Concept Association Bias of Vision-Language Models

Yutaro Yamada

NLP Colloquium, 2024/01/24

CLIP

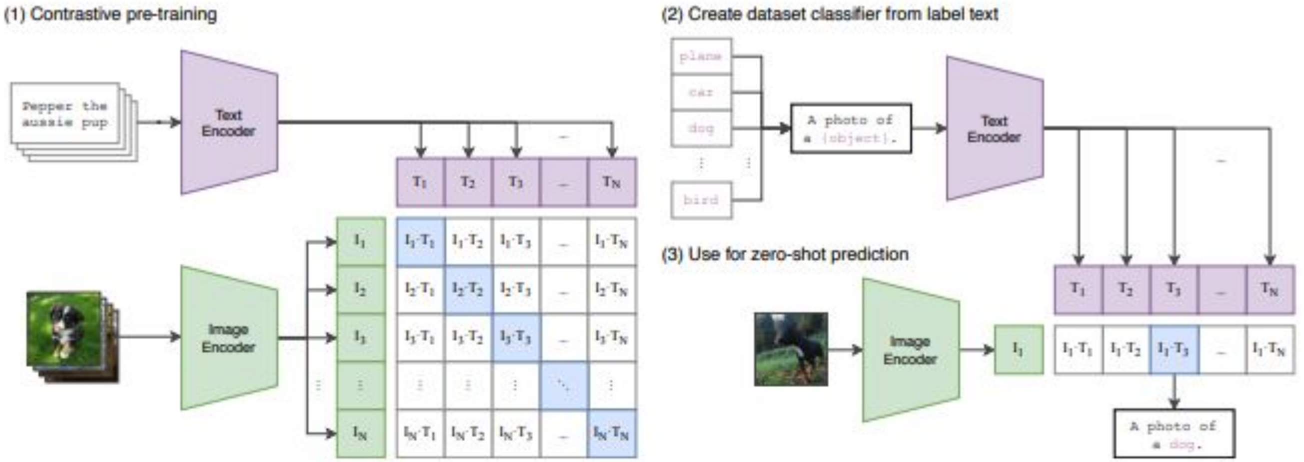
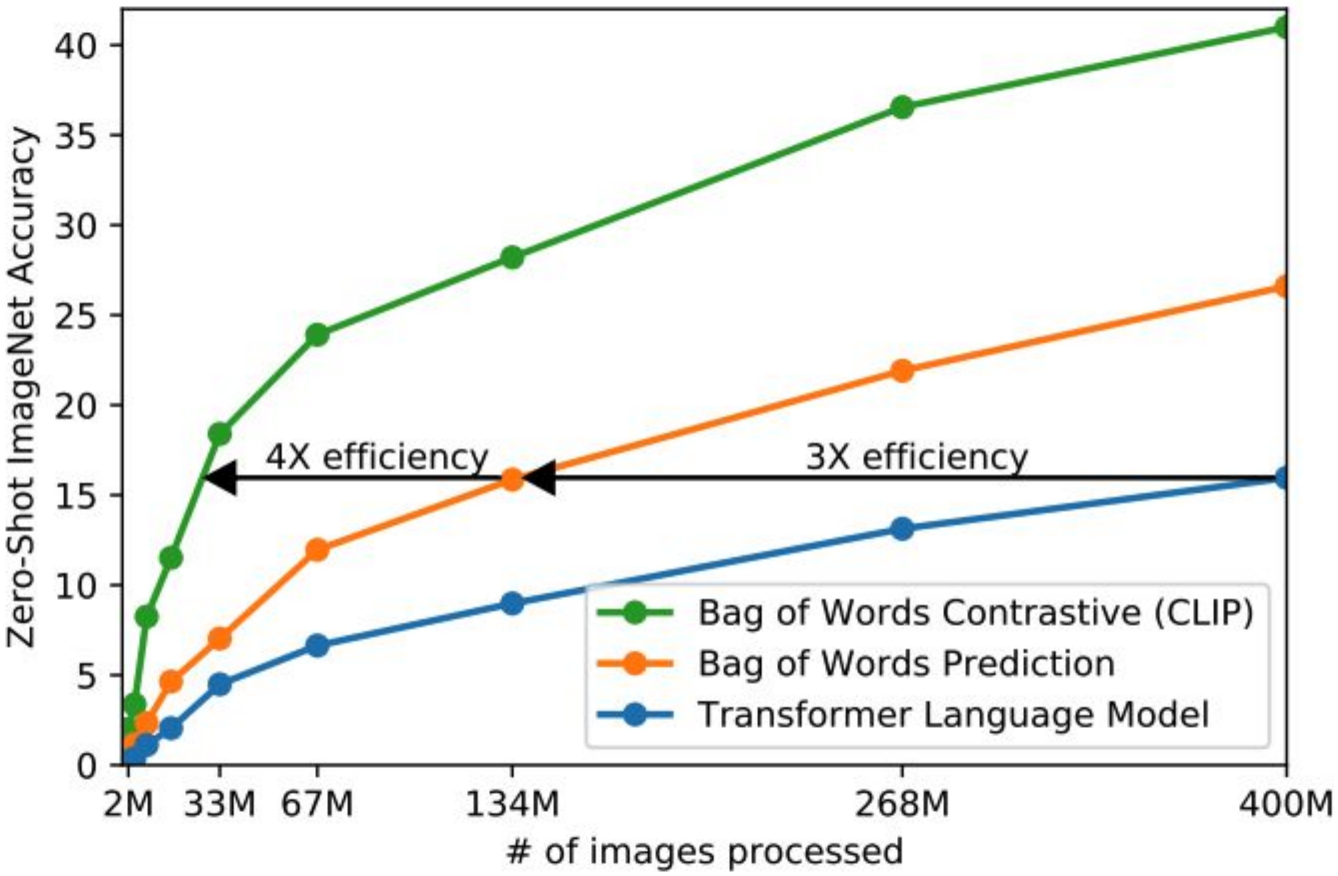


