

MDETR - Modulated Detection for End-to-End Multi-Modal Understanding

Докладчик — Седашов Данила

Рецензент — Григорьев Пётр

Практик-исследователь — Денисенко Наталья

Хакер — Кириллов Дмитрий

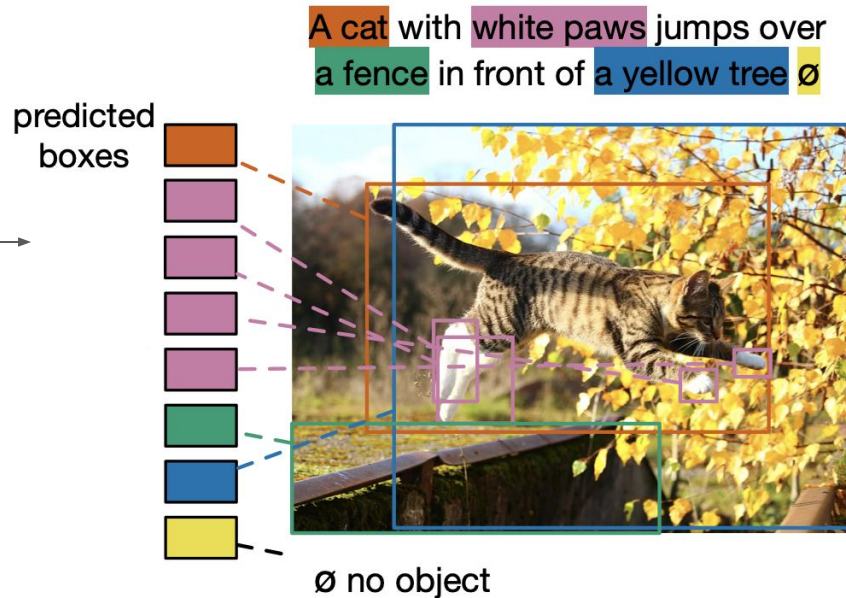
Докладчик

Седашов Данила

Задача modulated detection



“A cat with white paws jumps over
a fence in front of a yellow tree”



Мотивация

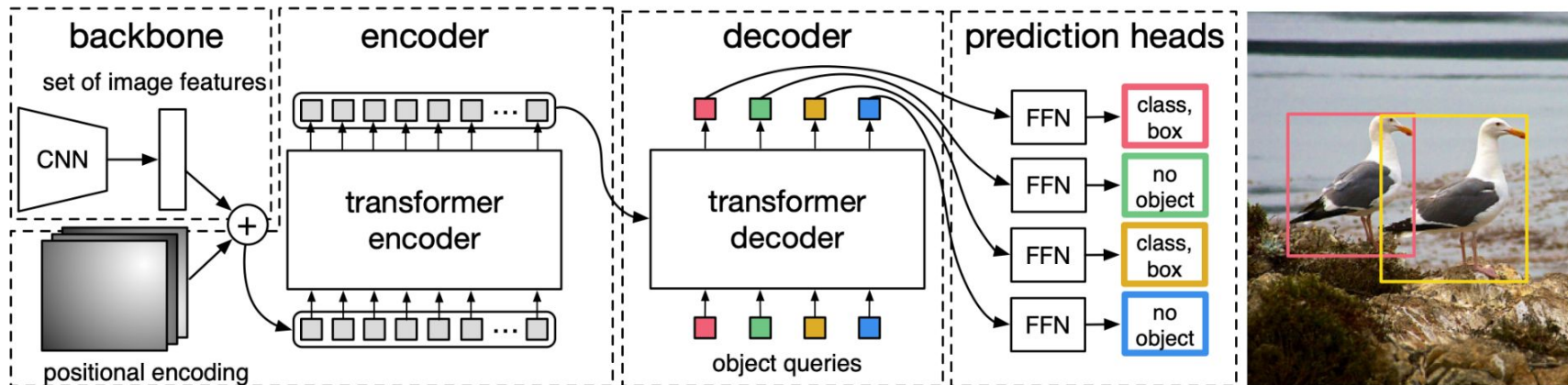
В большинстве подходов используется пайплайн детектор + выравнивание.
Детектор используется как blackbox

