Lecture 5: Vision-Language Representation Learners

Announcements

- A1 has been released, due Wed Oct 16 at 11:59pm.
- Looking ahead, project proposal will be due Wed Oct 23 at 11:59pm.

Course project

- Goal is to gain hands-on experience interacting with some of the large vision and vision-language models discussed in class
- Will have proposal, milestone, final report + presentation
- Can work individually or in teams of 2, grades will be calibrated by group size
- Policy for working on a project related to that of another class project:
  - “Your project may be related to that of another class project as long as permission is granted by instructors of both classes; however, you must clearly indicate in the project proposal, milestone, and final reports the exact portion of the project that is being counted for this course. In this case, you must prepare separate reports for each course, and submit your final report for the other course as well.”

Project options: Comparative analysis

Conduct a comparative or “red teaming” analysis of at least two large vision or vision-language models, considering a biomedical use case or motivation. Here, comparative analysis refers to probing the models to analyze where they work well or not, and identifying weaknesses, vulnerabilities, biases, or failure modes, in order to better understand limitations and suggest areas for further research. Your project should include comparative analysis of your selected models, which can be two or more completely different models, or multiple versions from one model family. You may use both biomedical and non-biomedical data, but the project should involve at least some biomedical data.

Project options: Implement an agentic system

Implement an agentic system that incorporates large vision or vision-language models to address a specific biomedical problem. Here, “agentic” refers to developing a system where model “agents” are used (potentially out-of-the-box, or in API-based fashion) in combination to achieve a specific goal. For this project, you must incorporate at least two model agents in your system, but only one must be a large vision or vision-language model (the other could be a language-only model, a specialized object detection model, or a search engine, for example).

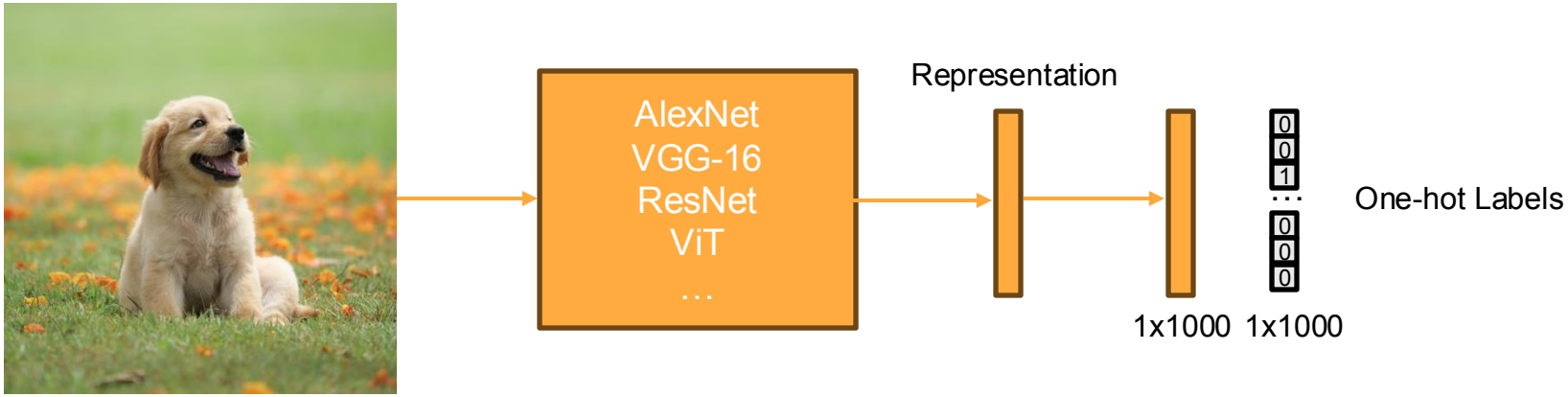
Project options: Explore a technical innovation

Explore and experimentally assess the effectiveness of a novel technical approach or innovation to enhance the capabilities of existing vision or vision-language models, considering a biomedical use case or motivation. For example, this could involve making model or architecture improvements, or curating a new instruction-tuning dataset. While the technical innovation does not need to be biomedicine-specific, at a minimum the potential utility of the innovation for specific biomedical applications must be discussed, and some biomedical data must be used to assess the effectiveness of the innovation.

Course project: more details

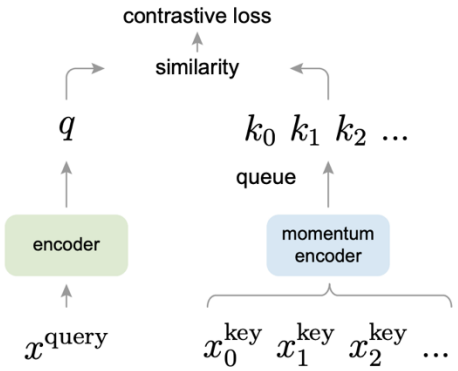
- Detailed project instructions and requirements for each project component will be posted on the class website and announced on Ed within the next day or so.
- Please do not hesitate to ask any questions or clarifications about potential project ideas.
- Not all projects will require GPU compute. For example, red teaming analysis projects can be done without this.
  - If you wish to pursue a project that requires GPU compute but do not have sufficient access, we have arranged with Google Cloud to provide compute credits of $50 / student.
  - It is possible these may be able to refreshed. We will share details as we have them. Please reach out to course staff for any questions or concerns about compute.
- We are also compiling and will share information about other compute resources more broadly in the coming days. If you are aware of other resources that your fellow students may benefit from (both GPUs and APIs), please let us know so that we can share the information.

ImageNet Classification

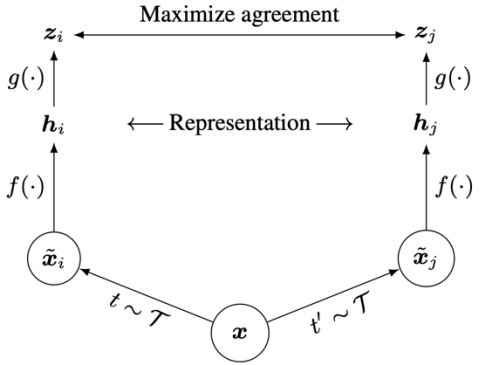


- Models are trained to predict a fixed set of predetermined object categories
- Require crowd-sourced labeling of “gold labels”

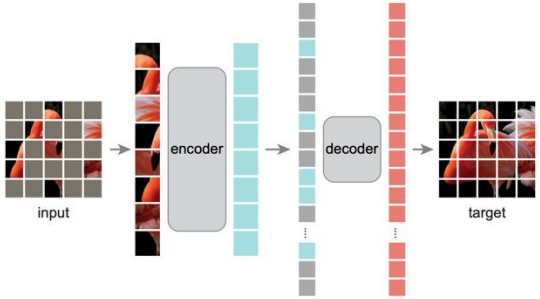
Self-Supervised Learning



MoCo (He et al. 2019)



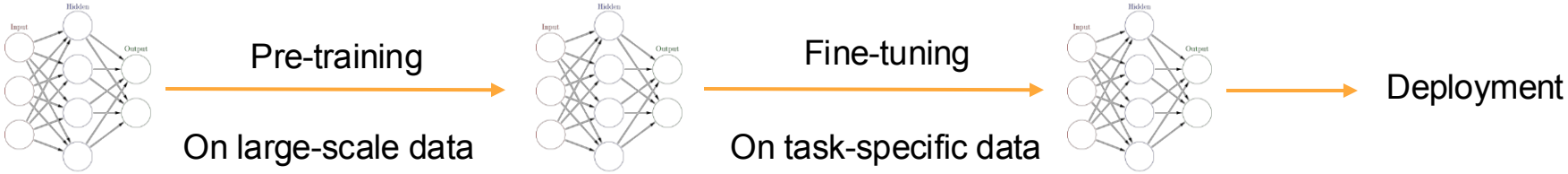
SimCLR (Chen et al. 2020)



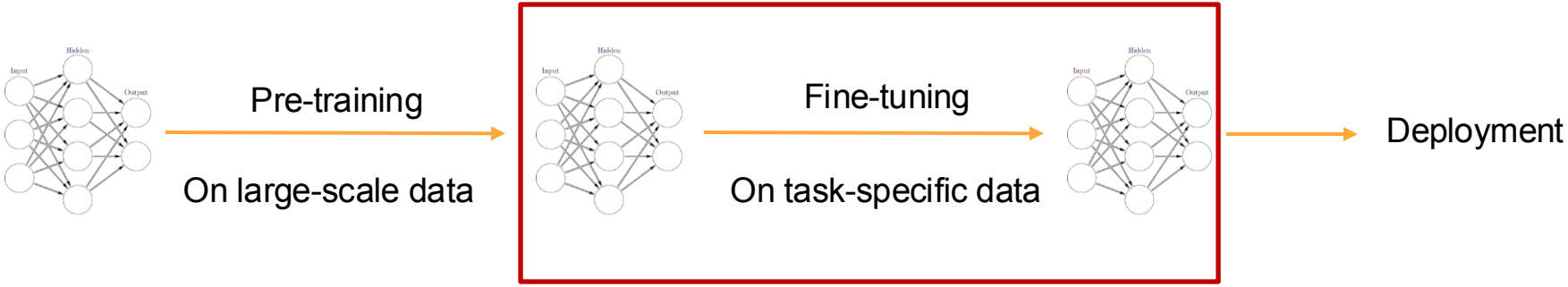
MAE (He et al. 2021)

learning to extract powerful representations without human annotated labels

Adapt to Downstream Task

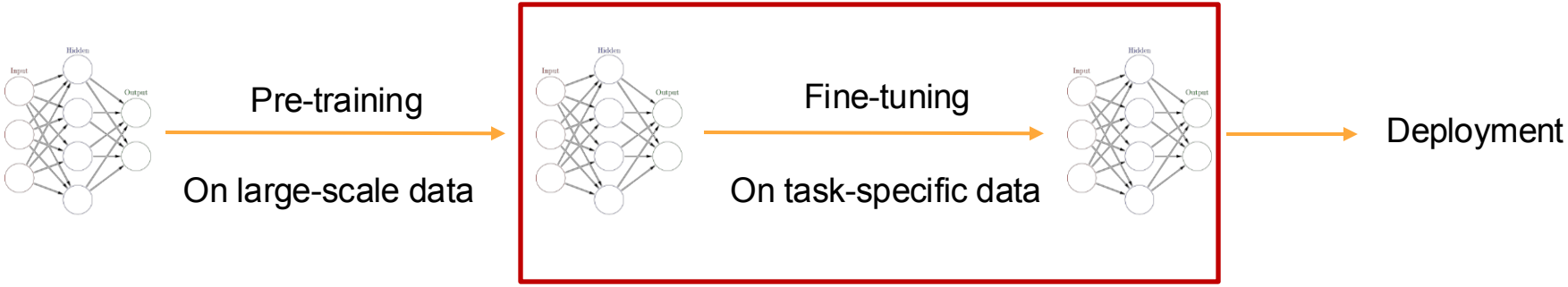


Adapt to Downstream Task



- Require task-specific annotations
- Reduce robustness

Adapt to Downstream Task



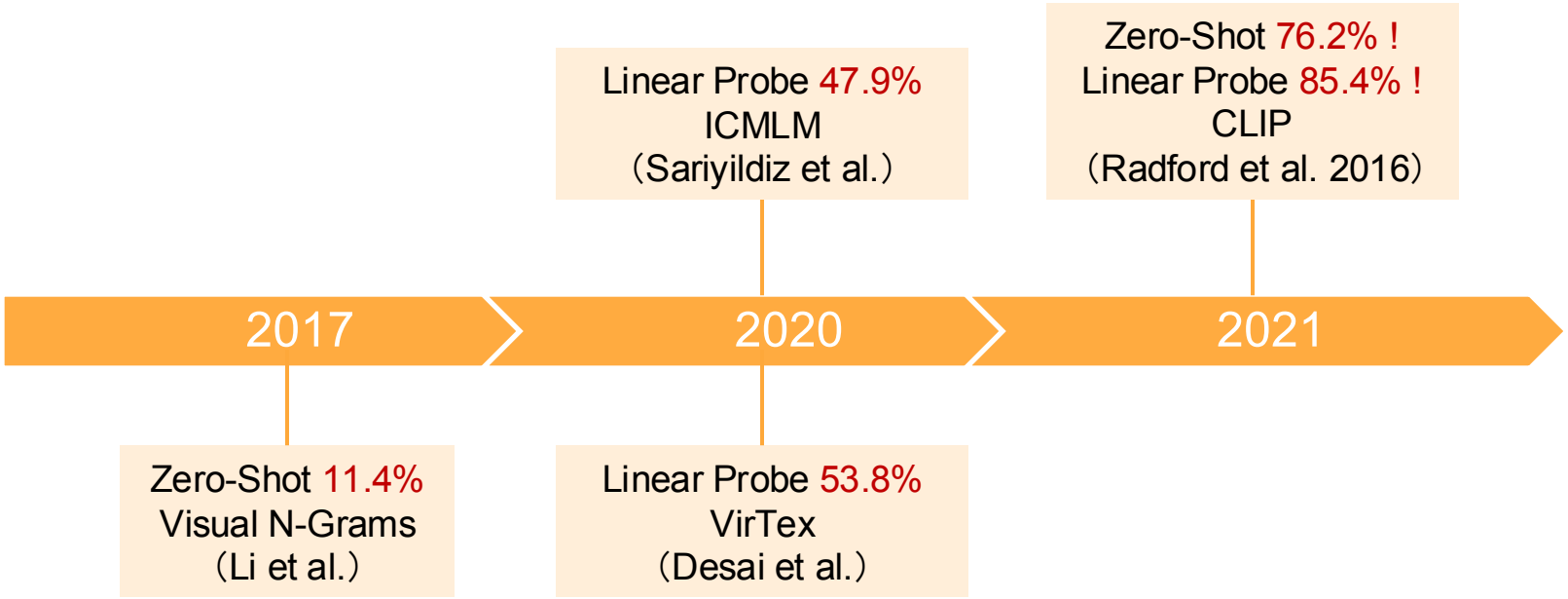
How to skip this step to enable zero-shot generalization?