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.


Zero-Shot ImageNet Accuracy
40
35
30
25
20
15
10
5
0
4X efficiency
3X efficiency
Bag of Words Contrastive (CLIP)
Bag of Words Prediction
Transformer Language Model
2M
33M
67M
134M
268M
400M
# of images processed

Applications of CLIP

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh*
OpenAI
aramesh@openai.com

CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields

Can Wang
Menglei Chai
Mingming He
USC Institute for Creative Technologies
com
hmm.lillian@gmail.com
Jing Liao*

PointCLIP: Point Cloud Understanding by CLIP

Renrui Zhang*[1], Ziyu Guo*[2], Wei Zhang[1]
Bin Cui[2], Yu Qiao[1], Peng Ga
[1]Shanghai AI Laboratory
[3]The Chinese University
{zhangrenrui, gaopeng, qi
2101210573@pku.edu.cn, h

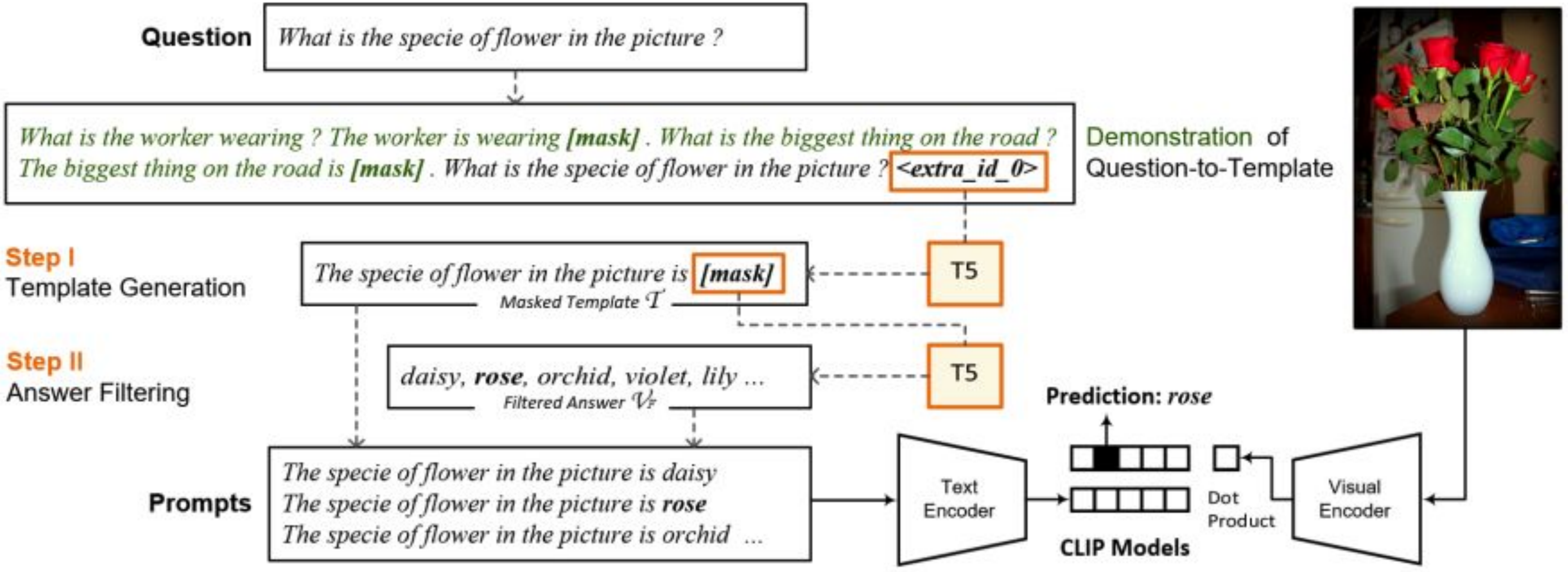
CLIPScore:
A Reference-free Evaluation Metric for Image Captioning

Jack Hessel[†] **Ari Holtzman**[‡] **Maxwell Forbes**[‡] **Ronan Le Bras**[†] **Yejin Choi**[†‡]
[†]Allen Institute for AI
[‡]Paul G. Allen School of Computer Science & Engineering, University of Washington
{jackh,ronanlb}@allenai.org {ahai,mbforbes,yejin}@cs.washington.edu

Zero-shot transfer of CLIP to ImageNet

DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	76.2%	76.2%
 ImageNet V2	64.3%	70.1%
 ImageNet Rendition	37.7%	88.9%
 ObjectNet	32.6%	72.3%
 ImageNet Sketch	25.2%	60.2%
 ImageNet Adversarial	2.7%	77.1%

