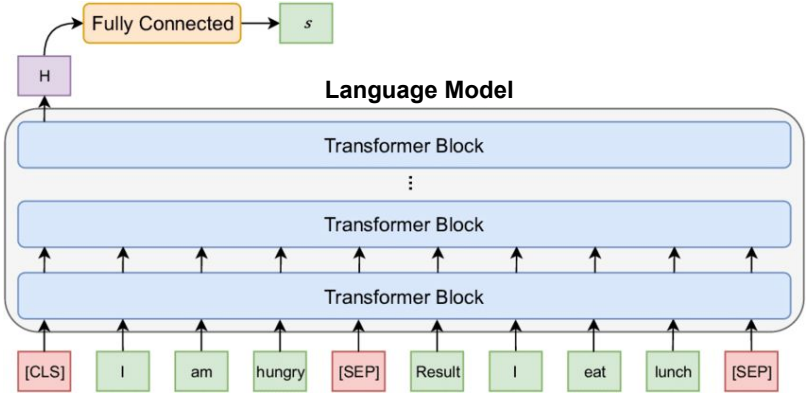
Foundation Models for Vision: CLIP: Contrastive Language-Image Pretraining

Équipe BRAIn

Vision

Text

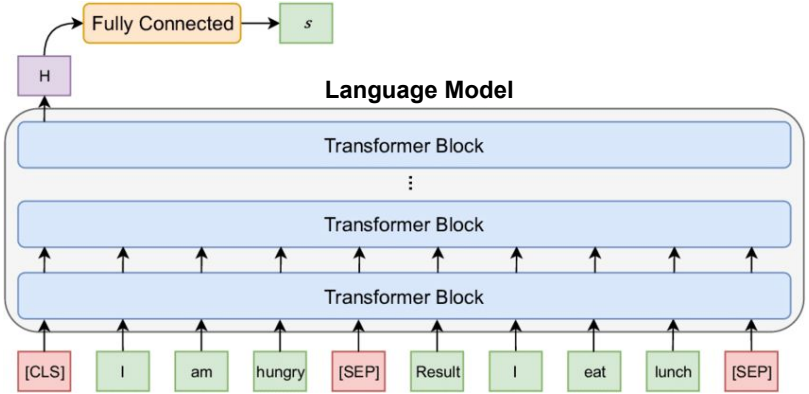
Uni-modal



Vision

Text

Uni-modal



Multi-modal

a teddy bear on a skateboard in times square

Text to Image generation

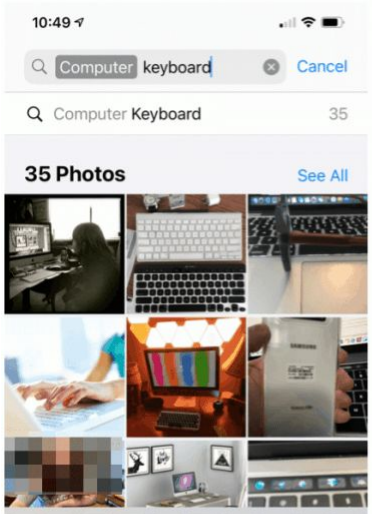
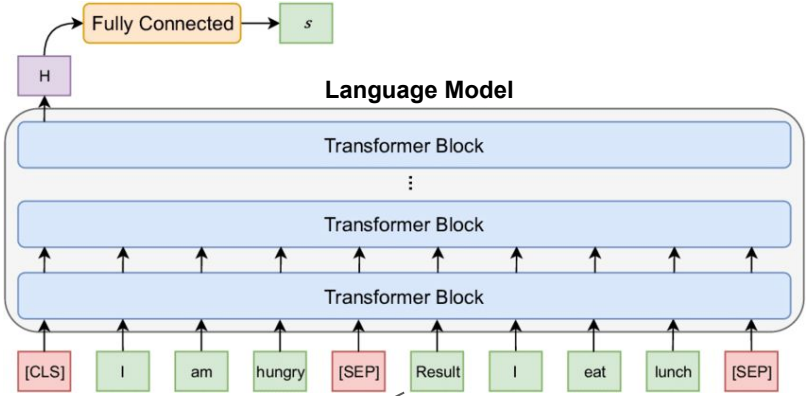