Natural Language Supervision

- Unlimited Data: vast amount of text-image pairs on the internet

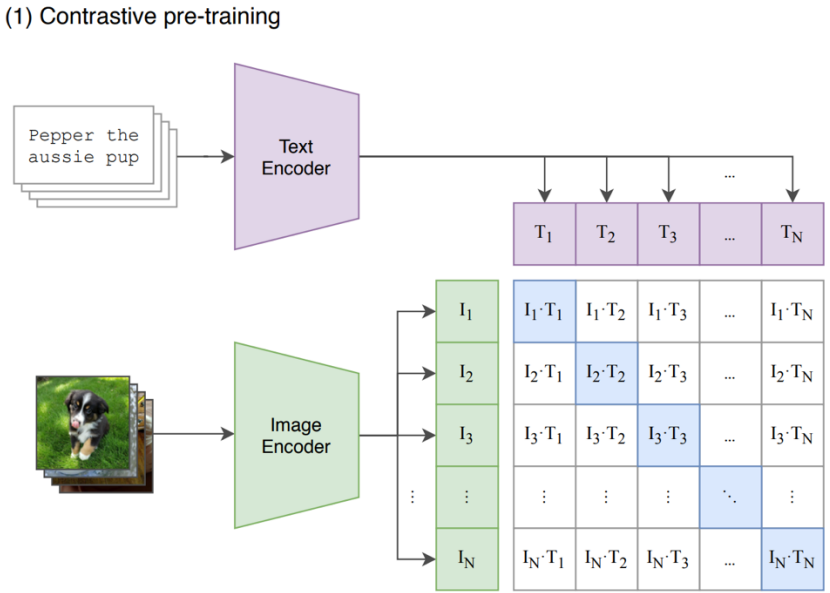
- Flexible zero-shot transfer: learn a representation that connects visual content to language

Natural Language Supervision



Models trained with natural language supervision (Acc on ImageNet)

Overview of CLIP

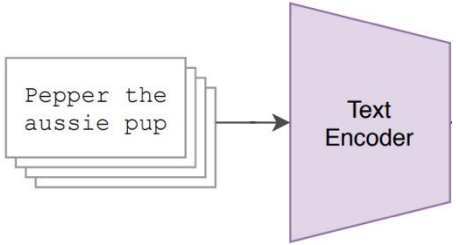


- Data: 400 million image-text pairs collected from the internet
- Model: ResNets and ViTs with up to 300M parameters
- Loss Function: Contrastive Loss
- Compute: Trained on 250 - 600 GPUs for up to 18 days.

CLIP: Contrastive Language-Image Pre-training

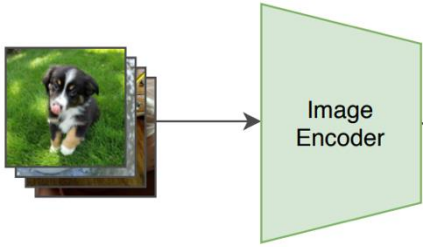
CLIP (Radford et al. 2021)

Encoders for CLIP



- Transformer
- 12- layer 512/768 wide model with 8/12 attention heads
- Max sequence length was capped at 76

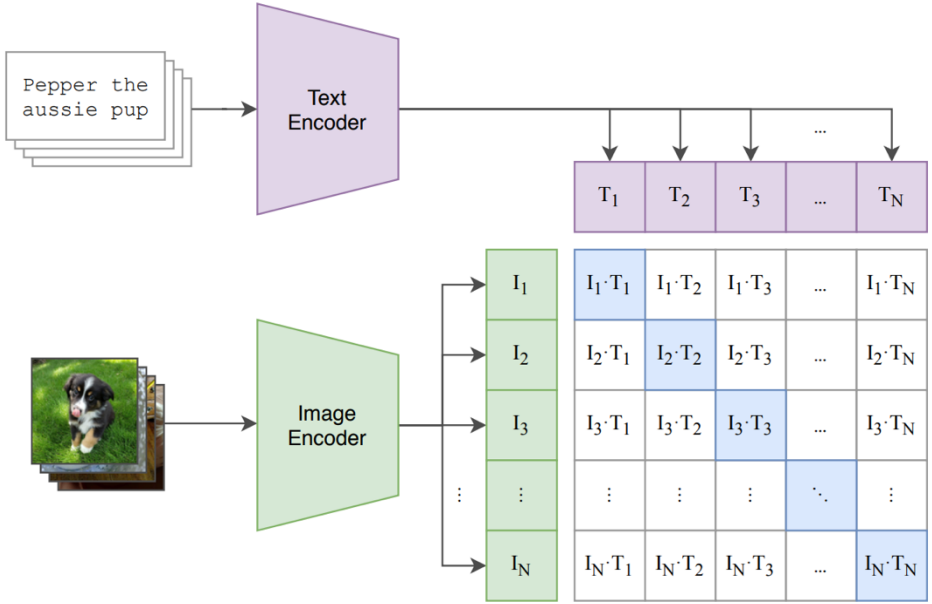
- Weights are trained from scratch



- ResNet-50, ResNet-101, ViT-B/32, ViT-B/16, ViT-L/14
- Pre-train at a higher resolution for one additional epoch ViT-L/14@336px (Best Model)
- 307M parameters

CLIP (Radford et al. 2021)

Contrastive Loss for CLIP



Postive Pairs:
Aligned text-images in a minibatch $N \times 1$

Negative Pairs:
Mis-aligned text-images in a minibatch $N \times N - N$

CLIP (Radford et al. 2021)

Contrastive Loss for CLIP

x_1

A photo of a golden retriever → y_1

x_2

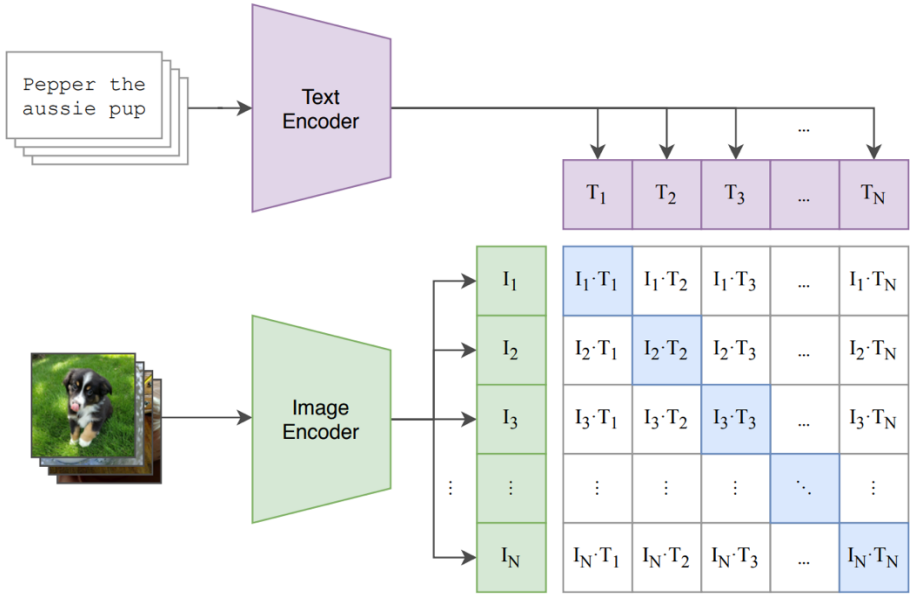
A lake surrounded by mountains and woods → y_2

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

$$L_{t2i} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\tau)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\tau)}$$

CLIP (Radford et al. 2021)