You can also ask about object attributes

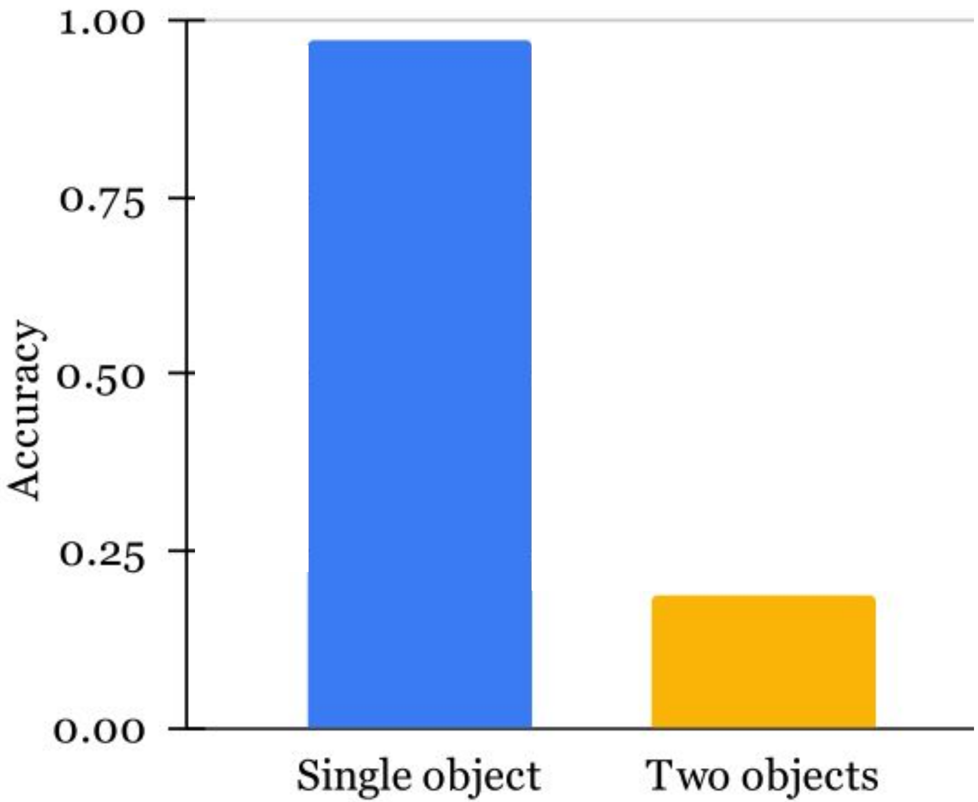


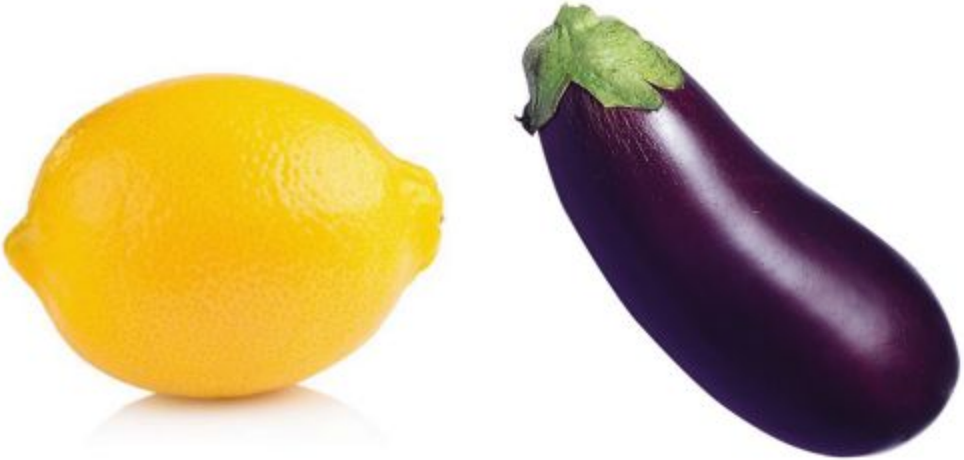
ACL 2022 Song et al. “CLIP Models are Few-shot Learners: Empirical Studies on VQA and Visual Entailment”

What if there are two objects?

Prompt: “The color of the eggplant is []”