- Словарь ограничен набором классов, который способен распознать детектор

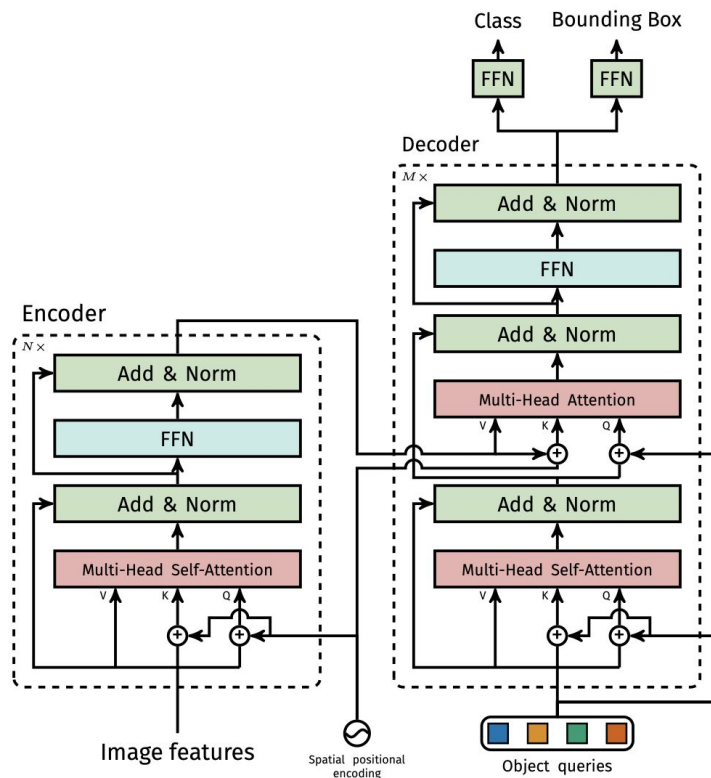
В недавних статьях были попытки решить данную проблему путём усложнения пайплайна, но не было показано, что такие методы способны улучшать качество на downstream задачах

DETR: DEtection TRansformer

- Детектор, основанный на трансформерах
- Визуальные признаки рассматриваются как последовательность и проходят через трансформер
- Трансформер на выходе выдаёт векторы, которые потом преобразуются в bounding boxes и метки классов
- Работает в end-to-end манере за счёт Hungarian Algorithm (поиск соответствия в двудольном графе)