Contrastive Loss for CLIP




```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

CLIP (Radford et al. 2021)

Data Curation for CLIP

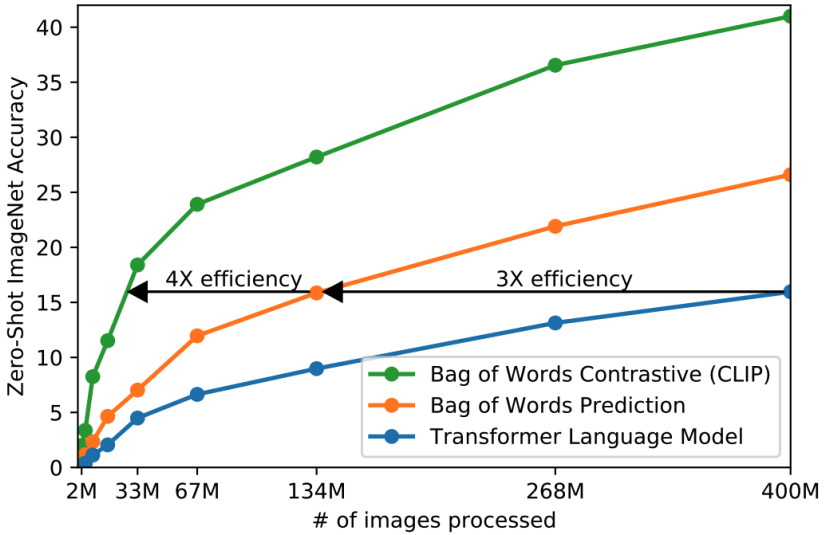
WebImageText (WIT) OpenAI Private Dataset

- 400M image-text data pairs
- search for (image, text) pairs as part of the construction process whose text includes one of a set of 500,000 queries
- balance the results by including up to 20,000 (image, text) pairs per query

Demystifying CLIP Data (Xu et al. ICLR 2024)

CLIP (Radford et al. 2021)

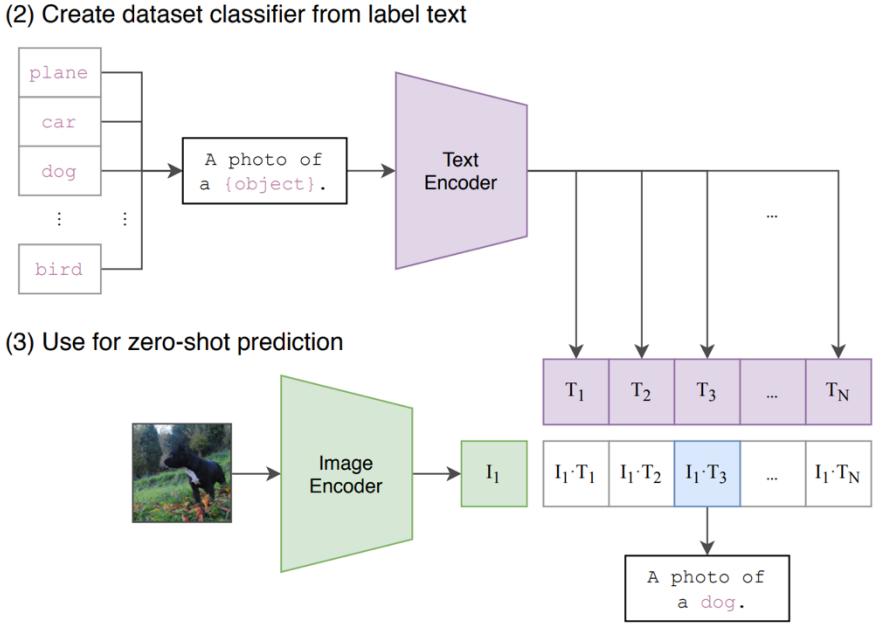
Efficiency of Contrastive Pretraining



Contrastive Learning is much more efficient at zero-shot transfer than the predictive baseline

CLIP (Radford et al. 2021)

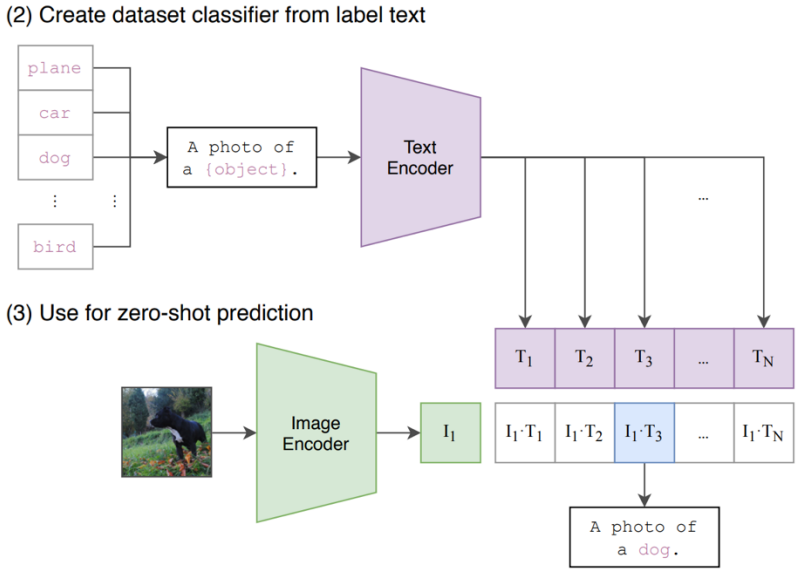
Zero-shot Prediction

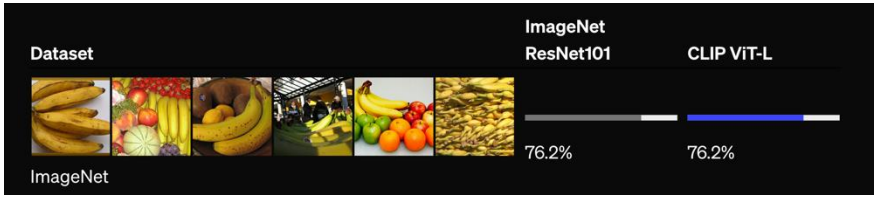


- Construct a prompt template "A photo of a {label}" rather than "{label}" (1.3% ↑)
  - Text with a single word relatively rare in the pre-training dataset
- Provide domain-specific context "A photo of a {label}, a type of pet." for Oxford Pets
- Prompt ensembling. 80 different context prompts ensemble for ImageNet (3.5% ↑)

CLIP (Radford et al. 2021)

Zero-shot Prediction





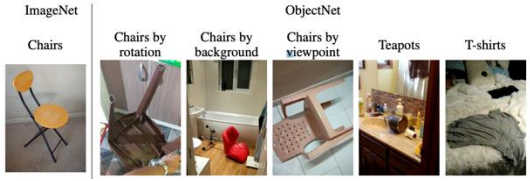
Comparable performance on ImageNet without training on that!

CLIP (Radford et al. 2021)

Evaluate the Robustness to Distribution Shift

ImageNetV2 (Recht et al. 2019)

ImageNet Sketch (Wang et al. 2019)

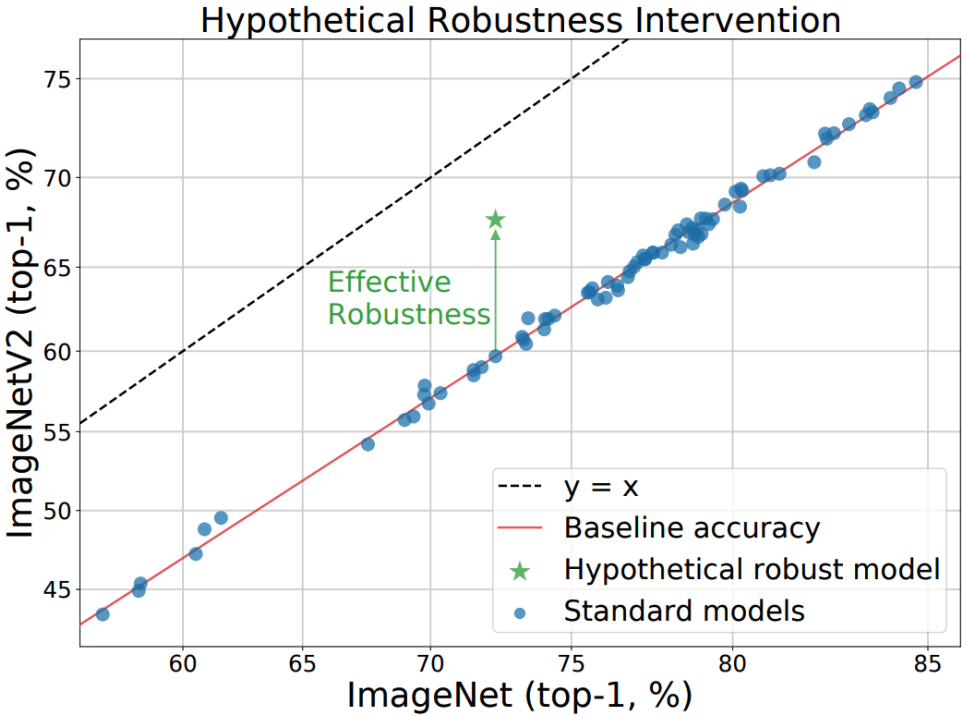


ObjectNet (Barbu et al. 2019)

ImageNet-R (Hendrycks et al. 2021)

ImageNet-A (Hendrycks et al. 2019)

Evaluate the Robustness to Distribution Shift

