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

Аюпов Шамиль
Степанов Никита
Данг Куинь Ньы
Еленик Константин

Мультимодальность

Ситуация, когда есть данные разной природы.

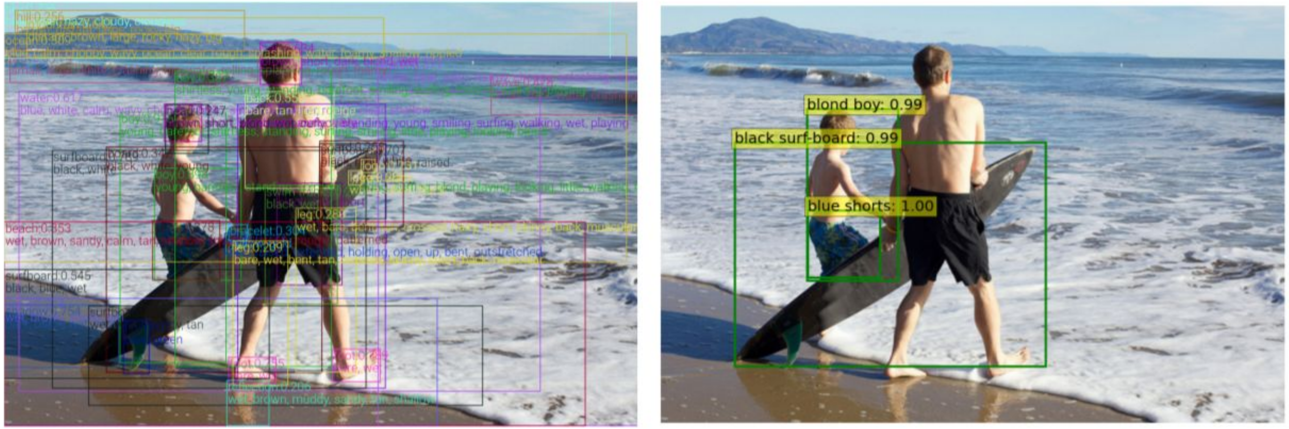
Скомбинировать данные разной природы – нетривиальная задача.

Задача MDETR: детекция, обусловленная на текст.

Мотивация

Мультимодальные системы часто используют для подзадач blackbox модели.

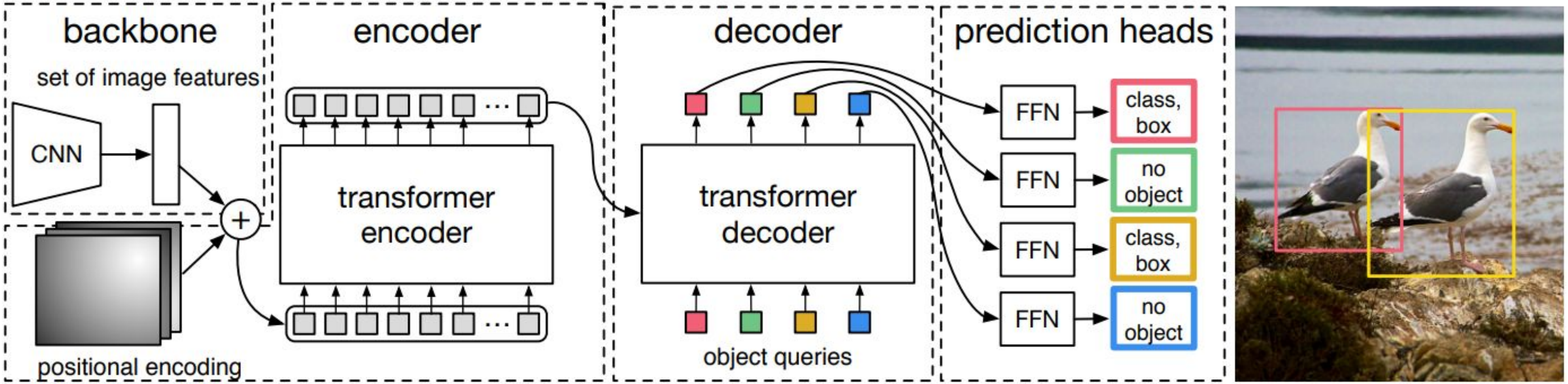
Можно ли это дело обучать end2end и улучшить за счет этого качество?



