Vision-Language Learning for Visual Recognition

Jianwei Yang
Microsoft Research

Microsoft

What is visual recognition?

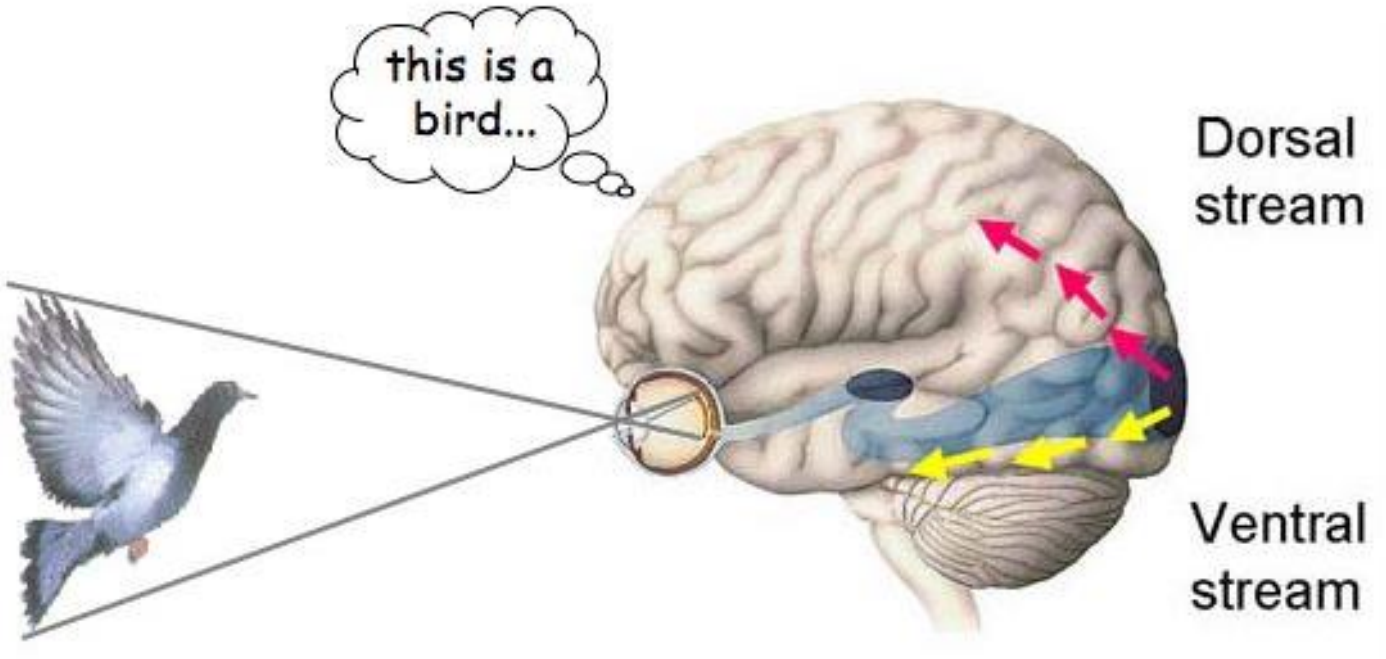
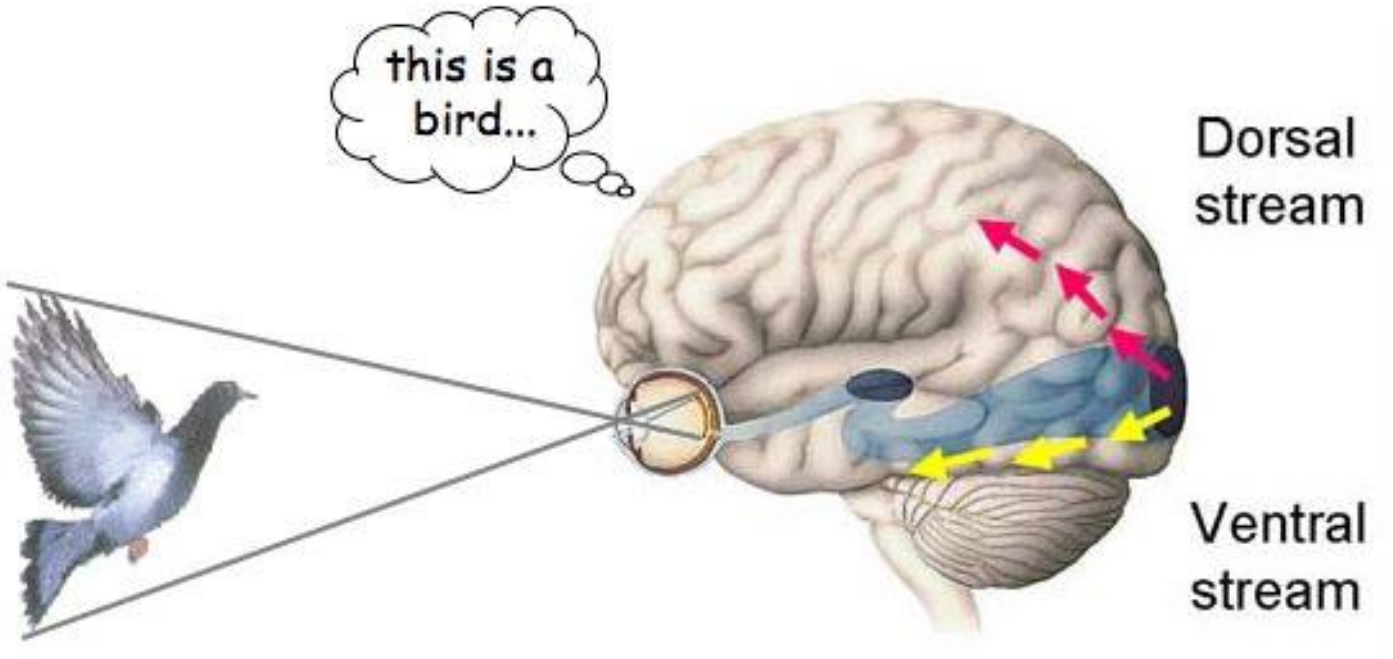


Figure Courtesy: [Perception Lecture Notes: Recognition (nyu.edu)]

What is visual recognition?

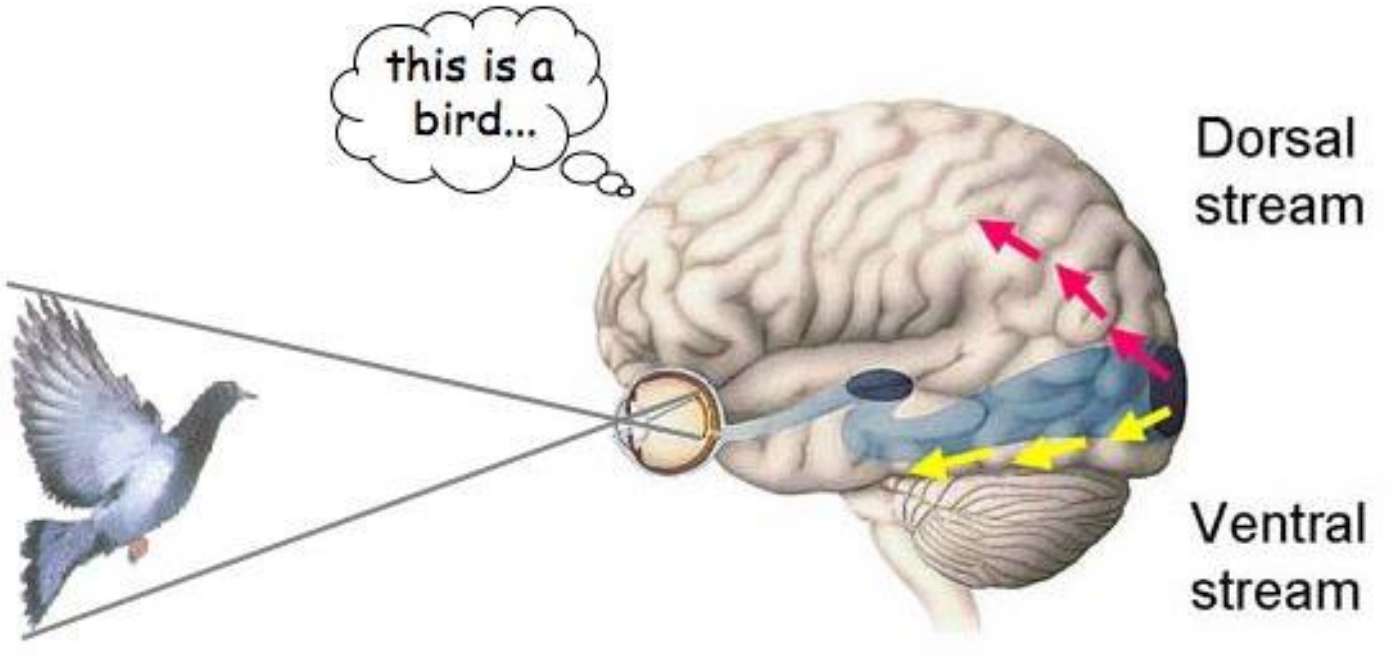


Visual object recognition serves as a gateway **from vision to cognitive processes such as categorization, language and reasoning**.

- ***Object Recognition***. *John E. Hummel*

Figure Courtesy: Perception Lecture Notes: Recognition (nyu.edu)

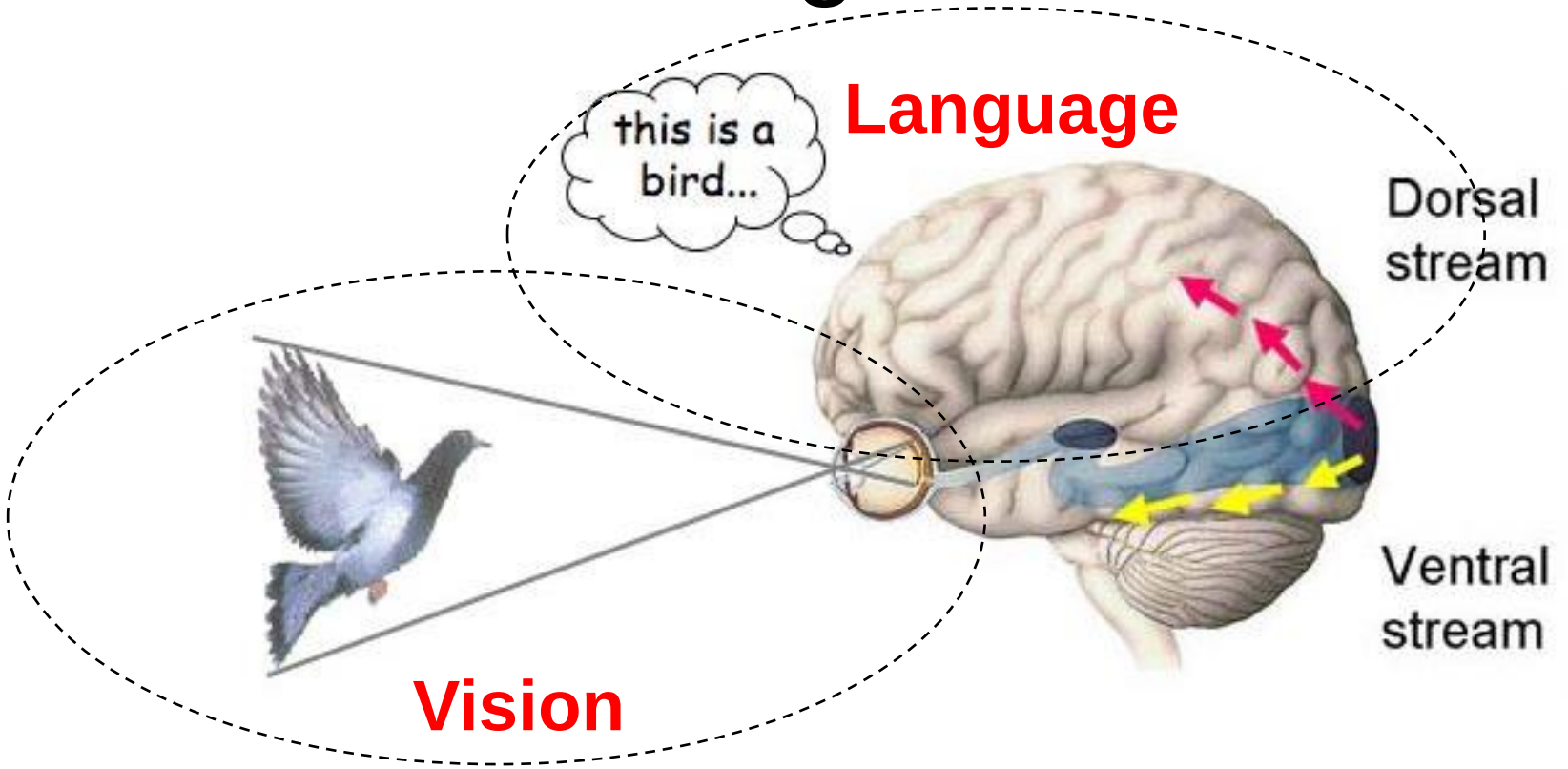
What is visual recognition?



Visual Recognition: Observe visual object and map it with semantic concept

Figure Courtesy: Perception Lecture Notes: Recognition (nyu.edu)

What is visual recognition?



Visual Recognition: Observe visual object and map it with semantic concept

Figure Courtesy: Perception Lecture Notes: Recognition (nyu.edu)

How model learns visual recognition?

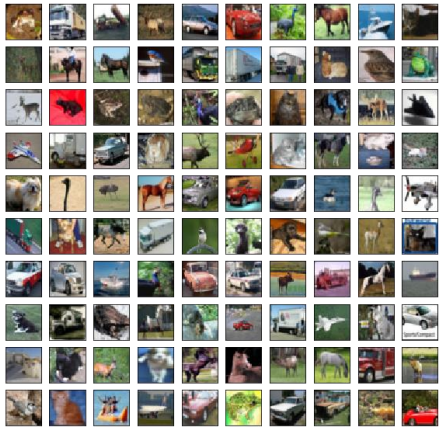
Supervised Learning

Map an image to a discrete label which is associated a visual concept

Image		Label (Concept)
()	⟶	“2” (Apple)

Supervised Learning

MNIST. LeCun *et al.*

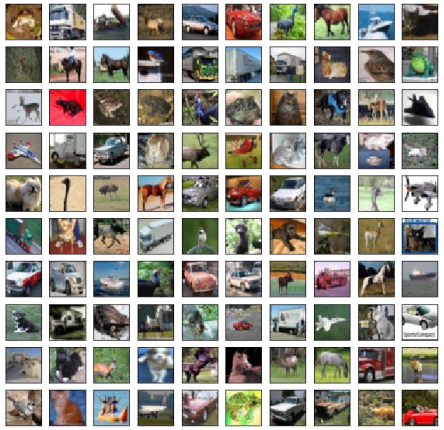


CIFAR-10. Krizhevsky *et al.*

ImageNet. Deng *et al.*

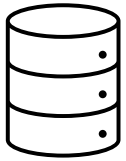
Supervised Learning

MNIST. LeCun *et al.*



CIFAR-10. Krizhevsky *et al.*

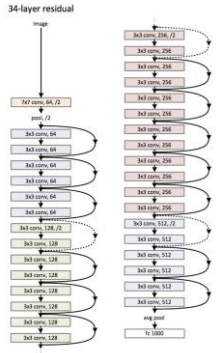
ImageNet. Deng *et al.*

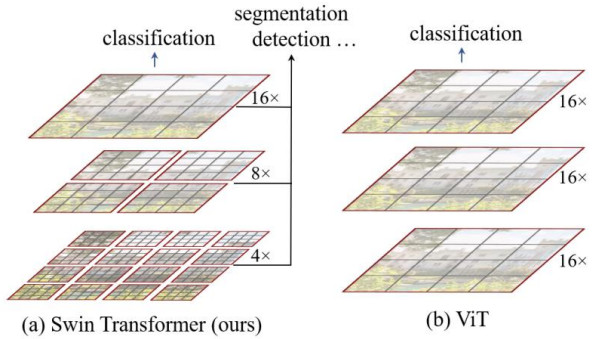


labels

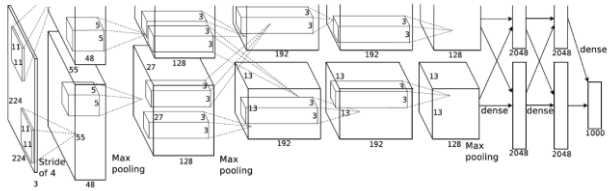
images

Ground-truth





ResNet. He *et al.* **Swin**. Liu *et al.*



AlexNet. Krizhevsky *et al.*

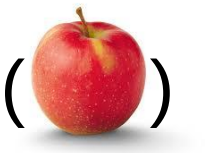
Supervised Learning

- Pros
  - Densely labeled samples for each category

- Cons
  - Requires a lot of human effort
  - Limited number of categories

Zero-Shot Learning (Canonical)

Map an image to description of a visual concept

Image	Descriptions (Concept)
() →	Fruit, Red, Sphere (Apple)
() →	Fruit, Yellow (Orange)

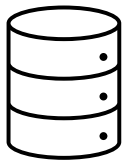
Zero-Shot Learning (Canonical)

CUB-200-2011. Wah *et al.*

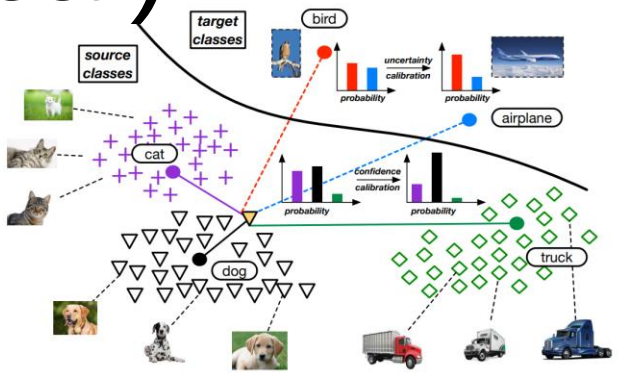
AwA2. Xian *et al.*

aPY. Farhadi *et al.*

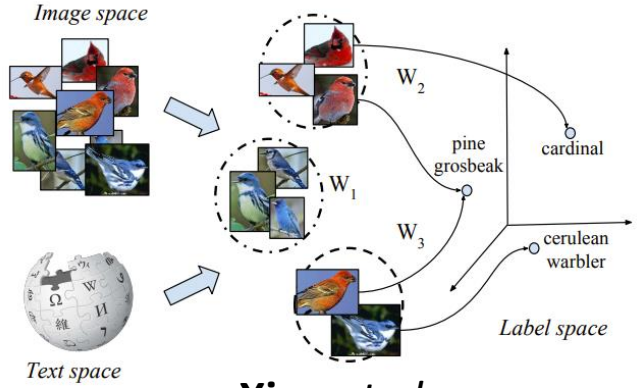
Label & descriptions



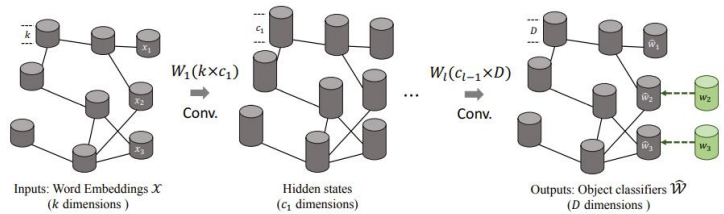
images



Liu *et al.*



Xian *et al.*



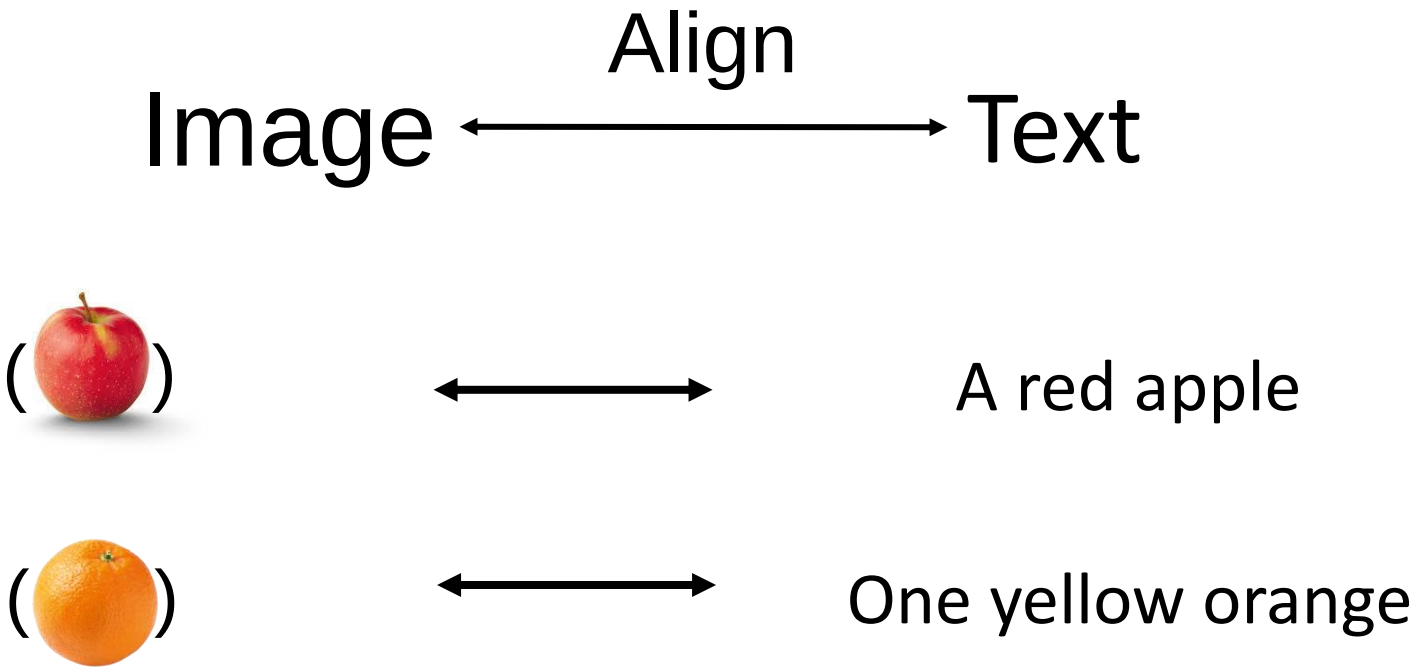
Wang *et al.*

Zero-Shot Learning (Canonical)

