CLIP: Contrastive Language-Image Pre-Training

by Klyuchnikova Ulyana

Visual ↔ language

Bananas lying on newspaper with some peas.

banana 85%

newspaper 10%

pea 5%

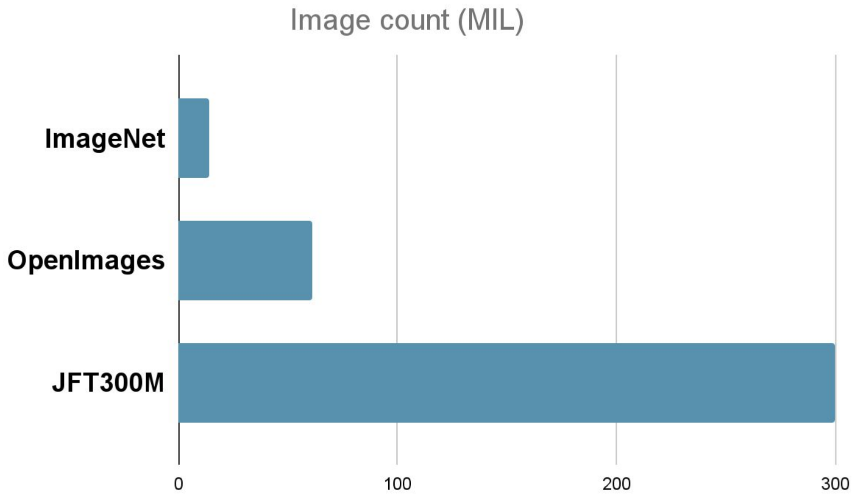
Fine-Grained Image Classification

Labrador Retriever 5%
French Bulldog 10%
Golden Retriever 8%
German Shepherd 7%
Poodle 4%
...

Supervised costs

~ **25K** slaves per **14MIL** images for 22K object categories

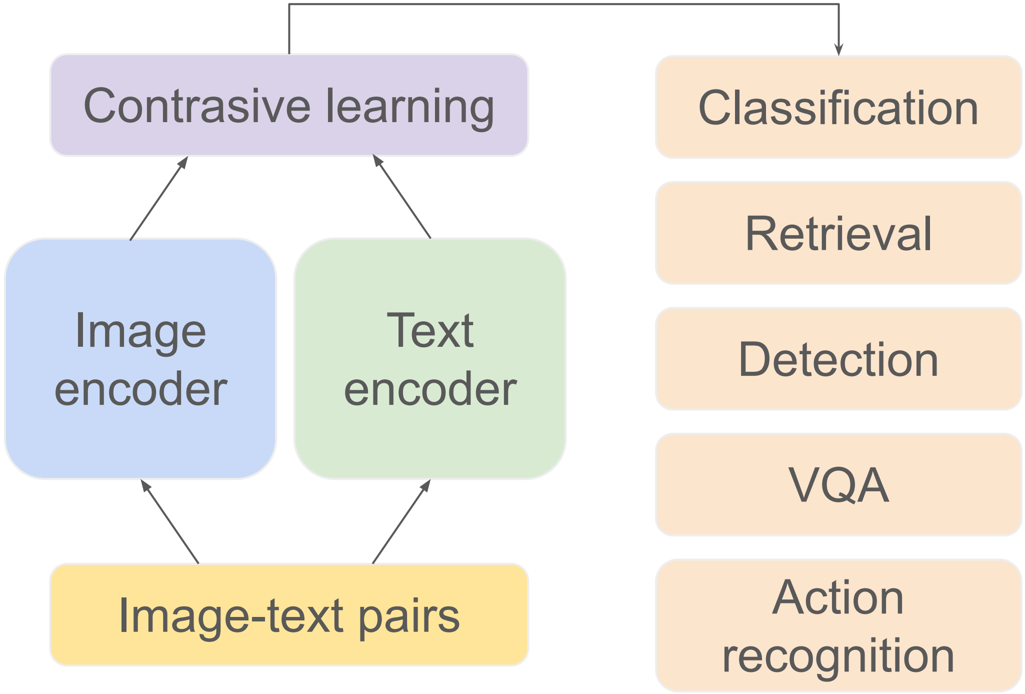
~ $0.025 to $10.00 per image



=> ImageNet costs around $70MIL ≈ 1 Island


ResNet, VGG,
Inception,
B-CNNs
VS
VLM

VLM: Vision-Language Models



VL-models

- learn joint representations of vision and language
- can use pre-trained models
- can be adapted for many tasks
- one/zero-shot tasks

CLIP Ideology

- who, why and when?

  OpenAI, Zero-shot, 2021

- why zero/one/few-shot is better?

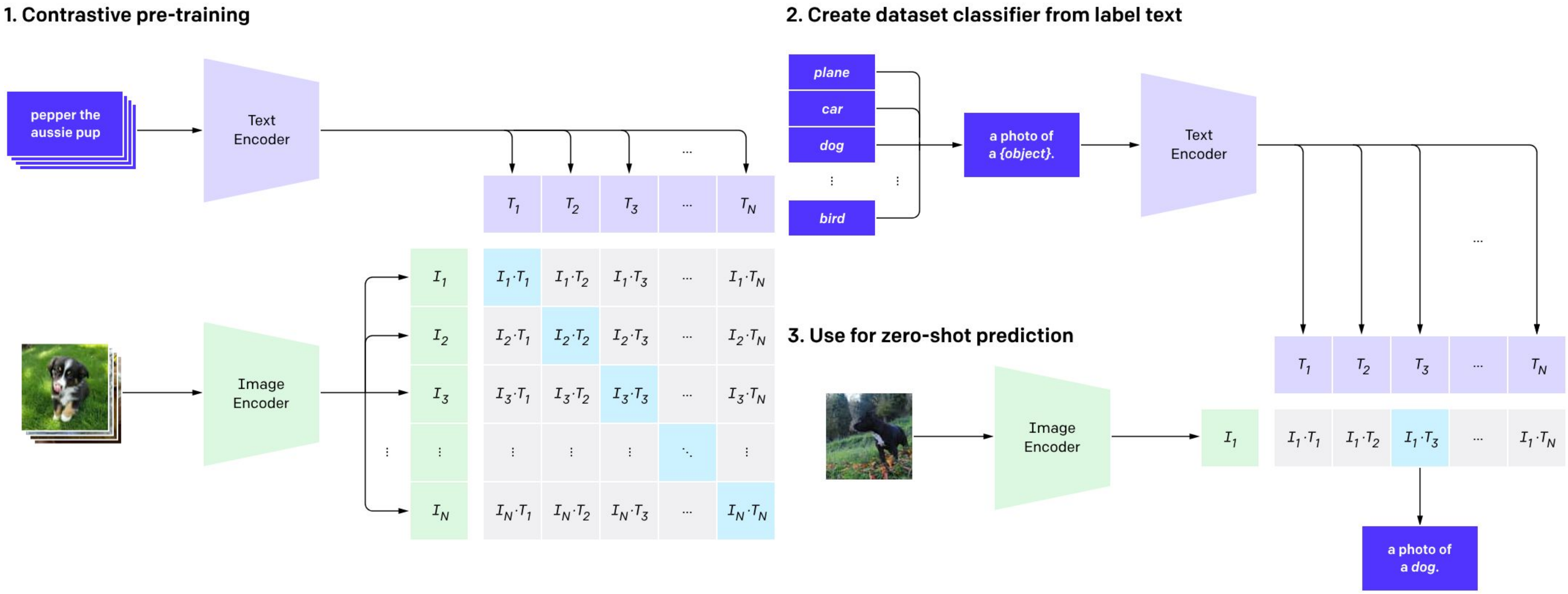
  Minimize costs, strong generalization ability