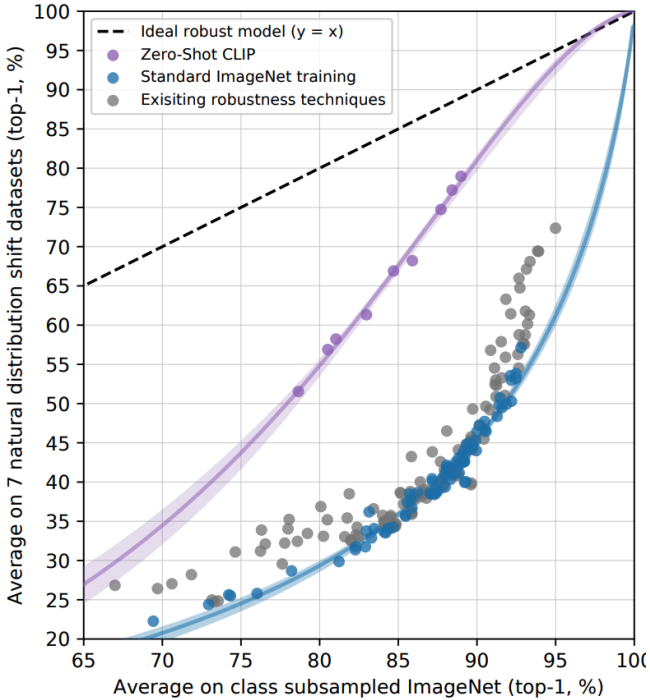


Measuring Robustness (Taori et al. 2020)

CLIP is More Robust to Distribution Shift

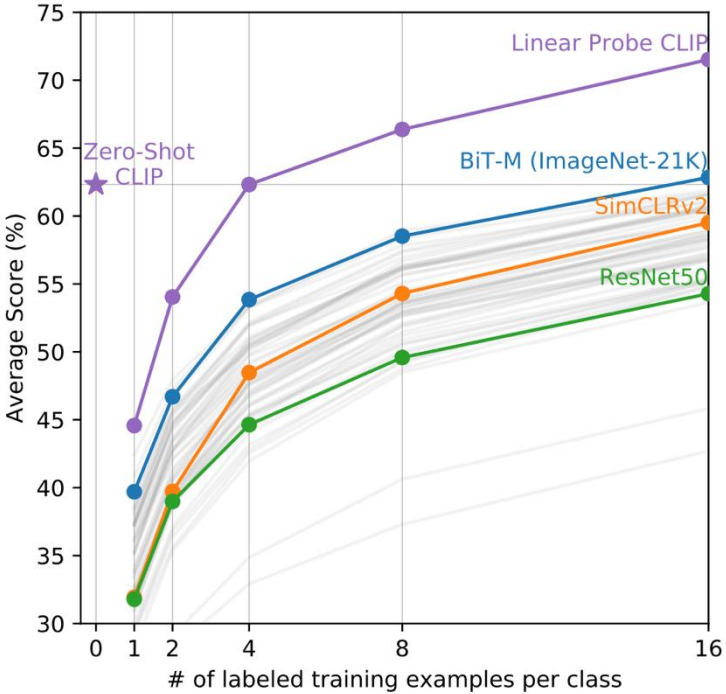
CLIP (Radford et al. 2021)

CLIP is More Robust to Distribution Shift



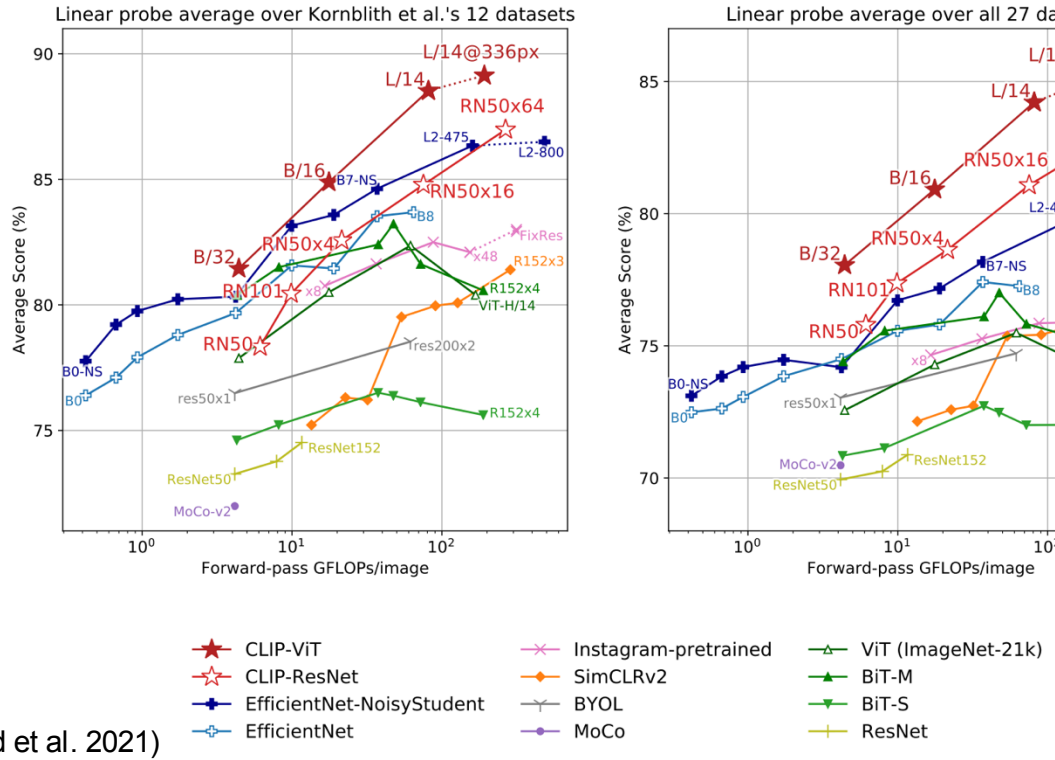
CLIP (Radford et al. 2021)

Zero-shot CLIP Outperforms Few-shot Probing



CLIP (Radford et al. 2021)

Linear Probe Performance of CLIP

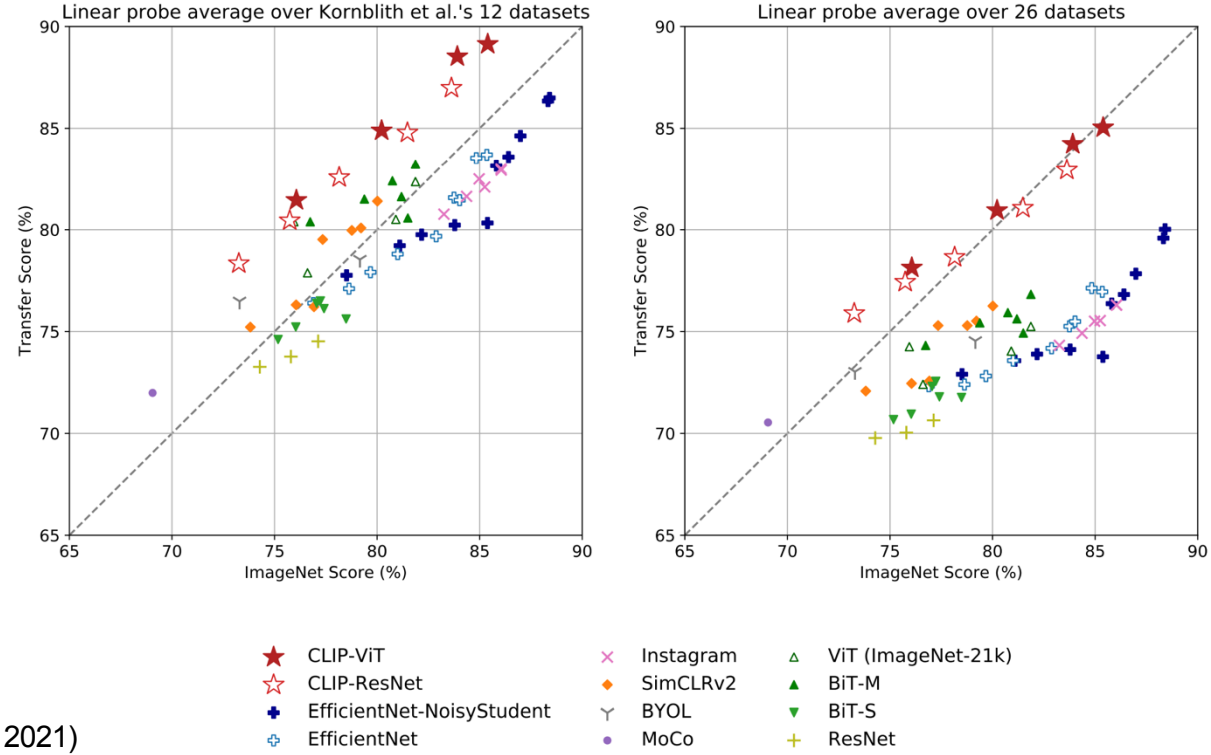


CLIP vision transformers are about 3x more compute efficient than CLIP ResNets

Outperforms the best performing model (a Noisy Student EfficientNet-L2)

CLIP (Radford et al. 2021)

CLIP is More Robust to Task Shift



CLIP (Radford et al. 2021)

Language Supervision is not New

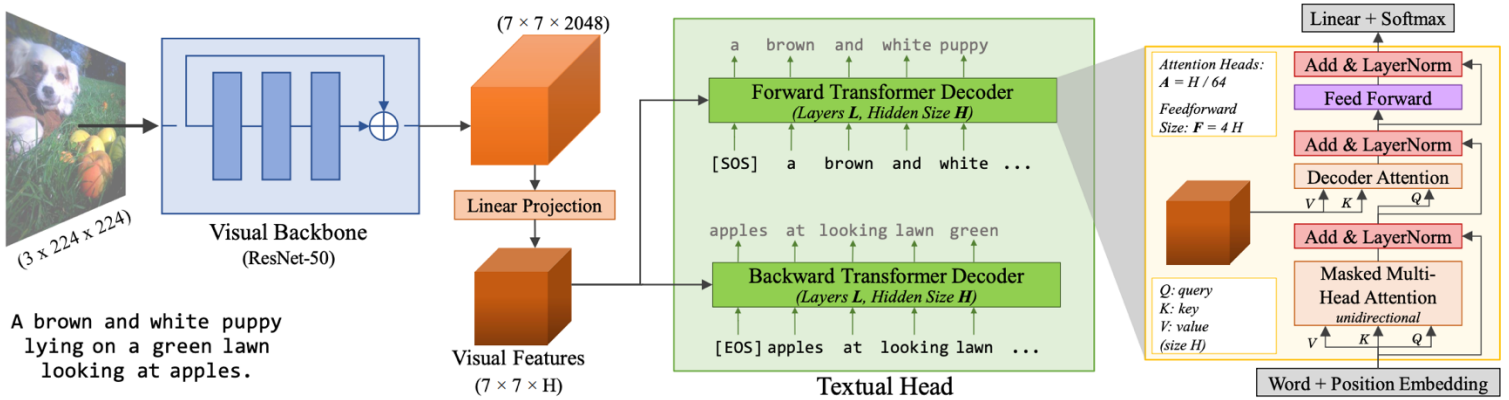
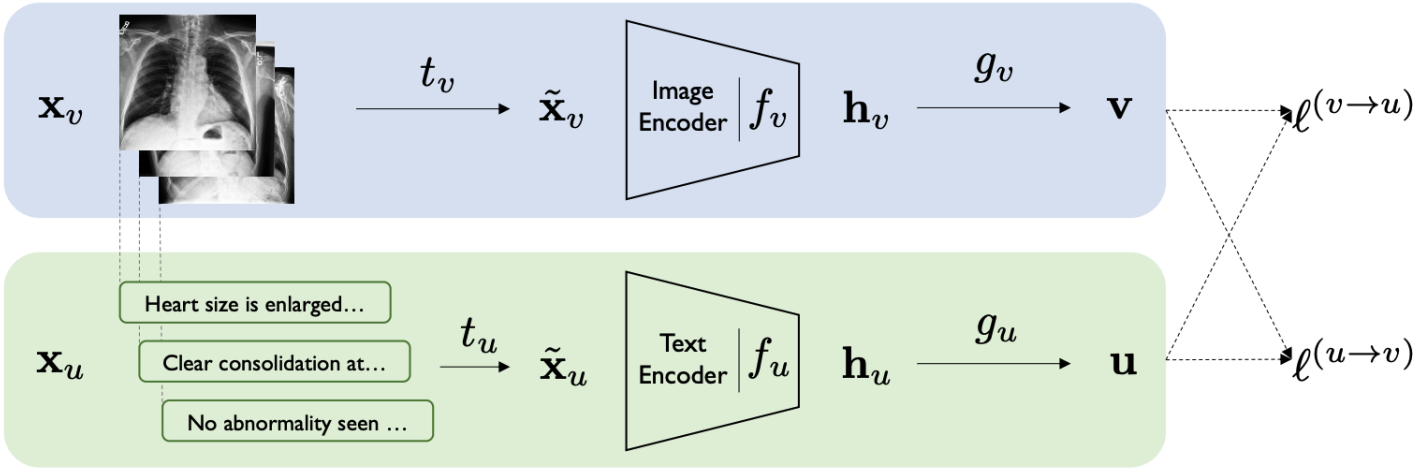


Image-conditioned Bi-directional Next Token Prediction

VirTex (Desai, Johnson 2020)

Language Supervision is not New



Contrastive Learning of Medical Visual Representations from Paired Images and Text

ConVIRT (Zhang et al. 2020)

Language Supervision is not New

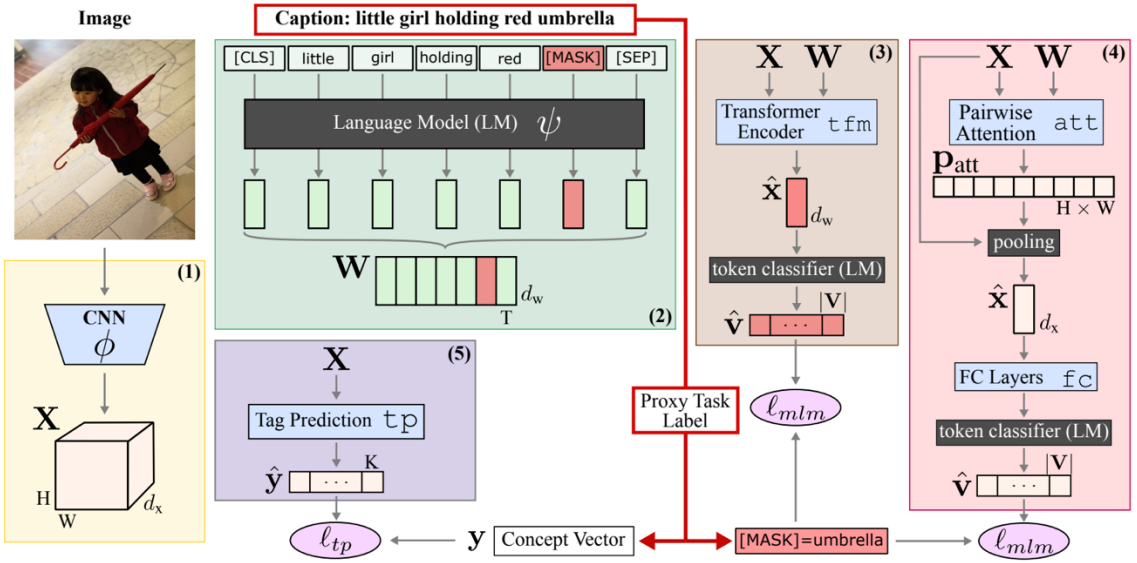


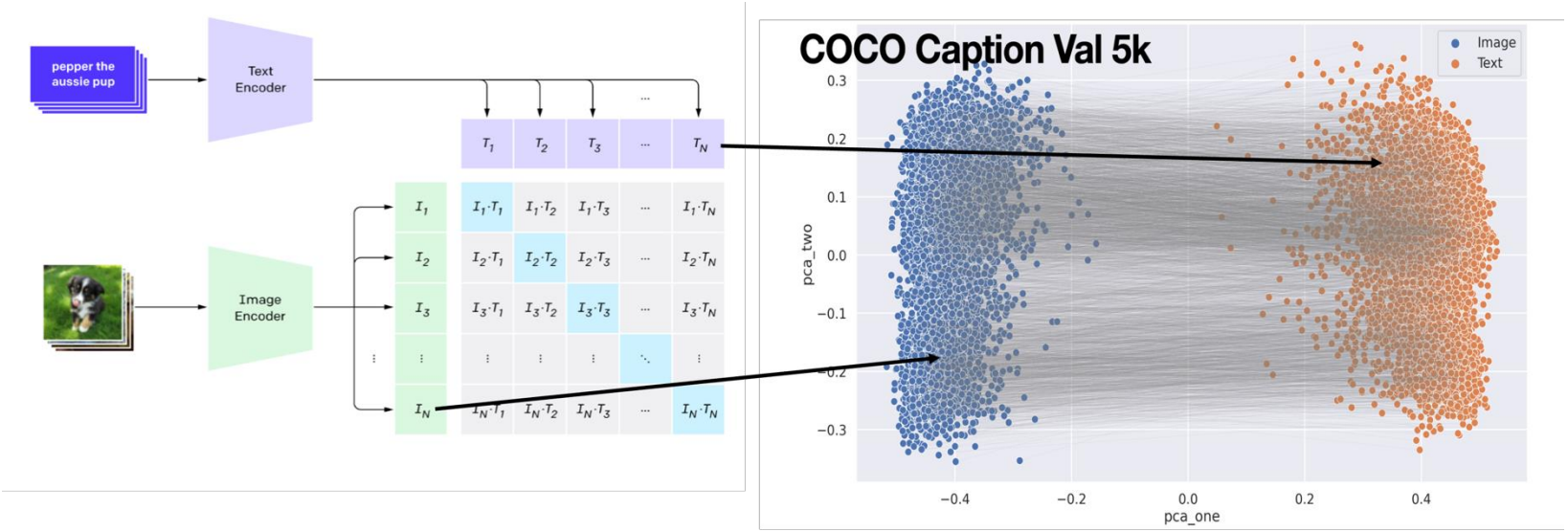
Image-conditioned masked language modeling

ICMLM (Sariyildiz et al. 2020)

Why CLIP Achieves Such Success

- Transferability:
  - Bridging Vision and Language
- Scalability:
  - Large-Scale Pretraining on 400M Data
- Simplicity:
  - Contrastive Learning Enables the Large-Scale Pre-Training

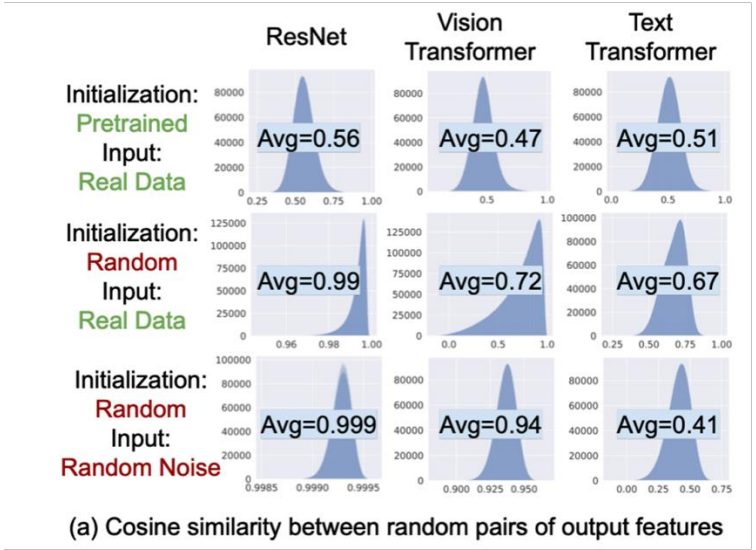
Understanding the Embedding Space of CLIP

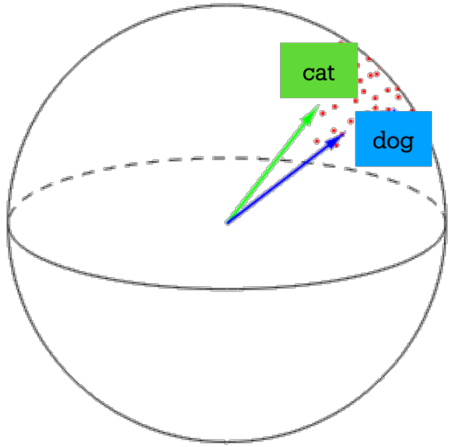