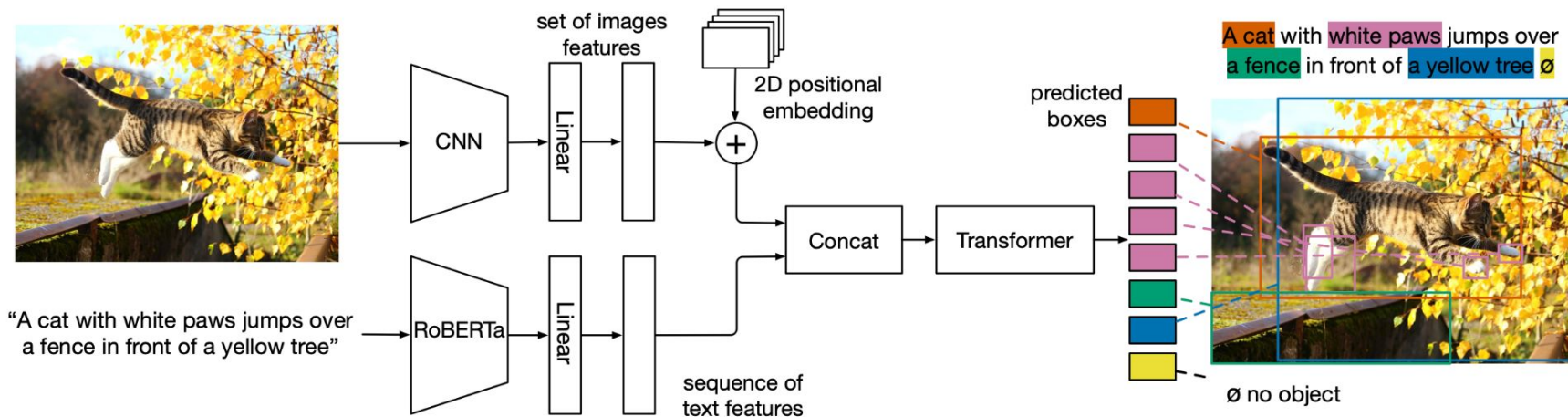


DETR: DEtection TRansformer



Архитектура MDETR

- очень похожа на DETR
- объединяет визуальные и текстовые признаки
- при обучении используются два специфичных лосса



Soft token prediction

- нужен для того, чтобы выравнивать предсказанные гексагоны с токенами в исходном тексте
- суть: учим модель предсказывать не класс (как в обычном детектировании), а soft распределение на порядковых номерах токенов



ø no object



Contrastive alignment

- завязываться исключительно на позиции токенов в тексте не очень хорошо
- хотим, чтобы общие эмбединги объектов были близко к текстовым эмбедингам этих же объектов
- считаем текстовыми признаками выход из энкодера, общими — из декодера

Пусть:

- L — количество токенов,
 N — объектов
- T_i^+ — множество токенов, соответствующих объекту O_i
- O_i^+ — множество объектов, соответствующих токenu t_i

$$l_o = \sum_{i=0}^{N-1} \frac{1}{|T_i^+|} \sum_{j \in T_i^+} -\log \left(\frac{\exp(o_i^\top t_j / \tau)}{\sum_{k=0}^{L-1} \exp(o_i^\top t_k / \tau)} \right)$$

$$l_t = \sum_{i=0}^{L-1} \frac{1}{|O_i^+|} \sum_{j \in O_i^+} -\log \left(\frac{\exp(t_i^\top o_j / \tau)}{\sum_{k=0}^{N-1} \exp(t_i^\top o_k / \tau)} \right)$$

Pre-training

- предобучение на наборе данных размера 1.3M
- задача phrase grounding
- при составлении набора данных комбинируют несколько датасетов (некоторые составлены для других задач multi-modal reasoning)
- SOTA на задаче phrase grounding!

Method	Val			Test		
	R@1	R@5	R@10	R@1	R@5	R@10
ANY-BOX-PROTOCOL						
BAN [22]	-	-	-	69.7	84.2	86.4
VisualBert[26]	68.1	84.0	86.2	-	-	-
VisualBert†[26]	70.4	84.5	86.3	71.3	85.0	86.5
MDETR-R101	78.9	88.8	90.8	-	-	-
MDETR-R101†*	82.5	92.9	94.9	83.4	93.5	95.3
MDETR-ENB3†*	82.9	93.2	95.2	84.0	93.8	95.6
MDETR-ENB5†*	83.6	93.4	95.1	84.3	93.9	95.8
MERGED-BOXES-PROTOCOL						
CITE [43]	-	-	-	61.9	-	-
FAOG [66]	-	-	-	68.7	-	-
SimNet-CCA [45]	-	-	-	71.9	-	-
DDPN [71]	72.8	-	-	73.5	-	-
MDETR-R101	79.0	86.7	88.6	-	-	-
MDETR-R101†*	82.3	91.8	93.7	83.8	92.7	94.4

Referring expression comprehension

- немного другая задача: нужно выделить один объект, про который говорится в тексте, а не все сразу
- Finetune 5 epochs
- SOTA на всех бенчмарках