- how did it perform in contrast to the previous

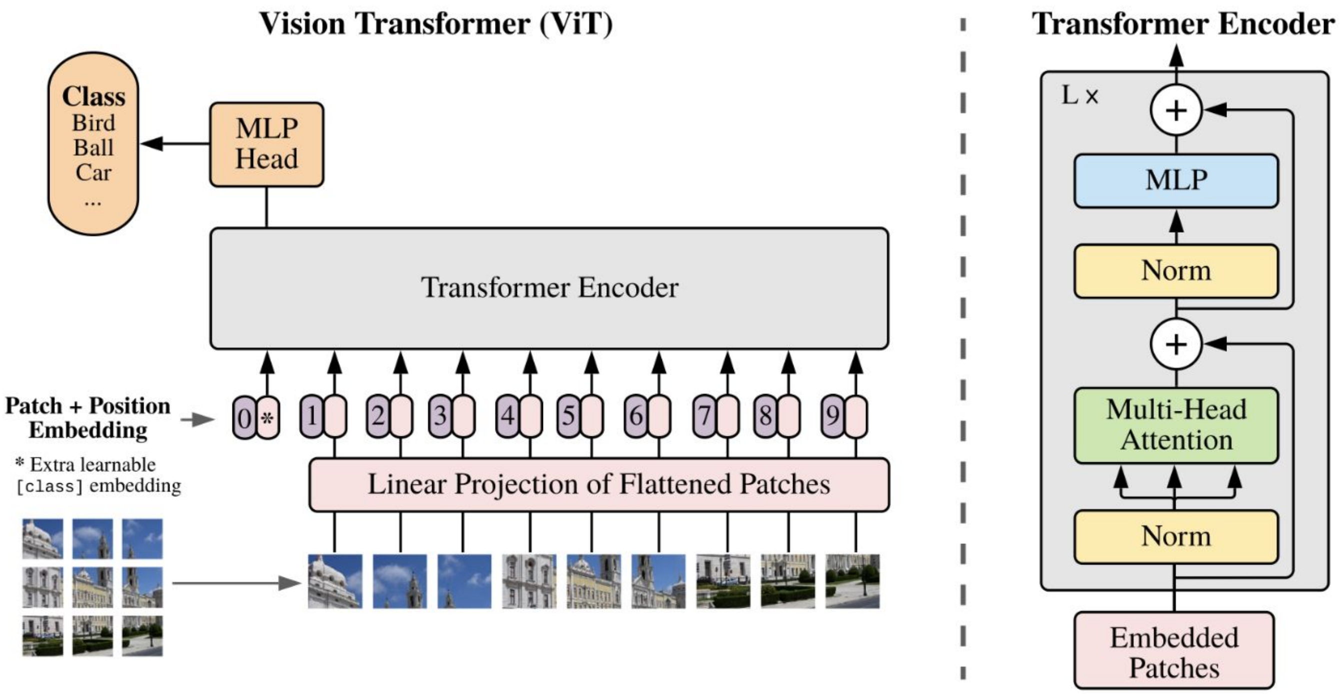
  Better on most datasets

- how did the model look like?

Basic architecture



VIT

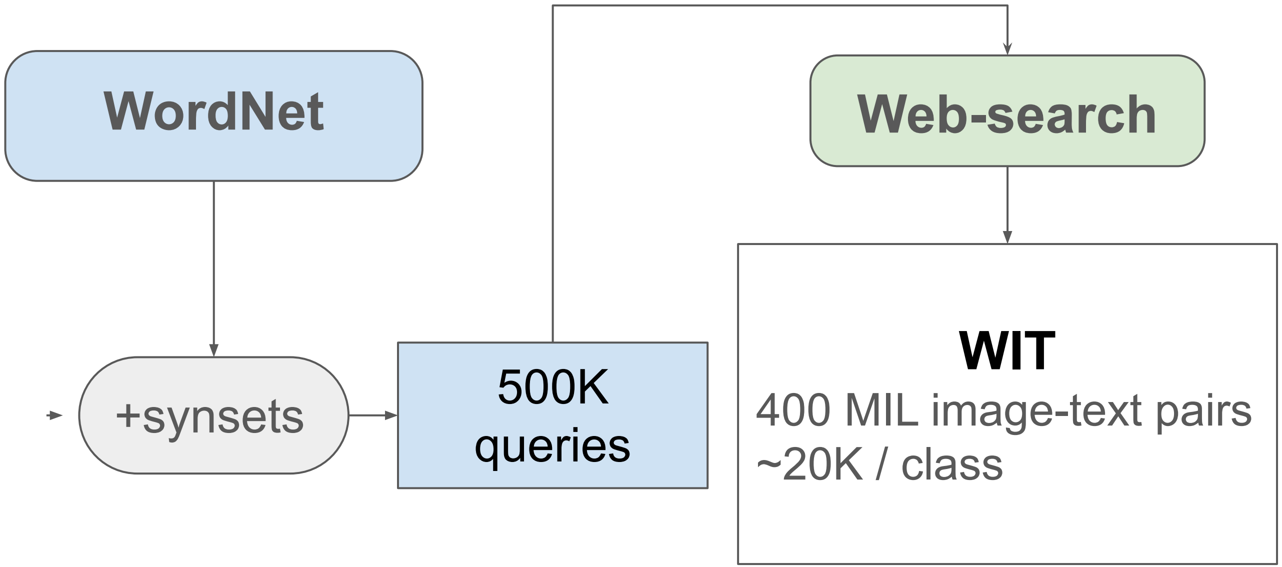


«An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale» A.Dosovitskiy et. al, 2020

Collecting data

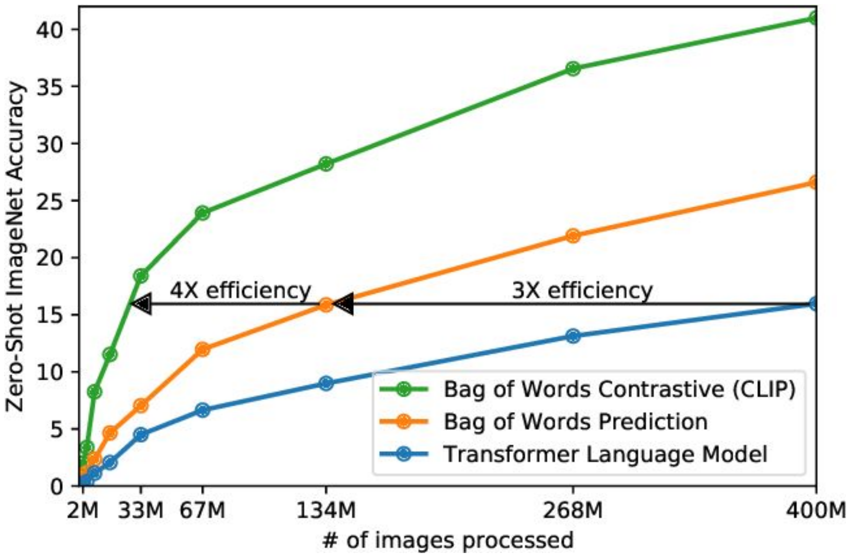
Wikipedia
The Free Encyclopedia

- names of popular articles
- words occuring > 100 times
- **bi-grams**



Contrasive pre-training

- image_encoder - ResNet or ViT
- text_encoder - CBOW or Transformer
- no cutting text



N x H x W x C → image_encoder → I = N x D_i

N x L → text_encoder → T = N x D_t

$$W_i \in R^{D_i, D_e}, W_t \in R^{D_t, D_e}$$

$$I_e = \frac{\mathbf{D_i W_i}}{\|\mathbf{D_i W_i}\|_2} \in R^{N, D_e}$$

$$T_e = \frac{\mathbf{D_t W_t}}{\|\mathbf{D_t W_t}\|_2} \in R^{N, D_e}$$

- linear projection from encoder to multi-modal embedding space
- τ - temperature is a trainable parameter

L2 - normalization

$$Logits = exp(t) I_e T_e^T \in R^{N,N}$$

$$L_{i2t} = -\frac{1}{N} \sum_{i}^{N} \log \frac{\exp(x_i^\top y_i / \sigma)}{\sum_{j=1}^{N} \exp(x_i^\top y_j / \sigma)}$$