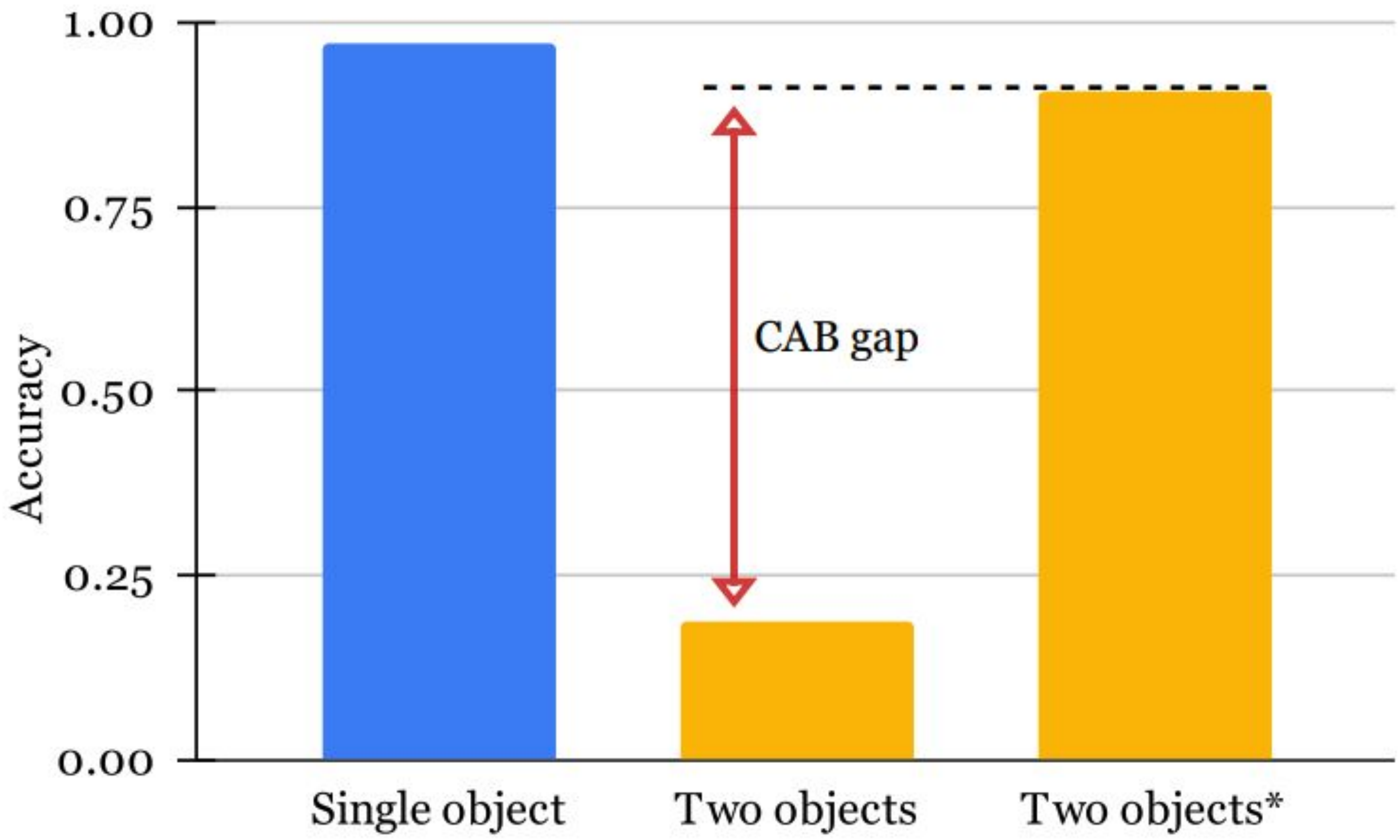
Zero-shot transfer of CLIP to color recognition