Image retrieval

Vision

Text

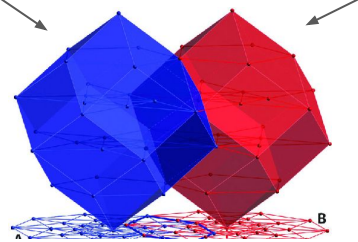
Uni-modal



Multi-modal

a teddy bear on a skateboard in times square

Text to Image generation



Alignement

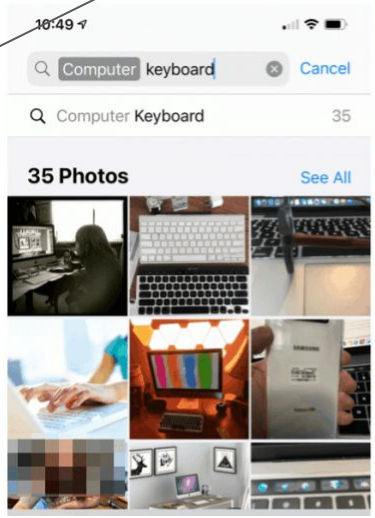


Image retrieval

CLIP: Connecting text and images

Cited by 12033

Web ImageText dataset: 400M paired image-text

The man at bat readies to swing at the pitch while the umpire looks on.

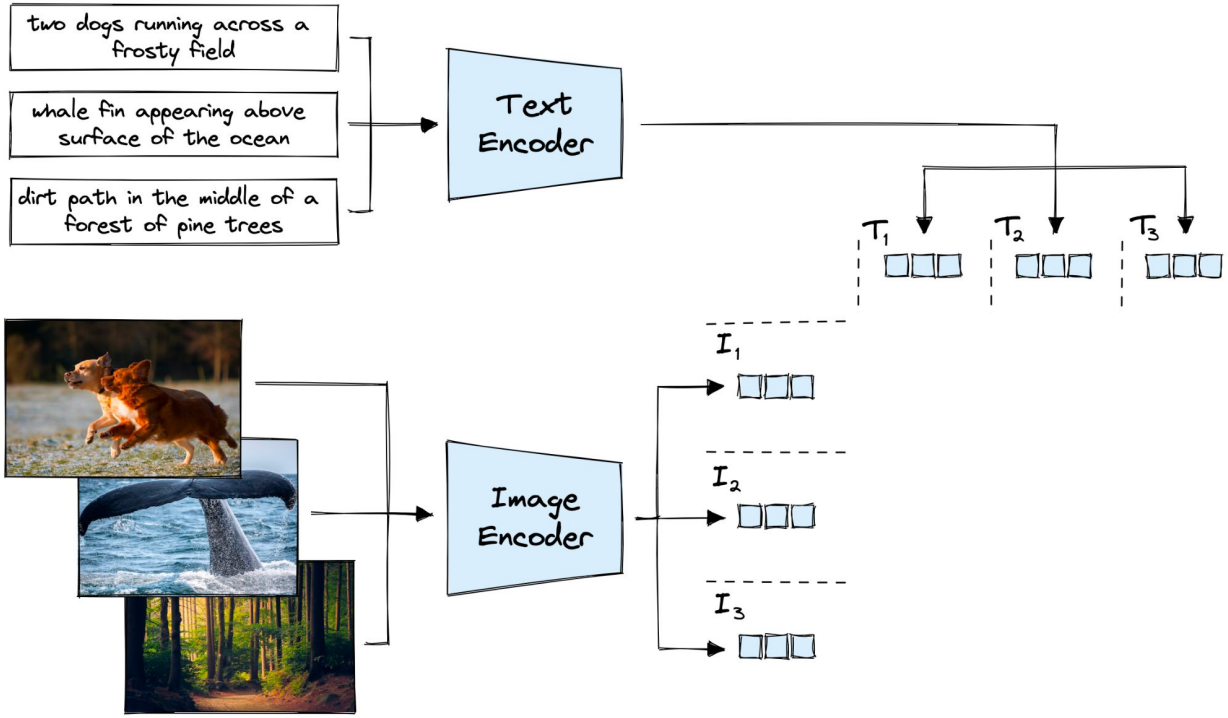
A large bus sitting next to a very tall building.