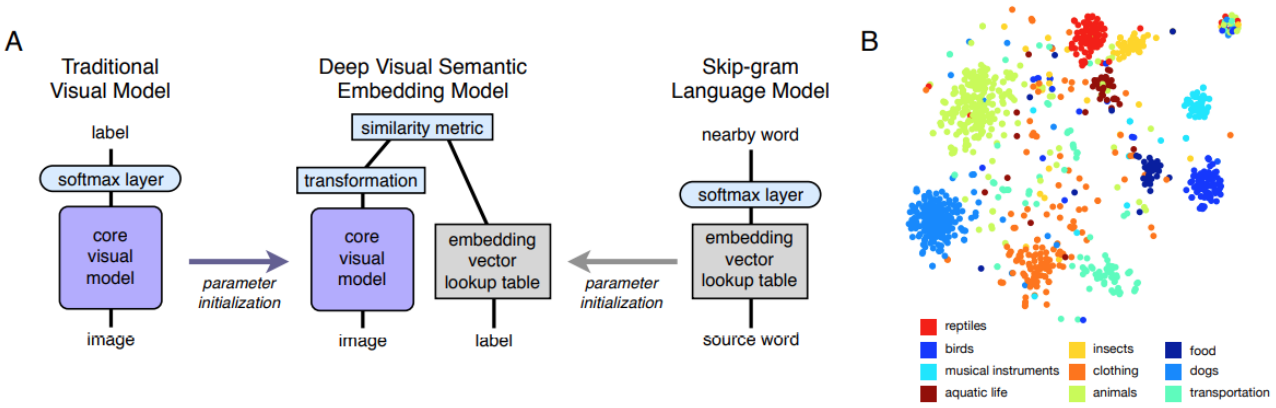
- Pros
  - Directly learn the visual-semantic matching

- Cons
  - Small scale with limited vocabulary
  - Fixed visual and text encoder

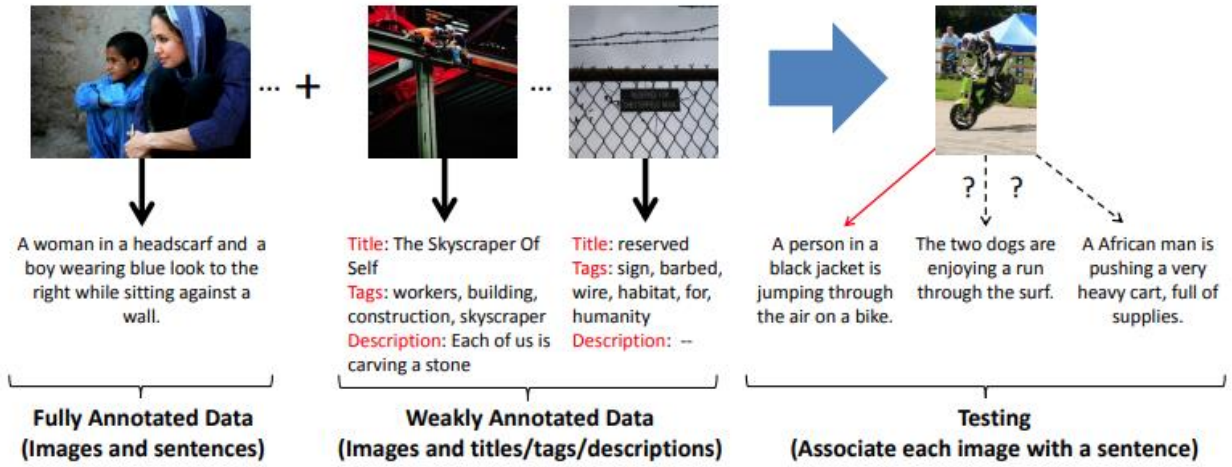
Visual-Semantic Space Learning



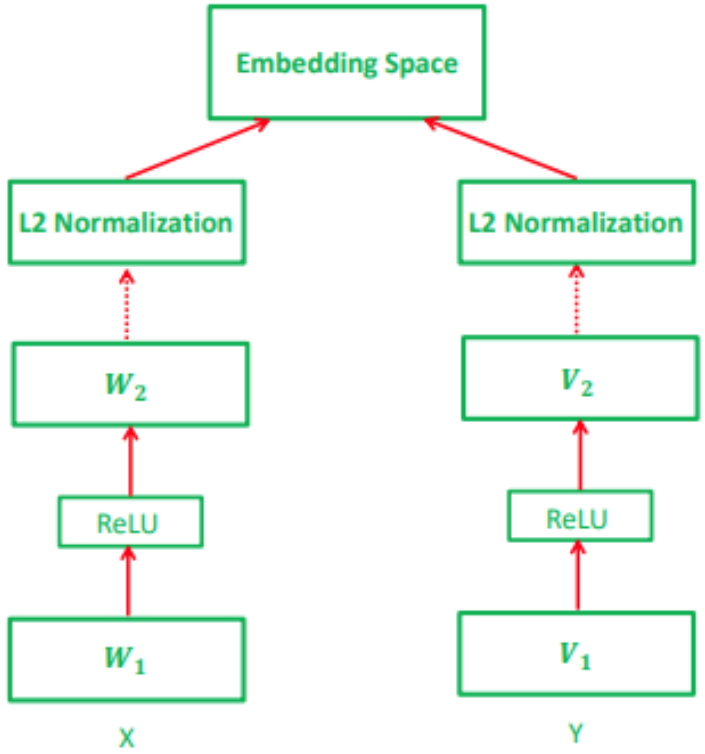
Visual-Semantic Space Learning



DeViSE. Frome *et al.*

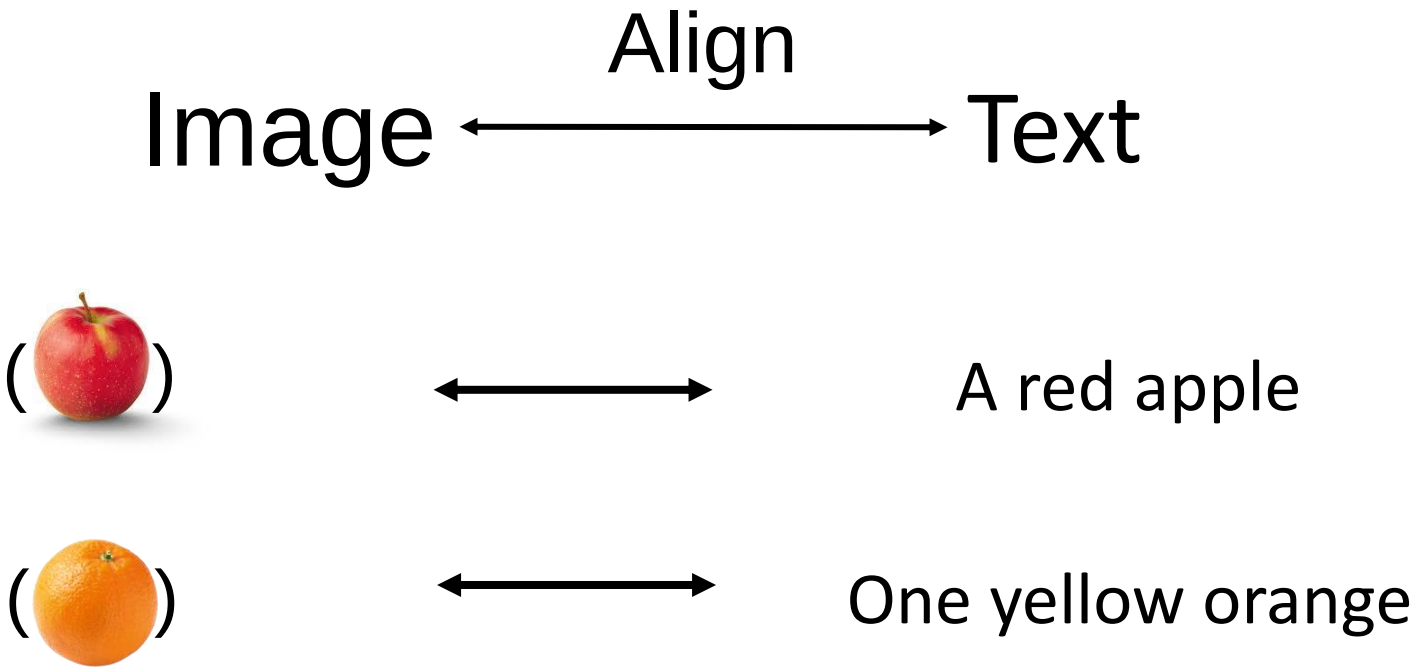


Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. Gong *et al.*



Learning Deep Structure-Preserving Image-Text Embeddings. Wang *et al.*

Visual-Semantic Space Learning

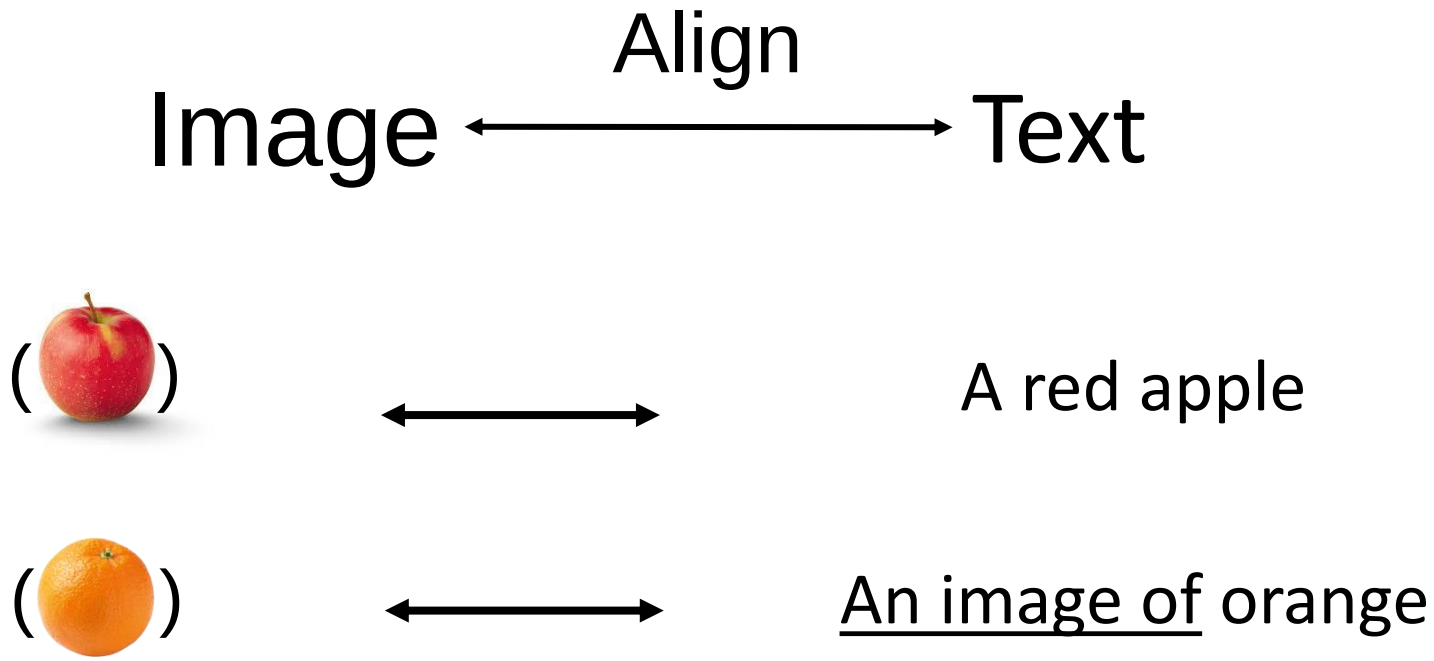


Visual-Semantic Space Learning

- Pros
  - Directly learn the visual-semantic matching

- Cons
  - Fixed visual and text encoder
  - Not showing the capacity for visual recognition

Contrastive Vision-Language Learning

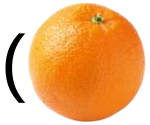


Contrastive Vision-Language Learning

Align

Image ⟷ Text

End-to-end learning on large-scale corpus

() ⟷ An image of orange

The most recent art

Contrastive Vision-Language Learning

A lot of research works come along the line of vision-language learning for vision

In this tutorial

- Vision-Language Learning for **Image-Level** Recognition

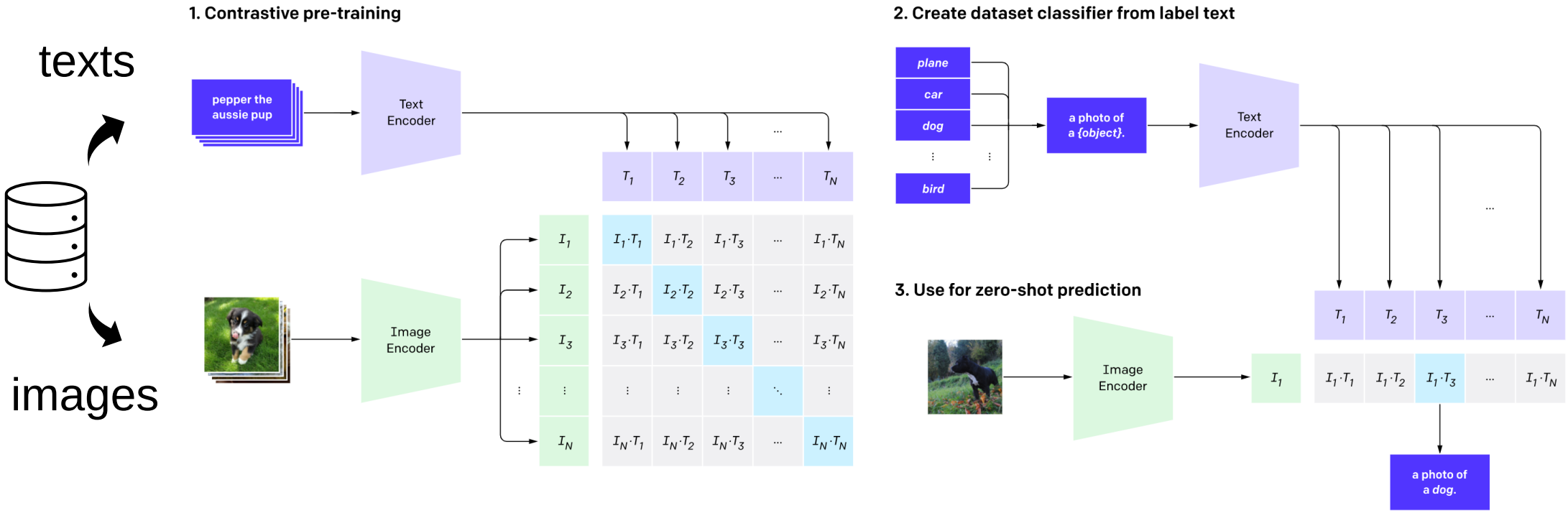
- Vision-Language Learning for Visual **Region-Level** Recognition

In this tutorial

- Vision-Language Learning for **Image-Level** Recognition

- Vision-Language Learning for Visual **Region-Level** Recognition

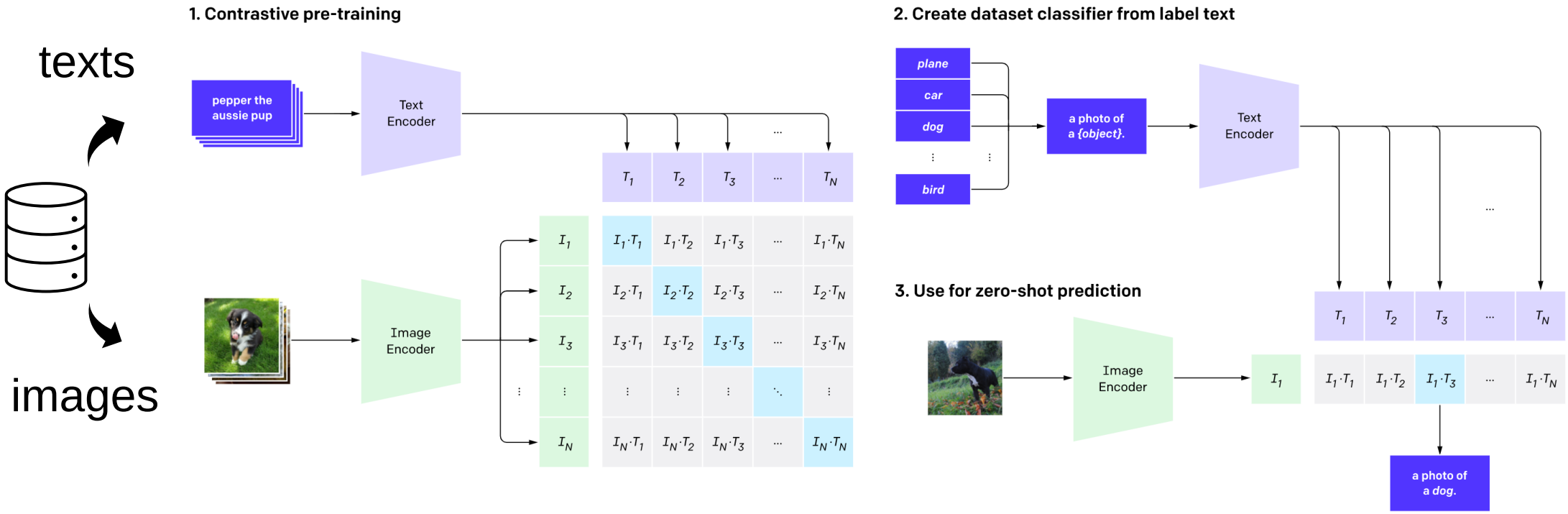
Contrastive Language-Image Learning (CLIP)



- Learning from 400M web-crawled image-text pairs
- Image recognition as an image-text matching problem