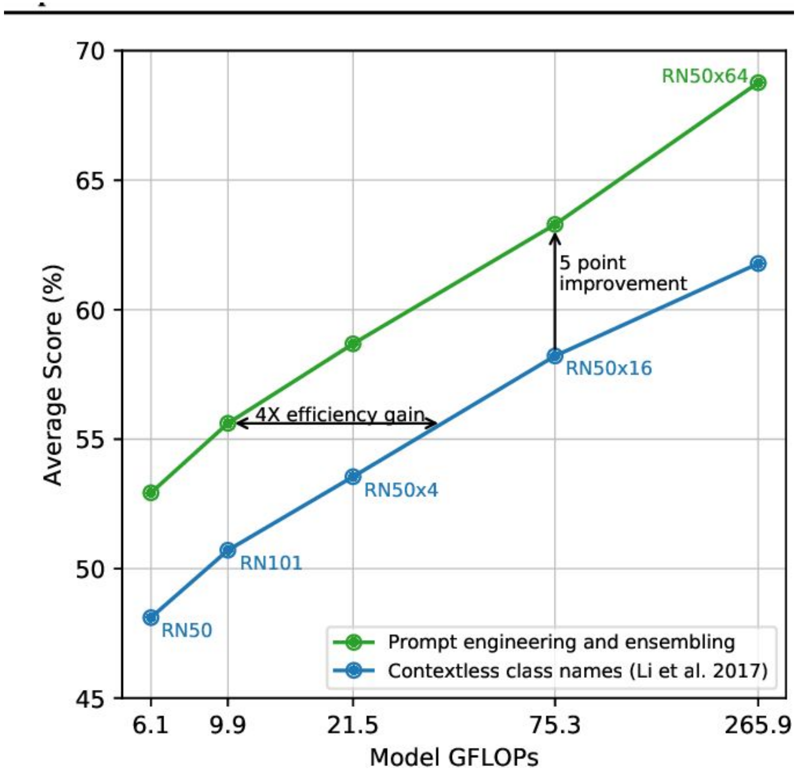
$$\boxed{Loss = \frac{L_{i2t} + L_{t2i}}{2}}$$

Promts

default: “A photo of a {label}.”

“A photo of a {label}, a type of pet.”

“a satellite photo of a {label}.”

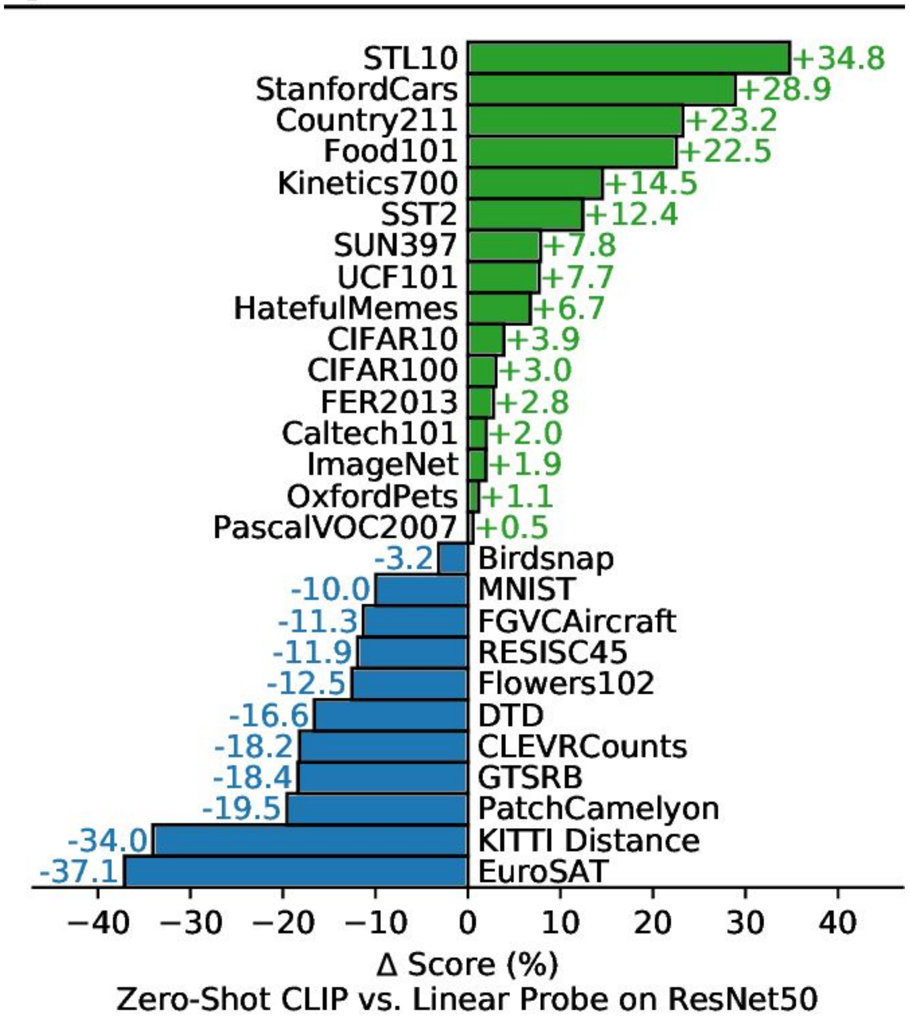


Zero-shot performance

general datasets: ImageNet, CIFAR10/100, STL10 🏆, PascalVOC2007

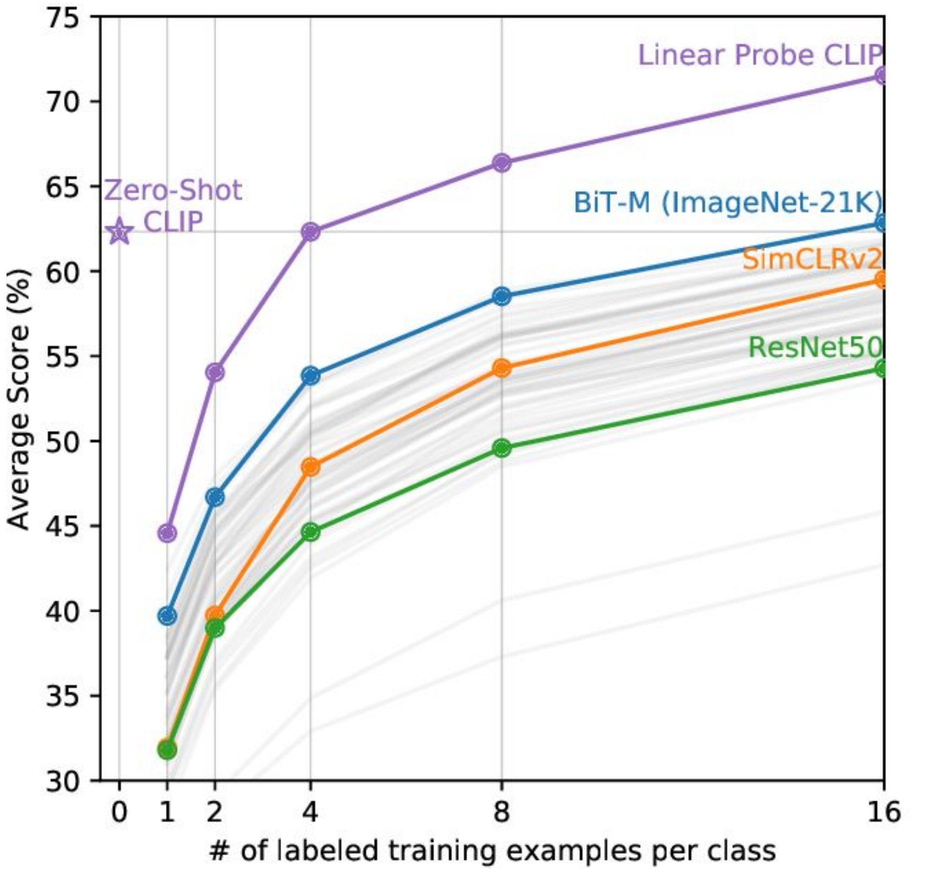
fine-grained: Stanford Cars, Food101, Flowers102, FGVCAircraft, OxfordPets, Birdsnap