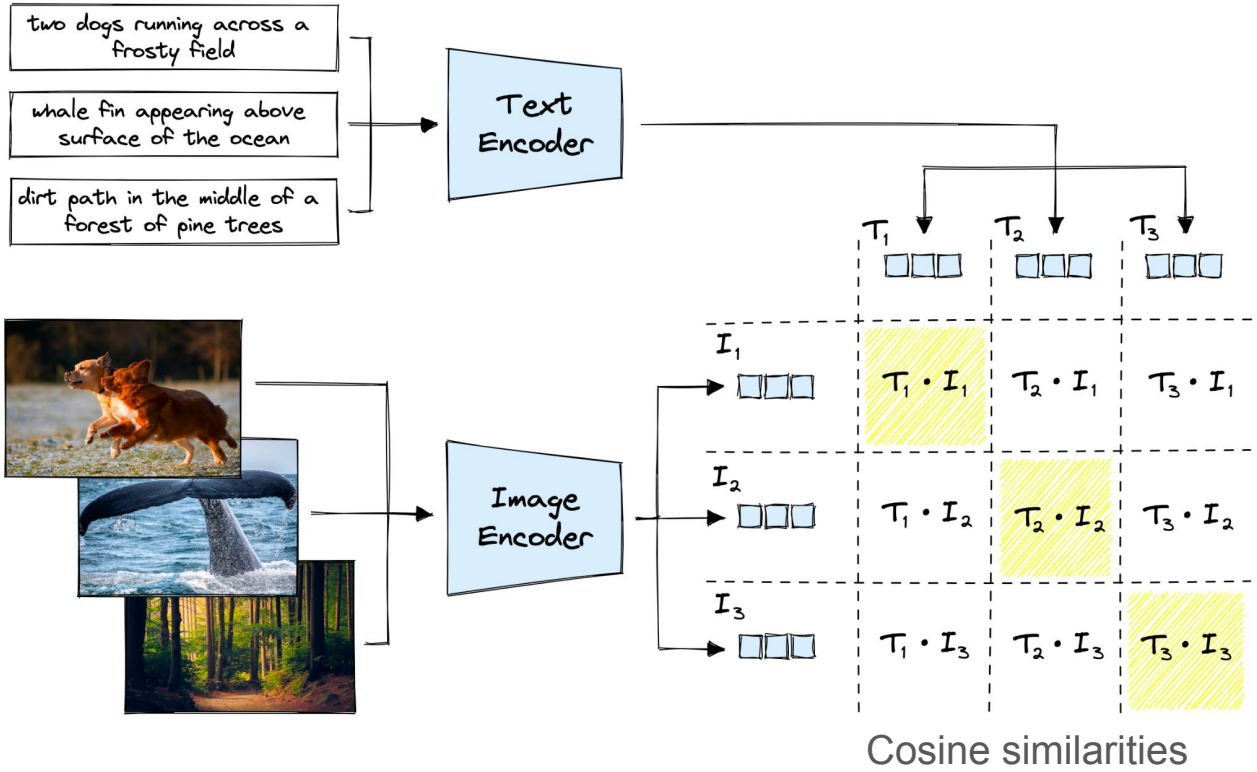
A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

