When and why vision-language models behave like bags-of-word, and what to do about it

6.11.2023, НИС, доклад подготовил: Потапов Ю.

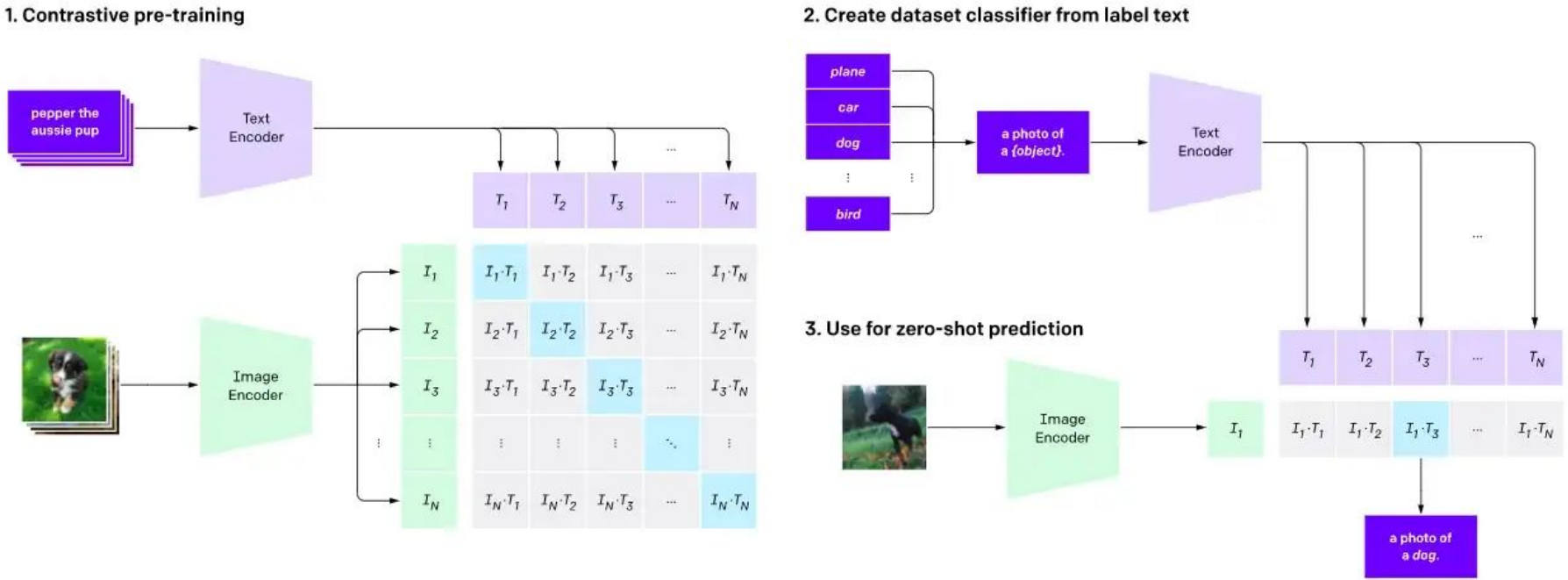
Vision-Language models

["Lion and cheetah",
"Cat and dog",
"Rabbit and bird"]

Vision-Language Model

[{'score': 0.9950, 'label': 'cat and dog'},
{'score': 0.0048, 'label': 'rabbit and lion'},
{'score': 8.81e-05, 'label': 'lion and cheetah'}]

Как обучается CLIP?



Pre-Training

1) взяли батч n*n
2) посчитали косинусную похожесть каждых двух
3) максимизируем диагональ и минизируем всё, что вне диагонали

Так ли хороша CLIP или нет?