Modality Gap Phenomenon: Paired text embeddings and visual embeddings are not exactly matched

Modality Gap (Zhang et al. 2022)

Understanding the Embedding Space of CLIP

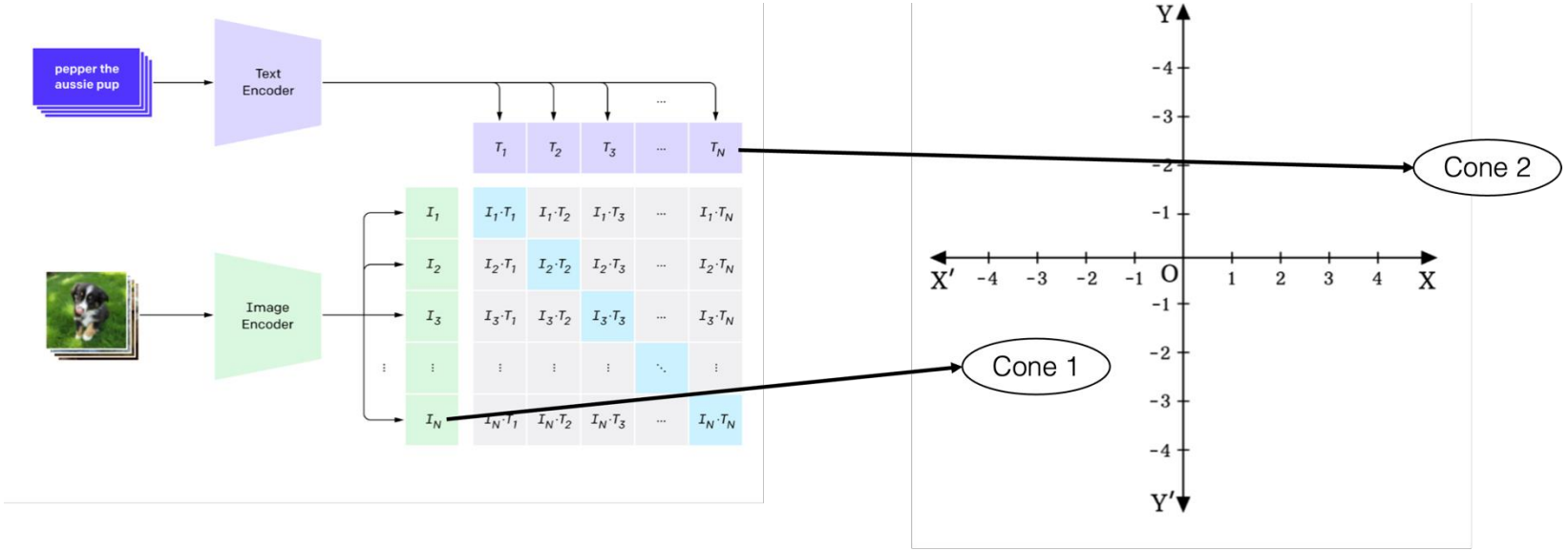




Cone Effect: General Phenomenon for Any Deep Neural Network

Modality Gap (Zhang et al. 2022)

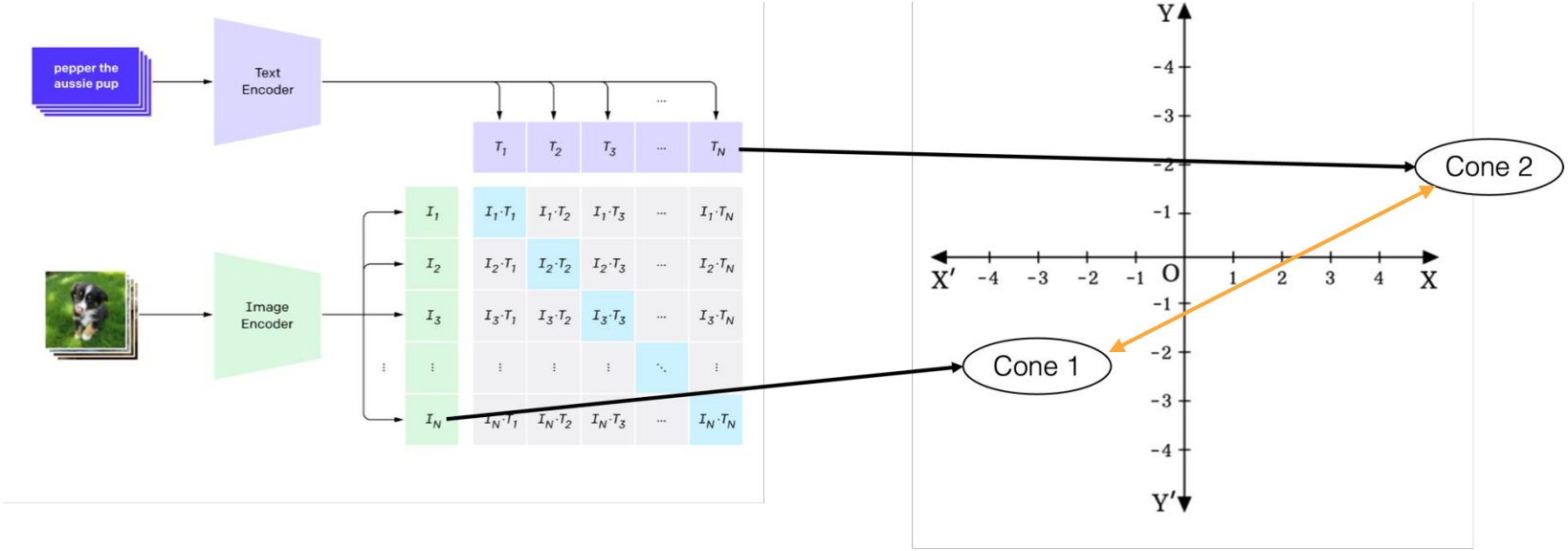
Understanding the Embedding Space of CLIP



Two encoders produce two cones

Modality Gap (Zhang et al. 2022)

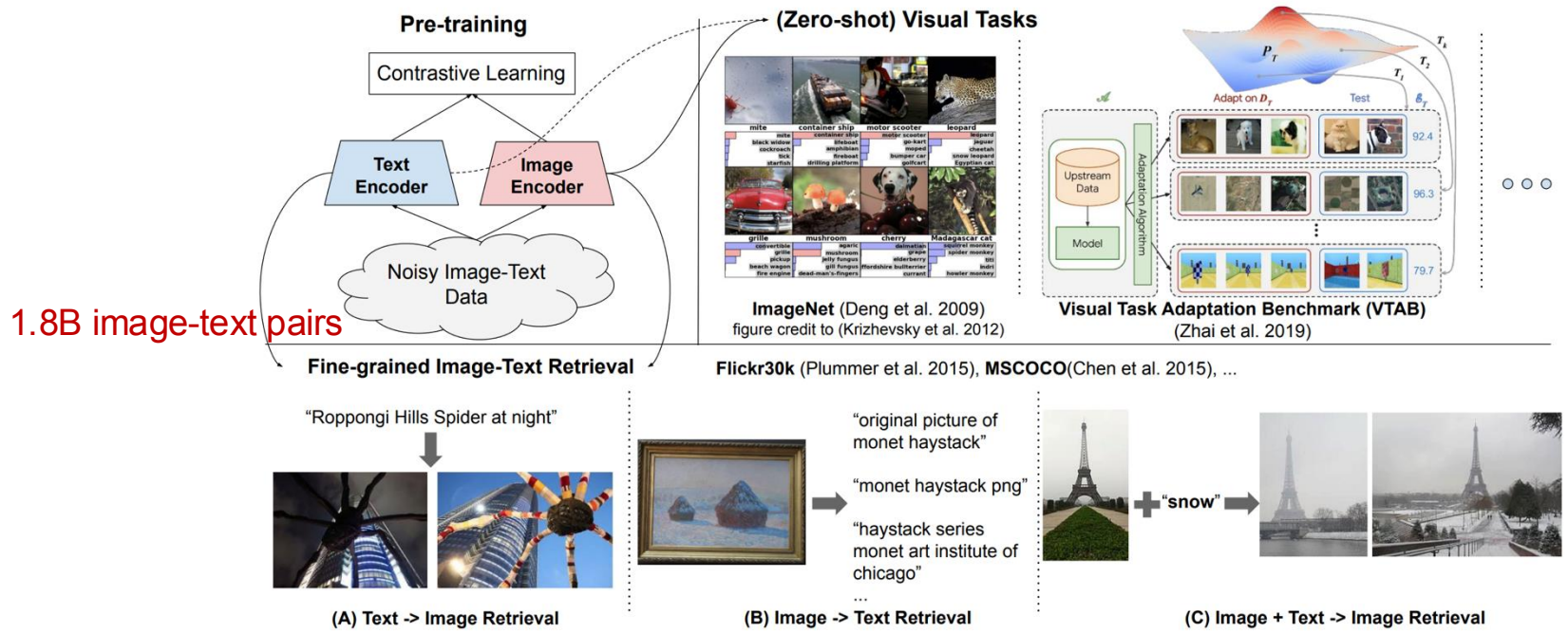
Understanding the Embedding Space of CLIP



Contrastive Learning Preserves the Gap

Modality Gap (Zhang et al. 2022)

Contrastive Learning with Noisy Text Supervision



ALIGN (Jia et al. 2021)

Contrastive Learning with Noisy Text Supervision

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**

Comparable Performance with CLIP on ImageNet

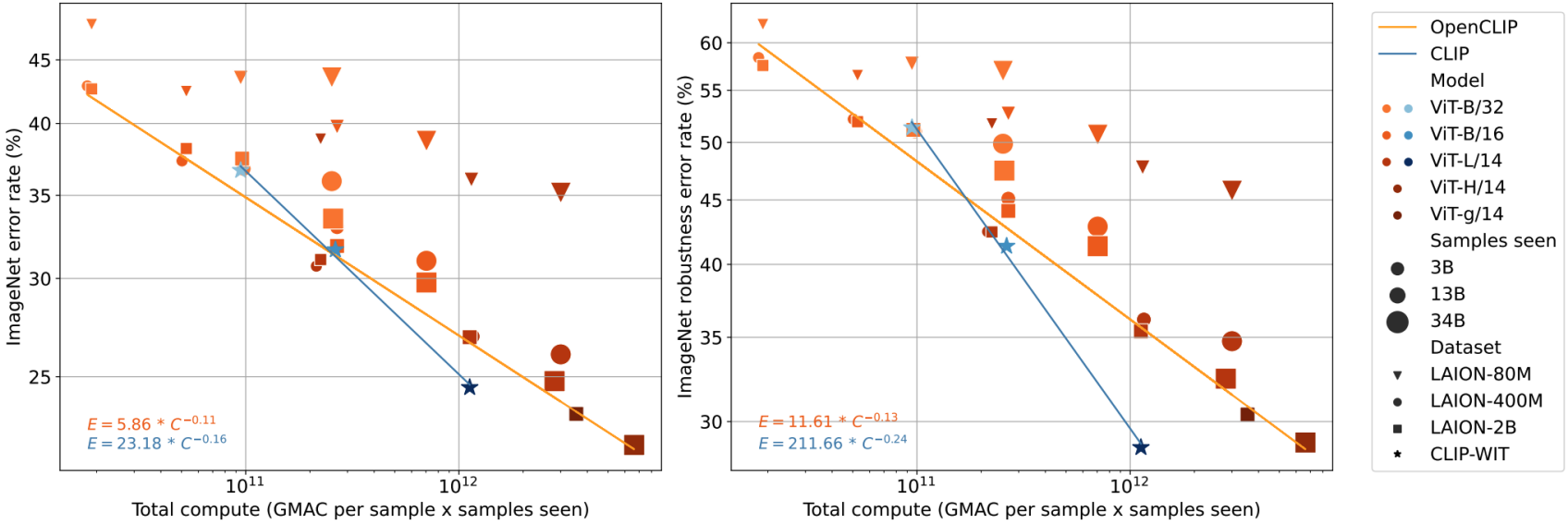
ALIGN (Jia et al. 2021)

How to Reproduce CLIP?

- Data: Open large-scale datasets LAION-400M/2B
- Model:
  - Same transformers as CLIP: ViT-B/32, ViT-B/16, ViT-L/14
  - Even larger transformers: ViT-H/14 and ViT-G/14
- Compute: up to 1520 NVIDIA A100 GPUs

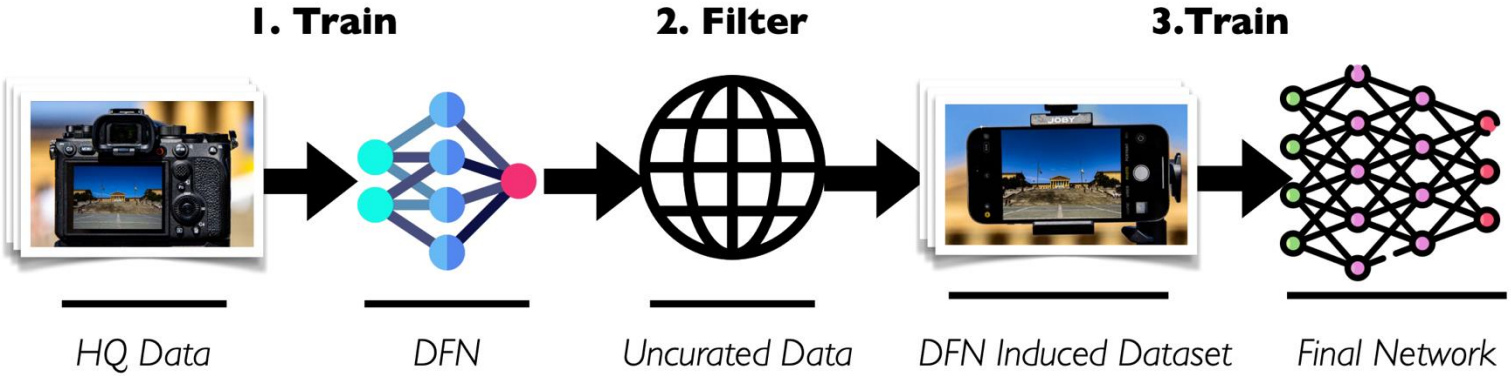
OpenCLIP (Cherti et al. 2022)

How to Reproduce CLIP?



OpenCLIP (Cherti et al. 2022)

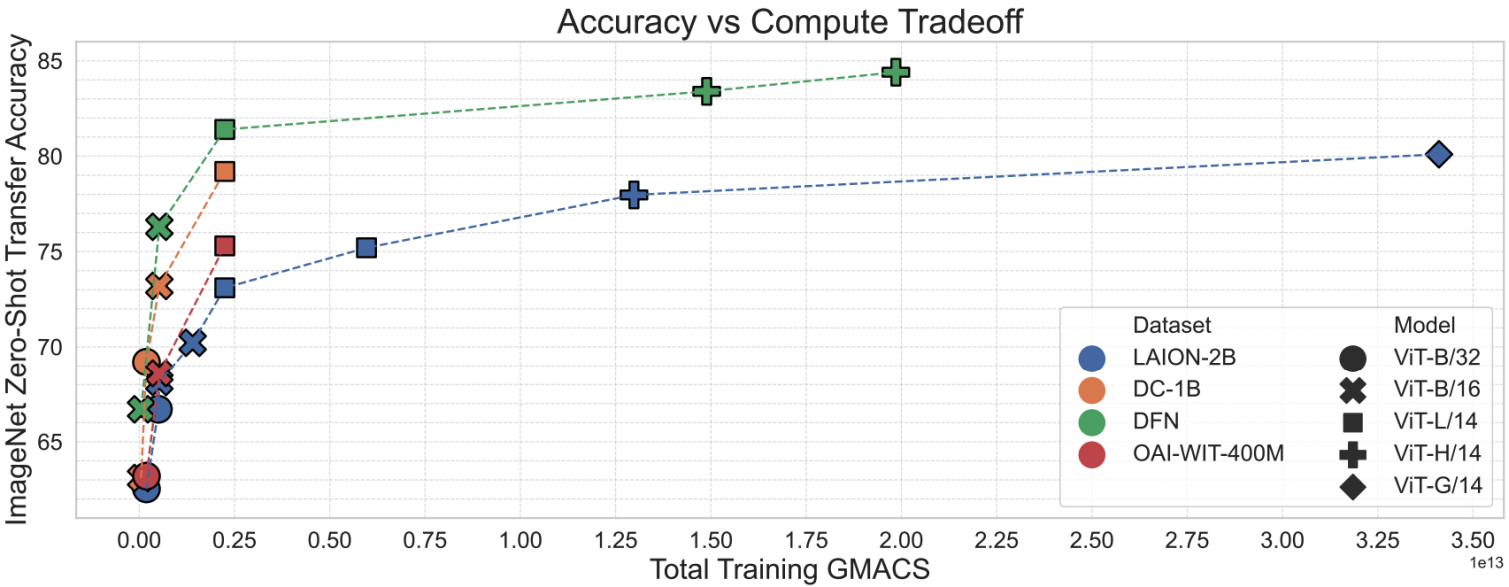
How to Improve CLIP?



Larger and Better Data

DFN (Fang et al. 2023)

How to Improve CLIP?



DFN (Fang et al. 2023)

How to Improve CLIP?

Modified Loss Function to decouples the batch size from the comparison

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

Larger Data: WebLI 10B/12B

Algorithm 1 Sigmoid loss pseudo-implementation.