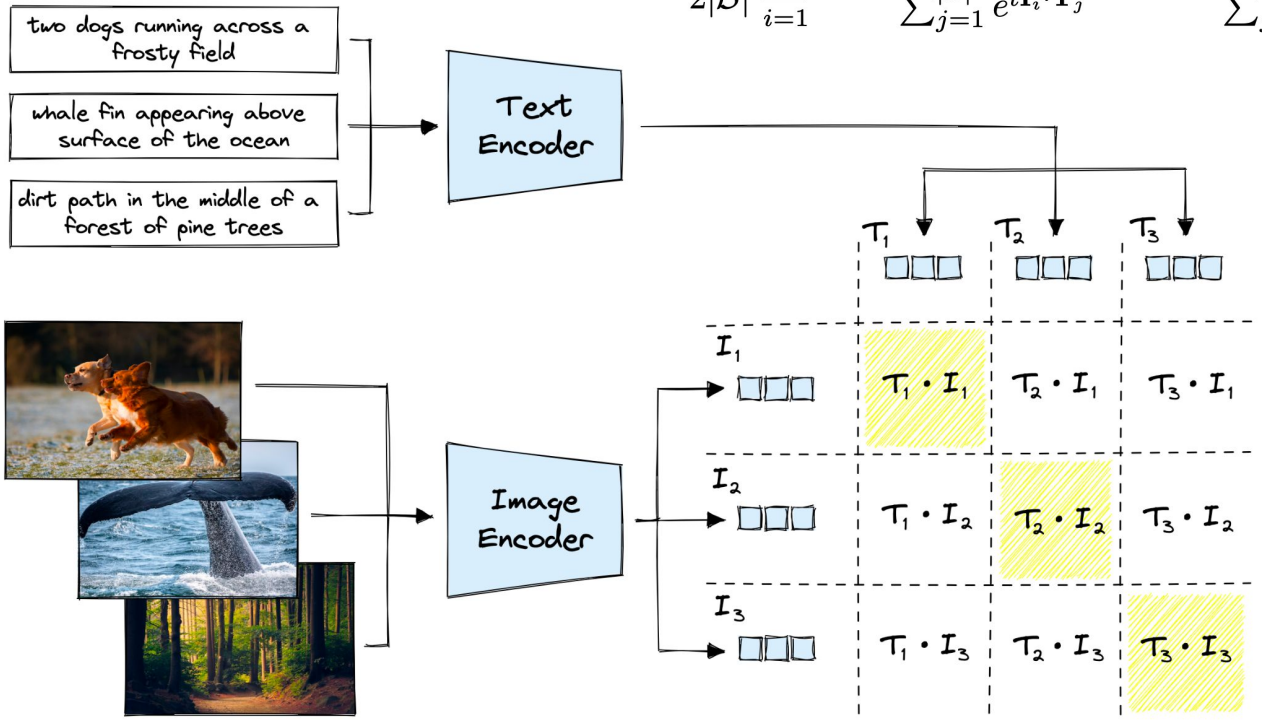
Two networks trained jointly

- Image encoder
- Text encoder

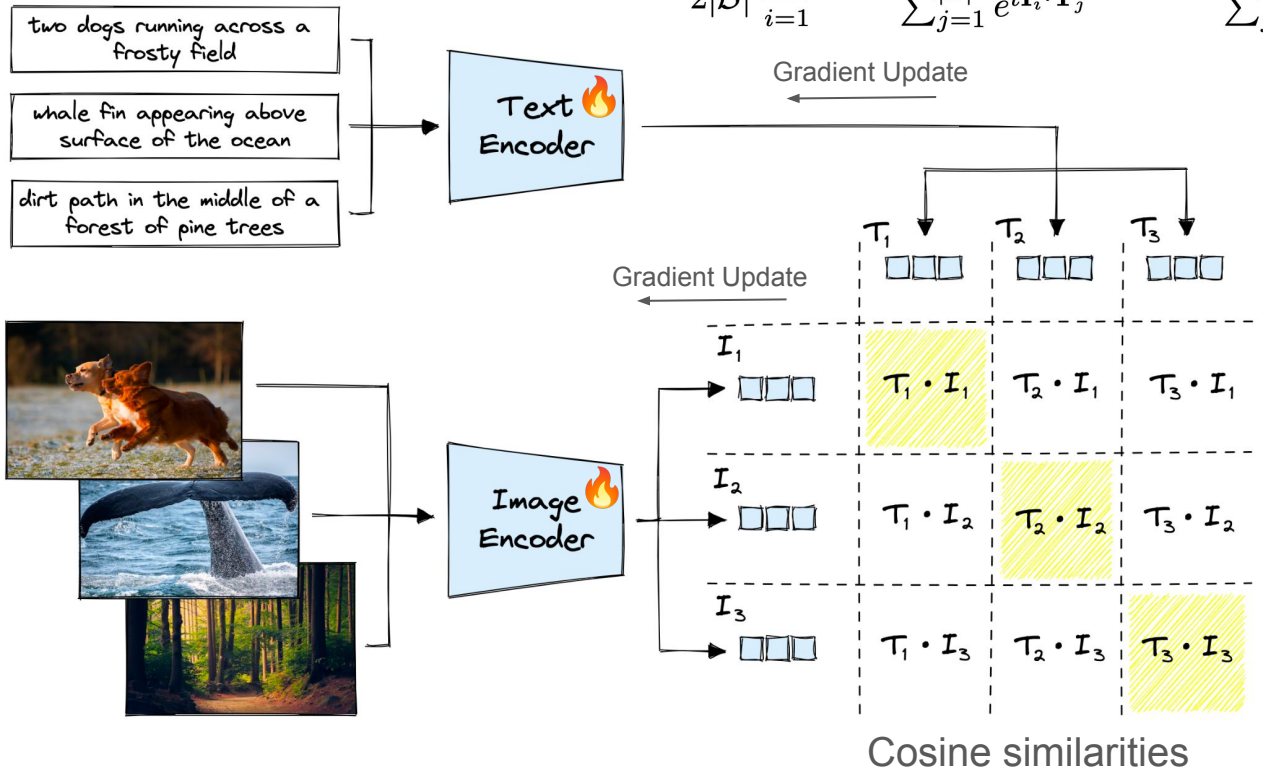

two dogs running across a frosty field
whale fin appearing above surface of the ocean
dirt path in the middle of a forest of pine trees
Text Encoder
T_1
T_2
T_3
Image Encoder
I_1
I_2
I_3


two dogs running across a frosty field
whale fin appearing above surface of the ocean
dirt path in the middle of a forest of pine trees
Text Encoder
Image Encoder
T_1
T_2
T_3
I_1
I_2
I_3
$T_1 \cdot I_1$
$T_2 \cdot I_1$
$T_3 \cdot I_1$
$T_1 \cdot I_2$
$T_2 \cdot I_2$
$T_3 \cdot I_2$
$T_1 \cdot I_3$
$T_2 \cdot I_3$
$T_3 \cdot I_3$
Cosine similarities