1) понимает ли CLIP разницу между: (порядок отношения между словами)

   “the horse is eating the grass” и “the grass is eating the horse”

2) понимает ли CLIP предложение композиционно? (спойлер – нет)

3) понимает ли CLIP разницу между (порядок отношения атрибутов слов): “the paved road and the white house”, the white road and the paved house”

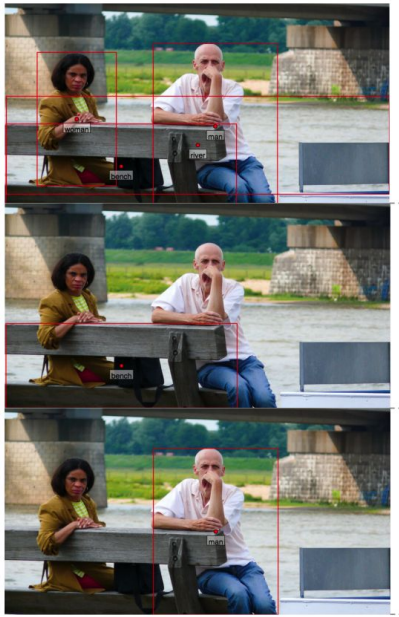
Attribution, Relation, and Order benchmark (ARO) for fine-grained evaluation of VLMs' relation, attribution, and order understanding.

авторы статьи предлагают использовать старые и новые датасеты для проверки **VLM**

1) **Visual Genome** – большой датасет с 100000 изображениями, аннотированных объектами, атрибутами и связями

   Авторы используют его, чтобы создать два новых датасета:
   Visual Genome Relation: x relation **y** != **y** relation **x**

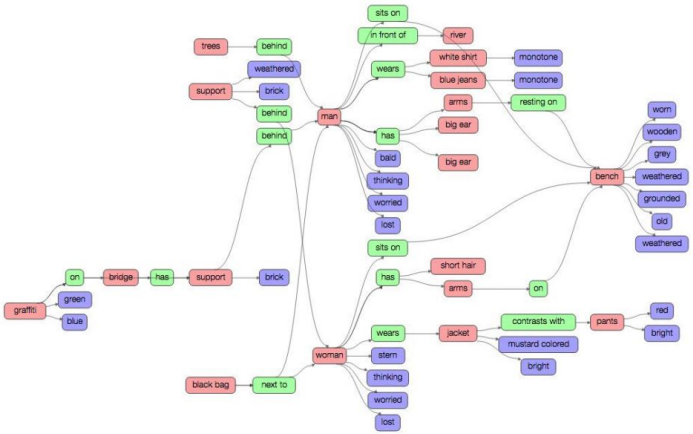
   Visual Genome Attribution: модели предоставляется выбор: “the **crouched cat** and the **open door**” and “the **open cat** and the **crouched door**”



A man and a woman sit on a park bench along a river.

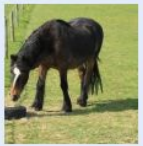
Park bench is made of gray weathered wood

The man is almost bald



Visual Genome Relation

Assessing relational understanding (23,937 test cases)

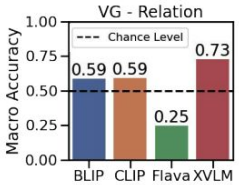


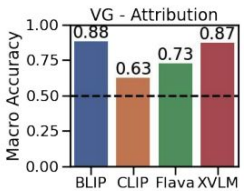
✔ *the horse is eating the grass*
✘ *the grass is eating the horse*

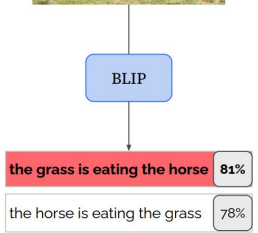
Visual Genome Attribution

Assessing attributive understanding (28,748 test cases)

✔ *the paved road and the white house*
✘ *the white road and the paved house*







COCO Order and Flickr Order

COCO Order and Flickr Order

Assessing sensitivity to order (6,000 test cases)

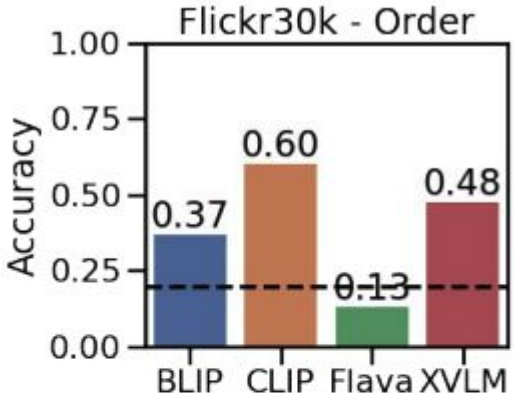
✔ a brown cat is looking at a gray dog and sitting in a white bathtub