```
# img_emb      : image model embedding [n, dim]
# txt_emb      : text model embedding [n, dim]
# t_prime, b   : learnable temperature and bias
# n            : mini-batch size

t = exp(t_prime)
zimg = l2_normalize(img_emb)
ztxt = l2_normalize(txt_emb)
logits = dot(zimg, ztxt.T) * t + b
labels = 2 * eye(n) - ones(n) # -1 with diagonal 1
l = -sum(log_sigmoid(labels * logits)) / n
```

SigLIP (Zhai et al. 2023)

How to Improve CLIP?

Method	Image Encoder		ImageNet-1k				COCO R@1	
	ViT size	# Patches	Validation	v2	ReaL	ObjectNet	I → T	T → I
CLIP	B	196	68.3	61.9	-	55.3	52.4	33.1
OpenCLIP	B	196	70.2	62.3	-	56.0	59.4	42.3
EVA-CLIP	B	196	74.7	67.0	-	62.3	58.7	42.2
SigLIP	B	196	**76.2**	**69.6**	82.8	**70.7**	**64.4**	**47.2**
SigLIP	B	256	76.7	70.0	83.1	71.3	65.1	47.4
SigLIP	B	576	78.6	72.1	84.5	73.8	67.5	49.7
SigLIP	B	1024	**79.2**	**73.0**	**84.9**	**74.7**	**67.6**	**50.4**
CLIP	L	256	75.5	69.0	-	69.9	56.3	36.5
OpenCLIP	L	256	74.0	61.1	-	66.4	62.1	46.1
CLIPA-v2	L	256	79.7	72.8	-	71.1	64.1	46.3
EVA-CLIP	L	256	79.8	72.9	-	75.3	63.7	47.5
SigLIP	L	256	**80.5**	**74.2**	**85.9**	**77.9**	**69.5**	**51.1**
CLIP	L	576	76.6	72.0	-	70.9	57.9	37.1
CLIPA-v2	L	576	80.3	73.5	-	73.1	65.5	47.2
EVA-CLIP	L	576	80.4	73.8	-	78.4	64.1	47.9
SigLIP	L	576	**82.1**	**75.9**	**87.0**	**81.0**	**70.6**	**52.7**
OpenCLIP	G (2B)	256	80.1	73.6	-	73.0	67.3	51.4
CLIPA-v2	H (630M)	576	81.8	75.6	-	77.4	67.2	49.2
EVA-CLIP	E (5B)	256	82.0	75.7	-	79.6	68.8	51.1
SigLIP	SO (400M)	729	**83.2**	**77.2**	**87.5**	**82.9**	**70.2**	**52.0**

The best CLIP model

2024/10

SigLIP (Zhai et al. 2023)

How to Improve CLIP?

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

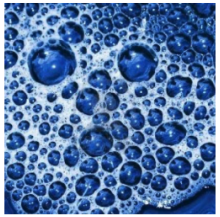
(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

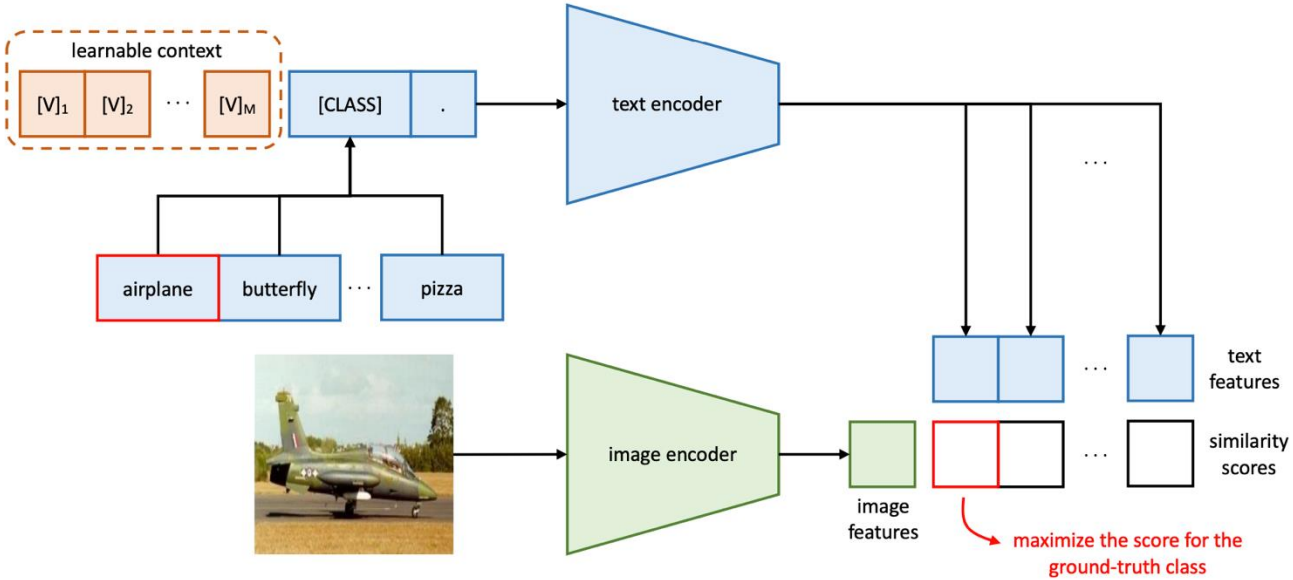
Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

Prompt engineering vs Context Optimization

CoOp (Zhou et al. 2021)

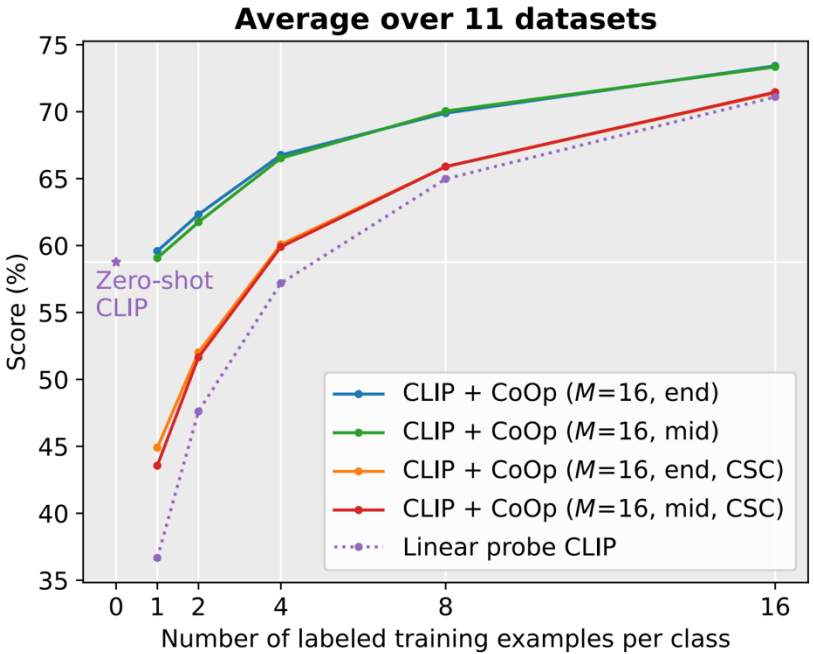
How to Improve CLIP?



Learning to prompt for few-shot classification

CoOp (Zhou et al. 2021)

How to Improve CLIP?

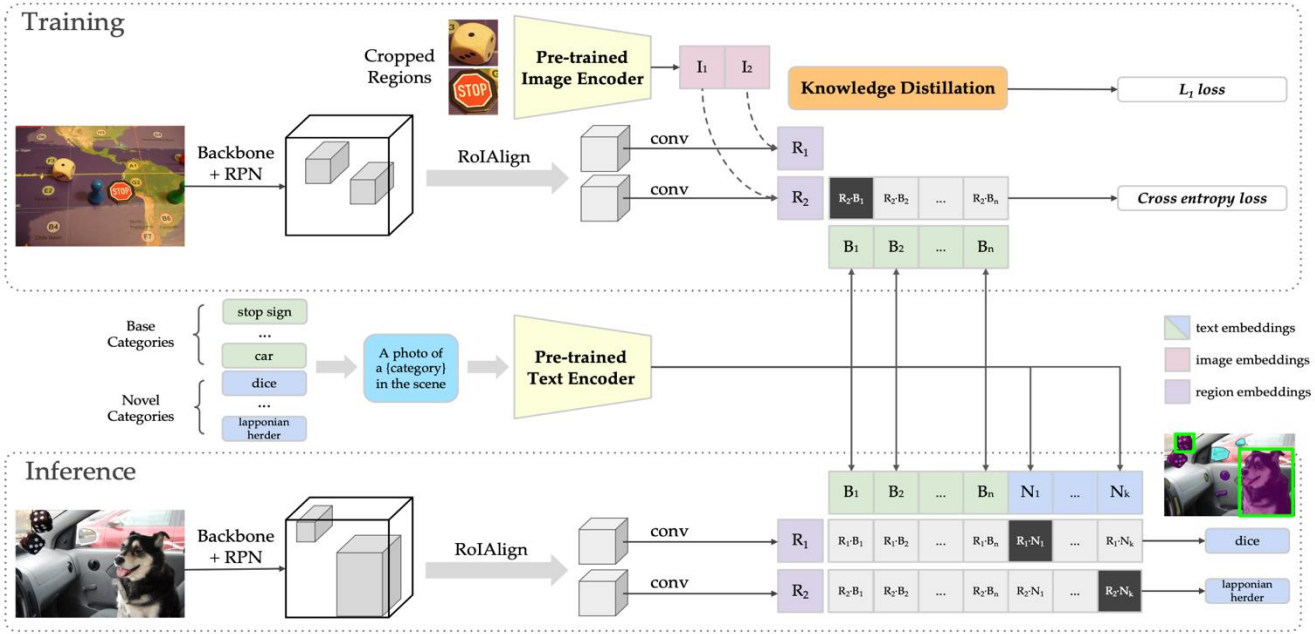


M denotes the context length. “end” or “mid” means putting the class token in the end or middle. CSC means class-specific context.

CoOp effectively turns CLIP into a strong few-shot learner

CoOp (Zhou et al. 2021)

Apply CLIP to Different Tasks



Open-Vocabulary Object Detection

ViLD (Gu et al. 2022)

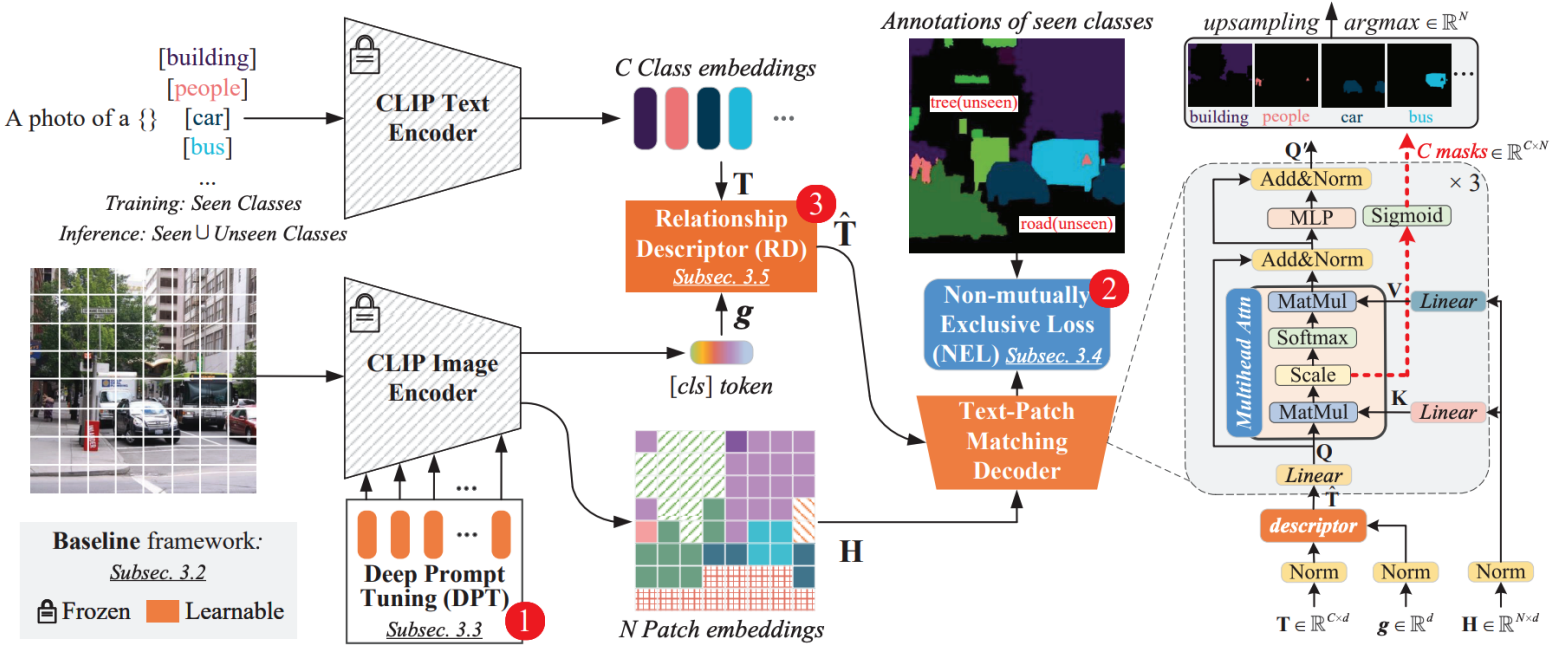
Apply CLIP to Different Tasks