Method	Detection backbone	Pre-training image data	RefCOCO			RefCOCO+			RefCOCOg	
			val	testA	testB	val	testA	testB	val	test
MAttNet[69]	R101	None	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
ViLBERT[34]	R101	CC (3.3M)	-	-	-	72.34	78.52	62.61	-	-
VL-BERT L [54]	R101	CC (3.3M)	-	-	-	72.59	78.57	62.30	-	-
UNITER.L[6]*	R101	CC, SBU, COCO, VG (4.6M)	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
VILLA.L[9]*	R101	CC, SBU, COCO, VG (4.6M)	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
ERNIE-ViL.L[68]	R101	CC, SBU (4.3M)	-	-	-	75.95	82.07	66.88	-	-
MDETR	R101	COCO, VG, Flickr30k (200k)	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
MDETR	ENB3	COCO, VG, Flickr30k (200k)	87.51	90.40	82.67	81.13	85.52	72.96	83.35	83.31

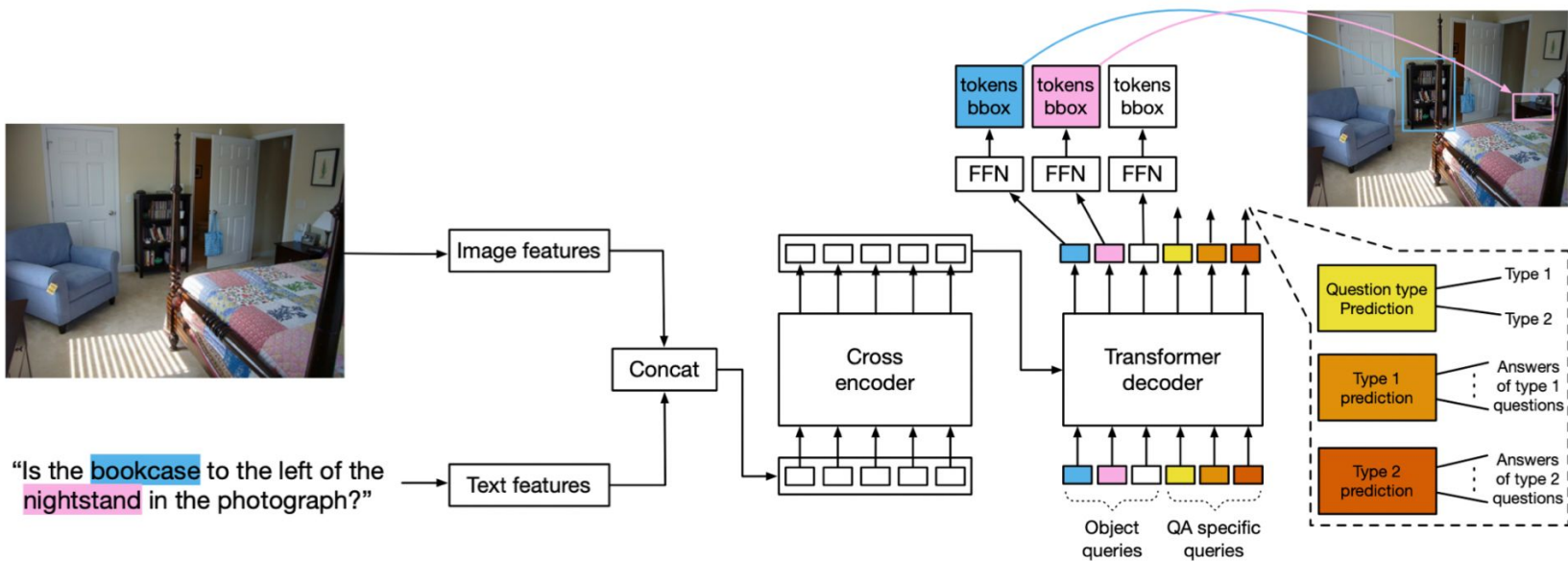
Referring expression segmentation

- Файнтюнинг в 2 шага: сначала дообучить на bounding boxes вокруг объектов, потом отдельно обучить голову для сегментации
- SOTA на одном из бенчмарков