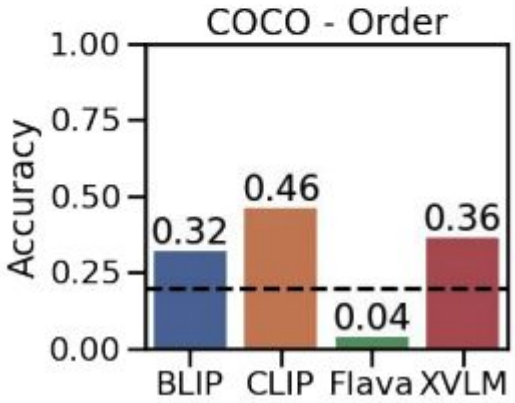
✗ (shuffle adjective/noun) a gray bathtub is looking at a white cat and sitting in a brown dog

✗ (shuffle all but adjective/noun) at brown cat a in looking a gray dog sitting is and a white bathtub

✗ (shuffle words within trigrams) cat brown a at is looking a gray dog in and sitting bathtub a white

✗ (shuffle trigrams) a brown cat a white bathtub is looking at a gray dog and sitting in





Результаты в Visual Genome:

	CLIP	NegCLIP	CLIP-FT	XVLM	BLIP	Flava	# Samples
Accuracy	0.59	0.8	0.64	0.73	0.59	0.24	
Spatial Relationships							
Accuracy	0.56	0.66	0.57	0.74	0.66	0.34	
above	0.48	0.60	0.54	0.80	0.64	0.55	269
at	0.59	0.93	0.71	0.72	0.49	0.15	75
behind	0.56	0.29	0.34	0.82	0.77	0.28	574
below	0.56	0.46	0.48	0.74	0.69	0.44	209
beneath	0.80	0.70	0.70	0.80	0.70	0.40	10
in	0.63	0.89	0.63	0.73	0.72	0.09	708
in front of	0.54	0.75	0.70	0.66	0.55	0.78	588
inside	0.50	0.91	0.67	0.69	0.72	0.12	58
on	0.52	0.86	0.58	0.86	0.76	0.12	1684
on top of	0.43	0.75	0.58	0.85	0.79	0.19	201
to the left of	0.49	0.50	0.50	0.52	0.51	0.50	7741
to the right of	0.49	0.50	0.50	0.52	0.49	0.51	7741
under	0.64	0.43	0.54	0.86	0.73	0.27	132
Verbs							
Accuracy	0.61	0.86	0.66	0.73	0.56	0.2	
carrying	0.33	0.83	0.75	0.75	0.67	0.08	12
covered by	0.47	0.36	0.36	0.61	0.58	0.56	36
covered in	0.79	0.50	0.50	0.14	0.29	0.14	14
covered with	0.56	0.56	0.50	0.56	0.50	0.19	16
covering	0.39	0.58	0.45	0.67	0.55	0.06	33
cutting	0.75	0.83	0.83	0.67	0.25	0.00	12
eating	0.57	1.00	0.67	0.62	0.52	0.00	21
feeding	0.90	0.80	0.80	0.60	0.30	0.20	10
grazing on	0.10	0.90	0.30	0.60	0.40	0.50	10
hanging on	0.79	1.00	0.93	0.93	0.79	0.00	14
holding	0.58	0.97	0.79	0.67	0.44	0.27	142
leaning on	0.67	1.00	1.00	0.75	0.58	0.08	12
looking at	0.84	1.00	0.68	0.68	0.55	0.26	31
lying in	0.47	1.00	0.60	0.87	0.67	0.00	15
lying on	0.60	0.88	0.50	0.93	0.75	0.17	60
parked on	0.67	0.86	0.38	0.76	0.86	0.00	21
reflected in	0.64	0.71	0.57	0.50	0.43	0.43	14
resting on	0.38	0.85	0.23	0.92	0.54	0.15	13
riding	0.71	0.98	0.78	0.82	0.41	0.02	51
sitting at	0.62	1.00	0.88	0.88	0.46	0.00	26
sitting in	0.57	0.96	0.78	0.87	0.83	0.30	23
sitting on	0.58	0.97	0.78	0.94	0.73	0.14	175
sitting on top of	0.50	0.90	0.80	1.00	0.80	0.10	10
standing by	0.67	0.92	0.67	0.83	0.67	0.67	12
standing in	0.73	0.98	0.69	0.69	0.49	0.05	59
standing on	0.60	1.00	0.63	0.83	0.73	0.06	52
surrounded by	0.64	0.71	0.64	0.71	0.64	0.79	14
using	0.84	1.00	1.00	0.68	0.58	0.00	19
walking in	0.70	1.00	0.70	0.60	0.50	0.00	10
walking on	0.79	1.00	0.79	0.84	0.42	0.05	19
watching	0.45	0.55	0.27	0.59	0.68	0.36	22
wearing	0.47	0.99	0.88	0.68	0.48	0.64	949