videos: Kinetics700, UCF101

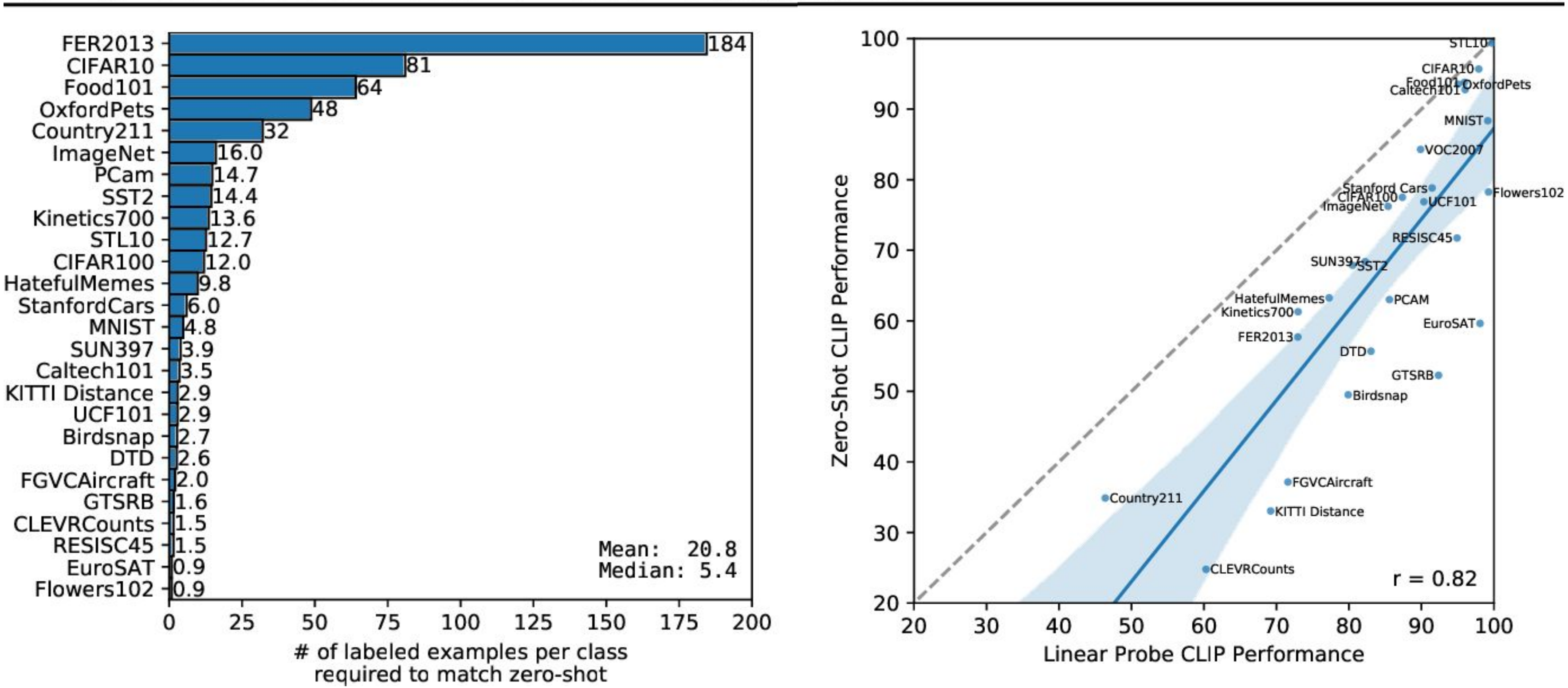


Few-shot performance

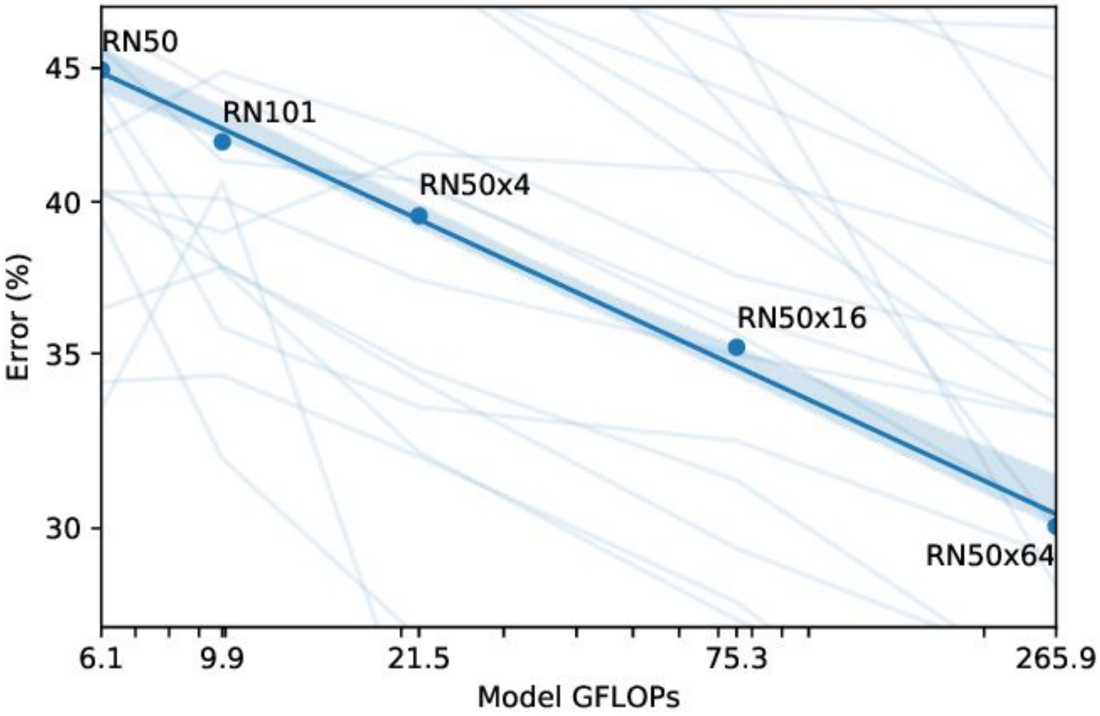
- Linear Probe CLIP outperformed best available ImageNet models
- 4-gram match zero-shot
- 0-1 skip