CLIP. Radford *et al.*

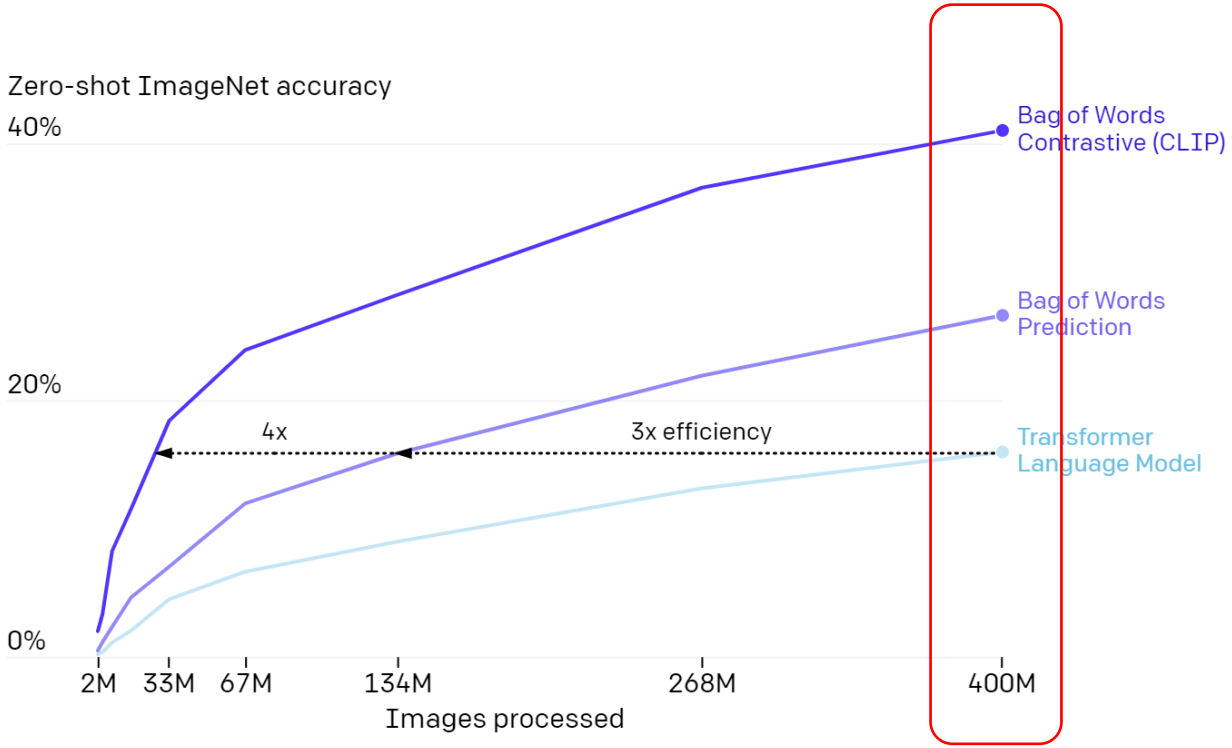
Contrastive Language-Image Learning (CLIP)



$$L_{i2t} = -\sum_j \log \frac{\exp I_j T_j^T}{\sum_k \exp I_j T_k^T} \qquad L_{t2i} = -\sum_j \log \frac{\exp I_j T_j^T}{\sum_k \exp I_k T_j^T}$$

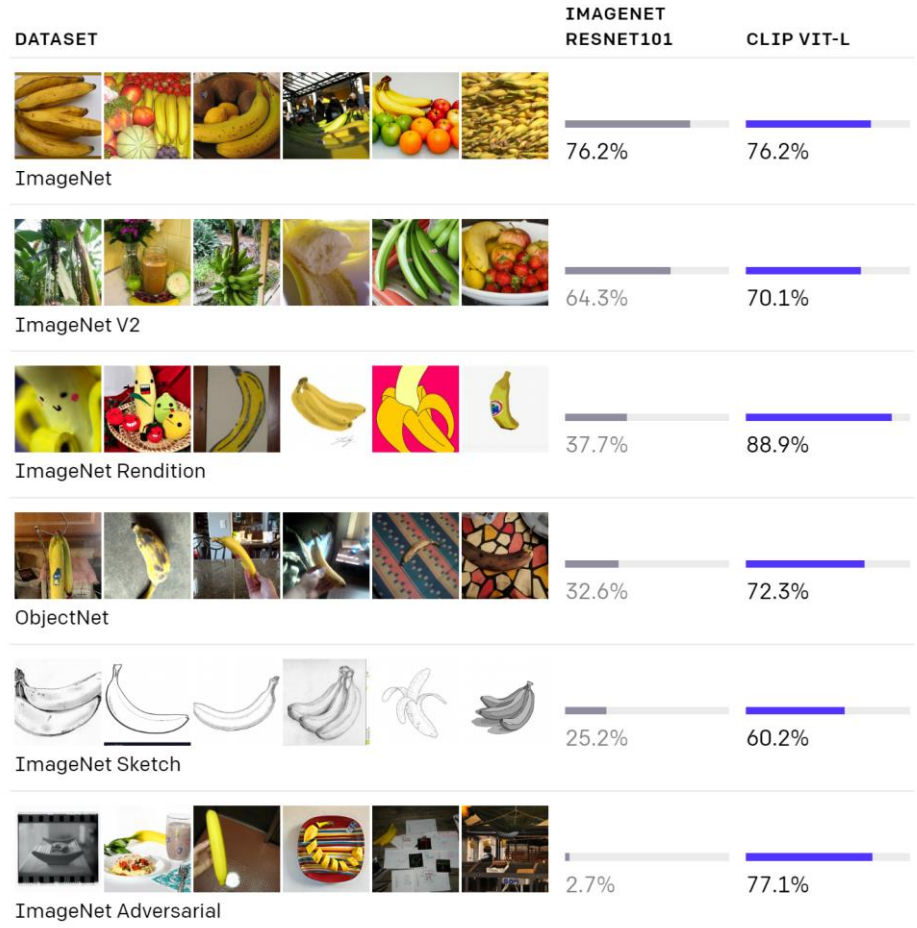
CLIP. Radford *et al.*

Contrastive Language-Image Learning (CLIP)



Contrastive learning is more effective than generative learning

Data really matter



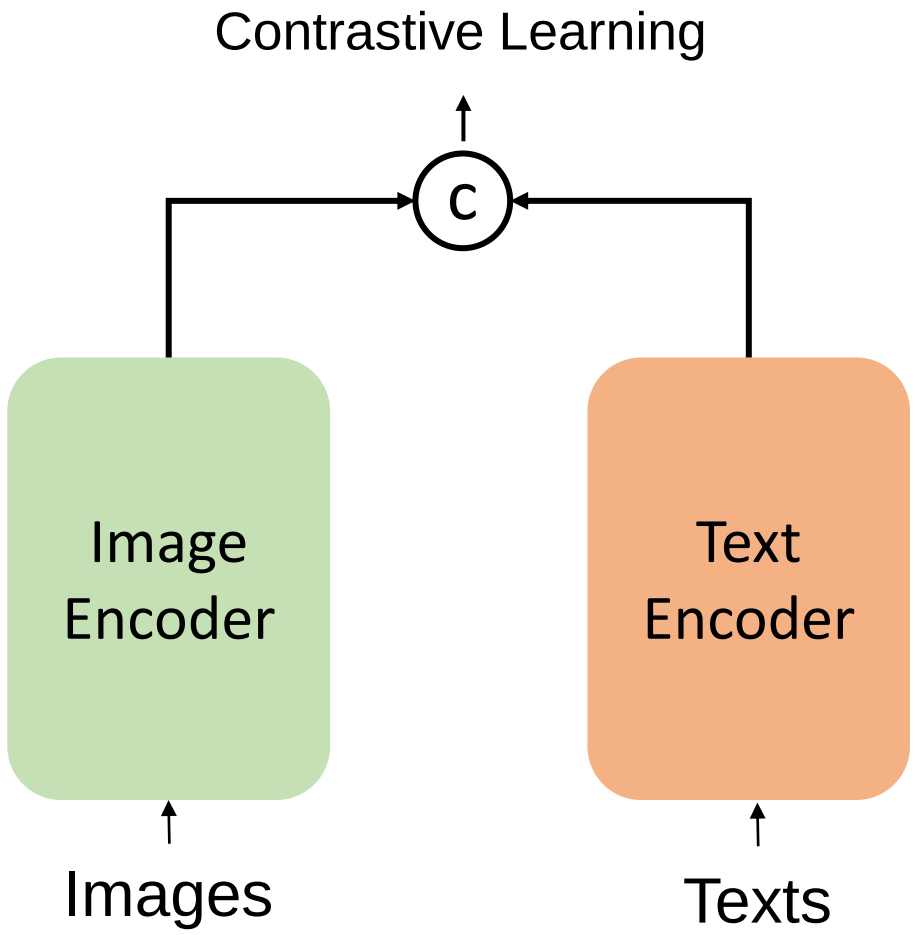
Zero-shot transferring

CLIP. Radford *et al.*

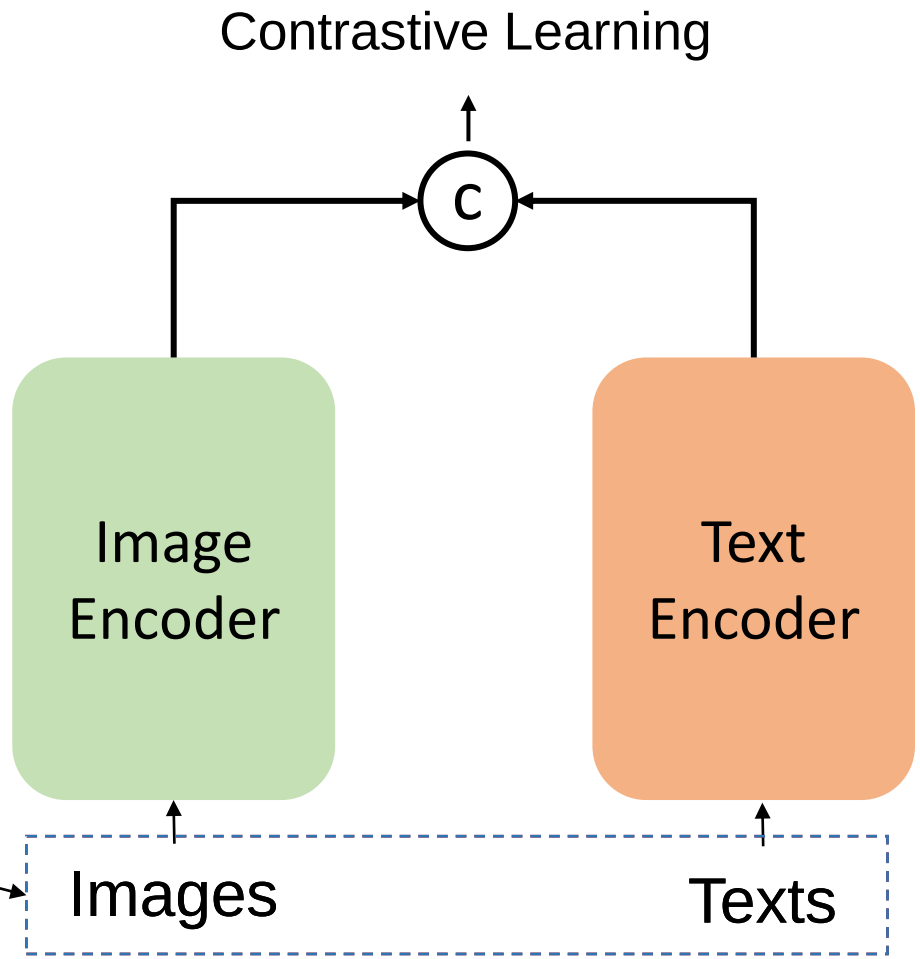
The Lesson from CLIP

- Image recognition can be formulated as an **image-text matching problem** instead of image-label mapping problem
- Image recognition does not require human-annotated image-label data but **huge amount of (noisy) image-text pairs**
- **Contrastive learning is a good learning objective** for multi-modal learning strategy compared with generative learning
- **Two-tower model without fusion** is sufficient to learn good and generic visual and language representations

The most recent art

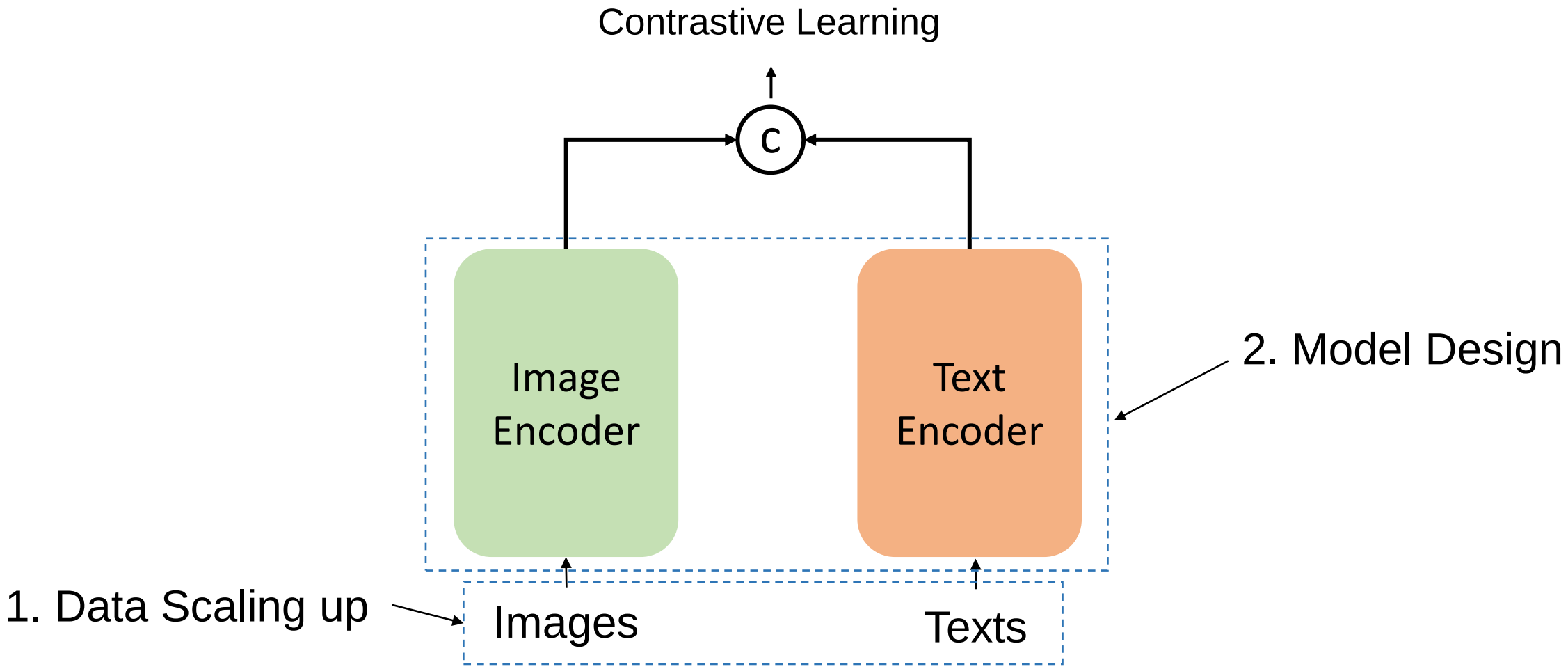


The most recent art

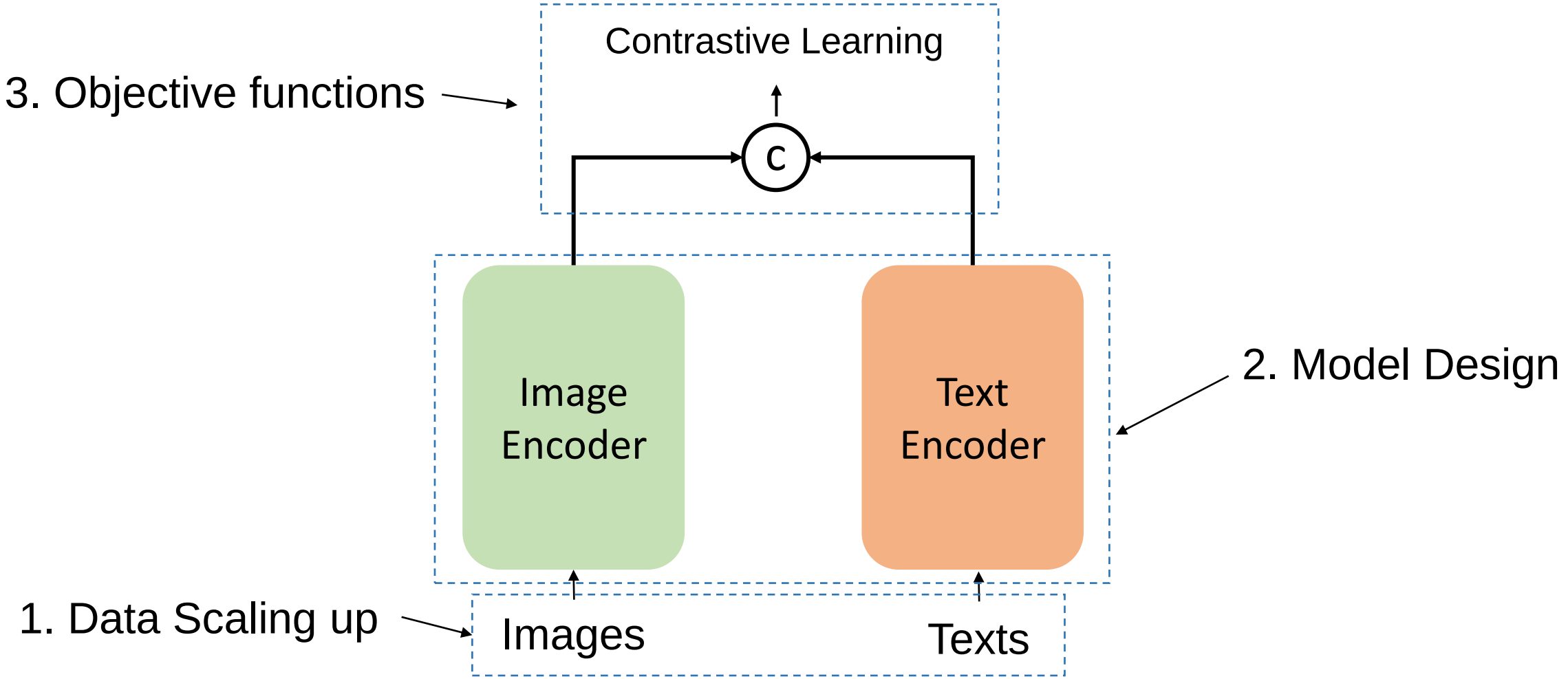


1. Data Scaling up

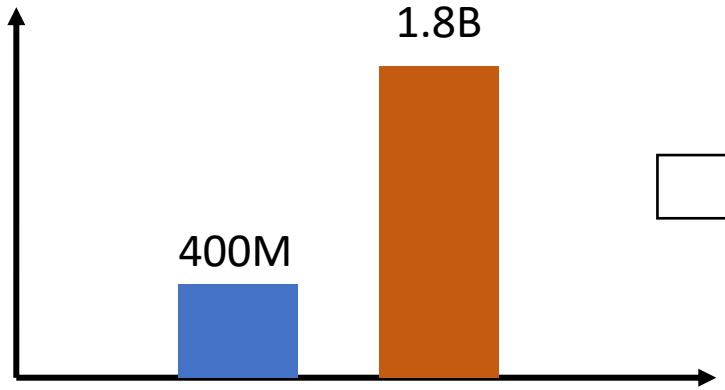
The most recent art



The most recent art



Data Scaling-Up



CLIP ALIGN. Jia *et al.*

Larger scale but more noisy data

"motorcycle front wheel"

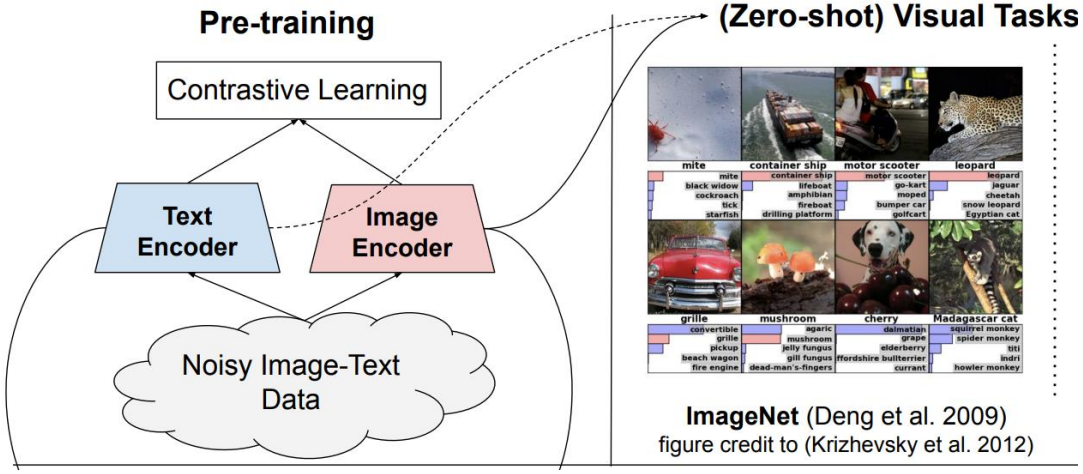
"*thumbnail for version as of 21 57 29 june 2010*"

"file frankfurt airport skyline 2017 05 jpg"

"file london barge race 2 jpg"

"moustache seamless wallpaper design"

"st oswalds way and shops"



ImageNet (Deng et al. 2009)
figure credit to (Krizhevsky et al. 2012)