(a) Current object detection pipeline outputs, predicting all possible objects in the image. This extensive annotation is essential to multi-modal understanding systems that treat detection as a black box.

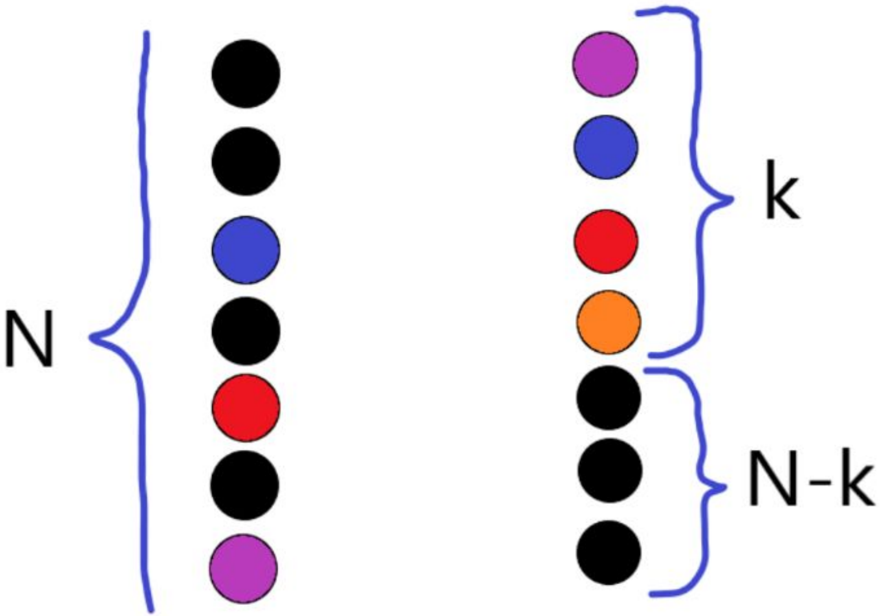
(b) MDETR predicts boxes relevant to the caption and labels them with the corresponding spans from the text. Here we use the caption: "blond boy wearing blue shorts. a black surf-board."

Recap: DETR



Фиксированное число классов (+ [no_object])

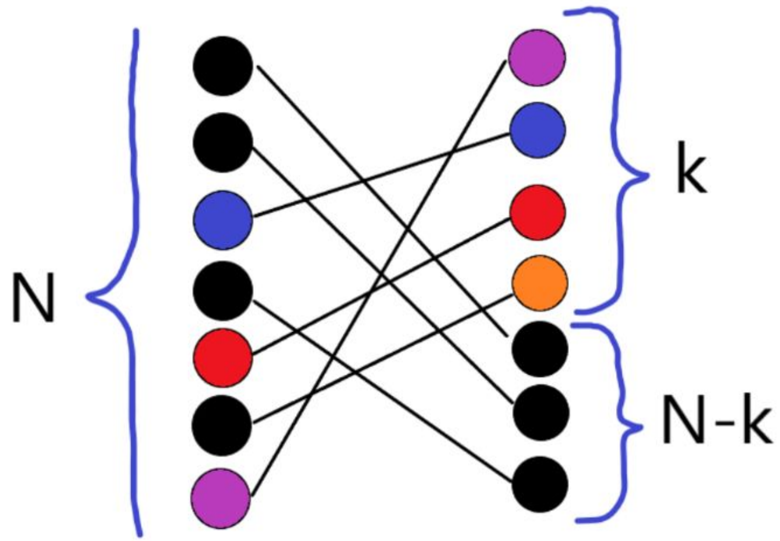
Recap: DETR Matching



$$w(\mathrm{u}, \mathrm{v}) = \begin{cases} 0, \text{ если v — фиктивная} \\ -\mathbb{P}[\mathrm{class}(\mathrm{u}) = \mathrm{class}(\mathrm{v})] + \mathcal{L}_{\mathrm{box}}(\mathrm{u}, \mathrm{v}) \end{cases}$$