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\Big(\overbrace{\log\frac{e^{t\mathbf{I}_i\cdot\mathbf{T}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{I}_i\cdot\mathbf{T}_j}}}^{\text{image}\,\rightarrow\,\text{text softmax}} + \overbrace{\log\frac{e^{t\mathbf{I}_i\cdot\mathbf{T}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{I}_j\cdot\mathbf{T}_i}}}^{\text{text}\,\rightarrow\,\text{image softmax}}\Big)$$

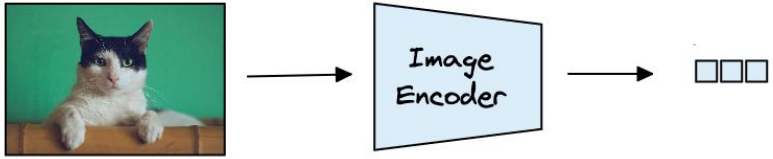


$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\Big(\overbrace{\log\frac{e^{t\mathbf{I}_i\cdot\mathbf{T}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{I}_i\cdot\mathbf{T}_j}}}^{\text{image}\,\rightarrow\,\text{text softmax}} + \overbrace{\log\frac{e^{t\mathbf{I}_i\cdot\mathbf{T}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{I}_j\cdot\mathbf{T}_i}}}^{\text{text}\,\rightarrow\,\text{image softmax}}\Big)$$