Результаты в COCO Order and Flickr Order

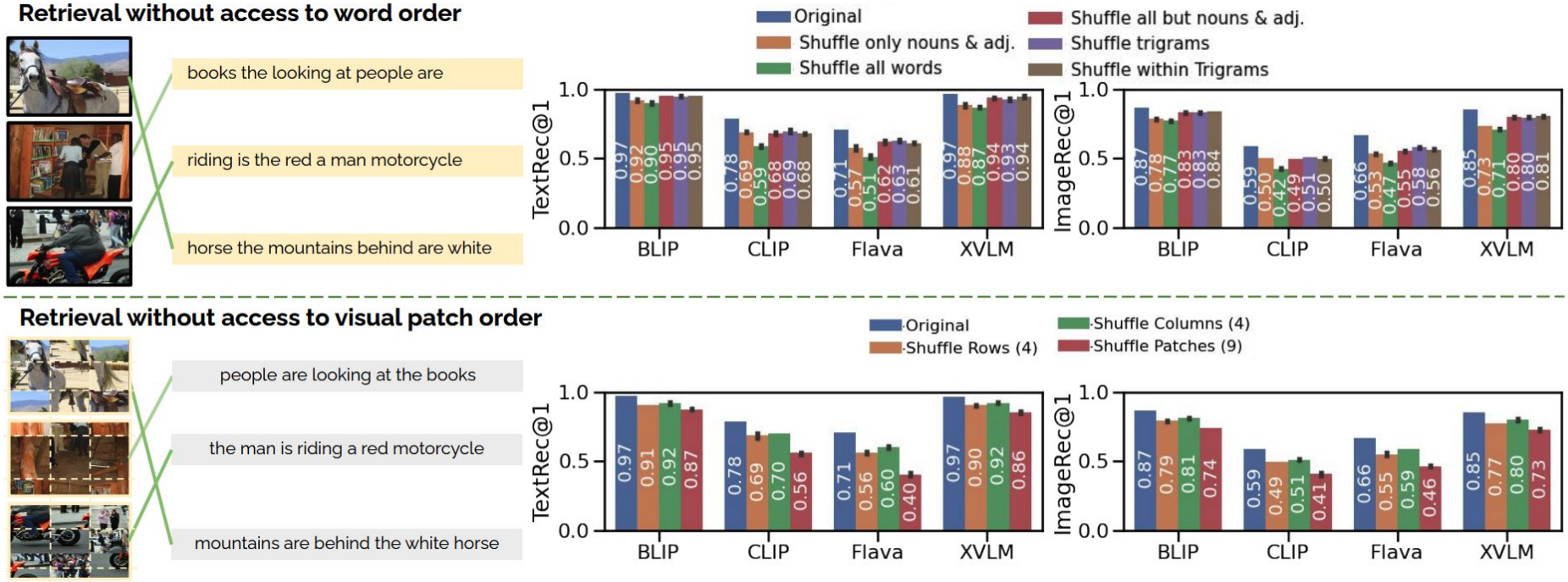
Strategy	Model	Text Recall@1	Text Recall@5	Image Recall@1	Image Recall@5
COCO Text Shuffle					
No Shuffling	BLIP	0.814	0.953	0.635	0.855
Shuffle only nouns and adj.	BLIP	0.710 ± 0.006	0.907 ± 0.001	0.513 ± 0.001	0.768 ± 0.001
Shuffle all words	BLIP	0.690 ± 0.004	0.899 ± 0.002	0.505 ± 0.002	0.763 ± 0.001
Shuffle all but nouns and adj.	BLIP	0.767 ± 0.002	0.935 ± 0.001	0.579 ± 0.000	0.817 ± 0.000
Shuffle trigrams	BLIP	0.762 ± 0.001	0.933 ± 0.000	0.581 ± 0.000	0.816 ± 0.000
Shuffle within Trigrams	BLIP	0.767 ± 0.003	0.934 ± 0.000	0.585 ± 0.000	0.820 ± 0.000
No Shuffling	CLIP	0.503	0.748	0.301	0.557
Shuffle only nouns and adj.	CLIP	0.420 ± 0.003	0.684 ± 0.007	0.244 ± 0.002	0.480 ± 0.001
Shuffle all words	CLIP	0.341 ± 0.001	0.608 ± 0.005	0.205 ± 0.001	0.422 ± 0.003
Shuffle all but nouns and adj.	CLIP	0.415 ± 0.002	0.671 ± 0.001	0.248 ± 0.002	0.483 ± 0.001
Shuffle trigrams	CLIP	0.411 ± 0.006	0.673 ± 0.004	0.251 ± 0.001	0.490 ± 0.002
Shuffle within Trigrams	CLIP	0.404 ± 0.002	0.661 ± 0.004	0.243 ± 0.001	0.478 ± 0.001
No Shuffling	Flava	0.454	0.788	0.388	0.682
Shuffle only nouns and adj.	Flava	0.335 ± 0.003	0.645 ± 0.002	0.287 ± 0.001	0.566 ± 0.000
Shuffle all words	Flava	0.338 ± 0.006	0.631 ± 0.001	0.260 ± 0.001	0.526 ± 0.002
Shuffle all but nouns and adj.	Flava	0.392 ± 0.005	0.692 ± 0.003	0.317 ± 0.002	0.601 ± 0.001
Shuffle trigrams	Flava	0.385 ± 0.006	0.698 ± 0.007	0.333 ± 0.002	0.621 ± 0.001
Shuffle within Trigrams	Flava	0.394 ± 0.005	0.698 ± 0.000	0.324 ± 0.001	0.606 ± 0.001
No Shuffling	XVLM	0.791	0.947	0.610	0.848
Shuffle only nouns and adj.	XVLM	0.655 ± 0.005	0.884 ± 0.000	0.462 ± 0.001	0.731 ± 0.000
Shuffle all words	XVLM	0.633 ± 0.006	0.879 ± 0.001	0.450 ± 0.002	0.723 ± 0.002
Shuffle all but nouns and adj.	XVLM	0.734 ± 0.004	0.928 ± 0.002	0.547 ± 0.003	0.803 ± 0.000
Shuffle trigrams	XVLM	0.727 ± 0.004	0.920 ± 0.002	0.544 ± 0.001	0.799 ± 0.004