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**

Zero-shot image classification on ImageNet

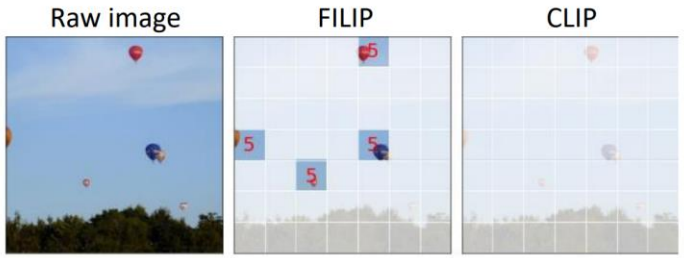
Model (backbone)	Acc@1 w/ frozen features	Acc@1	Acc@5
WSL (ResNeXt-101 32x48d)	83.6	85.4	97.6
CLIP (ViT-L/14)	85.4	-	-
BiT (ResNet152 x 4)	-	87.54	98.46
NoisyStudent (EfficientNet-L2)	-	88.4	98.7
ViT (ViT-H/14)	-	88.55	-
Meta-Pseudo-Labels (EfficientNet-L2)	-	**90.2**	**98.8**
ALIGN (EfficientNet-L2)	**85.5**	88.64	98.67

Image classification finetuning

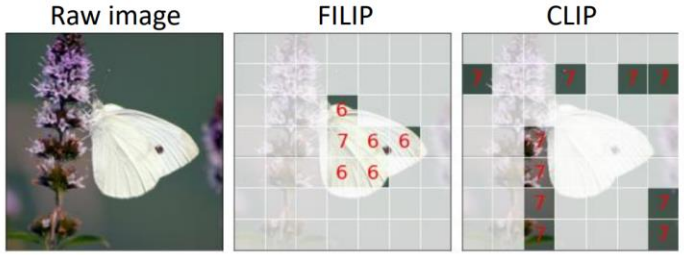
Model Design

From coarse-grain to fine-grained

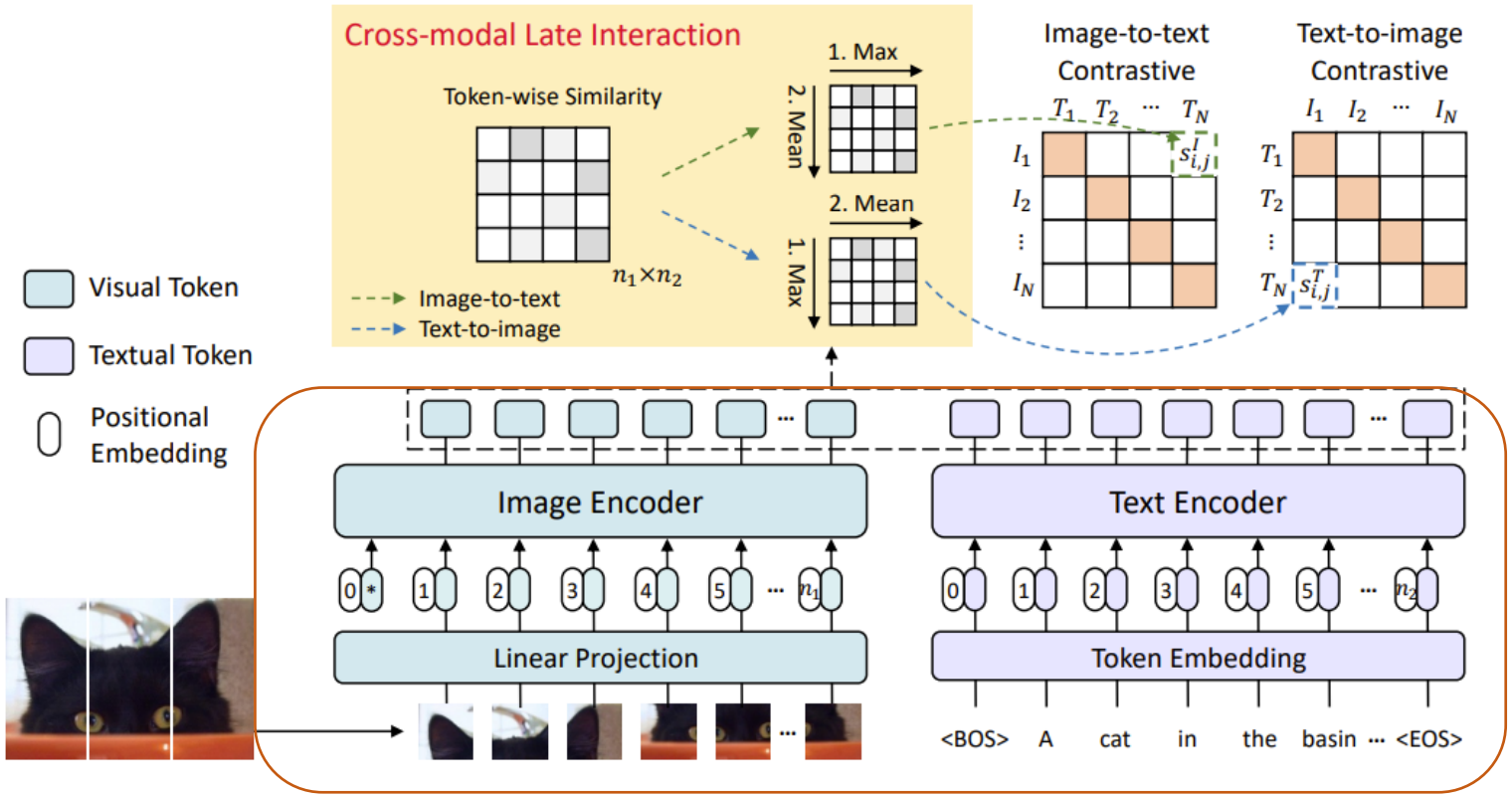
Intuition: computes the loss by first compute the token-wise similarity, and then aggregate the matrix by maxing.



(a) Balloon (5)



(c) Small white butterfly (5, 6, 7)



FILIP. Yao *et al.*

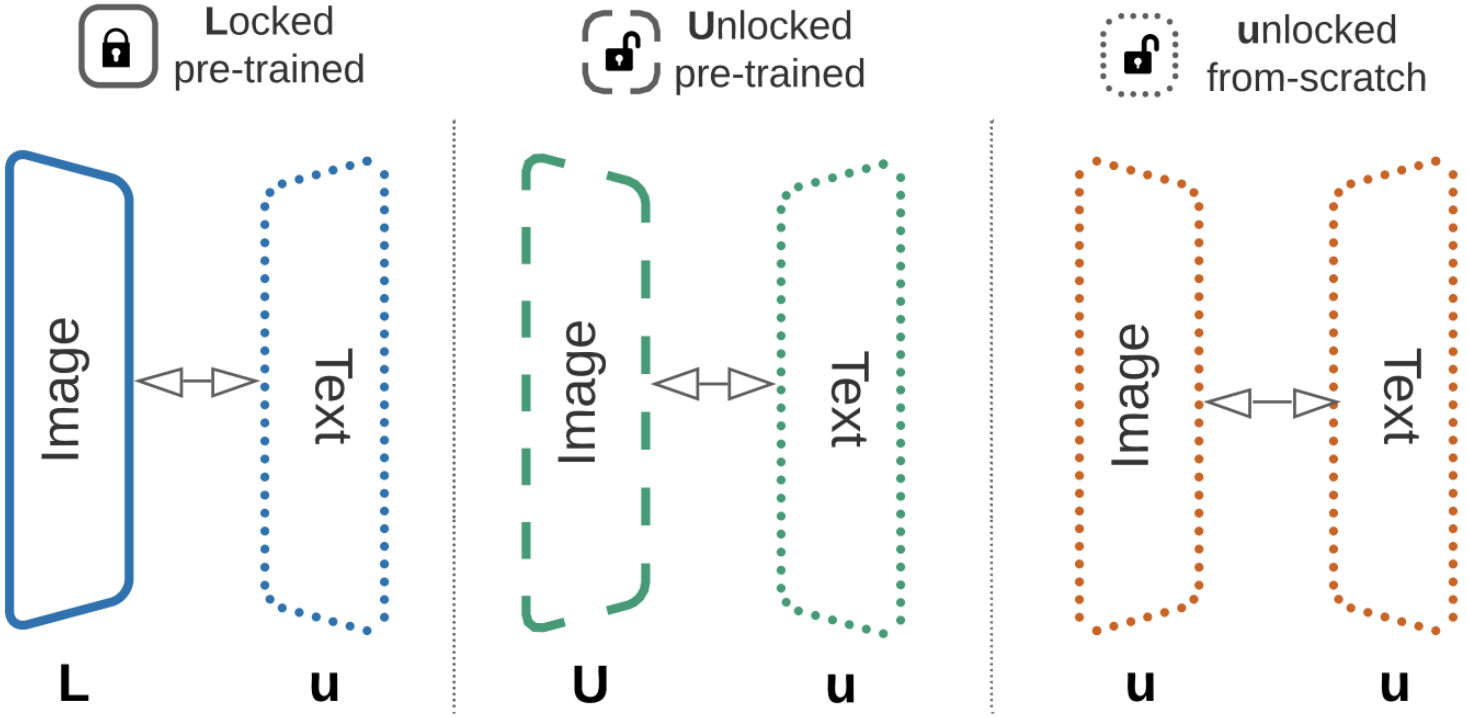
	CIFAR10	CIFAR100	Caltech101	StanfordCars	Flowers102	Food101	SUN397	DTD	Aircrafts	OxfordPets	EuroSAT	**ImageNet**	**Average**
CLIP-ViT-B/32	91.3	65.1	87.9	59.4	66.7	84.4	63.2	44.5	21.2	87.0	49.4	63.2	65.3
$\text{FILIP}_{\text{base}}$-ViT-B/32	86.9	65.5	91.9	55.4	85.3	82.8	69.1	49.3	57.2	88.1	49.9	68.8	**70.9**$^{+5.6}$
CLIP-ViT-L/14	96.2	77.9	92.6	77.3	78.7	92.9	67.7	55.3	36.1	93.5	59.9	75.3	75.3
$\text{FILIP}_{\text{large}}$-ViT-L/14	95.7	75.3	93.0	70.8	90.1	92.2	73.1	60.7	60.2	92	59.2	77.1	**78.3**$^{+3.0}$

Zero-shot recognition on downstream datasets

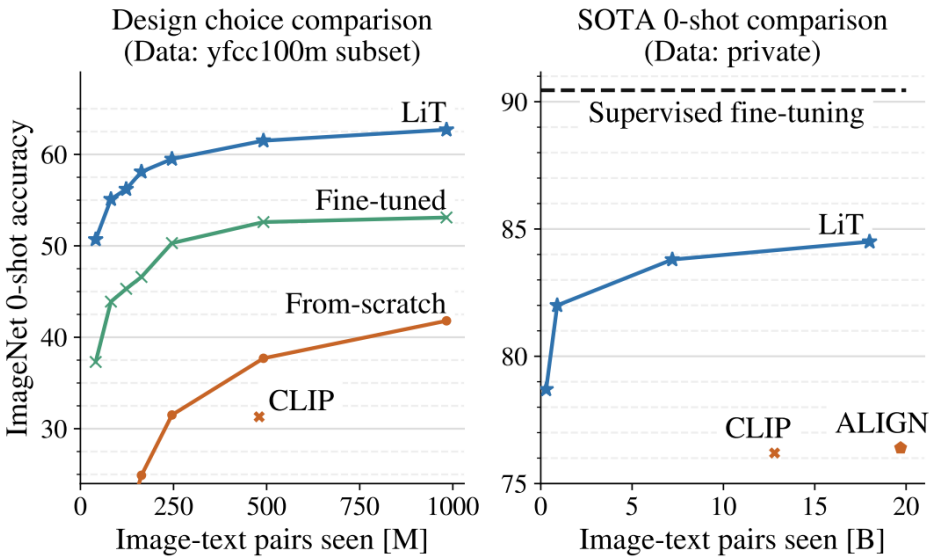
Model Design

From joint training to sequential training

Intuition: image and text encoder are not necessarily trained all from scratch



Locked Image-Tuning (LIT). Zhai *et al.*



Dataset	Method	INet	INet-v2	INet-R	INet-A	ObjNet	ReaL	VTAB-N
Private	CLIP [45]	76.2	70.1	88.9	77.2	72.3	-	-
	ALIGN [30]	76.4	70.1	92.2	75.8	-	-	-
	LiT	**84.5**	**78.7**	**93.9**	**79.4**	**81.1**	88.0	72.6
Public	CLIP [45]	31.3	-	-	-	-	-	-
	OpenCLIP [28]	34.8	30.0	-	-	-	-	-
	LiT	**75.7**	**66.6**	60.4	37.8	54.5	82.1	63.1
*	ResNet50 [25]	75.8	63.8	36.1	0.5	26.5	82.5	72.6

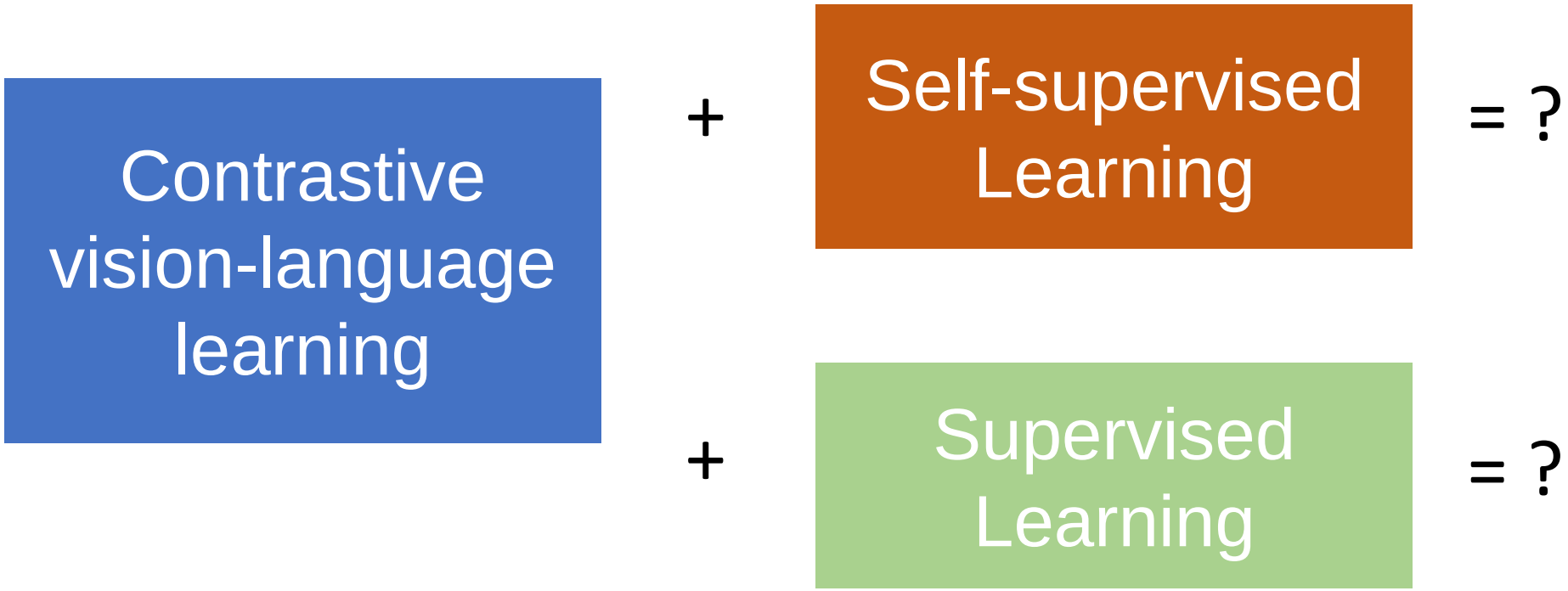
Zero-shot recognition

Model: ViT-B/16	Pre-training Dataset	Pre-training Labels?	Pre-training Full IN	Pre-training 10-shot	LiT 0-shot	LiT I→T	LiT T→I
MoCo-v3 [10]	IN	n	76.7	60.6	55.4	33.5	17.6
DINO [4]	IN	n	78.2	61.2	55.5	33.4	18.2
AugReg [54]	IN21k	y	77.4	63.9	55.9	30.3	17.2
AugReg [54]	IN	y	77.7	77.1	64.3	25.4	13.8
AugReg [54]	Places	y	-	22.5	28.5	25.1	12.9

Effect of pretraining

Learning Objectives

Combining contrastive learning with other learning objectives



Learning Objectives

Combining contrastive learning with other learning objectives

Contrastive vision-language learning

\+ Self-supervised Learning = ?

\+ Supervised Learning = ?

Learning Objectives

Self-supervised learning on each modality:

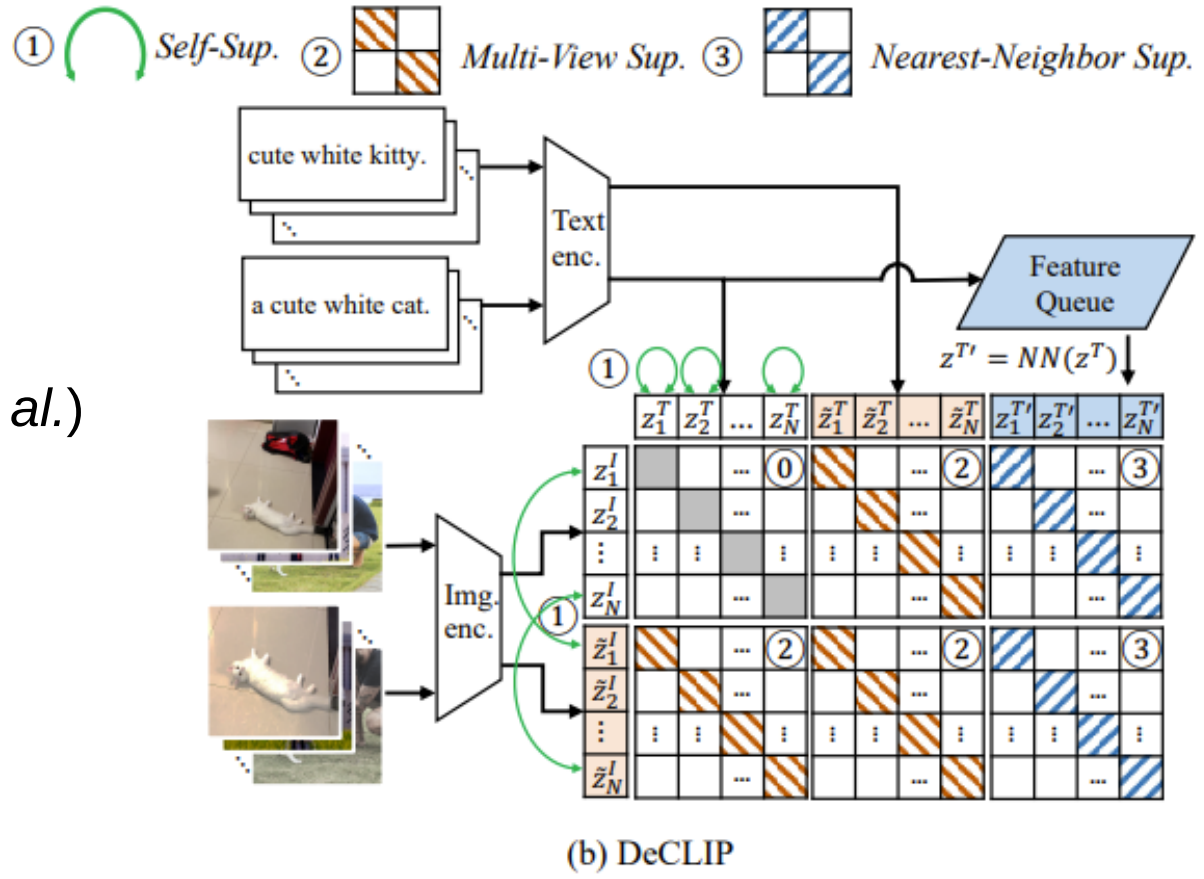
1. Image: SimSam (Chen *et al.***)**
2. Text: MLM (Devlin *et al*.)