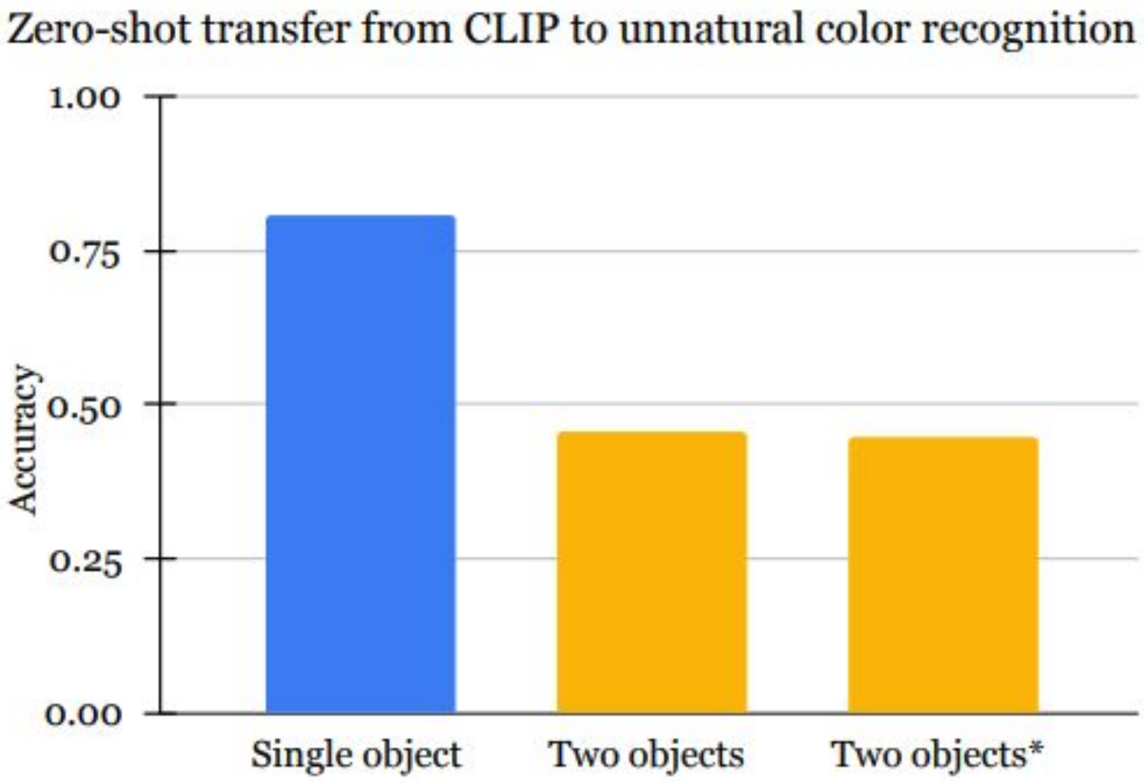


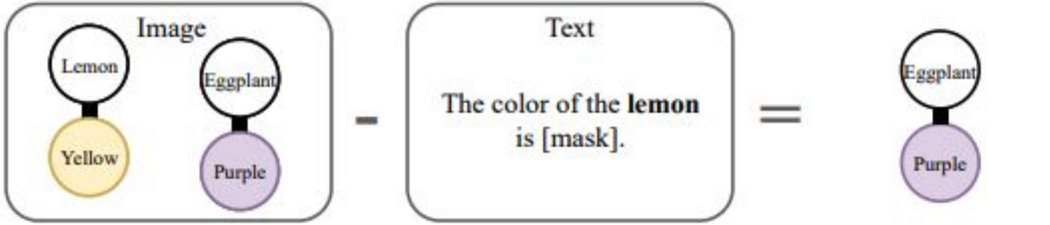


CLIP: "In this picture, the color of the lemon is purple."

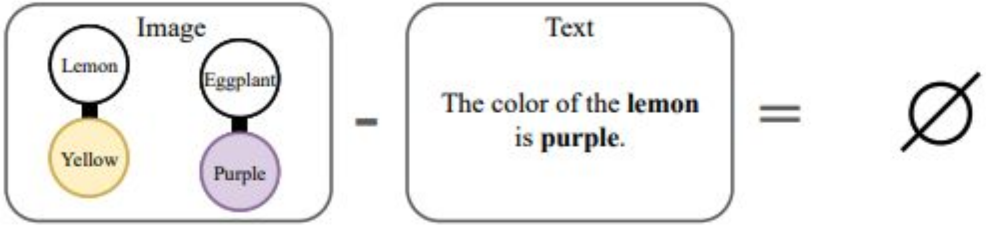
Zero-shot transfer of CLIP to color recognition



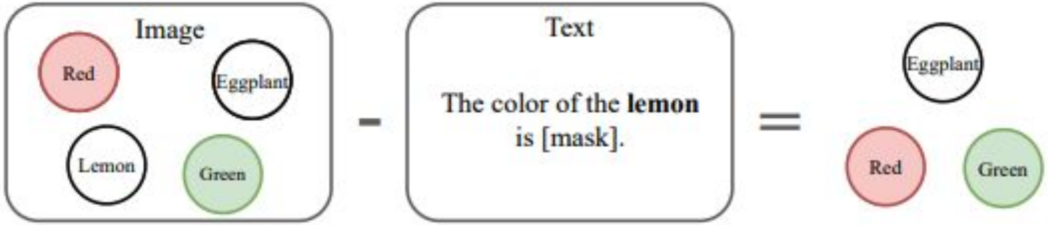

Zero-shot transfer from CLIP to unnatural color recognition
1.00
0.75
0.50
0.25
0.00
Accuracy
Single object
Two objects
Two objects*



(a) Natural color



(b) Natural color ([mask] = purple)