Shuffle within Trigrams	XVLM	0.739 ± 0.007	0.929 ± 0.004	0.554 ± 0.005	0.808 ± 0.001
Flickr30k Text Shuffle					
No Shuffling	BLIP	0.972	0.999	0.869	0.974
Shuffle only nouns and adj.	BLIP	0.919 ± 0.008	0.993 ± 0.004	0.786 ± 0.003	0.949 ± 0.001
Shuffle all words	BLIP	0.902 ± 0.008	0.988 ± 0.003	0.770 ± 0.002	0.939 ± 0.001
Shuffle all but nouns and adj.	BLIP	0.950 ± 0.003	0.996 ± 0.002	0.829 ± 0.006	0.961 ± 0.001
Shuffle trigrams	BLIP	0.948 ± 0.005	0.997 ± 0.001	0.828 ± 0.001	0.965 ± 0.002
Shuffle within Trigrams	BLIP	0.953 ± 0.004	0.997 ± 0.001	0.838 ± 0.002	0.964 ± 0.001
No Shuffling	CLIP	0.784	0.950	0.591	0.835
Shuffle only nouns and adj.	CLIP	0.690 ± 0.005	0.909 ± 0.001	0.501 ± 0.003	0.770 ± 0.003
Shuffle all words	CLIP	0.587 ± 0.007	0.854 ± 0.011	0.423 ± 0.006	0.694 ± 0.002
Shuffle all but nouns and adj.	CLIP	0.678 ± 0.010	0.904 ± 0.007	0.493 ± 0.003	0.764 ± 0.002
Shuffle trigrams	CLIP	0.698 ± 0.017	0.910 ± 0.005	0.509 ± 0.004	0.775 ± 0.002
Shuffle within Trigrams	CLIP	0.680 ± 0.006	0.903 ± 0.011	0.498 ± 0.006	0.766 ± 0.001
No Shuffling	Flava	0.707	0.941	0.664	0.900
Shuffle only nouns and adj.	Flava	0.573 ± 0.021	0.869 ± 0.009	0.532 ± 0.001	0.816 ± 0.002
Shuffle all words	Flava	0.504 ± 0.002	0.817 ± 0.009	0.467 ± 0.006	0.754 ± 0.006
Shuffle all but nouns and adj.	Flava	0.622 ± 0.016	0.888 ± 0.005	0.553 ± 0.004	0.823 ± 0.002
Shuffle trigrams	Flava	0.626 ± 0.016	0.895 ± 0.003	0.578 ± 0.002	0.849 ± 0.003
Shuffle within Trigrams	Flava	0.613 ± 0.007	0.889 ± 0.003	0.564 ± 0.006	0.834 ± 0.003
No Shuffling	XVLM	0.967	1.000	0.855	0.968
Shuffle only nouns and adj.	XVLM	0.881 ± 0.013	0.987 ± 0.001	0.733 ± 0.002	0.920 ± 0.002
Shuffle all words	XVLM	0.869 ± 0.003	0.982 ± 0.003	0.708 ± 0.001	0.908 ± 0.002
Shuffle all but nouns and adj.	XVLM	0.937 ± 0.003	0.994 ± 0.003	0.798 ± 0.003	0.949 ± 0.003
Shuffle trigrams	XVLM	0.924 ± 0.005	0.995 ± 0.002	0.795 ± 0.008	0.948 ± 0.001
Shuffle within Trigrams	XVLM	0.942 ± 0.013	0.996 ± 0.002	0.807 ± 0.003	0.953 ± 0.001