Multi-View Supervision:

1. Image + Augmented image (Caron *et al.*)
2. Text + Augmented text (Wei *et al.*)

Nearest-neighbor supervision:

1. Find the nearest text embeddings as the matched texts to an image



Supervision Exists Everywhere. Li *et al.*

Learning Objectives

Self-supervised learning on each modality:

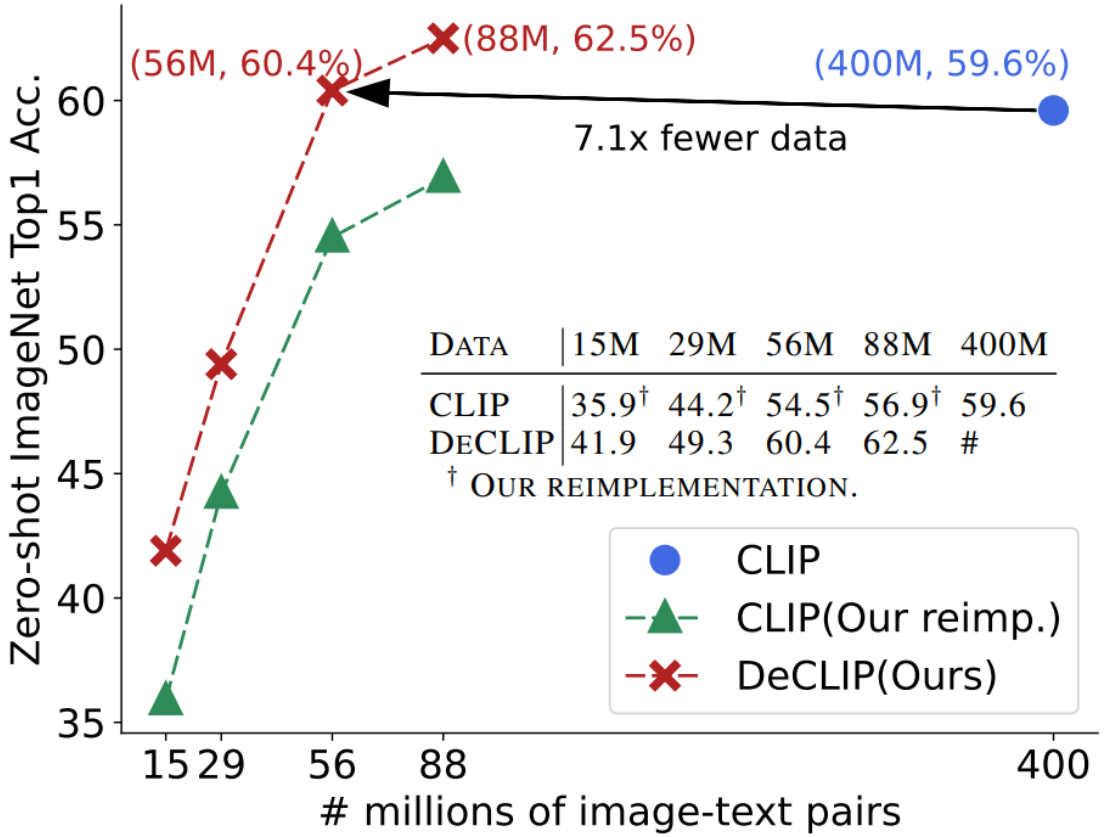
1. Image: SimSam (Chen *et al.***)**
2. Text: MLM (Devlin *et al*.)

Multi-View Supervision:

1. Image + Augmented image (Caron *et al.*)
2. Text + Augmented text (Wei *et al.*)

Nearest-neighbor supervision:

1. Find the nearest text embeddings as the matched texts to an image



Data	15M	29M	56M	88M	400M
CLIP	35.9†	44.2†	54.5†	56.9†	59.6
DeCLIP	41.9	49.3	60.4	62.5	#

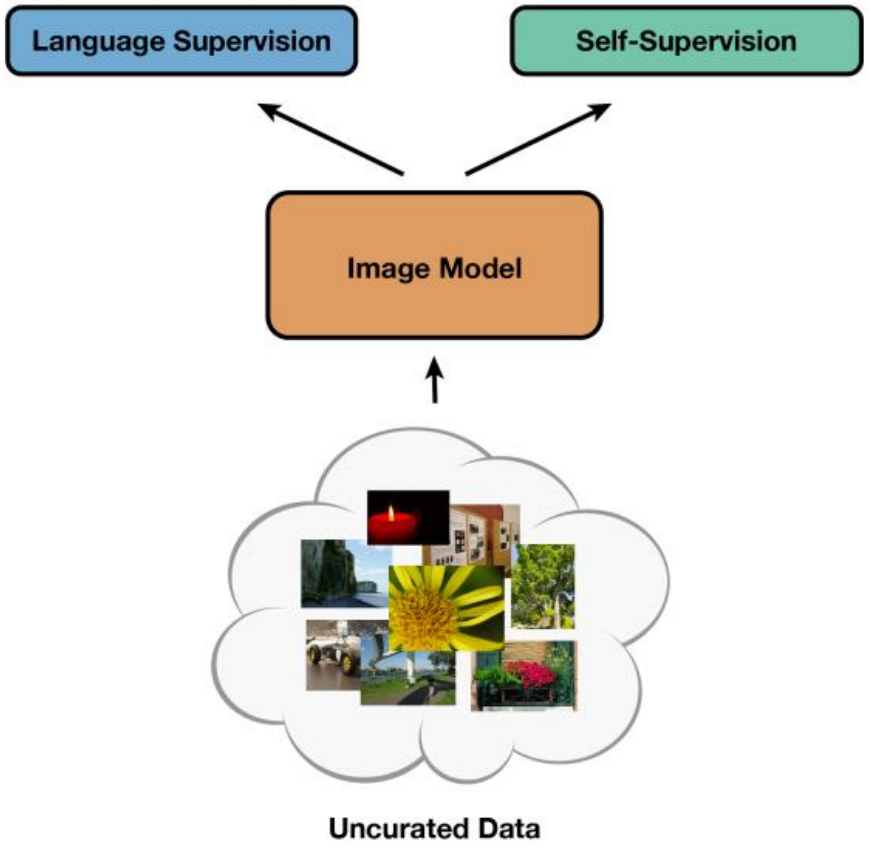
† Our reimplementation.

Combining vision-language and self-supervised learning improves data efficiency significantly

Supervision Exists Everywhere. Li *et al.*

Learning Objectives

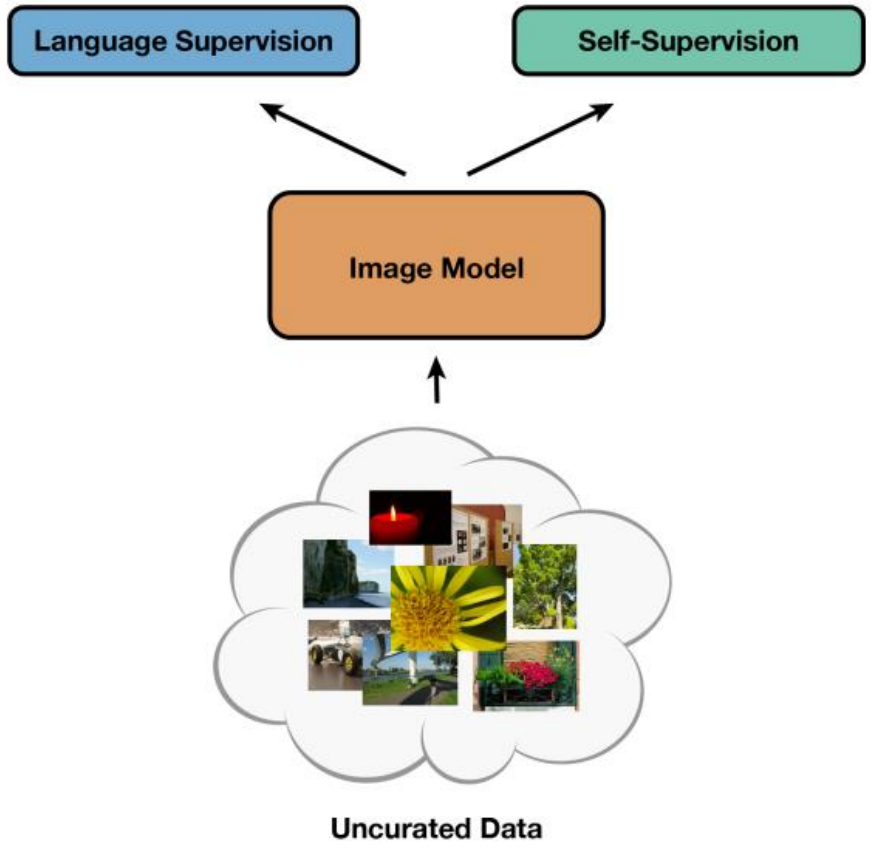
Simply combining contrastive language-image pretraining with self-supervised learning



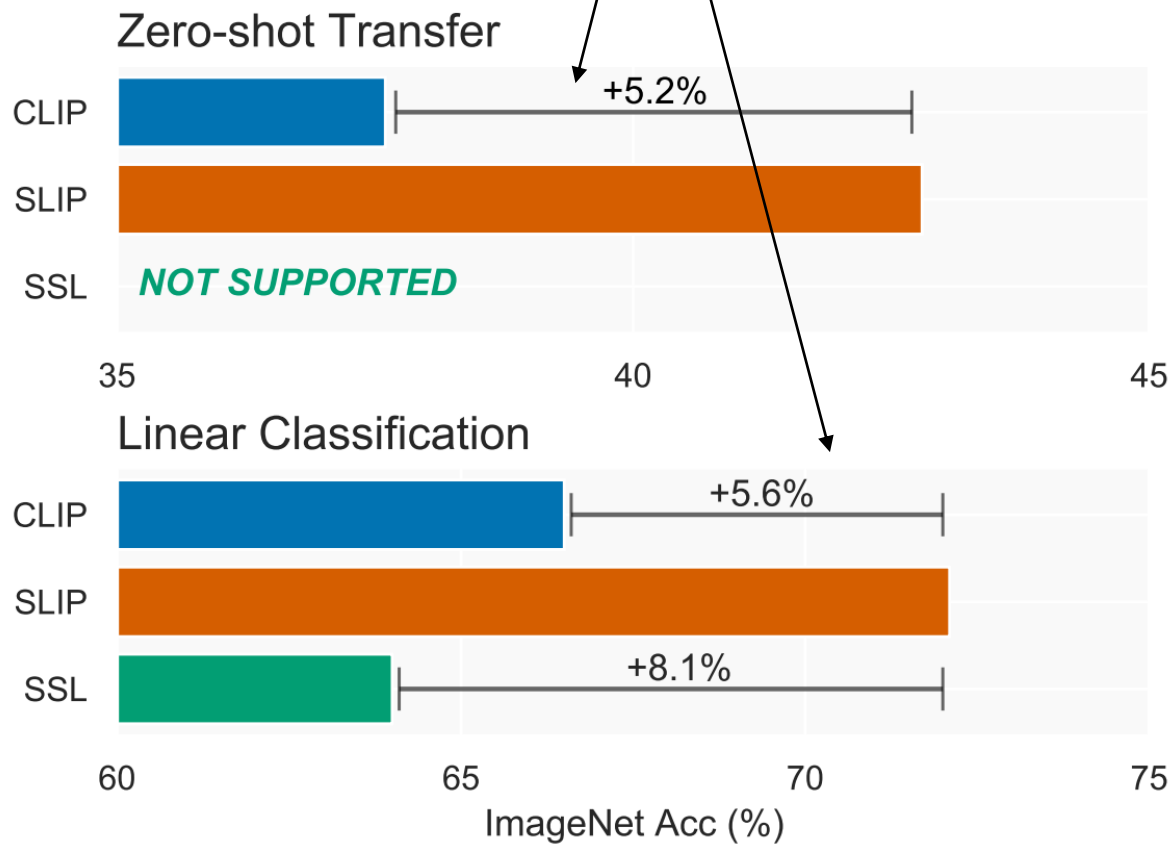
SLIP. Mu **et al.**

Learning Objectives

Simply combining contrastive language-image pretraining with self-supervised learning



SLIP outperforms CLIP on both zero-shot transfer and linear classification



SLIP. Mu **et al.**

Learning Objectives

Combining contrastive learning with other learning objectives

Contrastive vision-language learning

+ Self-supervised Learning = ?

+ Supervised Learning = ?

A new perspective

Supervised Learning

Label

Image

1 1 2 2




Chihuahua *Chihuahua* *Siberian Husky* *Siberian Husky*

Language-Image Learning

3 4 5 6




A white dog sitting on a couch

A dog looking bashfully to the side

A black cat is sleeping on a computer

Two parrots perching on a tree branch

Image

Text

1. Supervised learning use the categorical labels but with concepts associated
2. Language-image learning assume each image-text pair belongs to a unique category

Image-Text-Label Space

Label

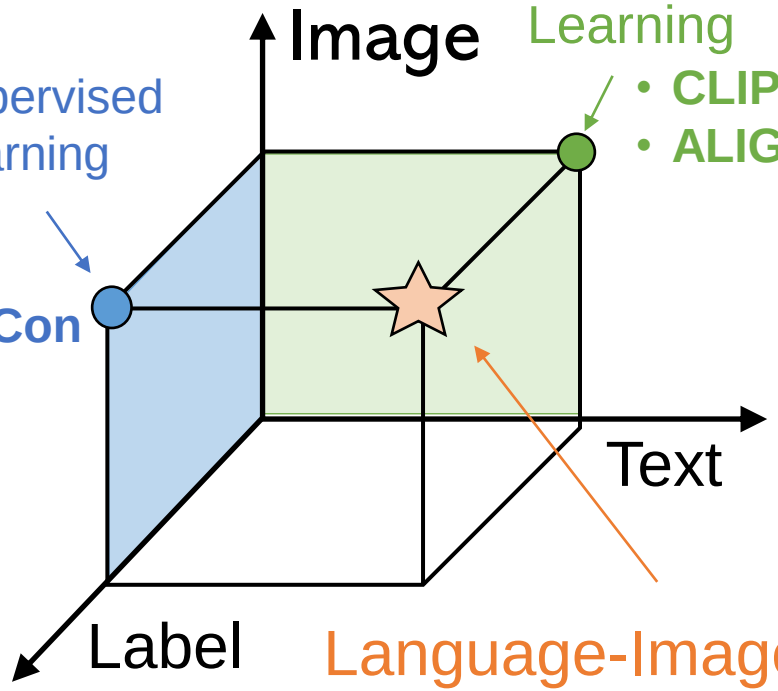
Image

Learning a mapping function from image to label

Supervised Learning

- **CE**
- **SupCon**

Pros: Discriminative visual representations
Cons: Limited visual concepts



Language-Image Learning

- **CLIP**
- **ALIGN**

Image

Text

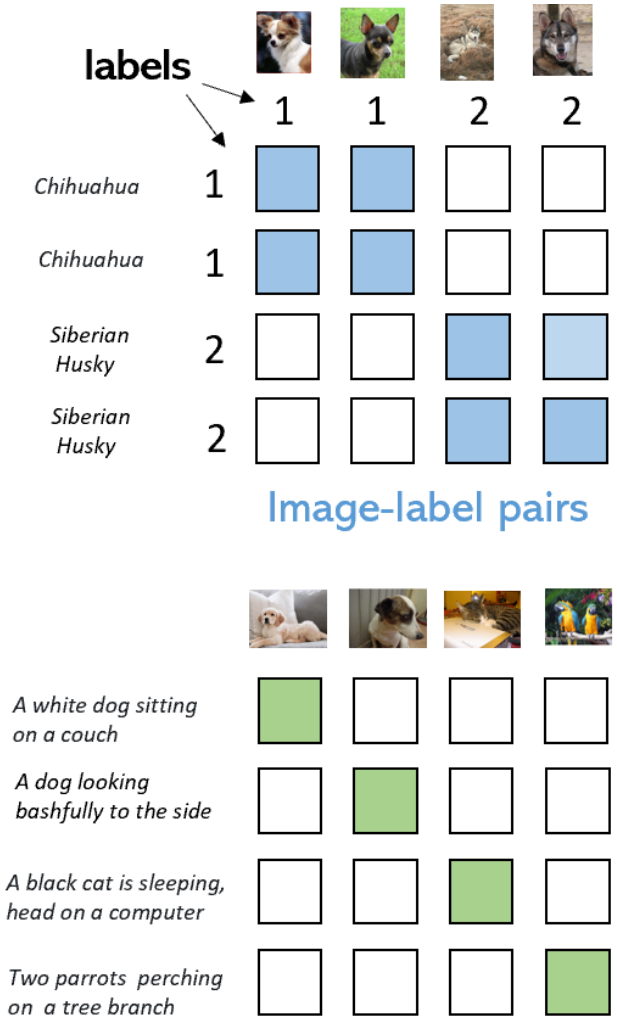
Pros: Broad semantic coverage
Cons: Representations are not discriminative enough

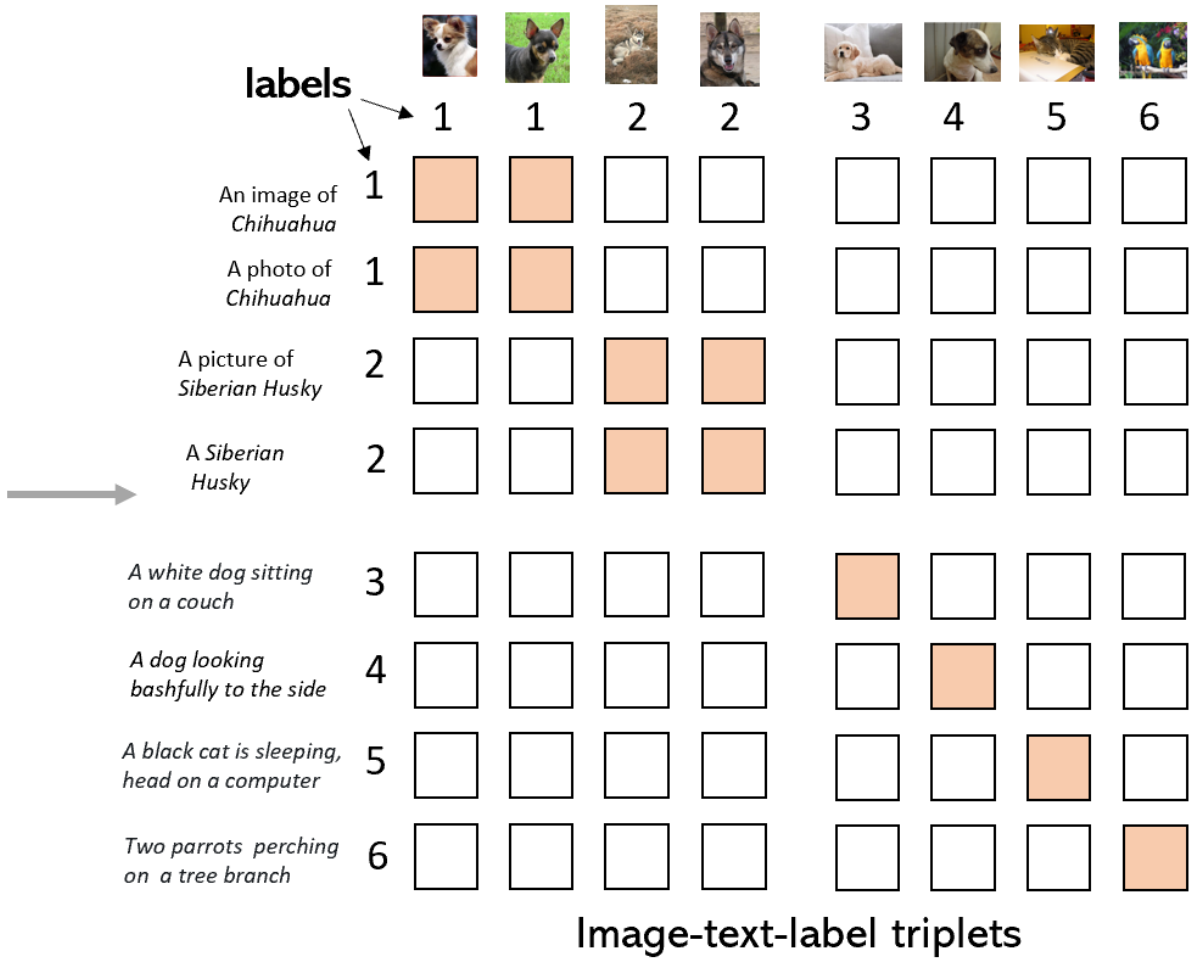
Language-Image-Label Learning

Best of both world:

Discriminative visual representations and Broad semantic coverage

Learning Objectives

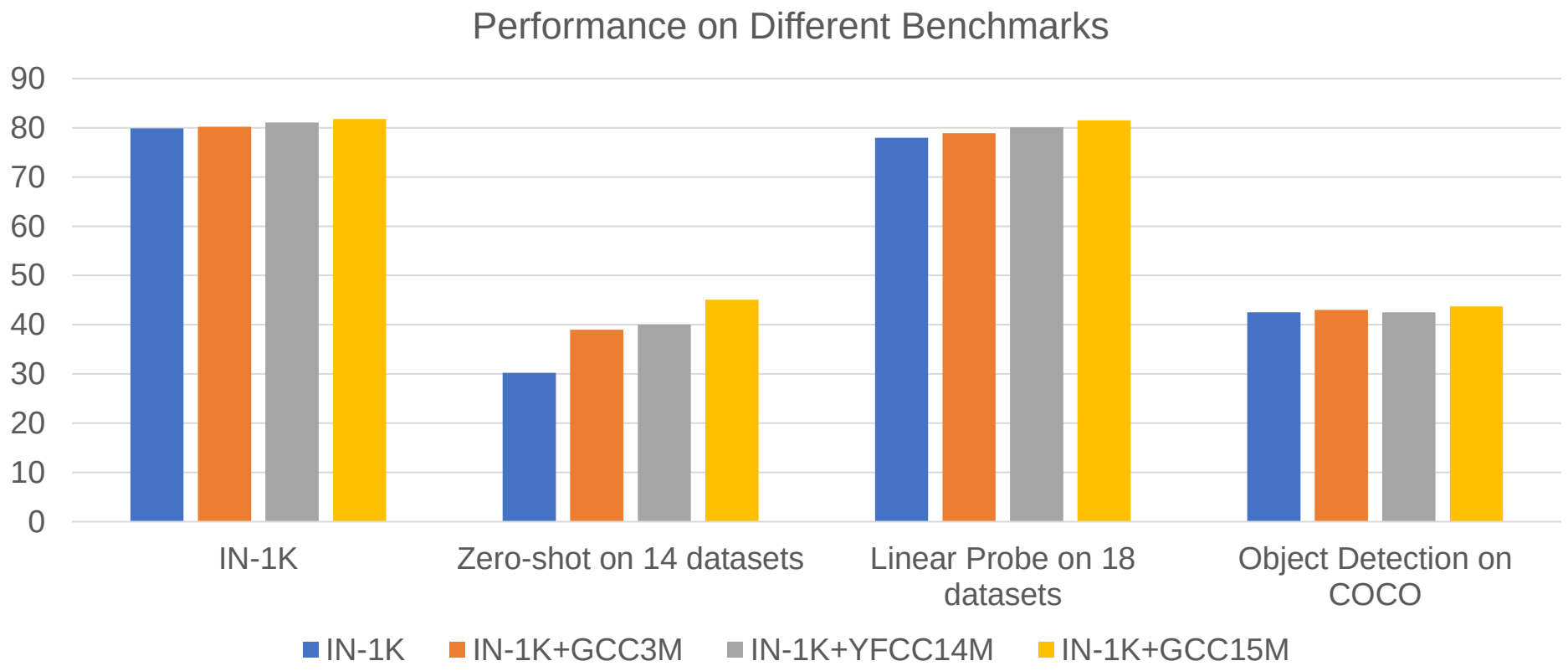




UniCL. Yang *et al.*

Learning Objectives

How image-text pairs benefits image classification on ImageNet-1K?