(c) Unnatural color

Zero-shot transfer to part-whole recognition

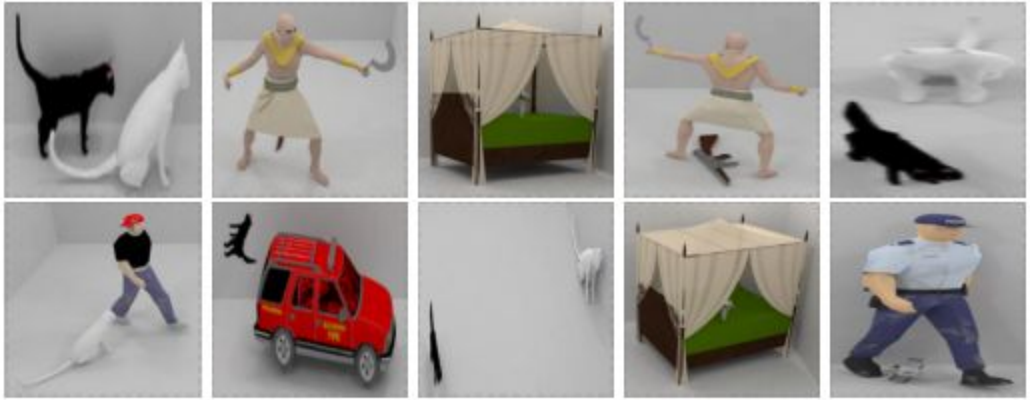
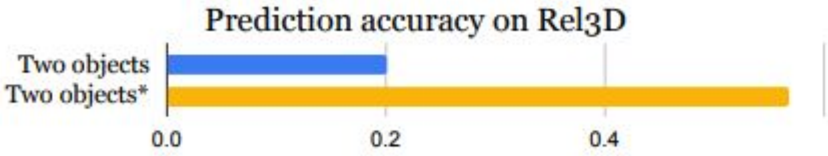
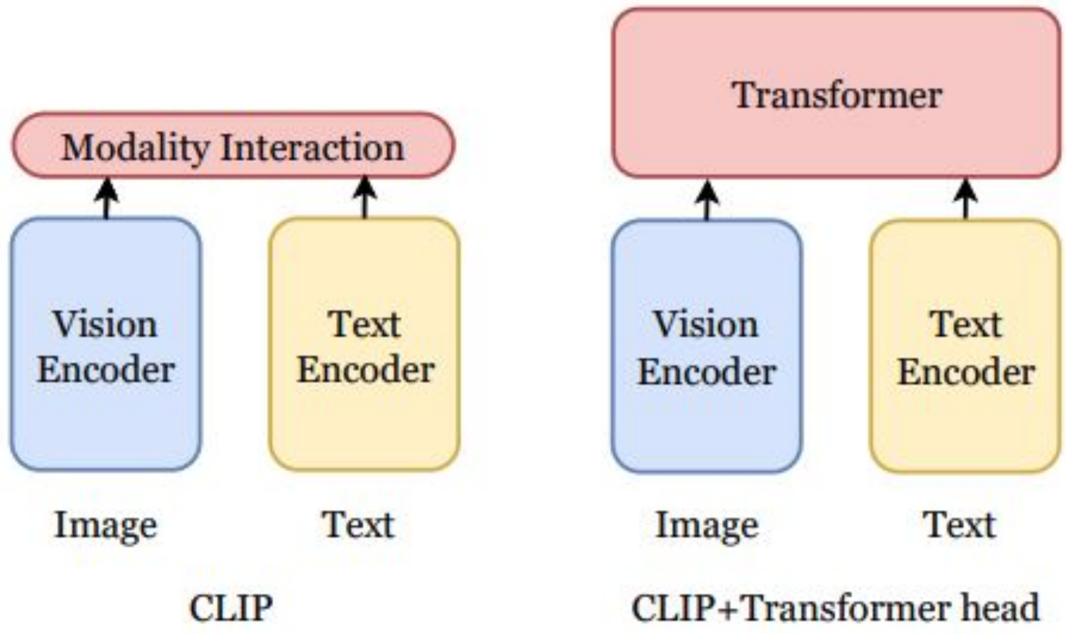


Figure 12. Example images from Rel3D [9].