Проблемы и гипотезы

1) Модели практически не отдают предпочтения правильно построенным предложениям (тест порядка)

2) CLIP не может определить правильный порядок составляющих в предложении (имеется в виду тест Attribution, Relation). Это согласуется с предыдущими наблюдениями, где авторы, использующие CLIP в качестве кодировщика с трудом генерируют изображения => CLIP не умеет понимать порядок и работает как bag-of-words

Визуализация проблемы



Почему так происходит?

CLIP обучаются с помощью Contrastive learning на обширных данных из интернета.

Гипотеза авторов заключается в том, что недостаток композиционного понимания связан с таким методом обучения.

Основная задача - идентификация соответствующих пар текста и изображения для поиска.

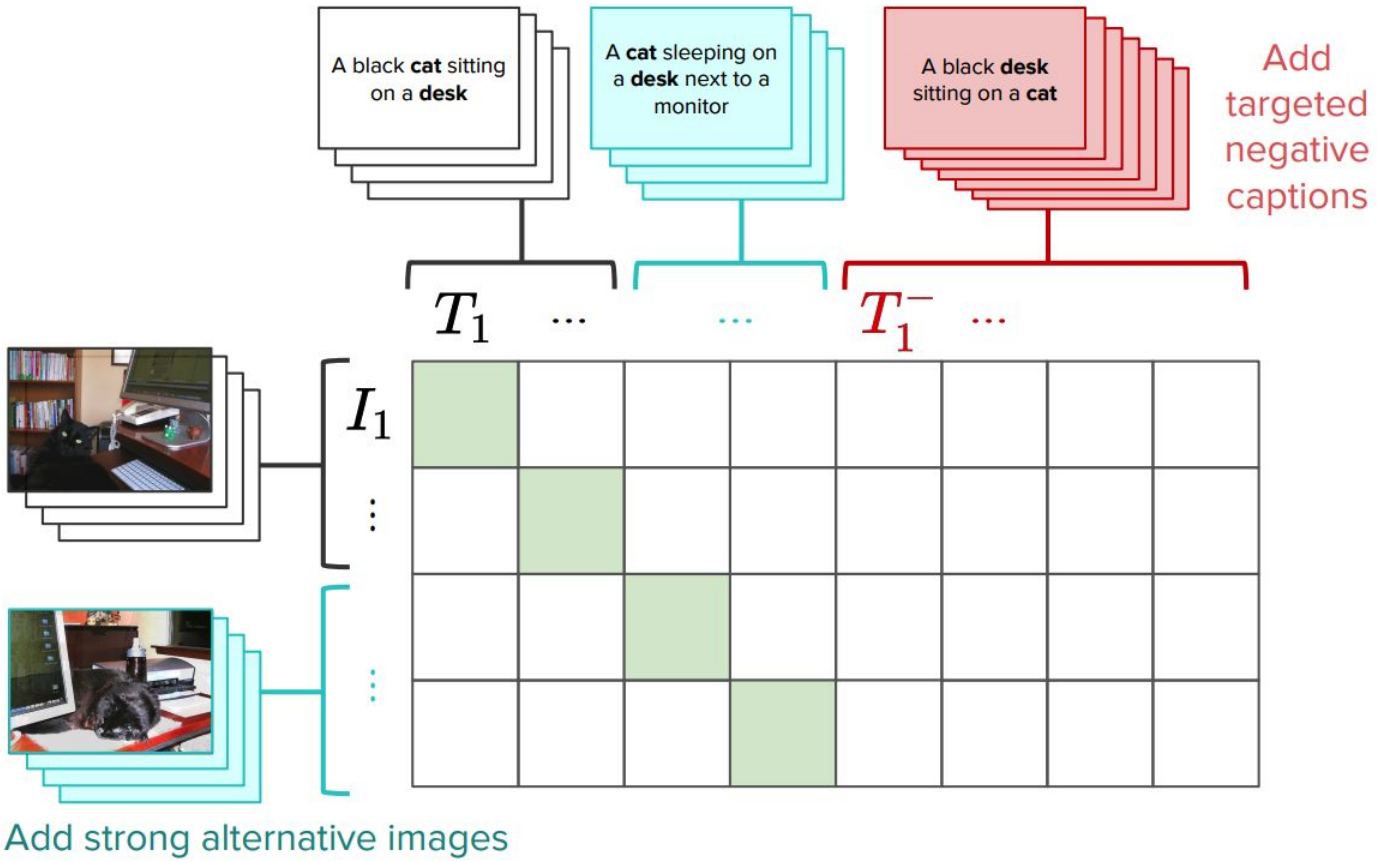
Экспериментально получено, что даже без учета порядка слов, модели достигают высокой производительности на длинных описаниях, говоря о возможности использования упрощенных стратегий для успешного решения задач.

Отсюда возникает вопрос, **как заставить модель обращать внимание на порядок слов**, если современные датасеты не заточены под это.

