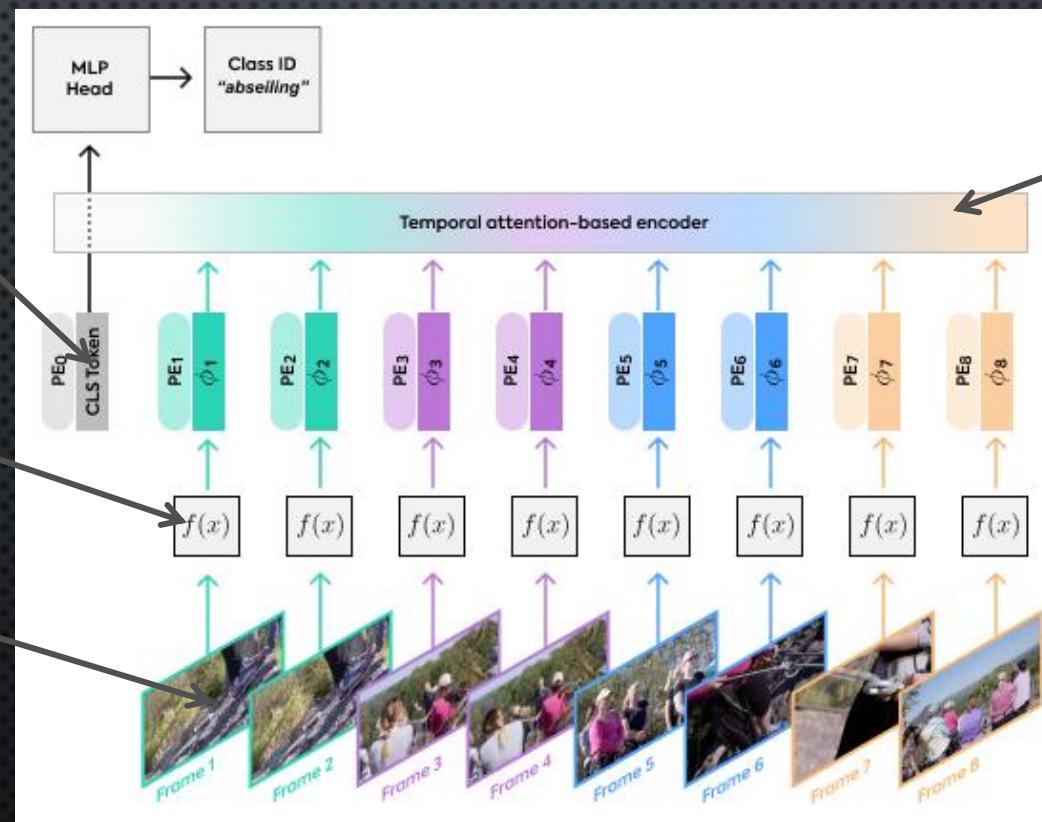
Удаление шума

Повышение
разрешения



Одна сеть трансформер
для всех задач

РАСПОЗНАВАНИЕ ДЕЙСТВИЙ. VIDEO TRANSFORMER NETWORK (VTN)



"Class token" is added to the start of the sequence

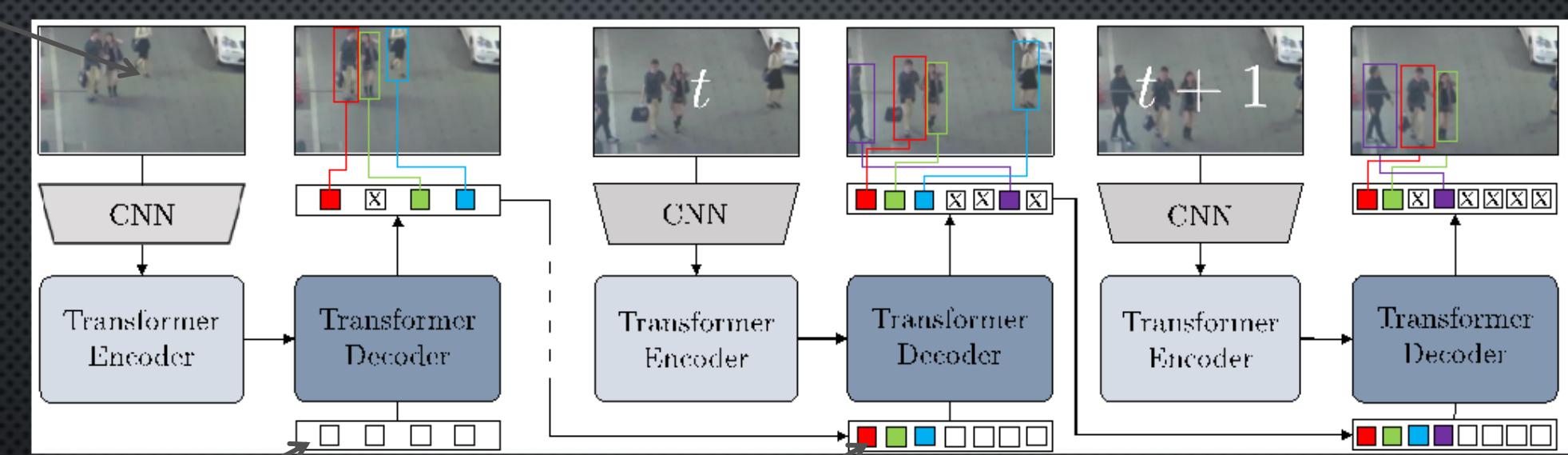
CNN используется для извлечения признаков

Кадры обрабатываются целиком

"Темпоральная" вариация трансформера

ОБРАБОТКА ВИДЕО. ОТСЛЕЖИВАНИЕ ОБЪЕКТОВ

Входной кадр



Объектные
запросы

Запросы
траектории

ВЫВОДЫ

Оригинальный подход

Высокая точность

End-to-end во многих задачах

Универсальный подход. Разнообразные задачи решаются одной архитектурой

Интерпретируемые результаты

Возможно предобучение на неразмеченных данных

Высокая вычислительная стоимость

Долгое обучение по сравнению с CNN



СПАСИБО ЗА
ВНИМАНИЕ