Models	Two objects	Two objects*	Single	CAB
CLIP	0.011	0.932	0.929	0.961
BLIP-contrast	0.086	0.879	0.846	0.896
BLIP-match	0.123	0.841	0.925	0.859
BLIP-2-contrast	0.138	0.840	0.844	0.851
BLIP-2-match	0.330	0.627	0.925	0.648
BLIP-2-caption	0.359	0.558	0.775	0.599
BLIP-caption	0.438	0.471	0.862	0.516
BLIP-2-FlanT5	0.604	0.377	0.984	0.386
OFA	0.855	0.078	0.879	0.111

How can we mitigate the CAB?



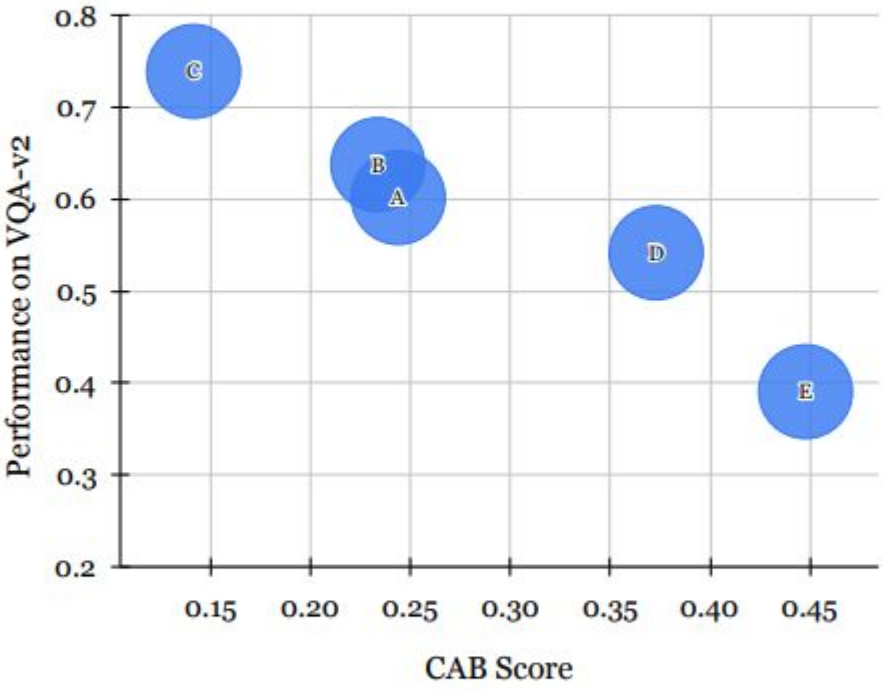

Performance on VQA-v2
CAB Score
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.15
0.20
0.25
0.30
0.35
0.40
0.45
C
B
A
D
E

Image Captioners Are Scalable Vision Learners Too

Michael Tschannen[∘,†] **Manoj Kumar**[∘] **Andreas Steiner**[∘]
Xiaohua Zhai **Neil Houlsby** **Lucas Beyer**[∘]
Google DeepMind

Abstract

Contrastive pretraining on image-text pairs from the web is one of the most popular large-scale pretraining strategies for vision backbones, especially in the context of large multimodal models. At the same time, image captioning on this type of data is commonly considered an inferior pretraining strategy. In this paper, we perform a fair comparison of these two pretraining strategies, carefully matching training data, compute, and model capacity. Using a standard encoder-decoder transformer, we find that captioning alone is surprisingly effective: on classification tasks, captioning produces vision encoders competitive with contrastively pretrained encoders, while surpassing them on vision & language tasks. We further analyze the effect of the model architecture and scale, as well as the pretraining data on the representation quality, and find that captioning exhibits the same or better scaling behavior along these axes. Overall our results show that plain image captioning is a more powerful pretraining strategy than was previously believed.

Prompt: "a chair with five legs"

DALL-E 3 ReStr