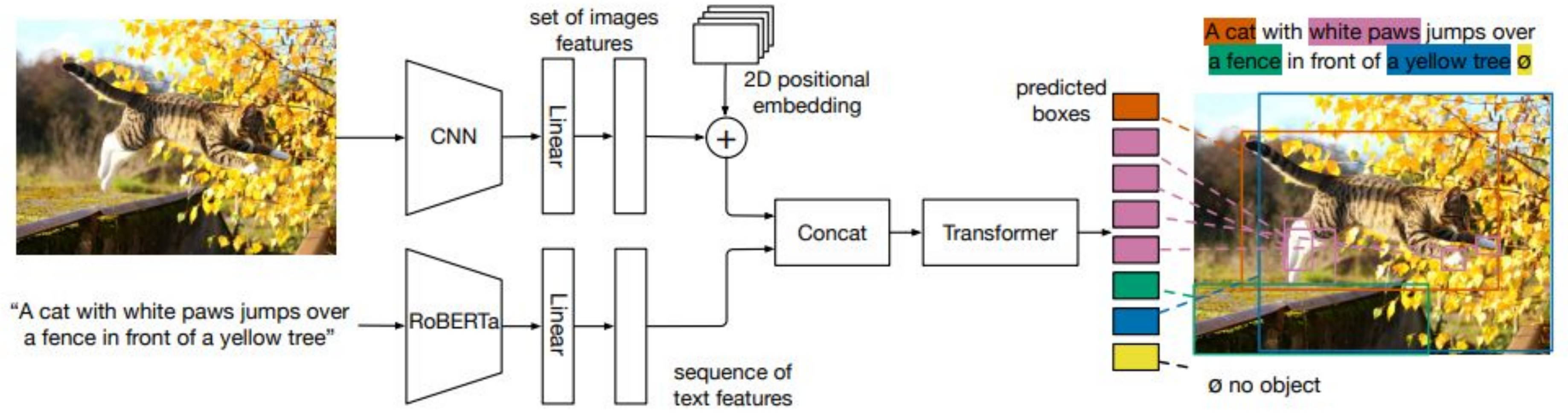
https://github.com/bayesgroup/HSE_ML_research_seminar/blob/master/2020-2021/182/40_Stepanov_Image_Transformers.pdf

Recap: DETR Loss



$$\mathcal{L} = \sum_{(\mathrm{u},\mathrm{v})\in \mathrm{M}} \left(-\ln \mathbb{P}[\text{class}(\mathrm{u}) = \text{class}(\mathrm{v})] + [\text{v не фиктивная}]\mathcal{L}_{\text{box}}(\mathrm{u},\mathrm{v}) \right)$$

MDETR. Архитектура



Нет фиксированных классов – как матчить сущности и считать Loss?

Soft Token Prediction

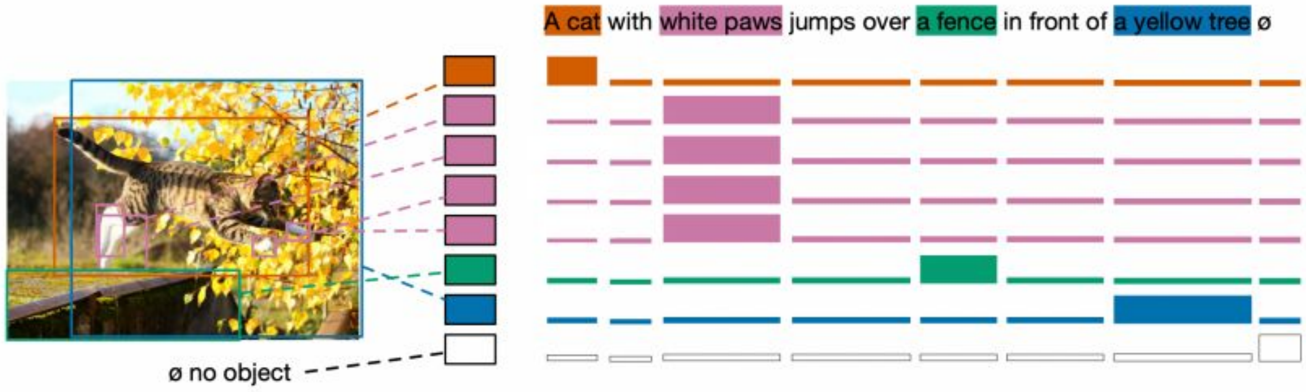


Figure 6: Illustration of the soft-token classification loss. For each object, the model predicts a distribution over the token positions in the input sequence. The weight of the distribution should be equally spread over all the tokens that refer to the predicted box.