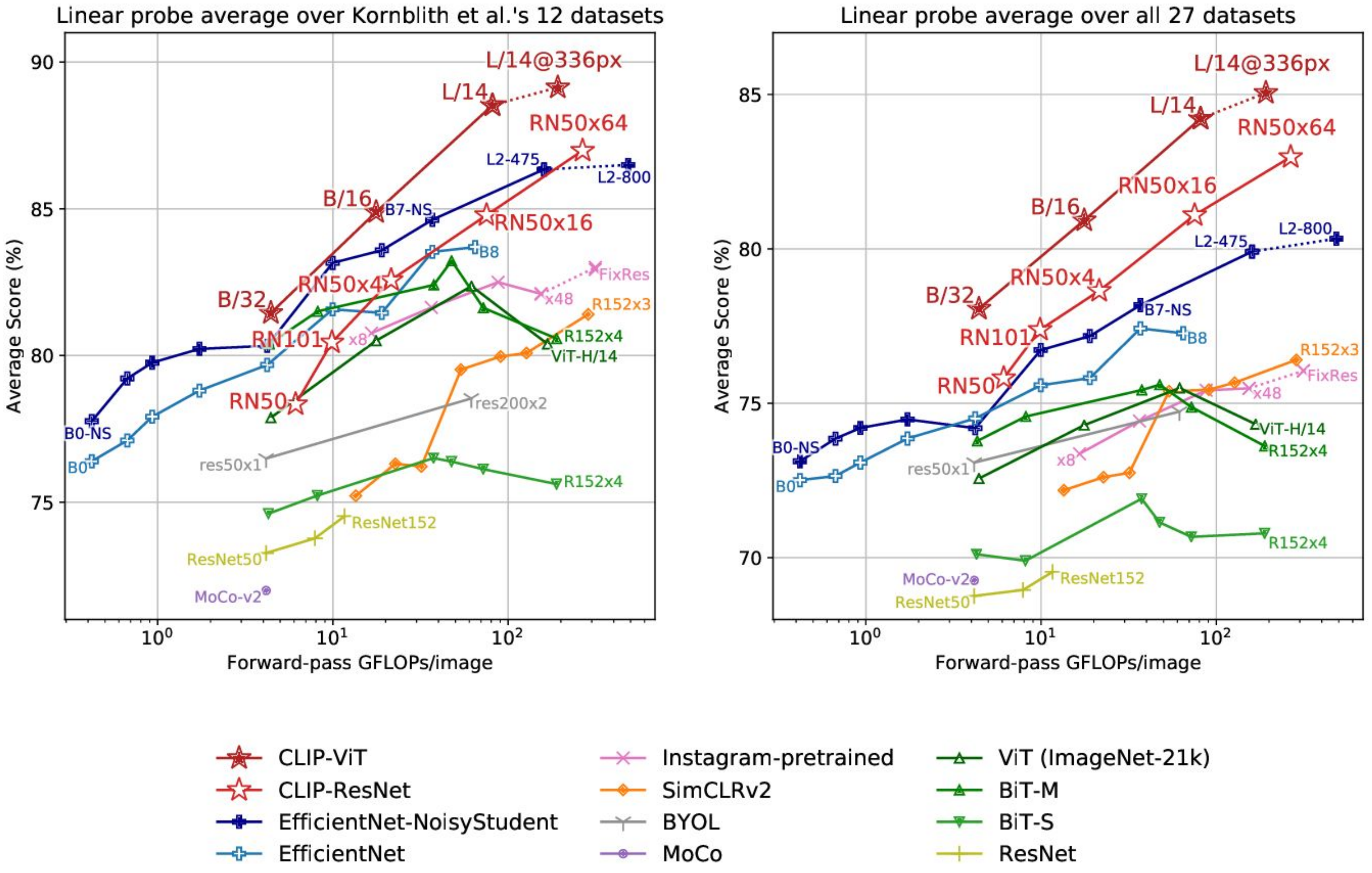
Zero-shot vs Few-shot / Line Probe



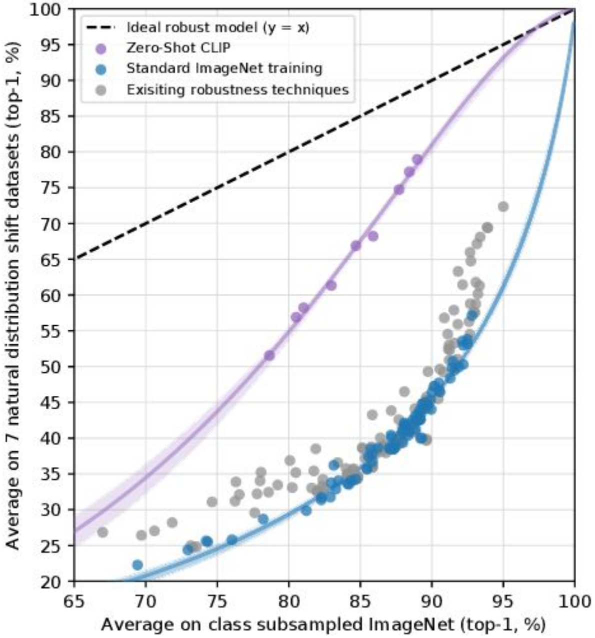
CLIP Scaling



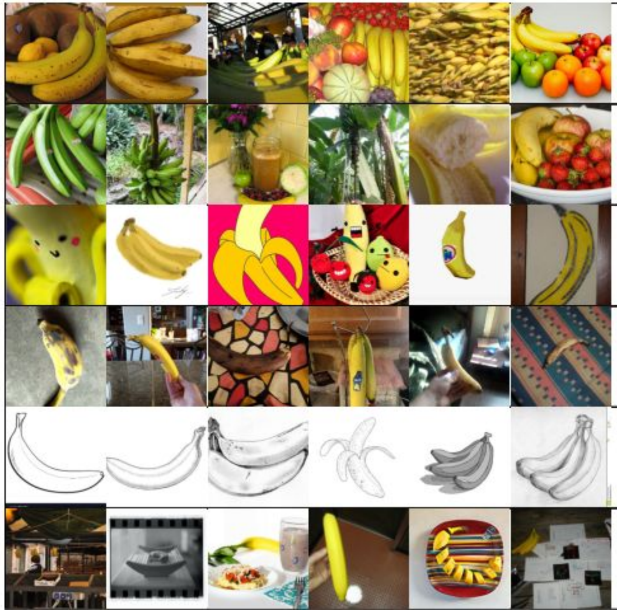
CLIP Embedding space



Zero-shot robustness

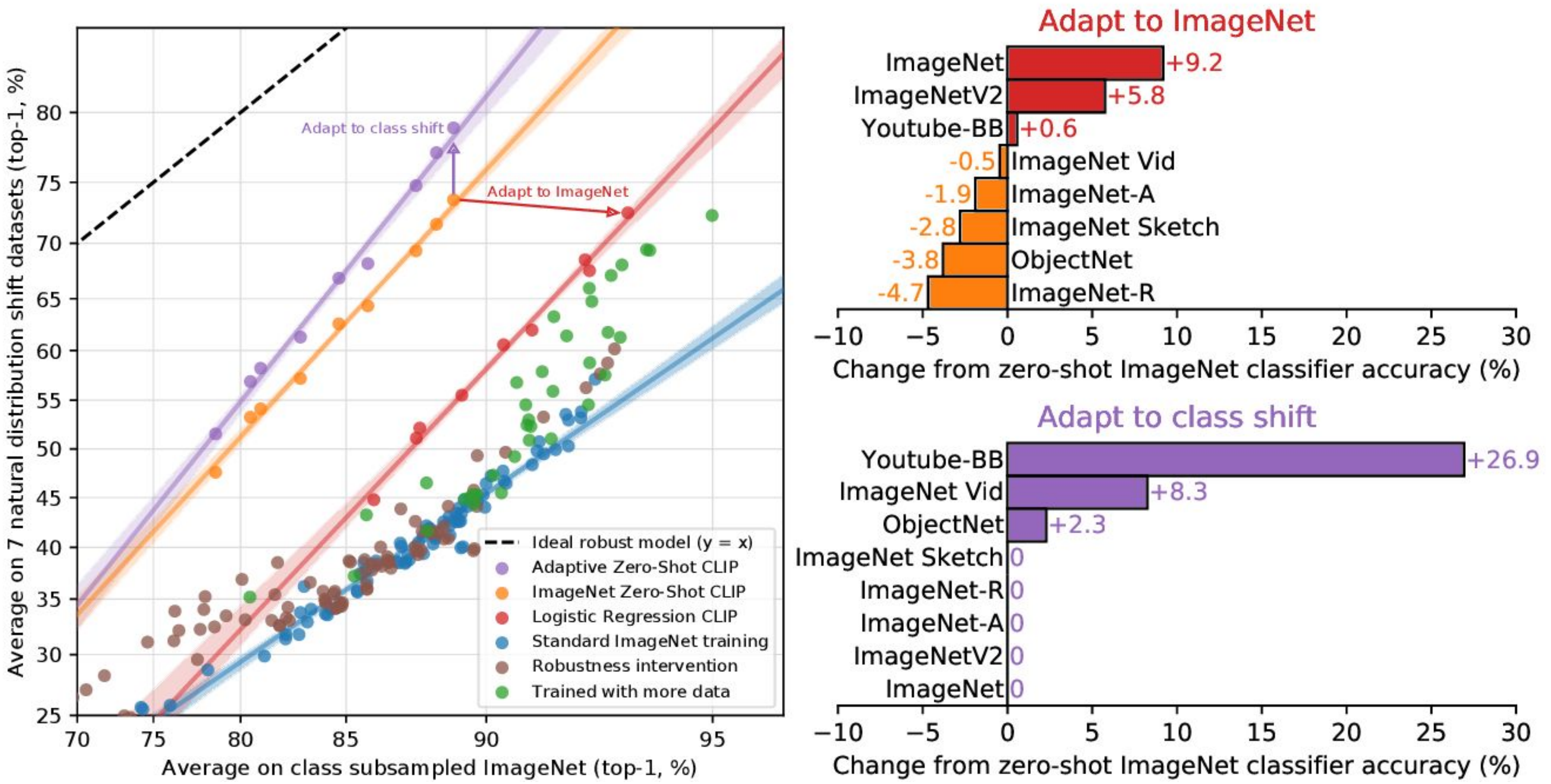


Dataset Examples



	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

Cheating hypothesis



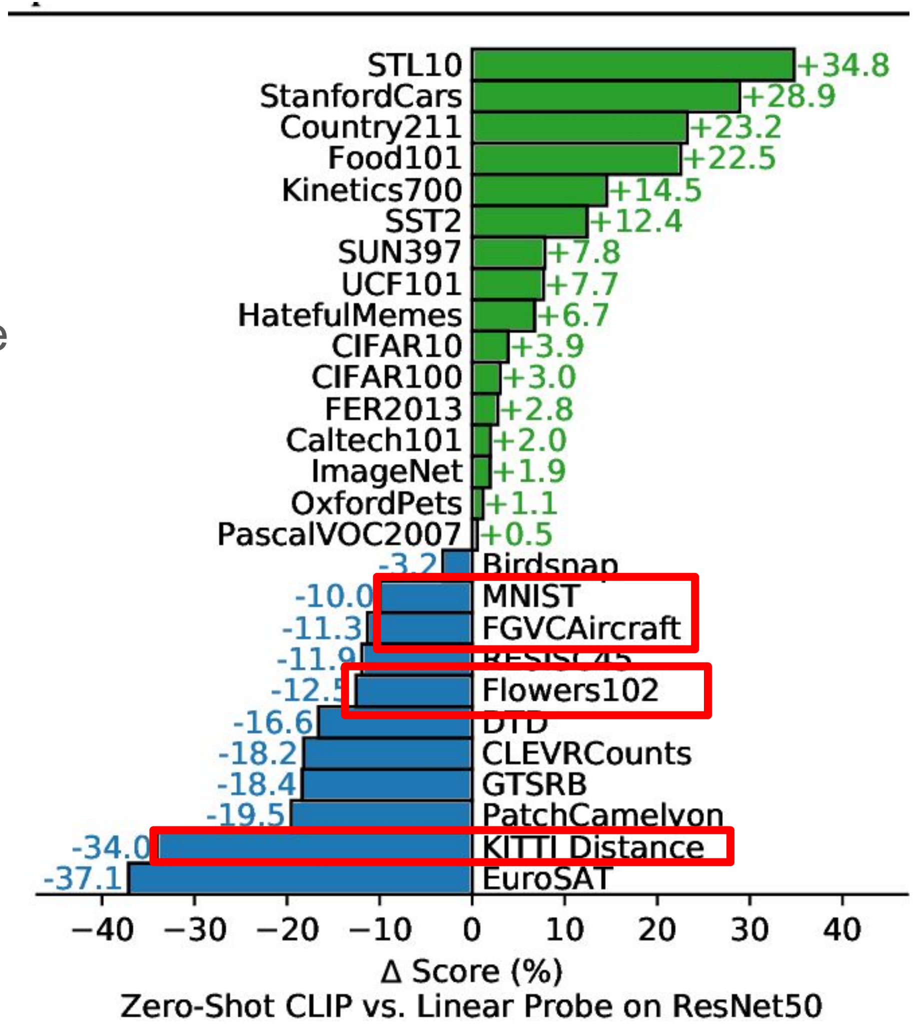
Limitations

abstract or systematic tasks: counting,

complex tasks: predicting how close the nearest car is

fine-grained: aircraft

MNIST: 88%



ALIGN: A Large-scale ImaGe and Noisy-text embedding

- 1.8 billion image-text pairs
- frequency-based filtering
- EfficientNet
- BERT + token embedding
- random cropping + horizontal flip
- batch size of 1024

"motorcycle front wheel"

"thumbnail for version as of 21 57 29 june 2010"

"file frankfurt airport skyline 2017 05 jpg"

"file london barge race 2 jpg"

"moustache seamless wallpaper design"

"st oswalds way and shops"

Performance

		Flickr30K (1K test set)						MSCOCO (5K test set)					
		image → text			text → image			image → text			text → image		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot	ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