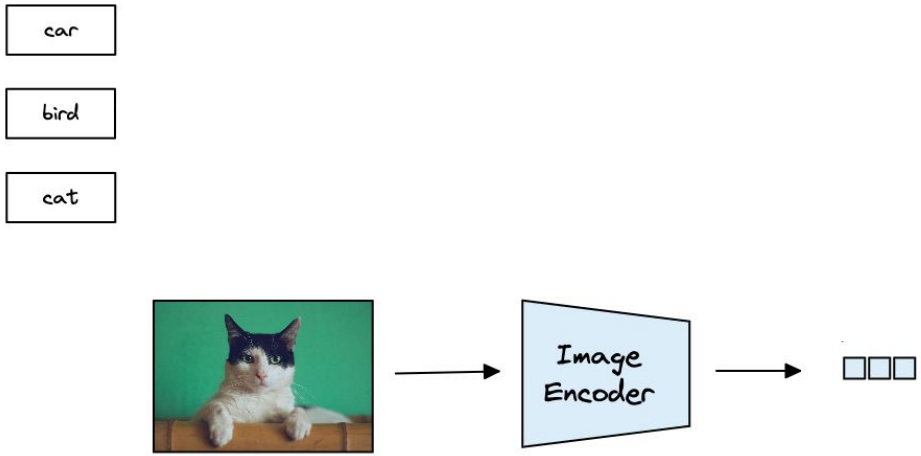


Free Zero-shot classification

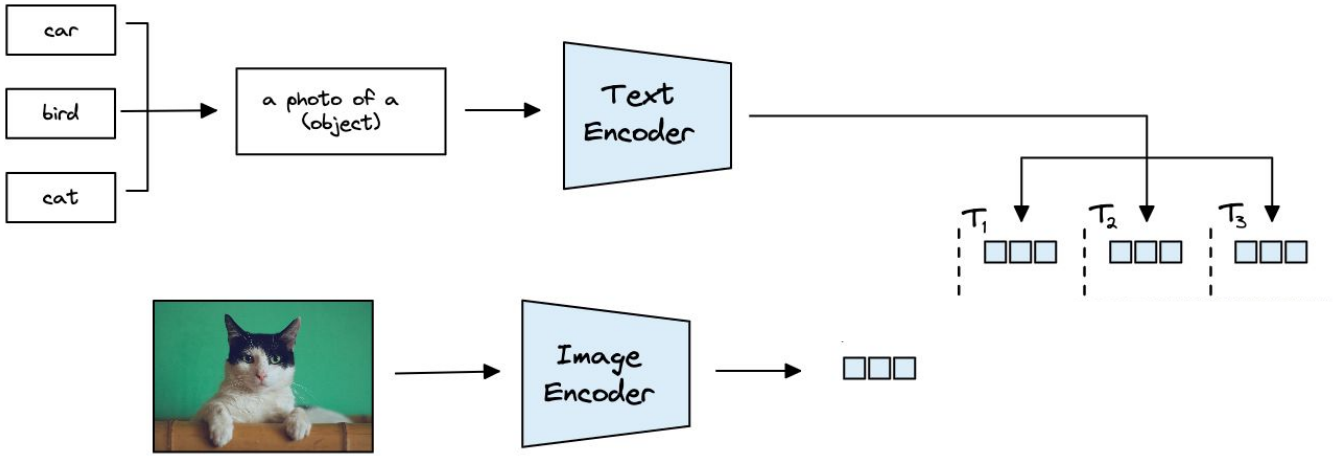
Free Zero-shot classification