Method	Backbone	PhraseCut			
		M-IoU	Pr@0.5	Pr@0.7	Pr@0.9
RMI[3]	R101	21.1	22.0	11.6	1.5
HULANet[62]	R101	41.3	42.4	27.0	5.7
MDETR	R101	53.1	56.1	38.9	11.9
MDETR	ENB3	53.7	57.5	39.9	11.9

Visual question answering

- добавляем новые object queries для типа вопроса и ответа на каждый тип



Visual question answering

Method	Pre-training img data	Test-dev	Test-std
MoVie [39]	-	-	57.10
LXMERT[55]	VG, COCO (180k)	60.0	60.33
VL-T5 [7]	VG, COCO (180k)	-	60.80
MMN [5]	-	-	60.83
OSCAR [28]	VG, COCO, Flickr, SBU (4.3M)	61.58	61.62
NSM [19]	-	-	63.17
VinVL [72]	VG, COCO, Objects365, SBU Flickr30k, CC, VQA, OpenImagesV5 (5.65M)	65.05	64.65
MDETR-R101	VG, COCO, Flickr30k (200k)	62.48	61.99
MDETR-ENB5	VG, COCO, Flickr30k (200k)	62.95	62.45

Table 5: Visual question answering on the GQA dataset.

CLEVR datasets

Method	CLEVR	CLEVR-Hu		CoGenT		CLEVR-Ref+
	Overall	- FT	+ FT	TestA	TestB	Acc
MAttNet[69]	-	-	-	-	-	60.9
MGA-Net[73]	-	-	-	-	-	80.1
FiLM[42]	97.7	56.6	75.9	98.3	78.8	-
MAC [17]	98.9	57.4	81.5	-	-	-
NS-VQA[67]*	99.8	-	67.8	99.8	63.9	-
OCCAM [60]	99.4	-	-	-	-	-
MDETR	99.7	59.9	81.7	99.8	76.7	100

Few-shot transfer

- детектор на заданном наборе классов из предобученной модели
- в pre-training нет случаев, когда нет ни одного соответствия между текстом и изображением; дообучаем на части размеченных данных
- LVIS dataset

Method	Data	AP	AP50	AP _r	AP _c	AP _f
Mask R-CNN	100%	33.3	51.1	26.3	34.0	33.9
DETR	1%	4.2	7.0	1.9	1.1	7.3
DETR	10%	13.7	21.7	4.1	13.2	15.9
DETR	100%	17.8	27.5	3.2	12.9	24.8
MDETR	1%	16.7	25.8	11.2	14.6	19.5
MDETR	10%	24.2	38.0	20.9	24.9	24.3
MDETR	100%	22.5	35.2	7.4	22.7	25.0

Итоги

- предложен end-to-end text-modulated detector на основе DETR
- модель применяется для двух задач — phrase grounding и referring expression comprehension — и устанавливает новую SOTA на этих задачах
- модель обобщается для других multimodal reasoning задач, показывает там конкурентные результаты

Рецензент

Григорьев Петр

Вклад

- Не ограничен предобученными визуальными детекторами
- Не ограничен фиксированным словарем
- Обнаруживает по любому запросу в свободной форме
- Непривычные комбинации объектов и атрибутов



Достоинства и недостатки

Качество: метрики показывают улучшение в качестве на представленных датасетах

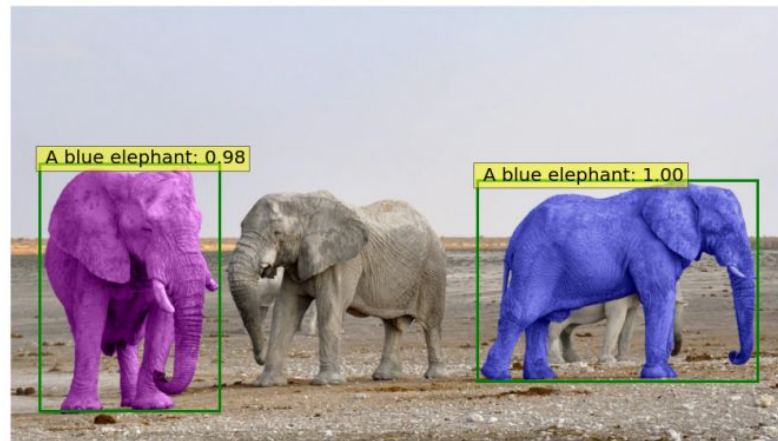
Новизна: исправлен существенный недостаток многих предыдущих моделей