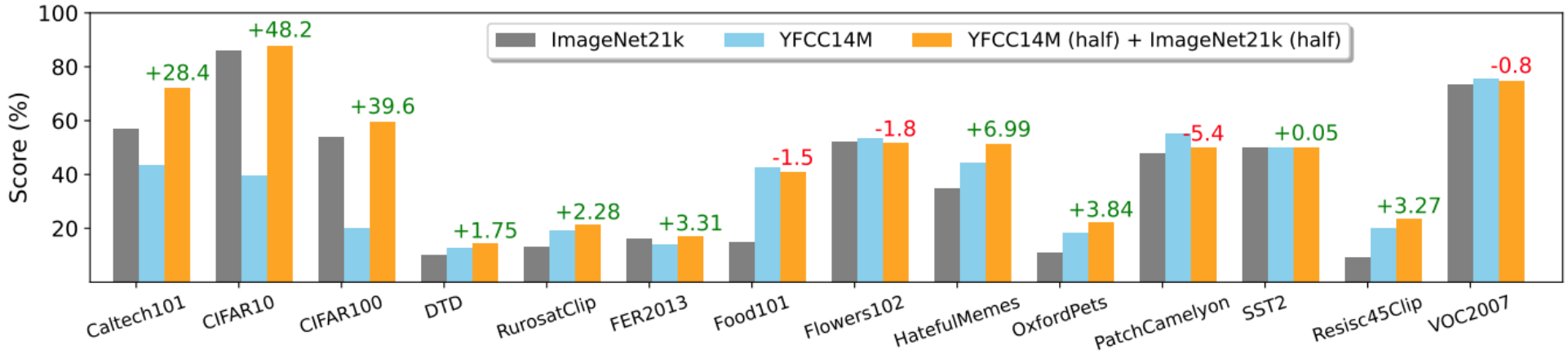


Adding image-text pairs particularly improve the zero-shot performance on various benchmarks

Learning Objectives

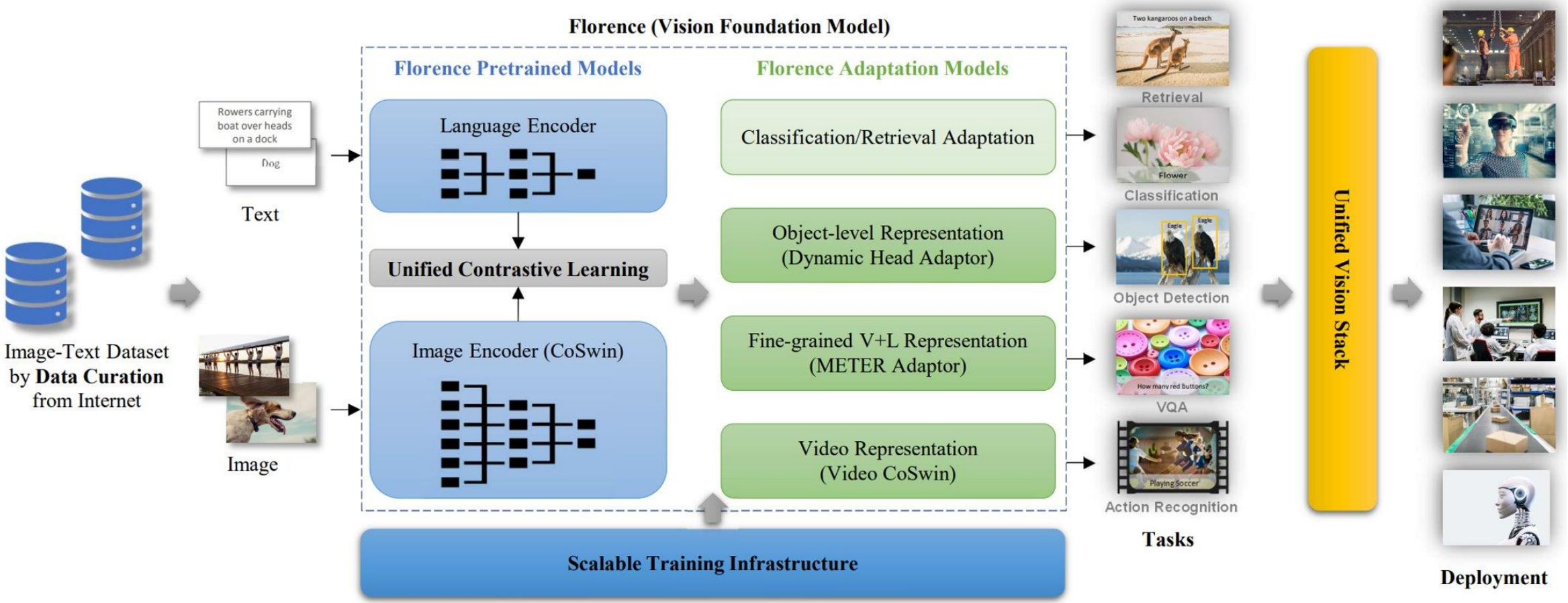
How image-label data benefits image-text pairs on low-shot recognition?

Training Data	Method	Metric			
		Zero-shot		ImageNet-1K Finetuning	Linear Probing 18 datasets
		ImageNet-1K	14 datasets		
YFCC-14M	CLIP	30.1	36.3	77.5	72.7
ImageNet-21K	UniCL	28.5	37.8	78.8	80.5
YFCC-14M(half) + ImageNet-21K(half)	UniCL	36.4	45.5	79.0	80.0
YFCC-14M + ImageNet-21K	UniCL	**40.5**	**49.1**	**80.2**	**81.6**



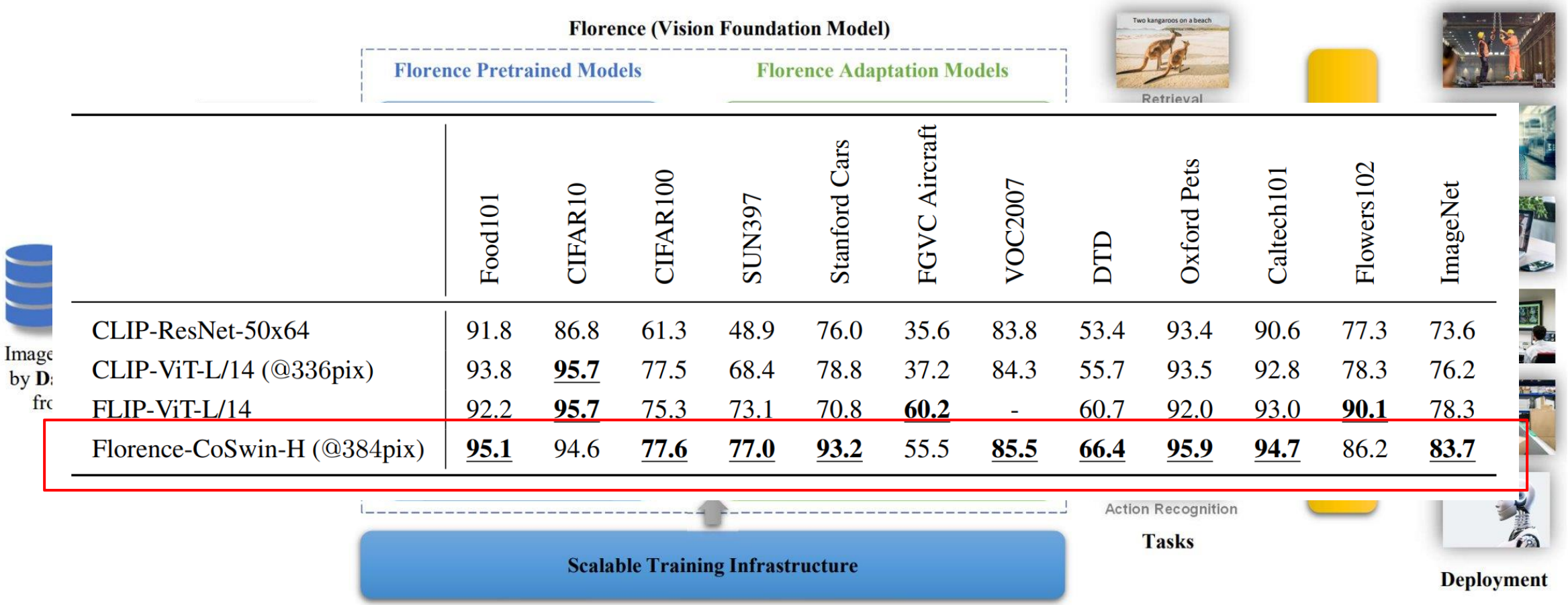
Adding image-label data to image-text pairs can significantly improve the zero-shot, few-shot recognition

Scaling Up



Florence. Yuan *et al.*

Scaling Up



	Food101	CIFAR10	CIFAR100	SUN397	Stanford Cars	FGVC Aircraft	VOC2007	DTD	Oxford Pets	Caltech101	Flowers102	ImageNet
CLIP-ResNet-50x64	91.8	86.8	61.3	48.9	76.0	35.6	83.8	53.4	93.4	90.6	77.3	73.6
CLIP-ViT-L/14 (@336pix)	93.8	**95.7**	77.5	68.4	78.8	37.2	84.3	55.7	93.5	92.8	78.3	76.2
FLIP-ViT-L/14	92.2	**95.7**	75.3	73.1	70.8	**60.2**	-	60.7	92.0	93.0	**90.1**	78.3
Florence-CoSwin-H (@384pix)	**95.1**	94.6	**77.6**	**77.0**	**93.2**	55.5	**85.5**	**66.4**	**95.9**	**94.7**	86.2	**83.7**

More at CVPR Keynote!

Florence. Yuan *et al.*

Learning Objectives

Combining contrastive learning with other learning objectives

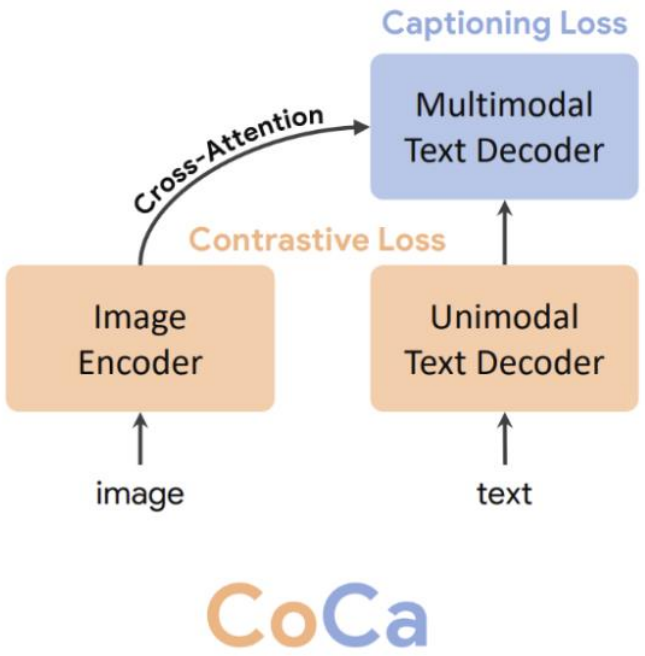
Contrastive vision-language learning

+

Generative Learning

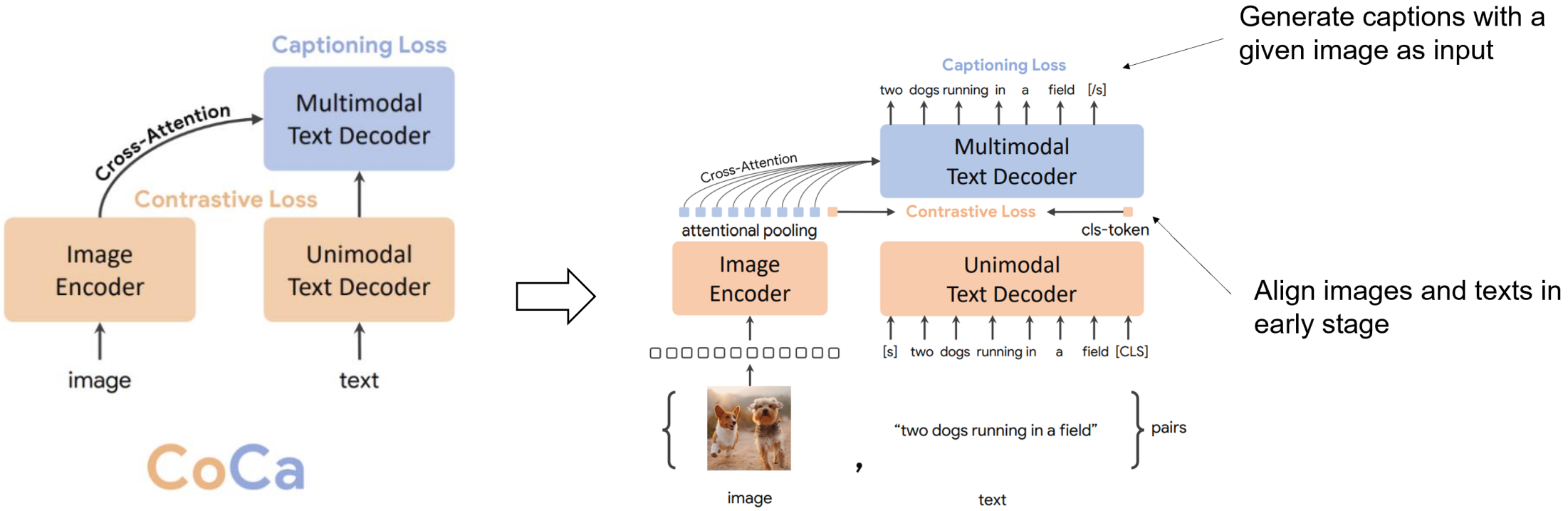
= ?

Learning Objectives



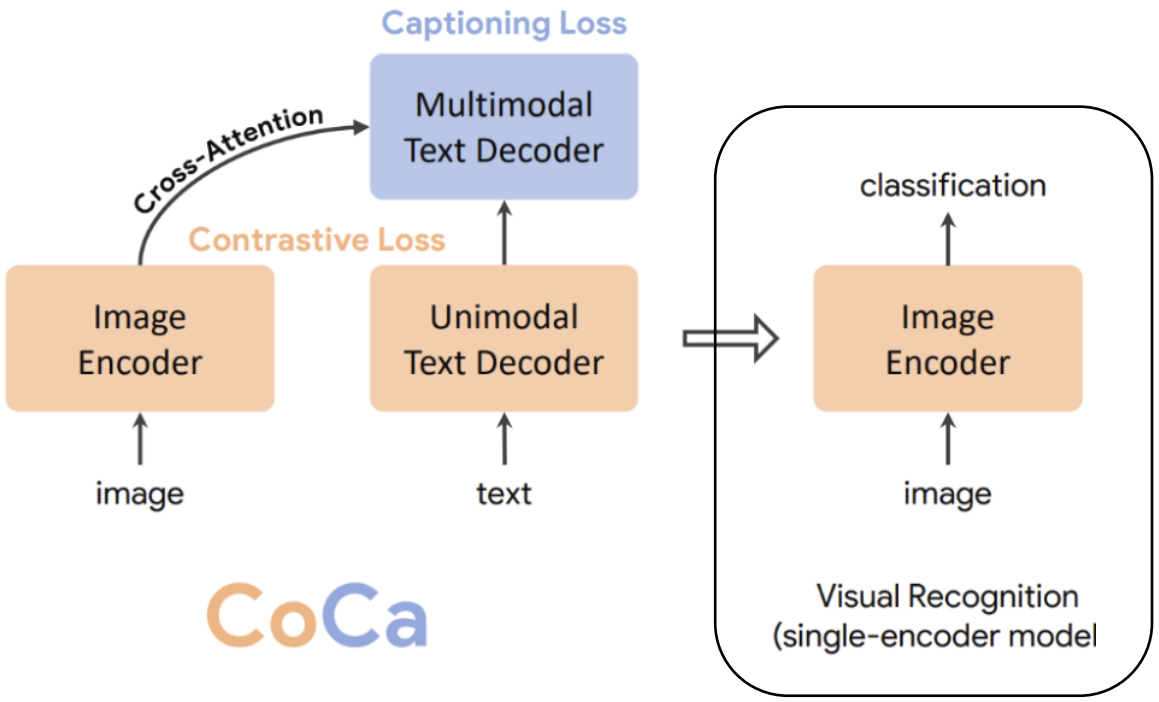
1. Training image encoder, unimodal text decoder, multimodal text decoder simultaneously
2. Training on both image-text pairs from ALIGN and JFT-3B form BigTransfer