Method	Training source	Novel AP	Base AP	Overall AP
Bilen & Vedaldi (2016)	image-level labels in $C_B \cup C_N$	19.7	19.6	19.6
Ye et al. (2019)		20.3	20.1	20.1
Bansal et al. (2018)	instance-level labels in C_B	0.31	29.2	24.9
Zhu et al. (2020)		3.41	13.8	13.0
Rahman et al. (2020)		4.12	35.9	27.9
Zareian et al. (2021)	image captions in $C_B \cup C_N$ instance-level labels in C_B	22.8	46.0	39.9
CLIP on cropped regions	image-text pairs from Internet (may contain $C_B \cup C_N$) instance-level labels in C_B	26.3	28.3	27.8
ViLD-text		5.9	61.8	47.2
ViLD-image		24.1	34.2	31.6
ViLD ($w = 0.5$)		**27.6**	59.5	51.3

Open-Vocabulary Object Detection 4.8% ↑

ViLD (Gu et al. 2022)

Apply CLIP to Different Tasks



CLIP for Zero-shot Semantic Segmentation

ZegCLIP (Zhou et al. 2021)

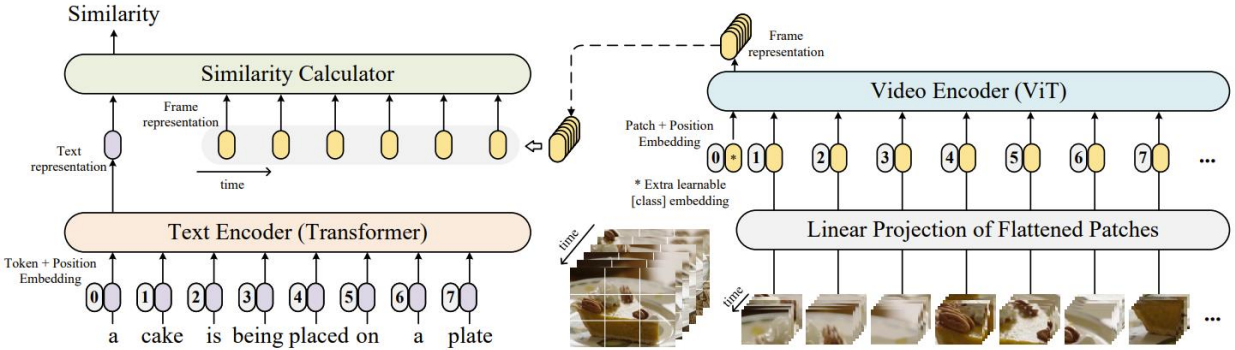
Apply CLIP to Different Tasks

Methods	PASCAL VOC 2012			
	pAcc	mIoU(S)	mIoU(U)	hIoU
Inductive				
SPNet [44]	-	78.0	15.6	26.1
ZS3 [3]	-	77.3	17.7	28.7
CaGNet [17]	80.7	78.4	26.6	39.7
SIGN [10]	-	75.4	28.9	41.7
Joint [1]	-	77.7	32.5	45.9
ZegFormer [12]	-	86.4	63.6	73.3
zsseg [49]	90.0	83.5	72.5	77.5
ZegCLIP (Ours)	**94.6**	**91.9**	**77.8**	**84.3**

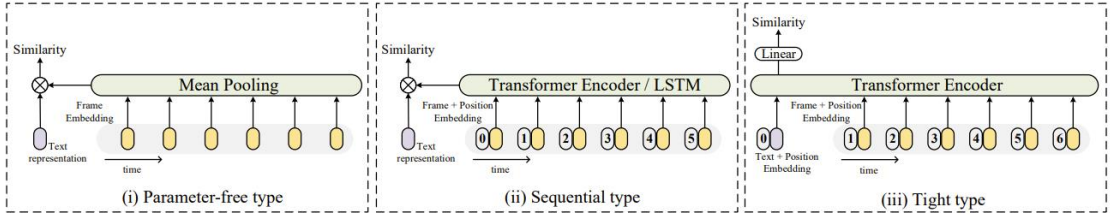
CLIP for Zero-shot Semantic Segmentation 5.3% ↑

ZegCLIP (Zhou et al. 2021)

Apply CLIP to Different Tasks



(a) Main structure



(b) Similarity calculator

CLIP for Video Retrieval

CLIP4Clip (Luo et al. 2021)

Apply CLIP to Different Tasks

Methods	TrainD	E2E	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
C+LSTM+SA[a]	M	✓	4.2	12.9	19.9	55	-
VSE[b]	M	✓	3.8	12.7	17.1	66	-
SNUVL[c]	M	✓	3.5	15.9	23.8	44	-
Kaufman et al.[d]	M	✓	4.7	16.6	24.1	41	-
CT-SAN[e]	M	✓	4.4	16.6	22.3	35	-
JSFusion[f]	M	✓	10.2	31.2	43.2	13	-
HowTo100M[g]	H+M	✓	14.9	40.2	52.8	9	-
ActBERT[h]	H+M		8.6	23.4	33.1	36	-
NoiseE[i]	H+M		17.4	41.6	53.6	8	-
UniVL[j]	H+M		21.2	49.6	63.1	6	-
HERO[k]	H+M		16.8	43.4	57.7	-	-
ClipBERT[l]	C+G+M	✓	22.0	46.8	59.9	6	-
(Ours)-meanP	W+M	✓	**42.1**	**71.9**	**81.4**	**2**	**15.7**
(Ours)-seqLSTM	W+M	✓	41.7	68.8	78.7	**2**	16.6
(Ours)-seqTransf	W+M	✓	42.0	68.6	78.7	**2**	16.2
(Ours)-tightTransf	W+M	✓	37.8	68.4	78.4	**2**	17.2

(a) Training on Training-7K

Methods	TrainD	E2E	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
MIL-NCE[m]	H	✓	9.9	24.0	32.4	29.5	-
CLIP-straight[n]	W	✓	31.2	53.7	64.2	4	-

(b) Zero-shot

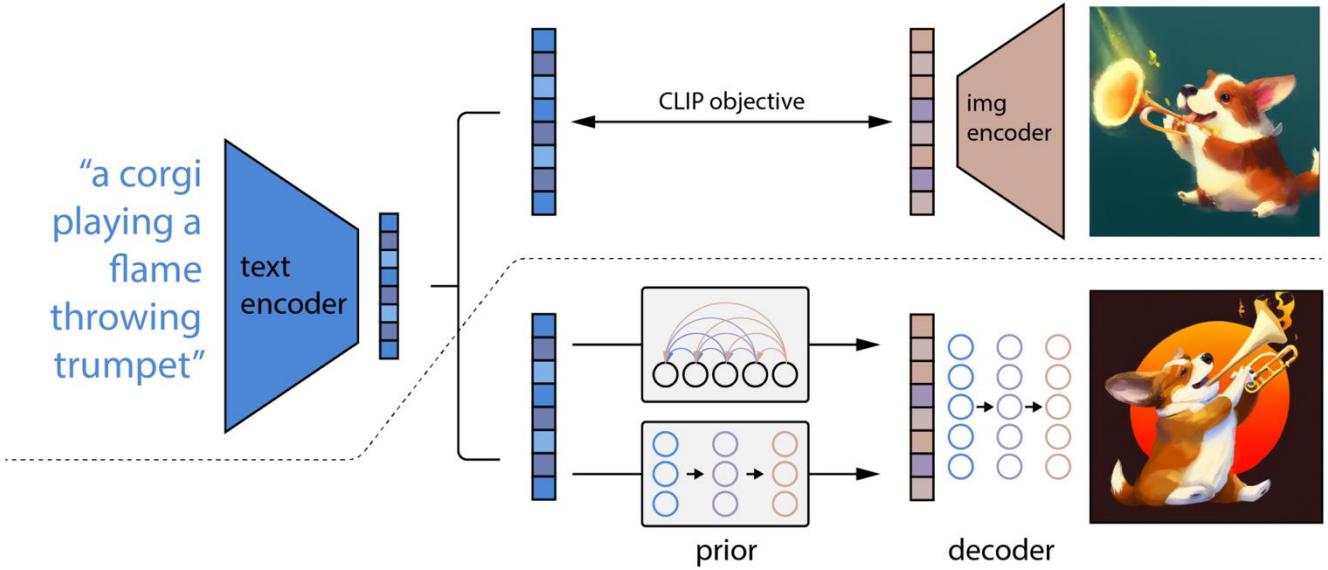
Methods	TrainD	E2E	R@1↑	R@5↑	R@10↑	MdR↓	MnR↓
CE[o]	M		20.9	48.8	62.4	6	28.2
MMT[p]	H+M		26.6	57.1	69.6	4	24.0