Доказательность: подробные объяснение, экспериментальные подтверждения

Рефлексия: подробно описаны недостатки метода

Сравнение: с некоторыми методами сравнение только на одном наборе данных (Oscar, VinVL)

Теряется смысл:



(b) Text prompt: “A blue elephant.”

Оформление

- Подробное описание методов-предшественников
- Понятные описания, проиллюстрированы примерами
- Есть выложенная запись доклада с ответами на вопросы
- Подробно описаны детали реализации

Воспроизводимость

- Подробно описаны детали реализации
- Есть выложенный код
- Есть Colab ноутбук с разобранными примерами
- Есть подробная инструкция по использованию

```
plot_inference_qa(im3, "What is tied to the bike?")
```



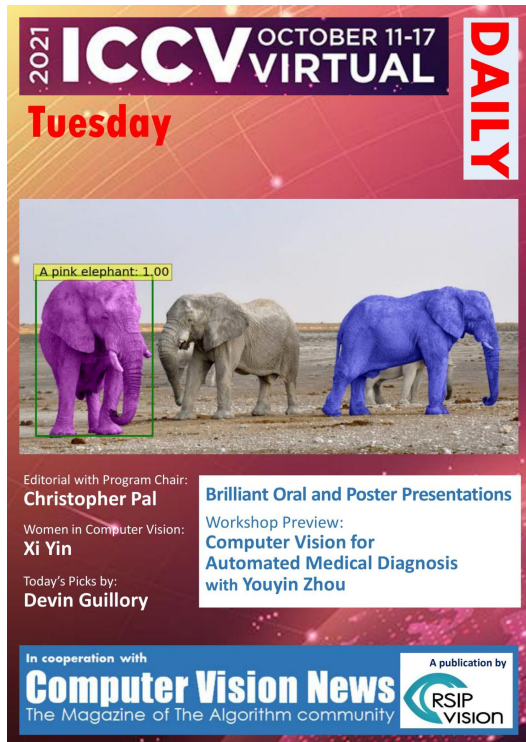
Predicted answer: bag confidence=48.01

Практик-исследователь

Денисенко Наталья

Публикация

- ICCV 2021
- oral



2021 **ICCV** OCTOBER 11-17 **VIRTUAL** **DAILY**

Tuesday

A pink elephant: 1.00

Editorial with Program Chair:
Christopher Pal

Women in Computer Vision:
Xi Yin

Today's Picks by:
Devin Guillory

Brilliant Oral and Poster Presentations

Workshop Preview:
**Computer Vision for
Automated Medical Diagnosis
with Youyin Zhou**

In cooperation with **Computer Vision News**
The Magazine of The Algorithm community