Learning Objectives



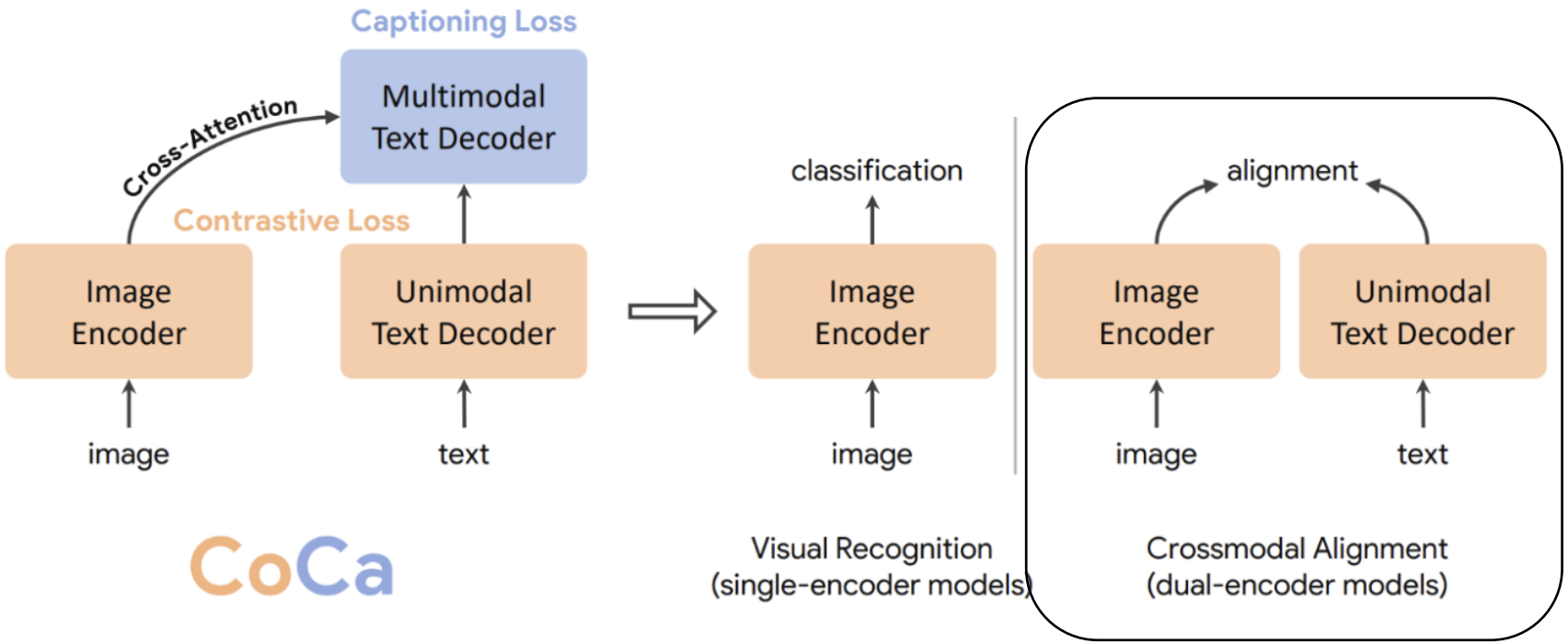
1. Training image encoder, unimodal text decoder, multimodal text decoder simultaneously
2. Training on both image-text pairs from ALIGN and JFT-3B form BigTransfer

Learning Objectives



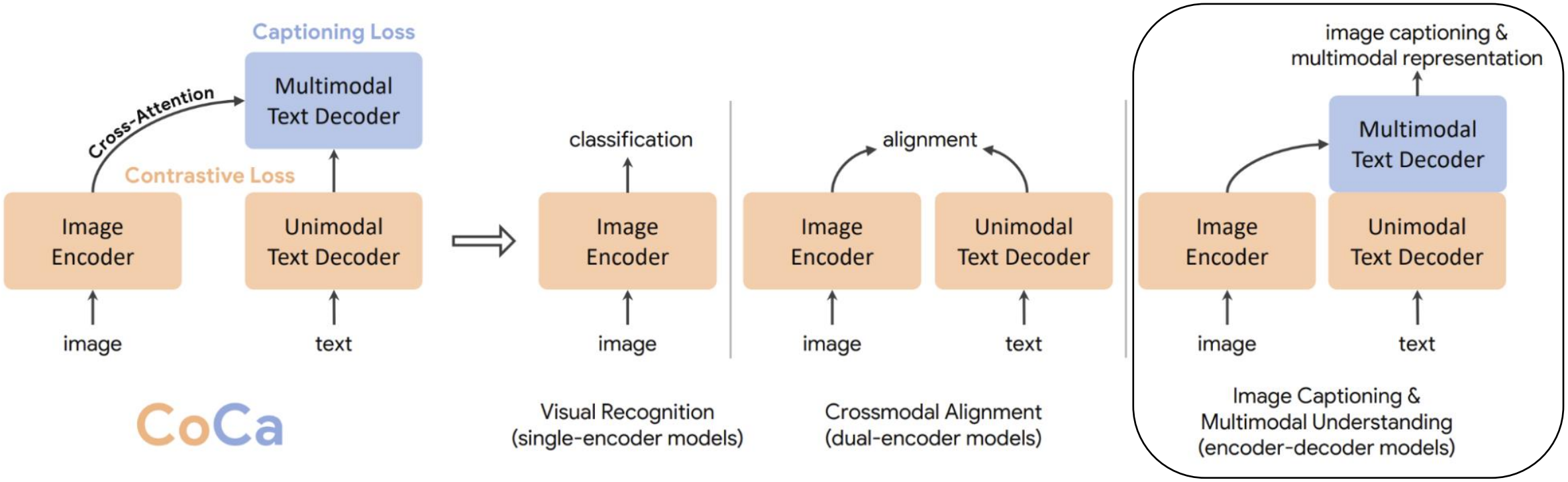
1. Training image encoder, unimodal text decoder, multimodal text decoder simultaneously
2. Training on both image-text pairs from ALIGN and JFT-3B form BigTransfer

Learning Objectives



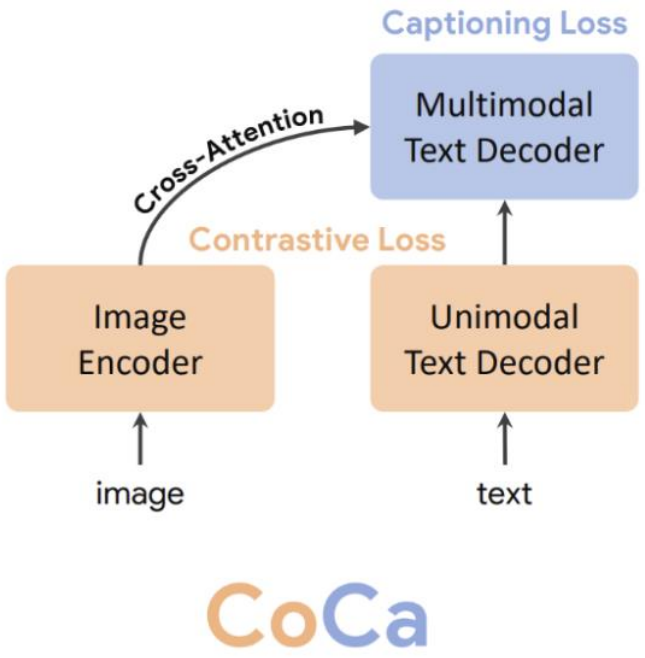
1. Training image encoder, unimodal text decoder, multimodal text decoder simultaneously
2. Training on both image-text pairs from ALIGN and JFT-3B form BigTransfer

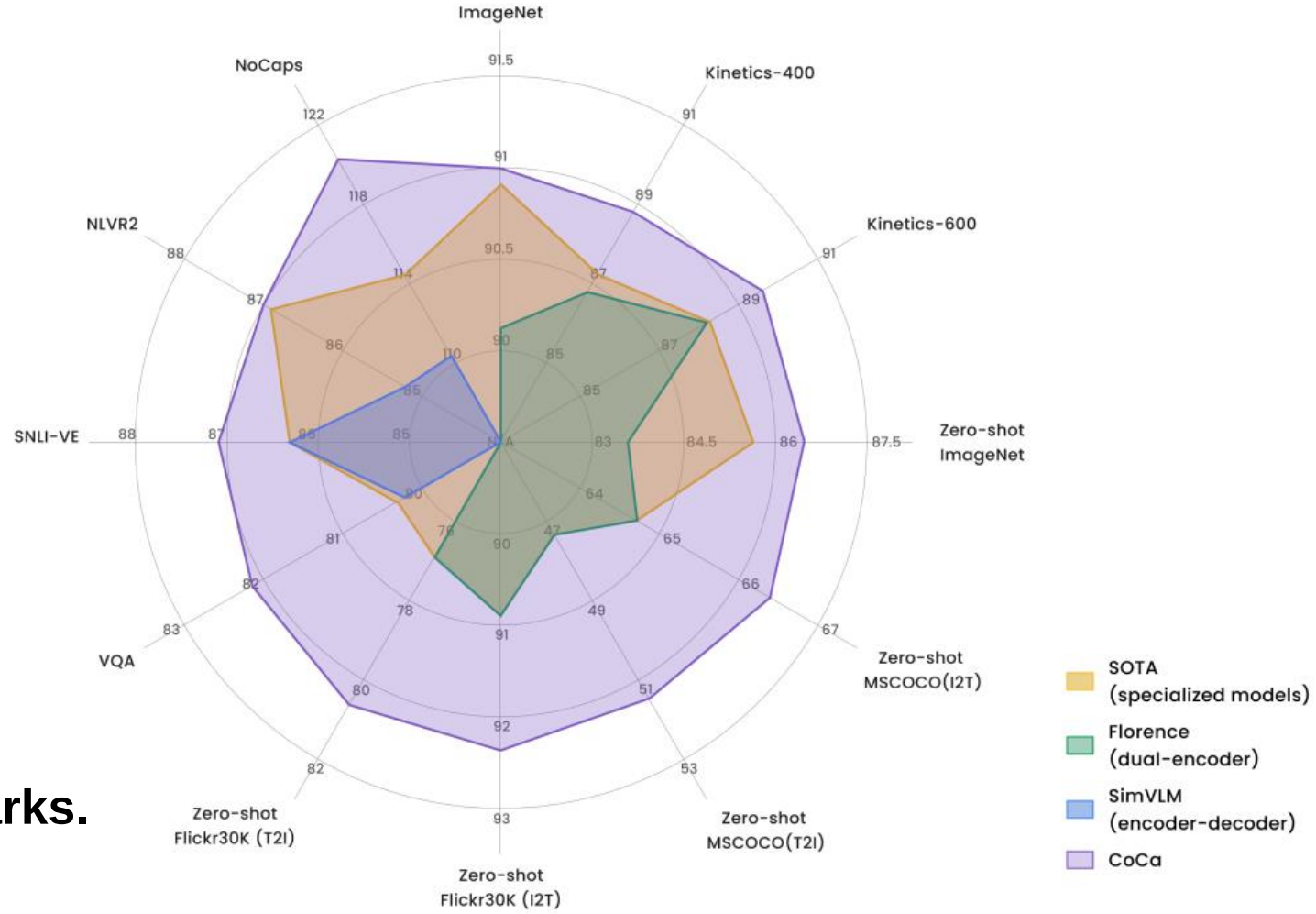
Learning Objectives



1. Training image encoder, unimodal text decoder, multimodal text decoder simultaneously
2. Training on both image-text pairs from ALIGN and JFT-3B form BigTransfer

Learning Objectives





CoCa achieved new SoTA on various image and video recognition benchmarks.

Why contrastive vision-language learning works well?

What kind of data the model is learning from?

Dataset	#Images	#Concepts	Vocab. Size	#Img/C.
CIFAR-10 [34]	50k	10	10	5000
CIFAR-100 [34]	50k	100	105	500
ImageNet-1K [10]	1.3M	1,000	1,233	1300
ImageNet-22k [10]	14.2M	21,841	14,733	650
GCC-3M [51]	3.3M	17,135	7,953	193
GCC-12M [5]	12M	584,261	98,347	21
YFCC-14M [56]	14M	650,236	214,380	22

1. Image classification data has limited concepts but dense annotations
2. Image-text pairs cover a huge amount of concepts, but relative sparse

Why contrastive vision-language learning works so well?

Concept coverage on downstream classification data for each pretraining data

Observations: pretraining corpus covers many concepts in the downstream datasets

Dataset			ImageNet-1K		ImageNet-21K		GCC-3M		GCC-12M		YFCC-14M	
Name	#Concepts	Vocab. Size	Cover.	#Img/C.	Cover.	#Img/C.	Cover.	#Img/C.	Cover.	#Img/C.	Cover.	#Img/C.
ImageNet-1K	1,000	1,233	**100%**	1300	0%	0	45.3%	247.0	**78.5%**	851.1	**69.3%**	1930.8
Food-101	102	139	4.0%	1300.0	20.8%	650.0	21.8%	39.8	**58.4%**	250.8	**67.3%**	408.8
CIFAR-10	10	10	0.0%	0.0	**90.0%**	650.0	**100.0%**	6175.4	**100.0%**	19969.8	**100.0%**	32998.9
CIFAR-100	100	100	24.0%	1300.0	**65.0%**	650.0	**95.0%**	3928.4	**99.0%**	15628.5	**99.0%**	18303.2
SUN397	397	432	5.0%	1300.0	28.5%	650.0	48.1%	818.9	**65.5%**	2355.4	**66.5%**	7043.2
Stanford Cars	196	291	0.0%	0.0	0.0%	0.0	0.0%	0.0	0.0%	0.0	0.0%	0.0
FGVC Aircraft (variants)	100	115	0.0%	0.0	0.0%	0.0	0.0%	0.0	22.0%	4.1	0.0%	0.0
VOC2007 classification	20	20	0.0%	0.0	75.0%	650.0	**85.0%**	14721.6	**85.0%**	19934.8	**85.0%**	31448.8
Describable Textures	47	47	0.0%	0.0	4.3%	650.0	14.9%	8.9	27.7%	53.2	36.2%	181.7
Oxford-IIIT Pets	37	53	5.4%	1300.0	13.5%	650.0	10.8%	80.9	**64.9%**	134.0	37.8%	169.0
Caltech-101	102	122	24.5%	1300.0	43.1%	650.0	**66.6%**	1633.8	**83.3%**	5249.7	**87.3%**	5017.7
Oxford Flowers 102	102	147	10.0%	1300.0	40.2%	650.0	17.6%	53.2	**50.0%**	194.3	**65.7%**	422.7
MNIST	10	10	0.0%	0.0	0.0%	0.0	40.0%	0.8	**100.0%**	46.0	**90.0%**	68.8
FER 2013 *	8	12	0.0%	0.0	8.3%	650.0	25.0%	5.9	41.7%	29.2	41.7%	11.5
STL10	10	10	0.0%	0.0	100%	650.0	**100.0%**	8778.6	**100.0%**	28547.6	**100.0%**	45587.5
GTSRB *	43	85	0.0%	0.0	0.0%	0.0	2.3%	12.7	2.3%	52.9	2.3%	551.3
PatchCamelyon	2	6	0.0%	0.0	0.0%	0.0	0.0%	0.0	**50.0%**	143.0	**50.0%**	15.0
UCF101 *	101	153	0.0%	0.0	0.0%	0.0	0.0%	0.0	**51.5%**	66.4	0.0%	0.0
Hateful Memes	2	2	0.0%	0.0	0.0%	0.0	**50.0%**	79.5	**50.0%**	2742.5	**50.0%**	321.5
EuroSAT	10	20	0.0%	0.0	0.0%	0.0	20.0%	2946.6	30.0%	5266.3	30.0%	15458.7
Resisc45	45	59	8.9%	1300.0	26.7%	650.0	**71.1%**	3688.6	**75.6%**	7572.0	**80.0%**	26317.6
Rendered-SST2	2	2	0.0%	0.0	**50.0%**	650.0	**50.0%**	1.0	**100.0%**	114.0	**100.0%**	1259.0

Why contrastive vision-language learning works so well?

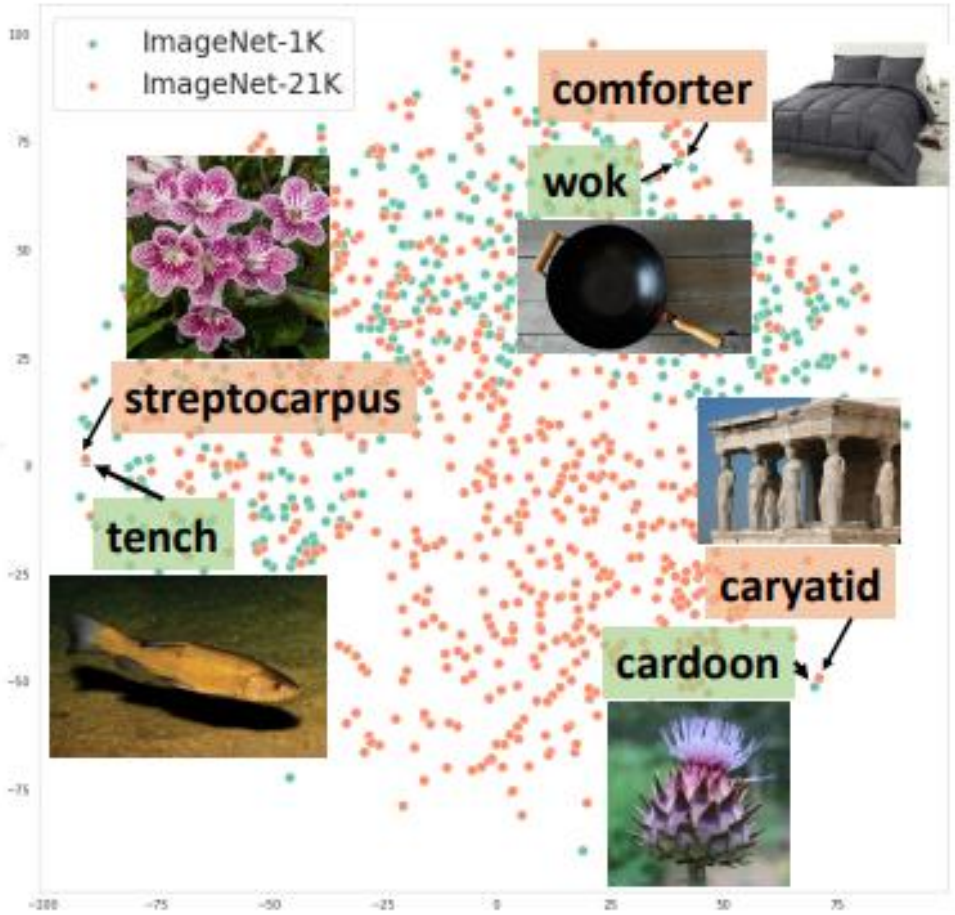
"Zero-shot" is not a real zero-shot

A realistic way to learn!