Хотим предсказывать положение объектов в тексте

Cross Entropy Loss.

Макс. L = 256 токенов.

Contrastive Alignment

Хотим, чтобы представления объектов и соответствующих токенов были близки. (InfoNCE)

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

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

o_i - представления объектов (выход декодера)

t_i – представления текстовых токенов (выход кросс-энкодера)

MDETR. Обучение

$\mathcal{L}_{box}$ – стандартные для детекции L1 и GIoU.

Matching: $\mathcal{L}_{ST} + \mathcal{L}_{box}$

Loss: $\mathcal{L}_{ST} + \mathcal{L}_{CA} + \mathcal{L}_{box}$

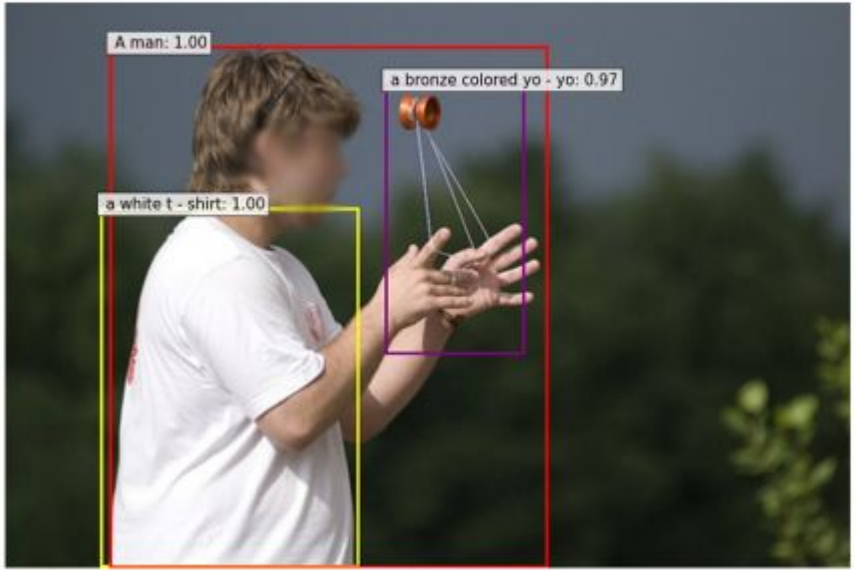
Pretrain 40 эпох на обработанной солянке из Flickr30k, COCO и Visual Genome

Downstream tasks

- Phrase Grounding
- Referring Expression Comprehension
- Referring Expression Segmentation
- Visual Question Answering

Везде, кроме Phrase Grounding, нетривиальная адаптация.

Phrase Grounding



(c) "A man in a white t-shirt does a trick with a bronze colored yo-yo"

Method	Val R@1	Val R@5	Val R@10	Test R@1	Test R@5	Test 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**

Table 3: Results on the phrase grounding task on Flickr30k enti-

Referring Expression Comprehension

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**

(b) "zebra facing away"

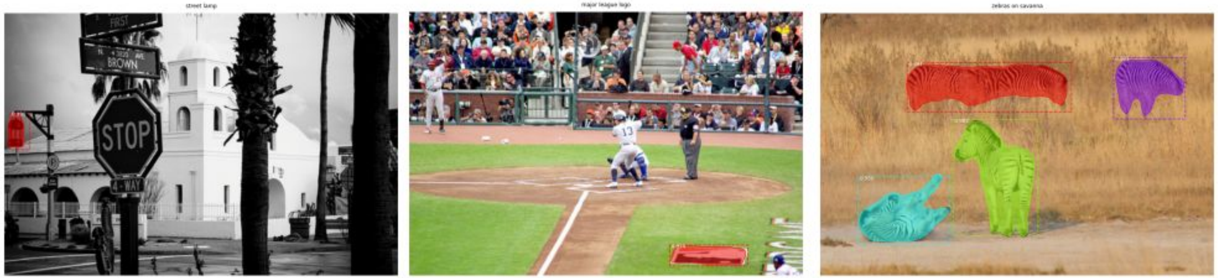
(c) "the man in the red shirt carrying baseball bats"

(d) "the front most cow to the right of the other cows"