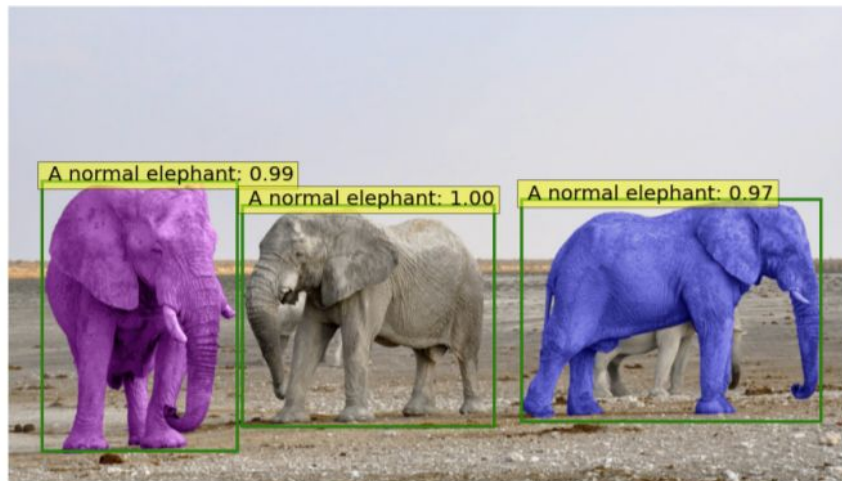
A publication by **RSIP vision**

Цитирования

22 цитирования

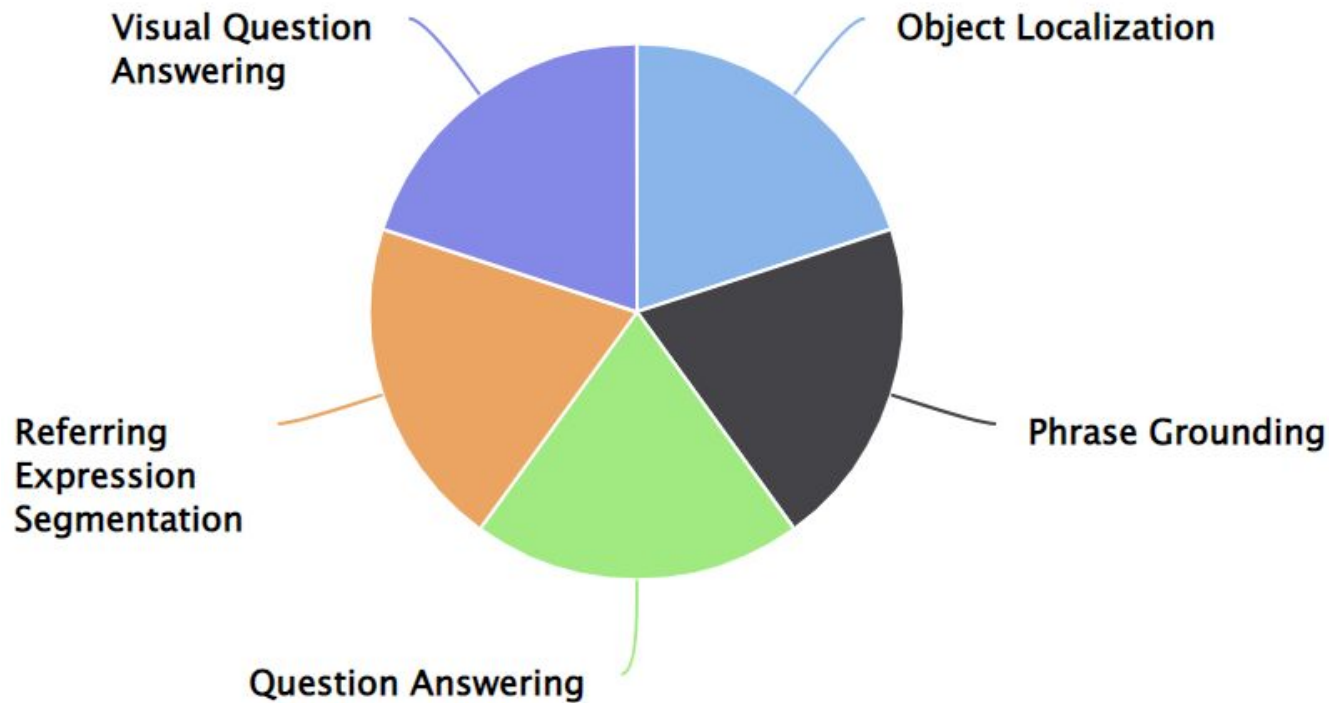
About 22 results (0.03 sec)

Развитие статьи



(c) Text prompt: "A normal elephant."

Применения



Авторы



Aishwarya Kamath



Mannat Singh



Yann LeCun



Gabriel Synnaeve



Ishan Misra



Nicolas Carion