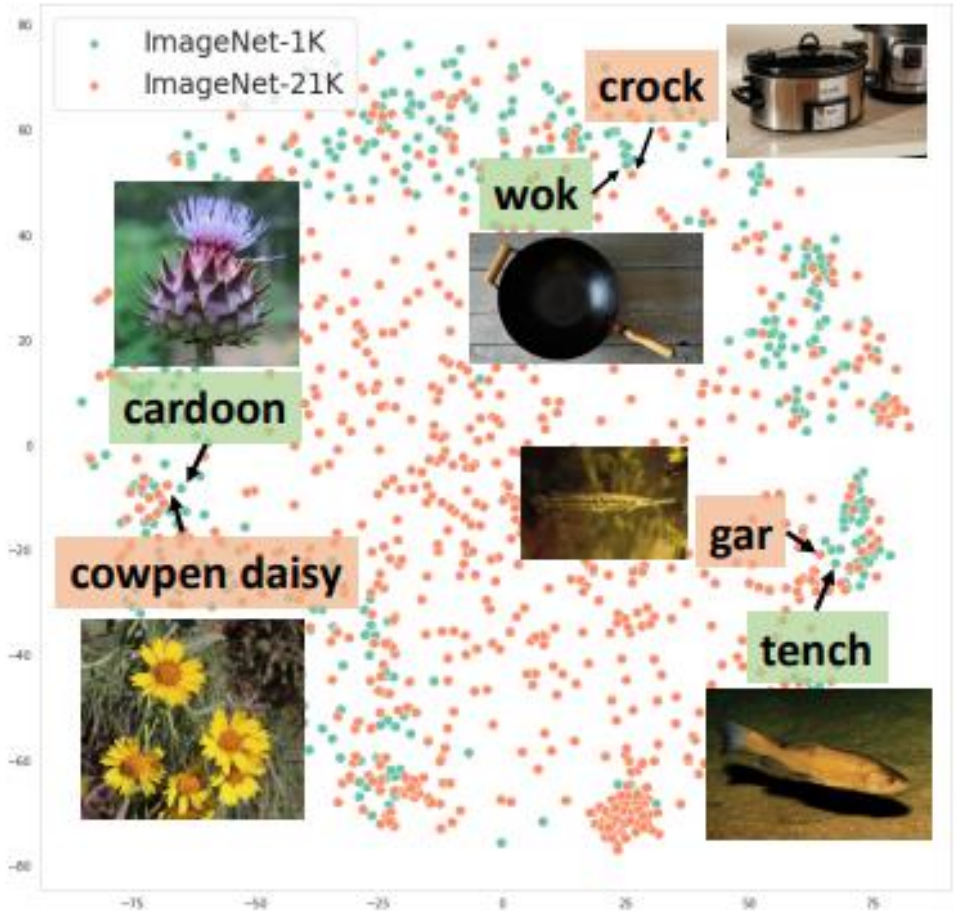
Dataset			ImageNet-1K		ImageNet-21K		GCC-3M		GCC-12M		YFCC-14M	
Name	#Concepts	Vocab. Size	Cover.	#Img/C.	Cover.	#Img/C.	Cover.	#Img/C.	Cover.	#Img/C.	Cover.	#Img/C.
ImageNet-1K	1,000	1,233	**100%**	1300	0%	0	45.3%	247.0	**78.5%**	851.1	**69.3%**	1930.8
Food-101	102	139	4.0%	1300.0	20.8%	650.0	21.8%	39.8	**58.4%**	250.8	**67.3%**	408.8
CIFAR-10	10	10	0.0%	0.0	**90.0%**	650.0	**100.0%**	6175.4	**100.0%**	19969.8	**100.0%**	32998.9
CIFAR-100	100	100	24.0%	1300.0	**65.0%**	650.0	**95.0%**	3928.4	**99.0%**	15628.5	**99.0%**	18303.2
SUN397	397	432	5.0%	1300.0	28.5%	650.0	48.1%	818.9	**65.5%**	2355.4	**66.5%**	7043.2
Stanford Cars	196	291	0.0%	0.0	0.0%	0.0	0.0%	0.0	0.0%	0.0	0.0%	0.0
FGVC Aircraft (variants)	100	115	0.0%	0.0	0.0%	0.0	0.0%	0.0	22.0%	4.1	0.0%	0.0
VOC2007 classification	20	20	0.0%	0.0	75.0%	650.0	**85.0%**	14721.6	**85.0%**	19934.8	**85.0%**	31448.8
Describable Textures	47	47	0.0%	0.0	4.3%	650.0	14.9%	8.9	27.7%	53.2	36.2%	181.7
Oxford-IIIT Pets	37	53	5.4%	1300.0	13.5%	650.0	10.8%	80.9	**64.9%**	134.0	37.8%	169.0
Caltech-101	102	122	24.5%	1300.0	43.1%	650.0	**66.6%**	1633.8	**83.3%**	5249.7	**87.3%**	5017.7
Oxford Flowers 102	102	147	10.0%	1300.0	40.2%	650.0	17.6%	53.2	**50.0%**	194.3	**65.7%**	422.7
MNIST	10	10	0.0%	0.0	0.0%	0.0	40.0%	0.8	**100.0%**	46.0	**90.0%**	68.8
FER 2013 *	8	12	0.0%	0.0	8.3%	650.0	25.0%	5.9	41.7%	29.2	41.7%	11.5
STL10	10	10	0.0%	0.0	100%	650.0	**100.0%**	8778.6	**100.0%**	28547.6	**100.0%**	45587.5
GTSRB *	43	85	0.0%	0.0	0.0%	0.0	2.3%	12.7	2.3%	52.9	2.3%	551.3
PatchCamelyon	2	6	0.0%	0.0	0.0%	0.0	0.0%	0.0	**50.0%**	143.0	**50.0%**	15.0
UCF101 *	101	153	0.0%	0.0	0.0%	0.0	0.0%	0.0	**51.5%**	66.4	0.0%	0.0
Hateful Memes	2	2	0.0%	0.0	0.0%	0.0	**50.0%**	79.5	**50.0%**	2742.5	**50.0%**	321.5
EuroSAT	10	20	0.0%	0.0	0.0%	0.0	20.0%	2946.6	30.0%	5266.3	30.0%	15458.7
Resisc45	45	59	8.9%	1300.0	26.7%	650.0	**71.1%**	3688.6	**75.6%**	7572.0	**80.0%**	26317.6
Rendered-SST2	2	2	0.0%	0.0	**50.0%**	650.0	**50.0%**	1.0	**100.0%**	114.0	**100.0%**	1259.0

Visualizations

Visualization of learned text concept embeddings in IN-1K and IN-21K



Left: learning purely on imagenet-1k

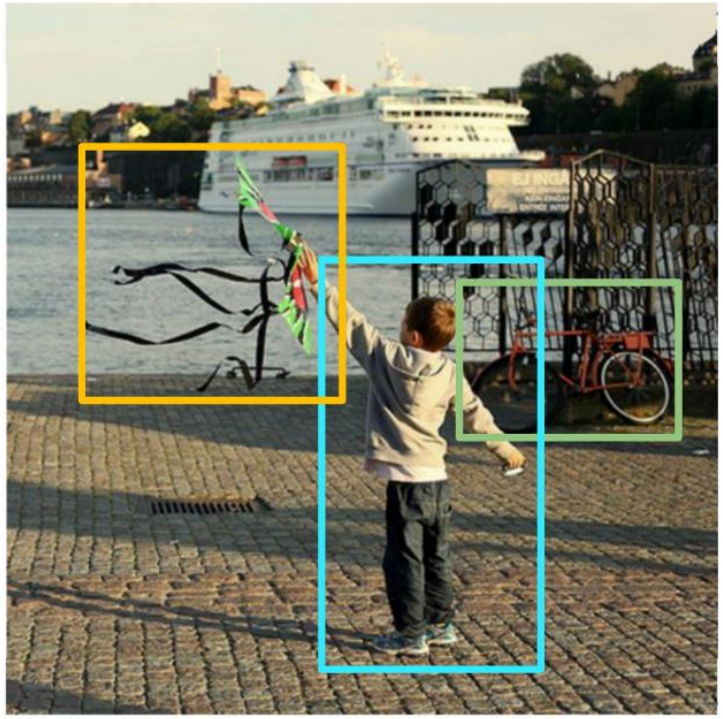


Right: learning on imagenet-1k plus image-text pairs

Take-aways

- Visual recognition is essentially a vision-language problem, i.e., aligning images with semantic concepts/descriptions.
- Large-scale image-text pairs can help model learn the understanding of rich semantics
- Contrastive vision-language learning can be combined with other learning methods, such as self-supervised, supervised or even generative learning.

Visual Region Recognition

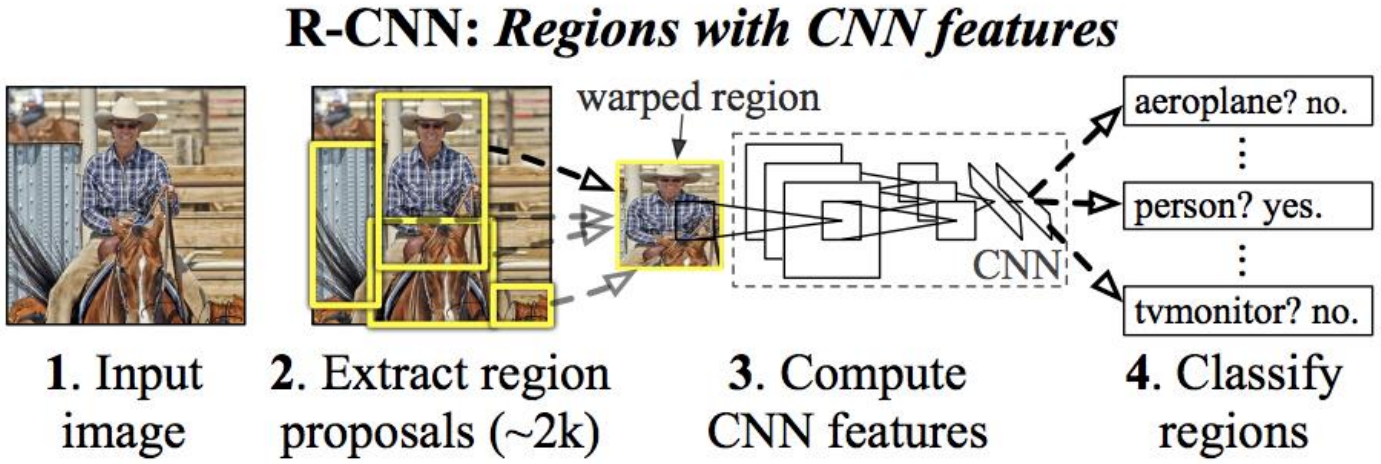


Objects: boy, kite, bicycle

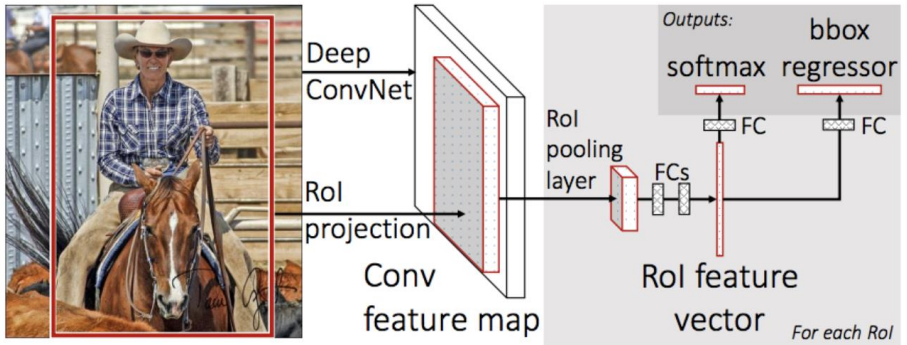
- Attributes: little, green, orange
- Relationships: flying, near
- Caption: “A little boy is flying a green kite near an orange bicycle.”
- Question: What is the boy next to the bicycle doing?

Objects play a major role in visual understanding.

Visual Region Recognition



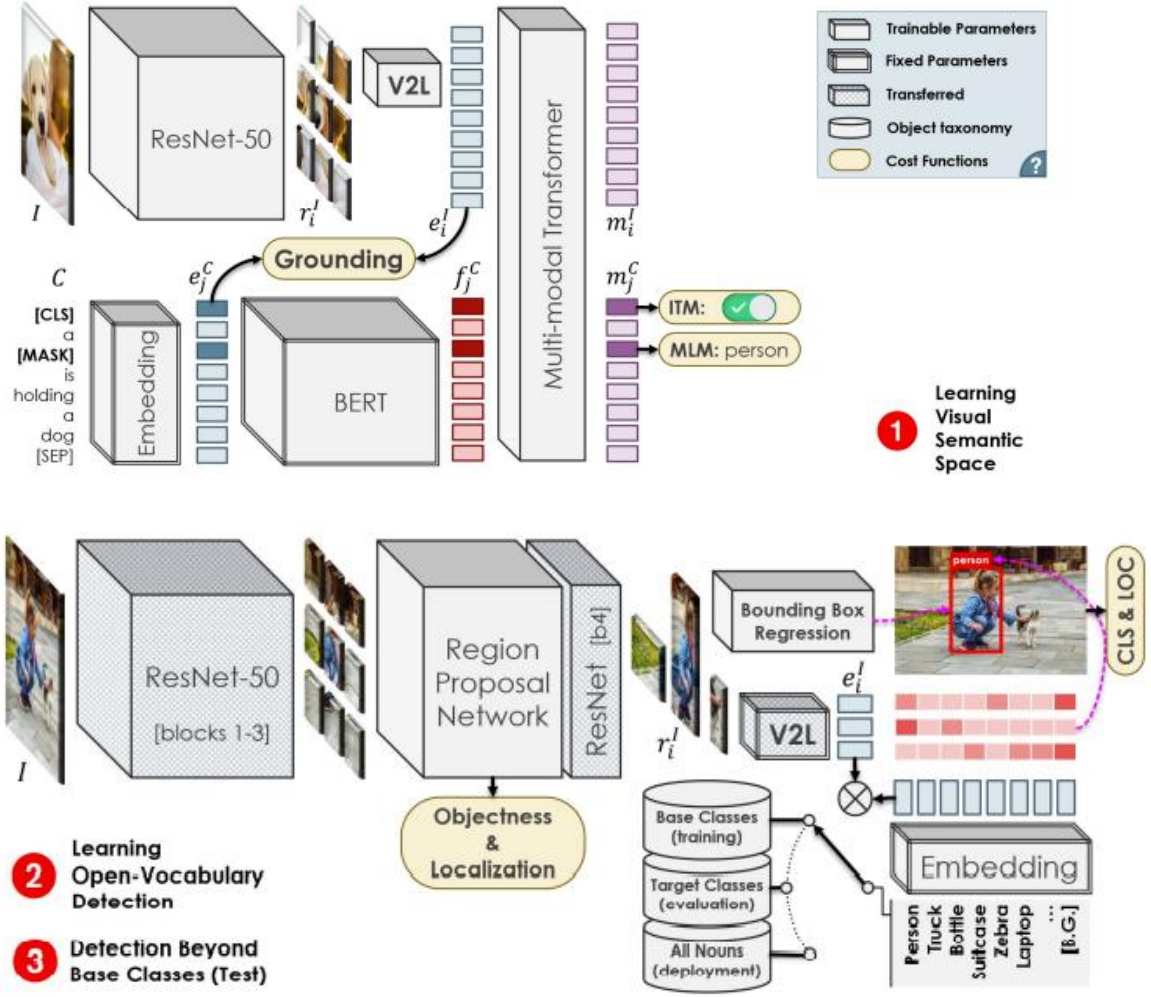
R-CNN. Girshick *et al.*



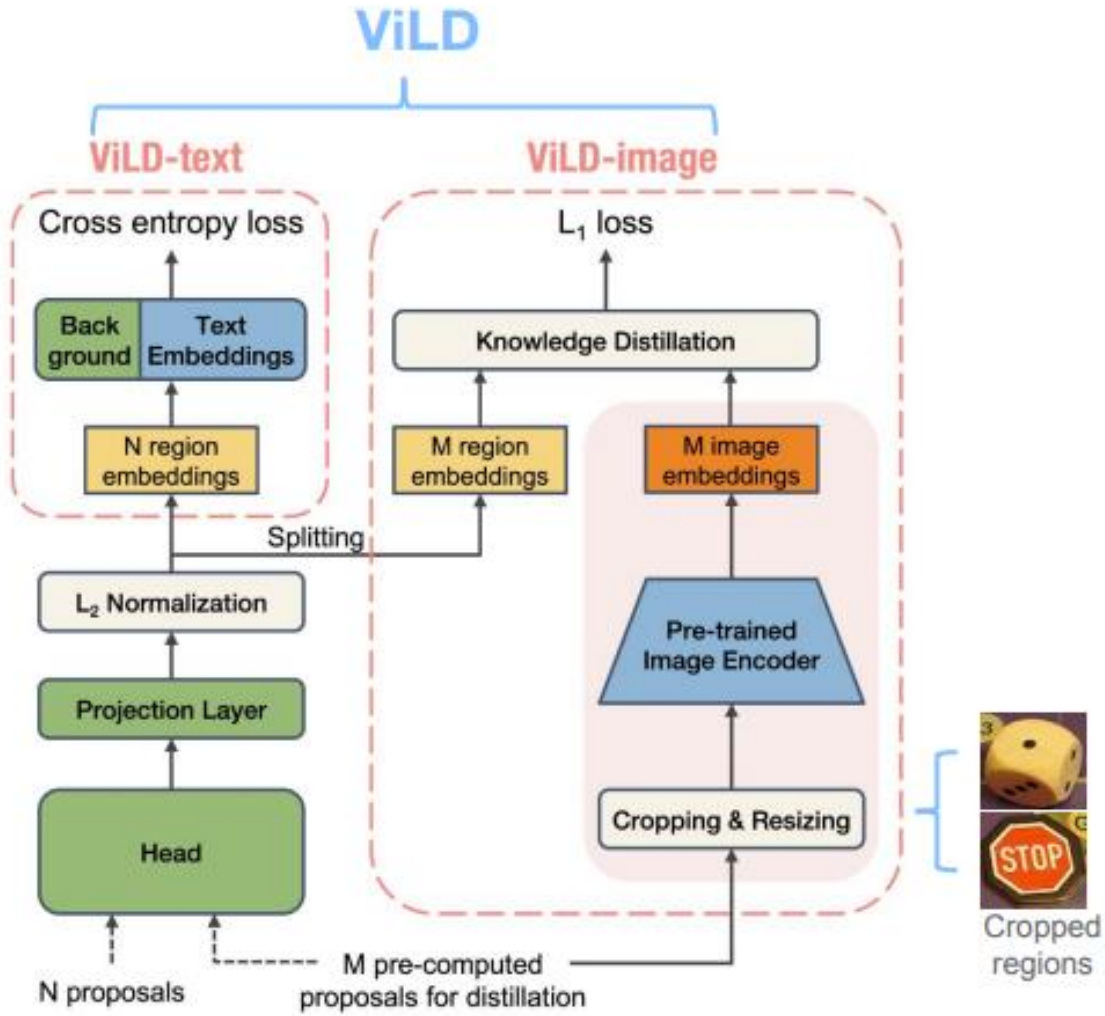
Fast R-CNN. Girshick *et al.*

Can we localize the regions and then recognize them with vision-language models?

Open-Vocabulary R-CNN

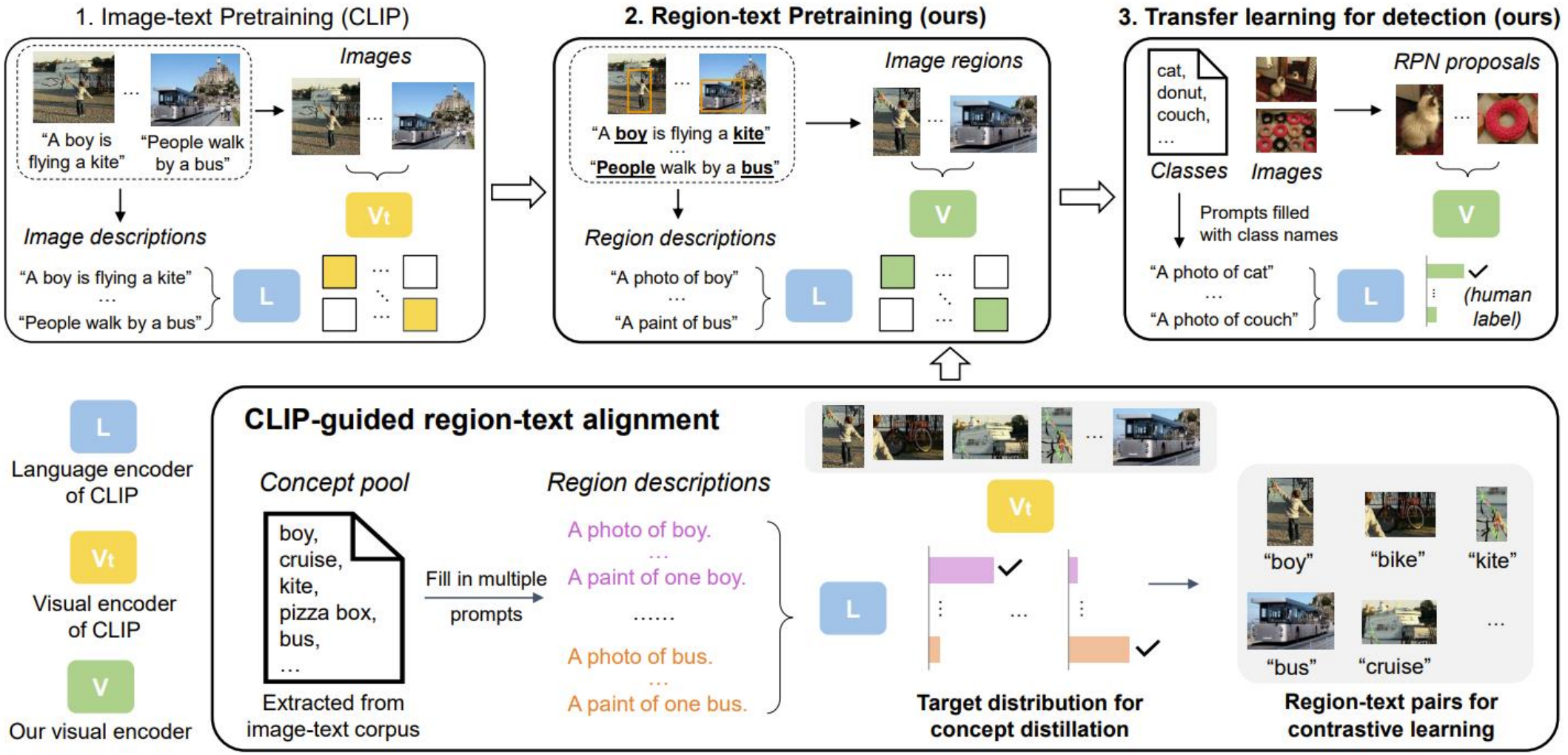


OVR-CNN. Zareian *et al.*



ViLD. Gu *et al.*

Open-Vocabulary R-CNN



RegionCLIP. Zhong *et al.*

Conclusion

- **Summary**
  - Vision-language learning can help vision significantly, including image-level and region-level vision recognition.
  - Vision-language learning leverage image-text pairs, but can goes beyond by integrating supervised learning, self-supervised learning, generative learning, etc.
  - Vision-language learning can serves as a generic learning paradigm for zero-shot, few-shot, transfer learning.

- **Challenge**
  - How to learn from the large-scale image-text pairs more efficiently?
  - How to learn more discriminative visual representations for downstream tasks?
  - How to enable fine-grained understanding of visual contents?

Thanks for your Attention!