Referring Expression Segmentation

Segmentation head
Dice/F1 loss + Focal loss

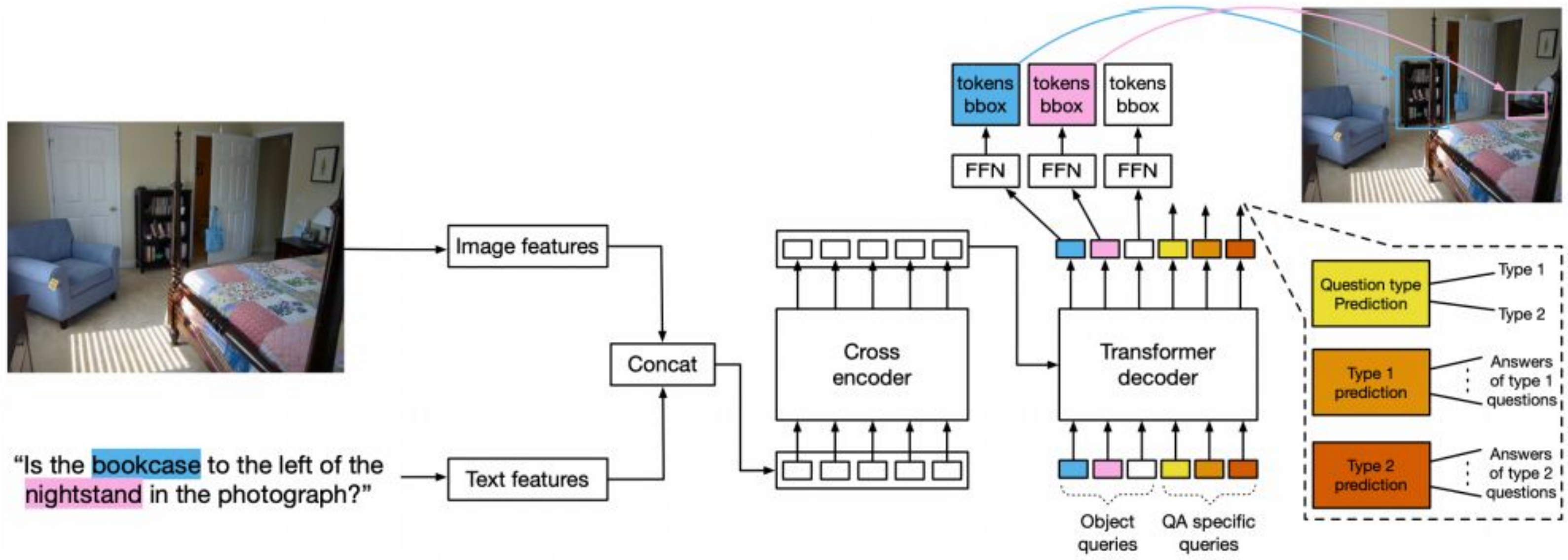
Method	Backbone	PhraseCut			
		M-IoU	Pr@0.5	Pr@0.7	Pr@0.9
RMI[3]	R101	21.1	22.0	11.6	1.5
HULANet[62]	R101	41.3	42.4	27.0	5.7
MDETR	R101	**53.1**	**56.1**	**38.9**	**11.9**
MDETR	ENB3	**53.7**	**57.5**	**39.9**	**11.9**



(a) Query: "street lamp" **(b)** Query: "major league logo" **(c)** Query: "zebras on savanna"

Figure 8: Qualitative segmentation examples on the phrasecut dataset

Visual Question Answering



Visual Question Answering 2

Figure 5: MDETR provides interpretable predictions as seen here. For the question "What is on the table?", MDETR fine-tuned on GQA predicts boxes for key words in the question, and is able to provide the correct answer as "laptop". Image from COCO val set.

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.

Заключение

- MDETR – мультимодальная система (изображение + текст)
- Предобучается на задачу детекции, обусловленной на текст
- Хорошо дообучается на downstream задачи

Список источников

- MDETR -- Modulated Detection for End-to-End Multi-Modal Understanding
- End-to-End Object Detection with Transformers
- https://github.com/bayesgroup/HSE_ML_research_seminar/blob/master/2020-2021/182/40_Stepanov_Image_Transformers.pdf