	UNITER	83.6	95.7	97.7	68.7	89.2	93.9	-	-	-	-	-	-
	CLIP	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
	ALIGN	**88.6**	**98.7**	**99.7**	**75.7**	**93.8**	**96.8**	**58.6**	**83.0**	**89.7**	**45.6**	**69.8**	**78.6**
Fine-tuned	GPO	88.7	98.9	99.8	76.1	94.5	97.1	68.1	90.2	-	52.7	80.2	-
	UNITER	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
	ERNIE-ViL	88.1	98.0	99.2	76.7	93.6	96.4	-	-	-	-	-	-
	VILLA	87.9	97.5	98.8	76.3	94.2	96.8	-	-	-	-	-	-
	Oscar	-	-	-	-	-	-	73.5	92.2	96.0	57.5	82.8	**89.8**
	ALIGN	**95.3**	**99.8**	**100.0**	**84.9**	**97.4**	**98.6**	**77.0**	**93.5**	**96.9**	**59.9**	**83.3**	**89.8**

Model	ImageNet	ImageNet-R	ImageNet-A	ImageNet-V2
CLIP	76.2	88.9	**77.2**	**70.1**
ALIGN	**76.4**	**92.2**	75.8	**70.1**

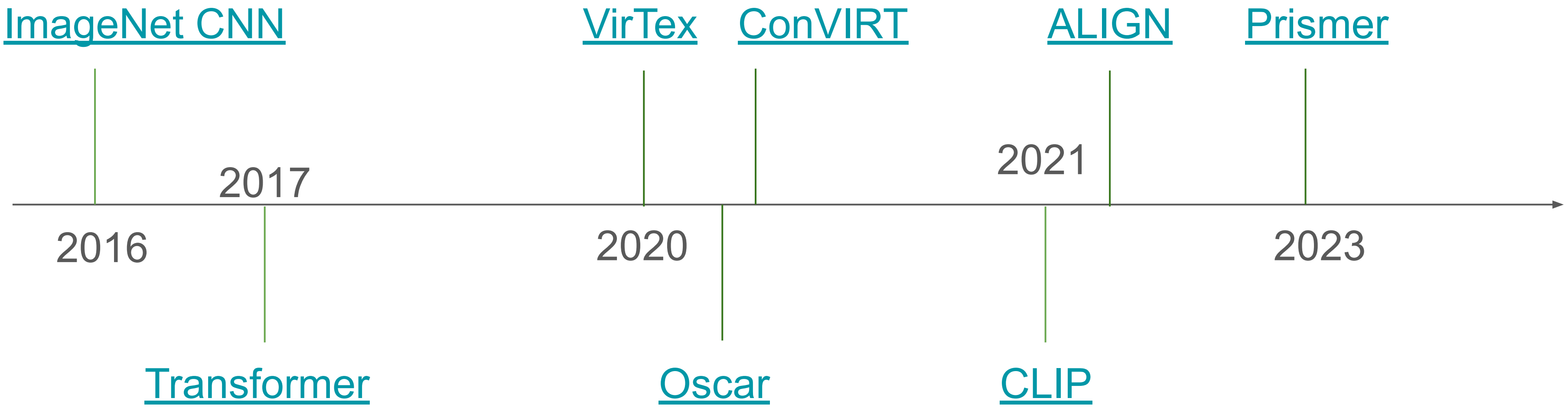
What if we allow other languages?

metric is the mean Recall (mR).

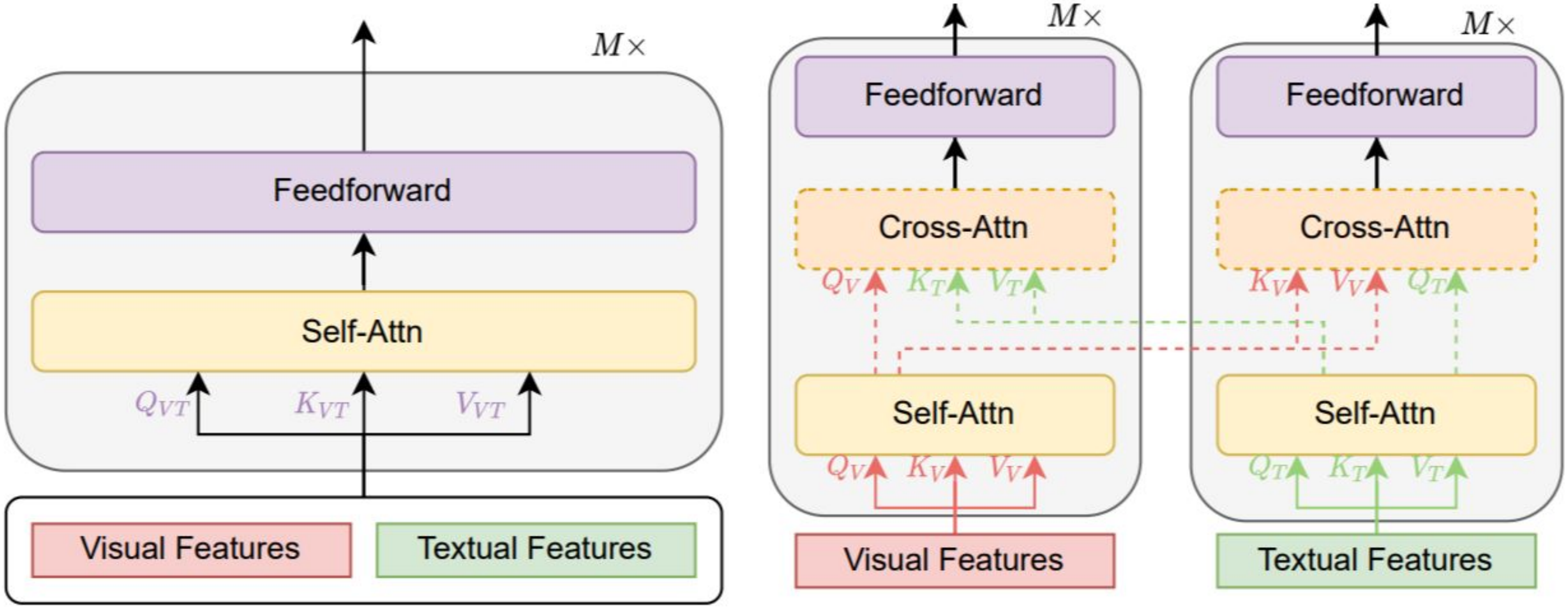
Model	en	de	fr	cs
zero-shot				
M^3P	57.9	36.8	27.1	20.4
$\mathbf{ALIGN_{EN}}$	**92.2**	-	-	-
$\mathbf{ALIGN_{mling}}$	90.2	84.1	**84.9**	63.2
w/ fine-tuning				
M^3P	87.7	82.7	73.9	72.2
UC2	88.2	**84.5**	83.9	**81.2**