AVLnet[q]	H+M		27.1	55.6	66.6	4	-
SSB[r]	H+M		30.1	58.5	69.3	3	-
MDMMT[s]	MD+M		38.9	69.0	79.7	**2**	16.5
Frozen[t]	CW+M	✓	31.0	59.5	70.5	3	-
HiT[u]	H+M		30.7	60.9	73.2	2.6	-
TT-CE+[v]	M		29.6	61.6	74.2	3	-
(Ours)-meanP	W+M	✓	43.1	70.4	80.8	**2**	16.2
(Ours)-seqLSTM	W+M	✓	42.5	70.8	80.7	**2**	16.7
(Ours)-seqTransf	W+M	✓	**44.5**	71.4	**81.6**	**2**	15.3
(Ours)-tightTransf	W+M	✓	40.2	**71.5**	80.5	**2**	**13.4**

(c) Training on Training-9K

CLIP for Video Retrieval 20.1% ↑

CLIP4Clip (Luo et al. 2021)

Apply CLIP to Different Tasks



Hierarchical Text-Conditional Image Generation with CLIP Latents

DALL-E 2 (Ramesh et al. 2022)

Apply CLIP to Different Tasks

Lecture 7

DALL-E 2 (Ramesh et al. 2022) Hierarchical Text-Conditional Image Generation with CLIP Latents

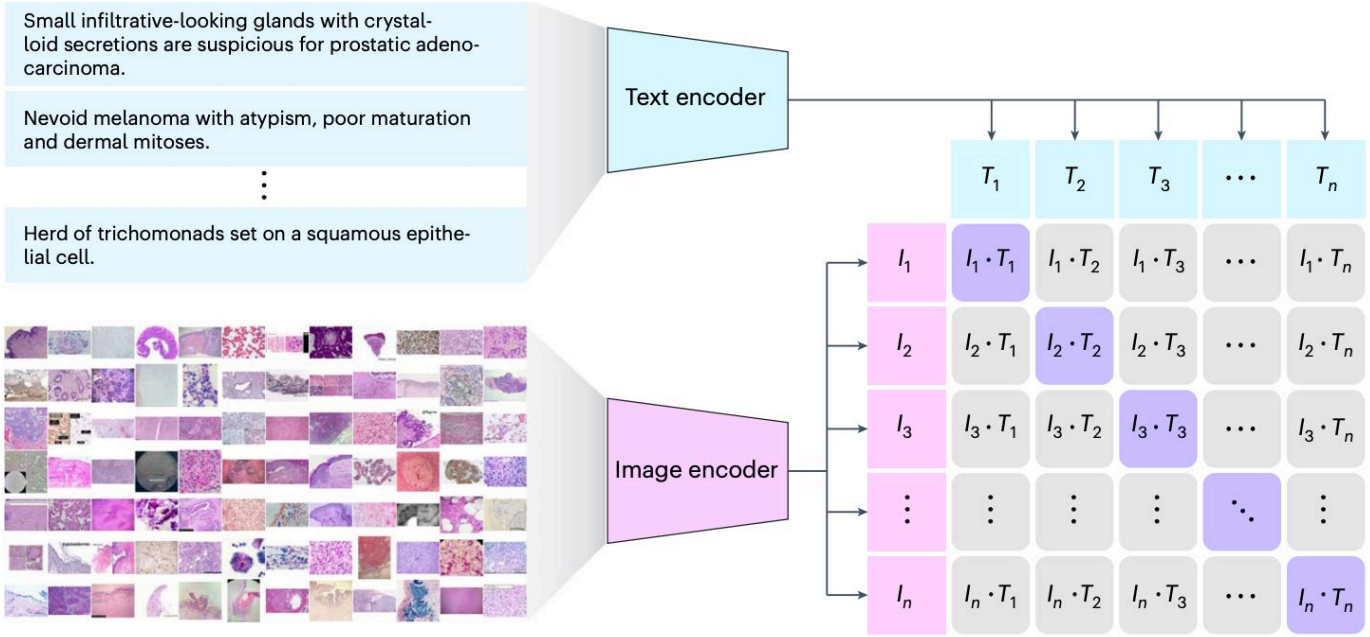
Apply CLIP to Different Tasks

Meta MovieGen 2024

Text input summary: A red-faced monkey with white fur is bathing in a natural hot spring. The monkey is playing in the water with a miniature sail ship in front of it, made of wood with a white sail and a small rudder. The hot spring is surrounded by lush greenery, with rocks and trees.

Apply CLIP to Different Tasks

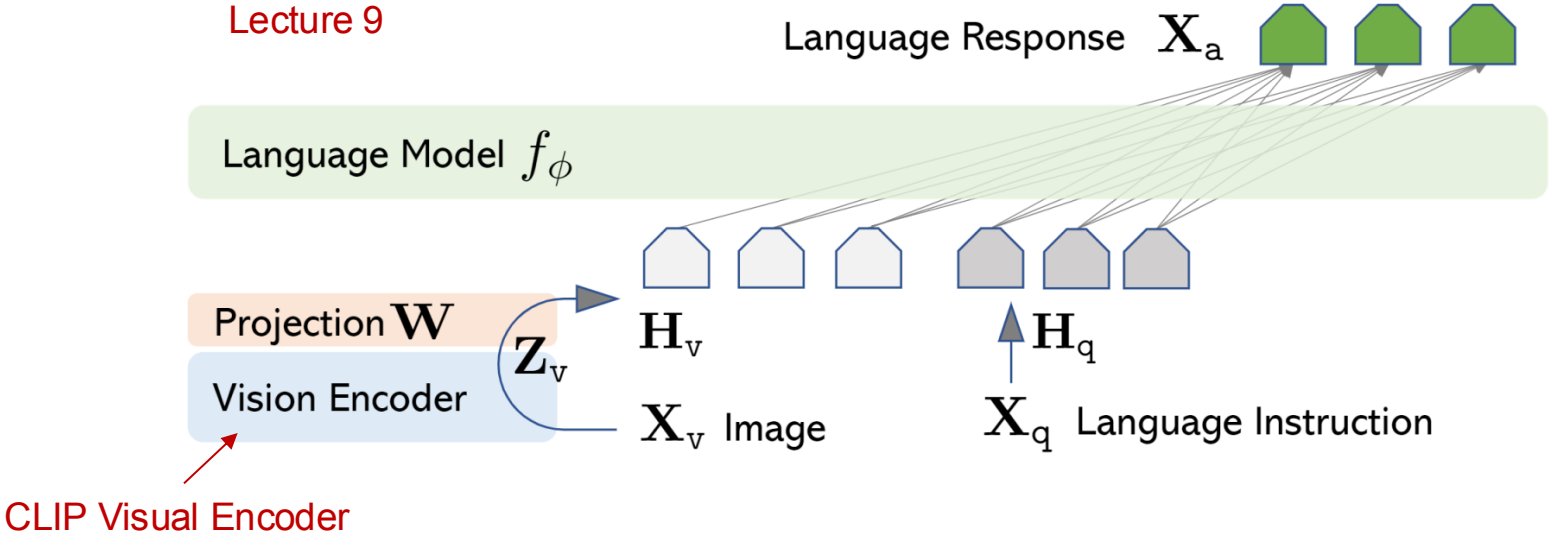
Lecture 8



Fine-tuning CLIP on Pathology Image-Text Pairs

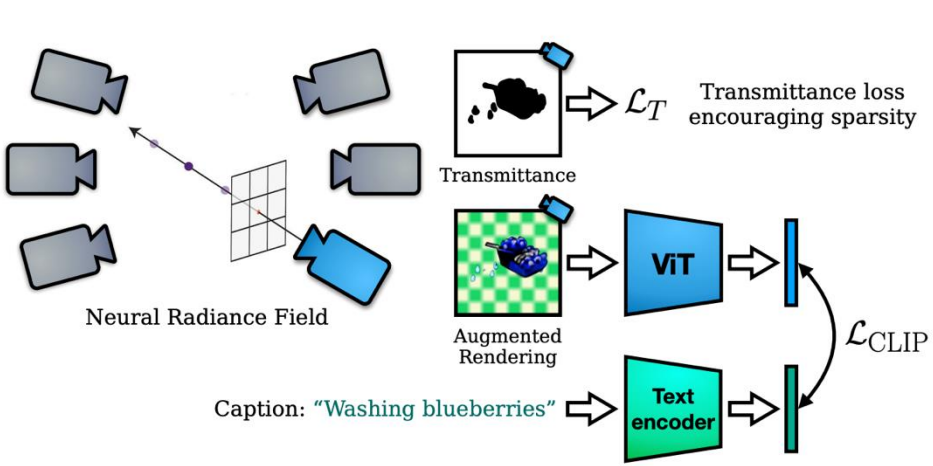
PILP (Huang et al. 2023)

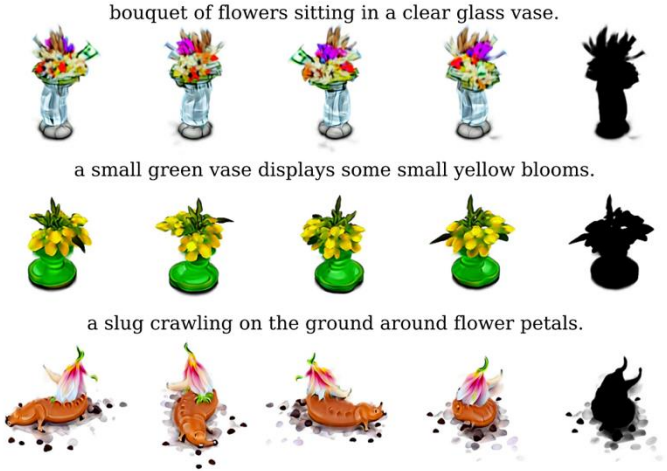
Apply CLIP to Different Tasks



LLaVA (Liu et al. 2023)

Apply CLIP to Different Tasks

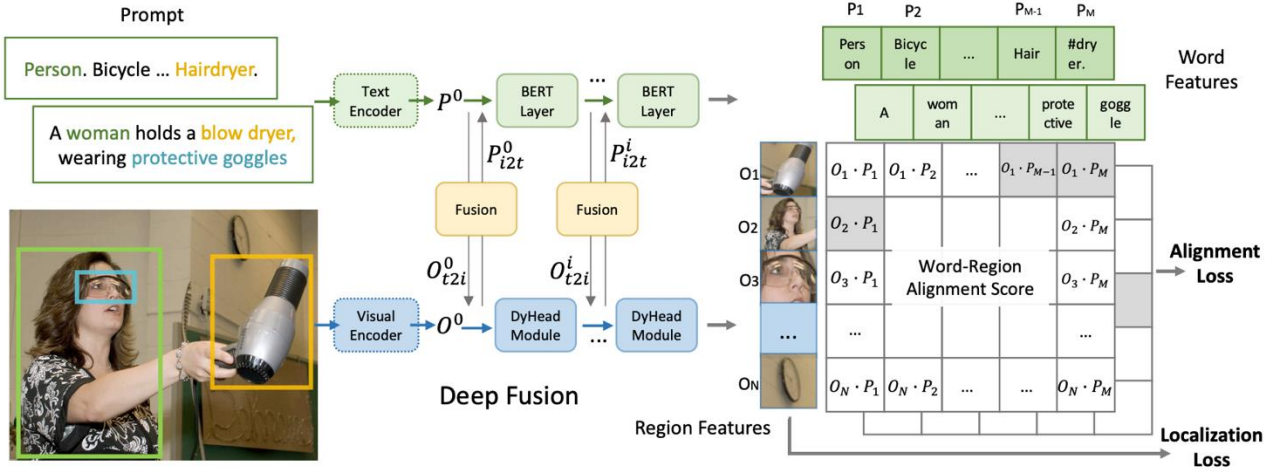


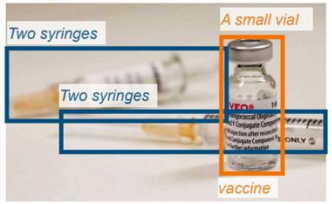


CLIP for Zero-shot 3D Object Generation

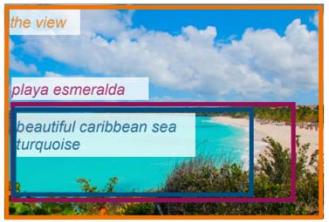
DreamField (Jain et al. 2022)

Region-based CLIP





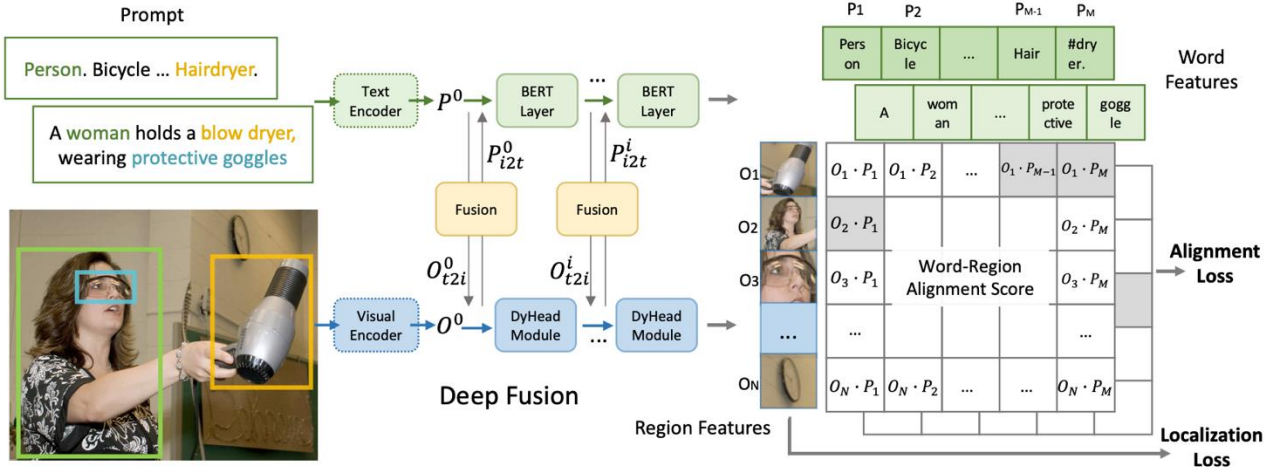
Two syringes and a small vial of vaccine.

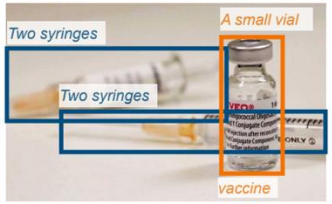


playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

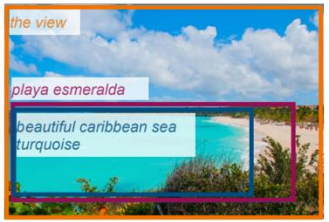
GLIP (Li et al. 2022)

Region-based CLIP





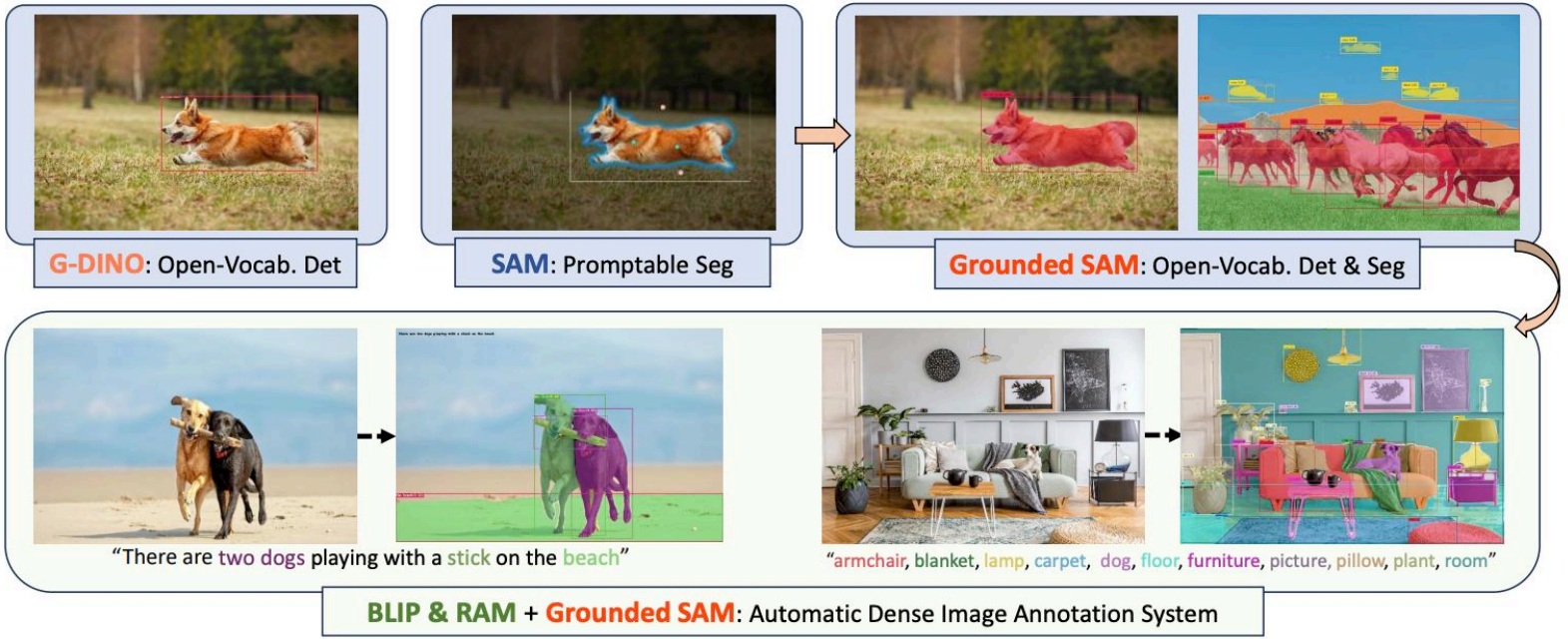
Two syringes and a small vial of vaccine.



playa esmeralda in holguin, cuba. the view from the top of the beach. beautiful caribbean sea turquoise

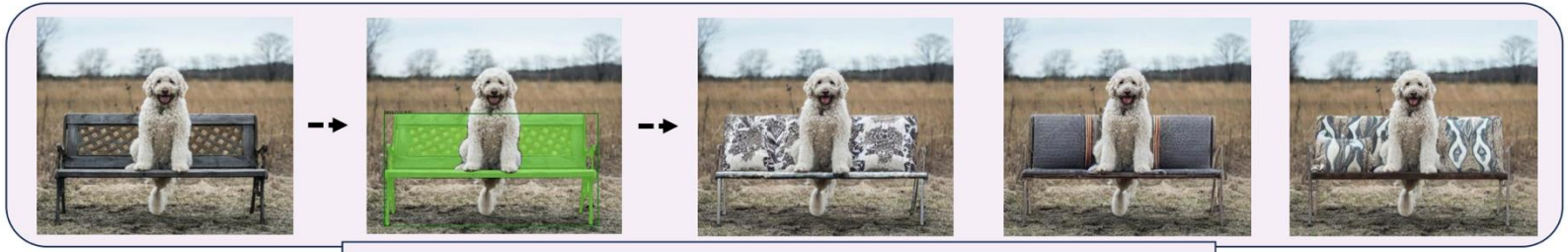
GLIP (Li et al. 2022)

Assembling Open-World Models for Diverse Visual Tasks



Grounded SAM (Ren et al. 2022)

Assembling Open-World Models for Diverse Visual Tasks



Grounded SAM (Ren et al. 2022)

Next time

- Vision-language Representation Learners in Biomedicine