Решение проблемы

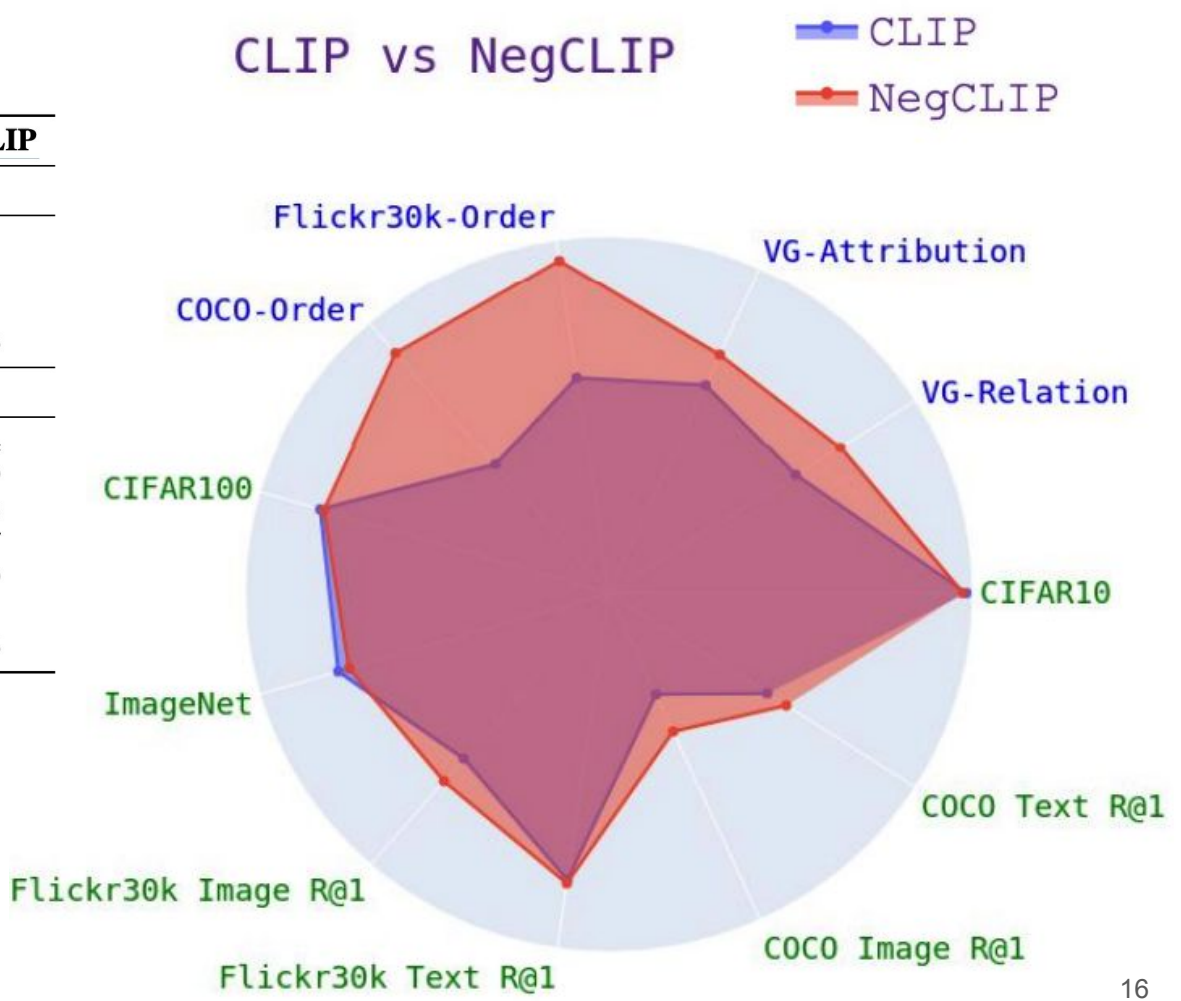
1) **генерация негативов, меняя местами различные слова:**
   For example, the caption “The horse is eating the grass and the zebra is drinking the water” either becomes “The zebra is eating the grass and the horse is drinking the water” (noun swapping) or “The horse is drinking the grass and the zebra is eating the water” (verb phrase swapping).
2) **выбор сильных альтернатив:**
   используем clip для пар описание-картинка, находим для каждого изображения ближайшего соседа и добавляем его подписи и негативные подписи из п.1

Схема обучения



Результат NegCLIP

	CLIP	CLIP-FT	NegCLIP
Compositional Tasks			
VG-Relation	0.59	0.63	0.81
VG-Attribution	0.62	0.65	0.71
Flickr30k-PRC	0.59	0.50	0.91
COCO-PRC	0.46	0.36	0.86
Downstream Tasks			
CIFAR10	0.95	0.95	0.94
CIFAR100	0.80	0.80	0.79
ImageNet	0.75	0.74	0.72
Flickr30k Image R@1	0.59	0.67	0.67
Flickr30k Text R@1	0.78	0.83	0.79
COCO Image R@1	0.30	0.42	0.41
COCO Text R@1	0.50	0.59	0.56



Выводы

1) основная идея статьи заключается в том, что модель, которую обучали под одну общую задачу, можно использовать очень аккуратно в задачах, которые ей не предлагались (более тонких и сложных). Зачастую это применение модели ни к чему не приведет.
2) было показано, что текущие VLM (not transformer based) плохо улавливают композицию и порядок
3) были разработаны датасеты для оценивания этой тонкой языковой задачи
4) был предложен метод, расширяющий CLIP для более корректной работы с этими датасетами, который также превзошел baseline, который назвали NegClip