Let’s talk about the universe…

History of it all



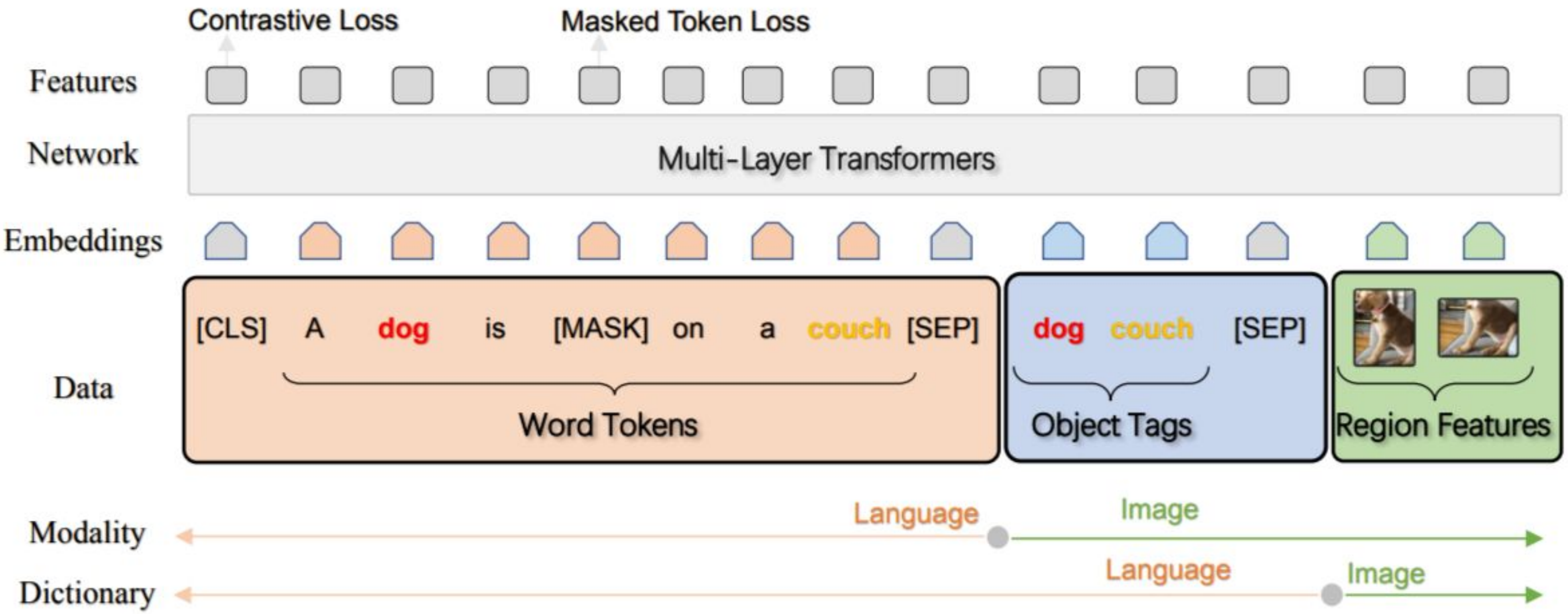
VLM types



(a) Single-Stream Architecture

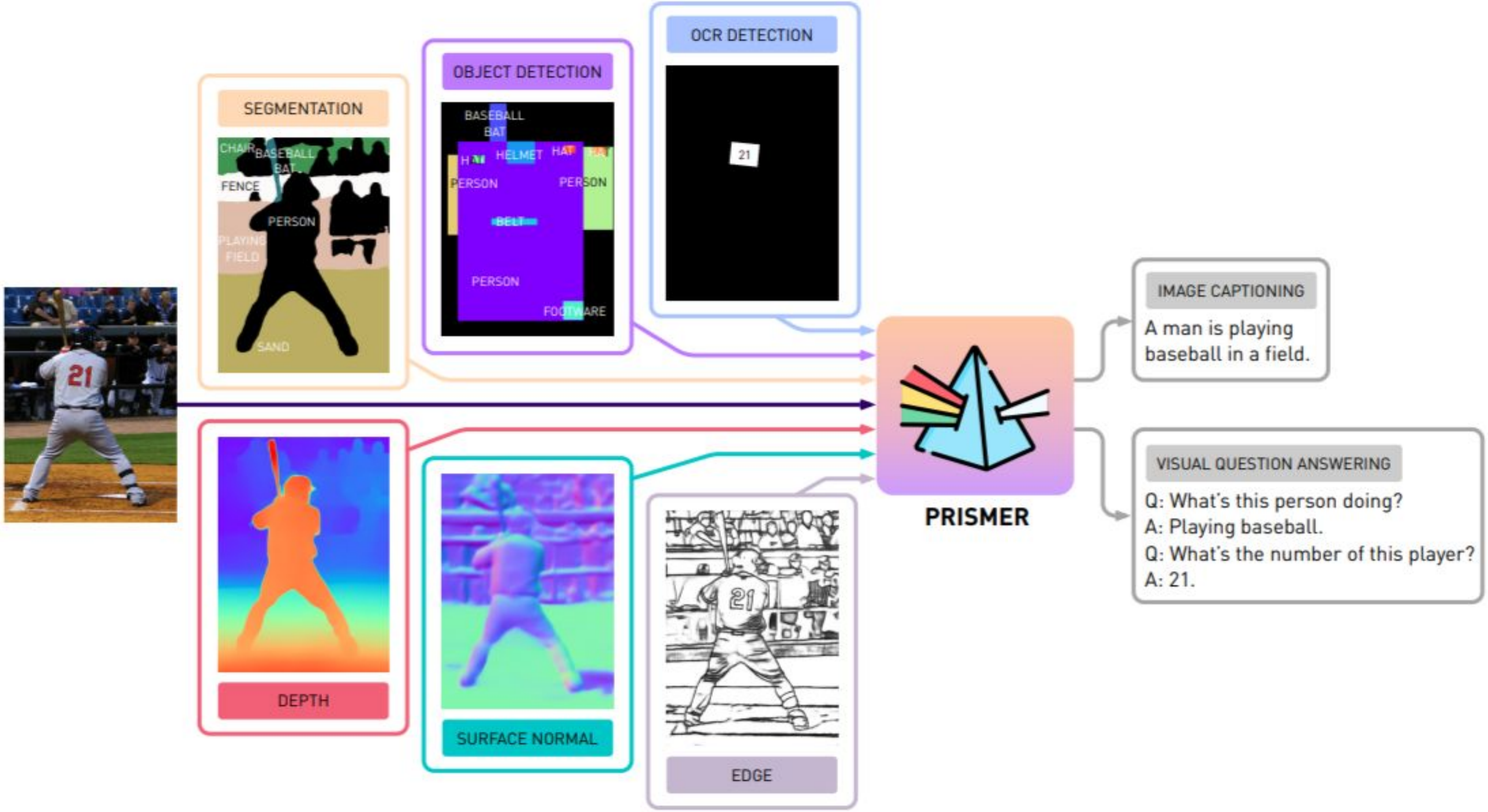
(b) Dual-Stream Architecture

Oscar

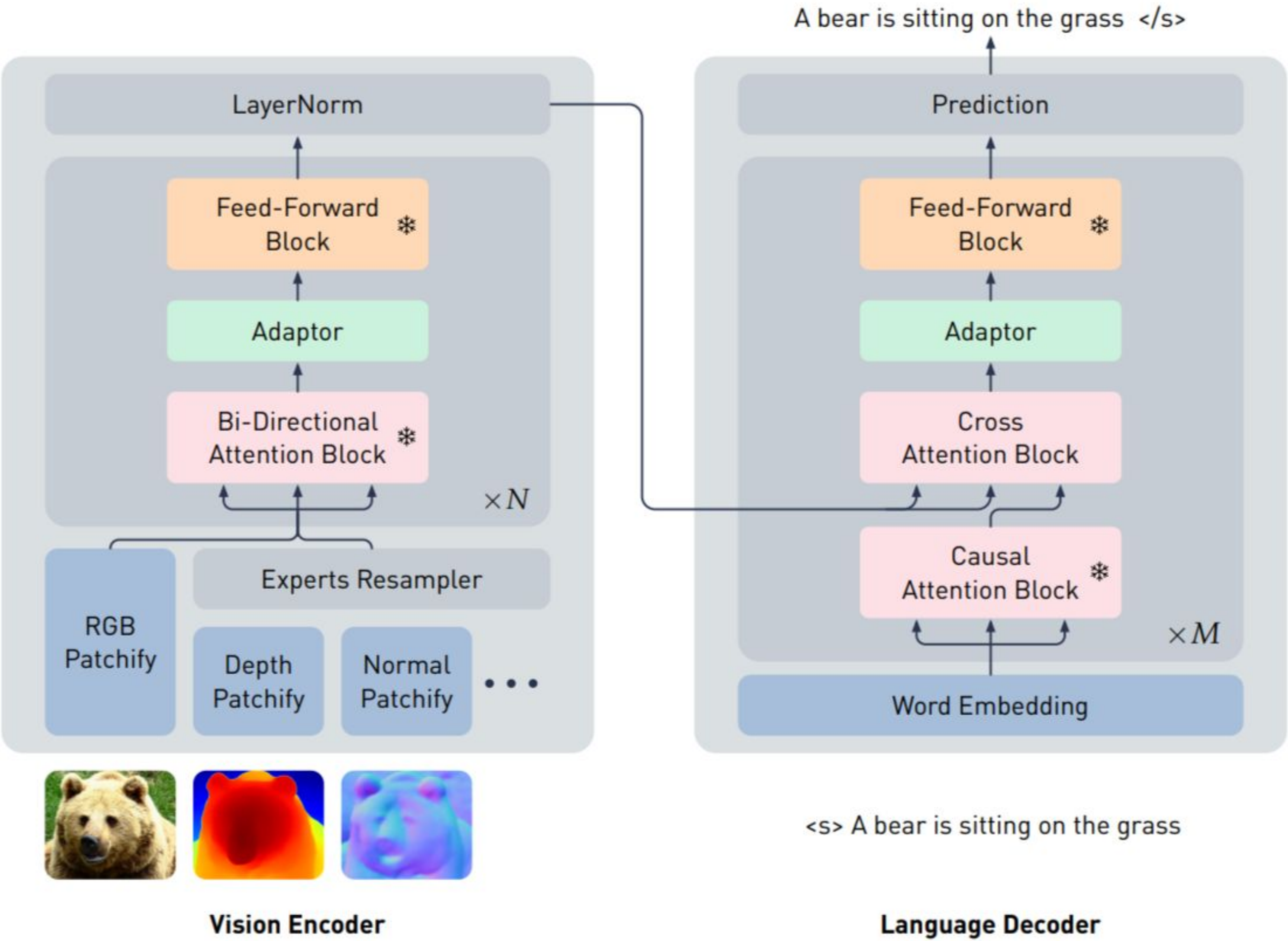


Contrastive Loss
Masked Token Loss
Features
Network
Multi-Layer Transformers
Embeddings
Data
[CLS] A dog is [MASK] on a couch [SEP]
Word Tokens
dog couch [SEP]
Object Tags
Region Features
Modality
Language
Image
Dictionary
Language
Image

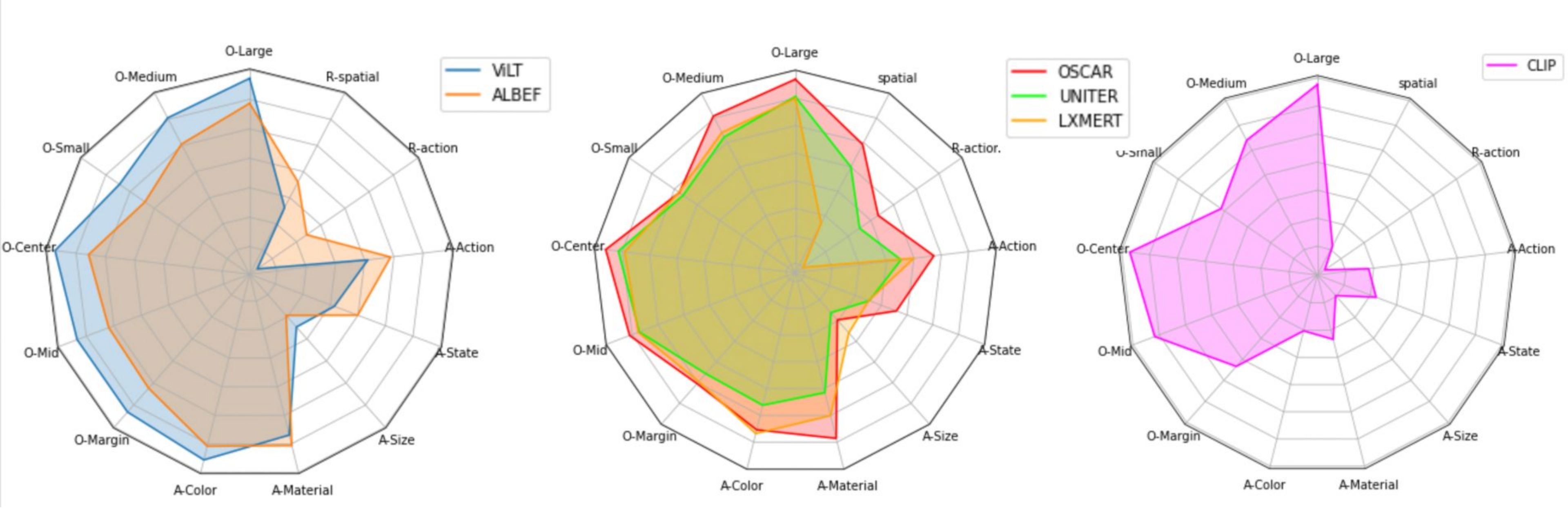
Prismer


SEGMENTATION
CHAIR
BASEBALL BAT
FENCE
PERSON
PLAYING FIELD
SAND
OBJECT DETECTION
BASEBALL BAT
HELMET
HAT
PERSON
PERSON
BELT
PERSON
FOOTWARE
OCR DETECTION
21
DEPTH
SURFACE NORMAL
EDGE
PRISMER
IMAGE CAPTIONING
A man is playing baseball in a field.
VISUAL QUESTION ANSWERING
Q: What's this person doing?
A: Playing baseball.
Q: What's the number of this player?
A: 21.

Prismer


A bear is sitting on the grass </s>
LayerNorm
Feed-Forward Block
Adaptor
Bi-Directional Attention Block
×N
RGB Patchify
Experts Resampler
Depth Patchify
Normal Patchify
Prediction
Feed-Forward Block
Adaptor
Cross Attention Block
Causal Attention Block
×M
Word Embedding
<s> A bear is sitting on the grass
Vision Encoder
Language Decoder

Modifications



Sources

Learning Transferable Visual Models From Natural Language Supervision

Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision

VL-CheckList: Evaluating Pre-trained Vision-Language Models with Objects, Attributes and Relations

Vision-and-Language Pretraining

Prismer: A Vision-Language Model with Multi-Task Experts