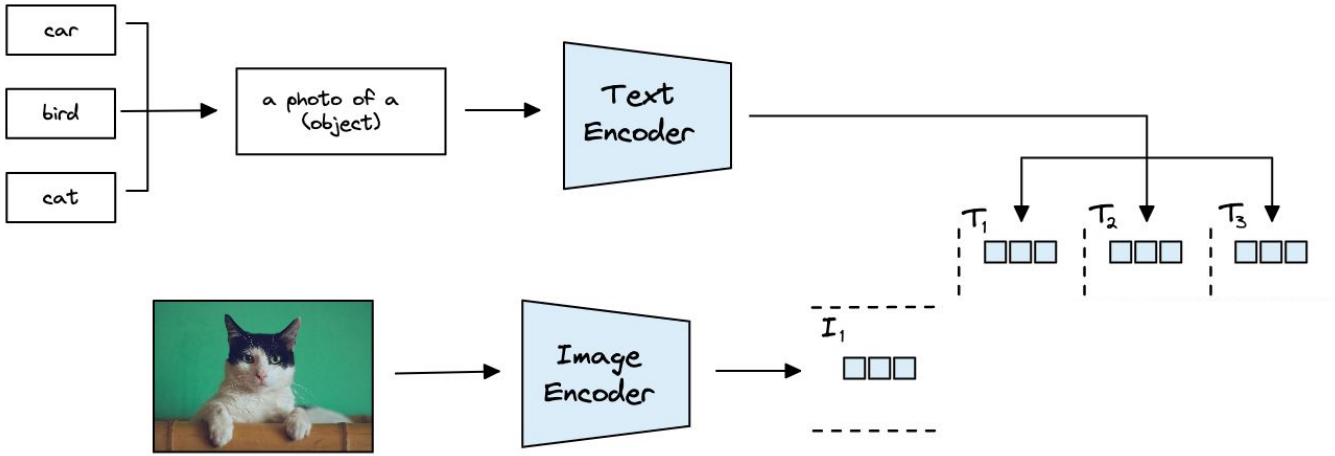
Free Zero-shot classification



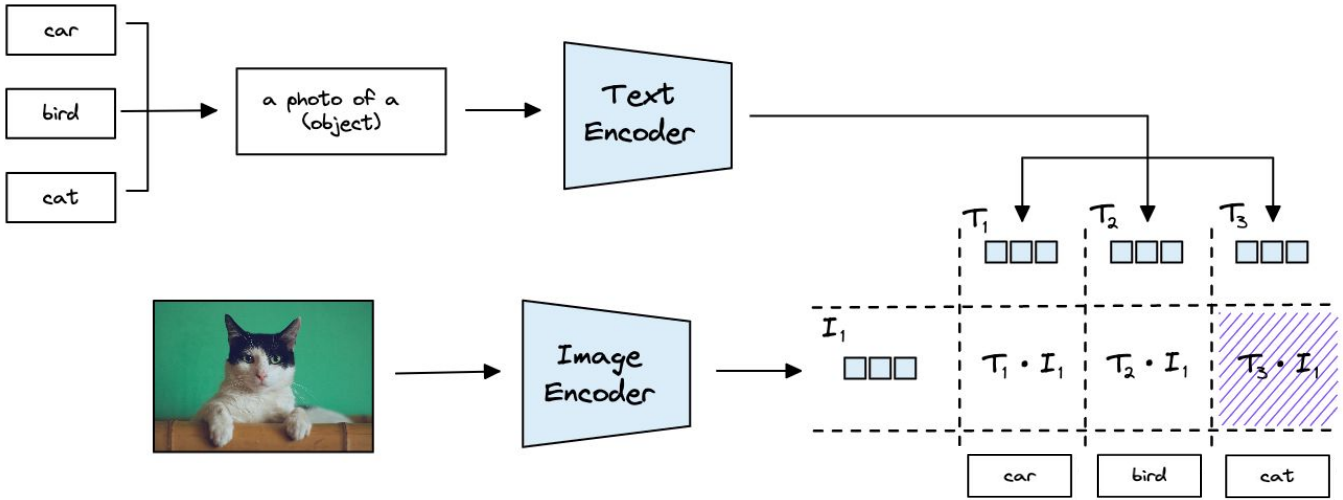
Free Zero-shot classification



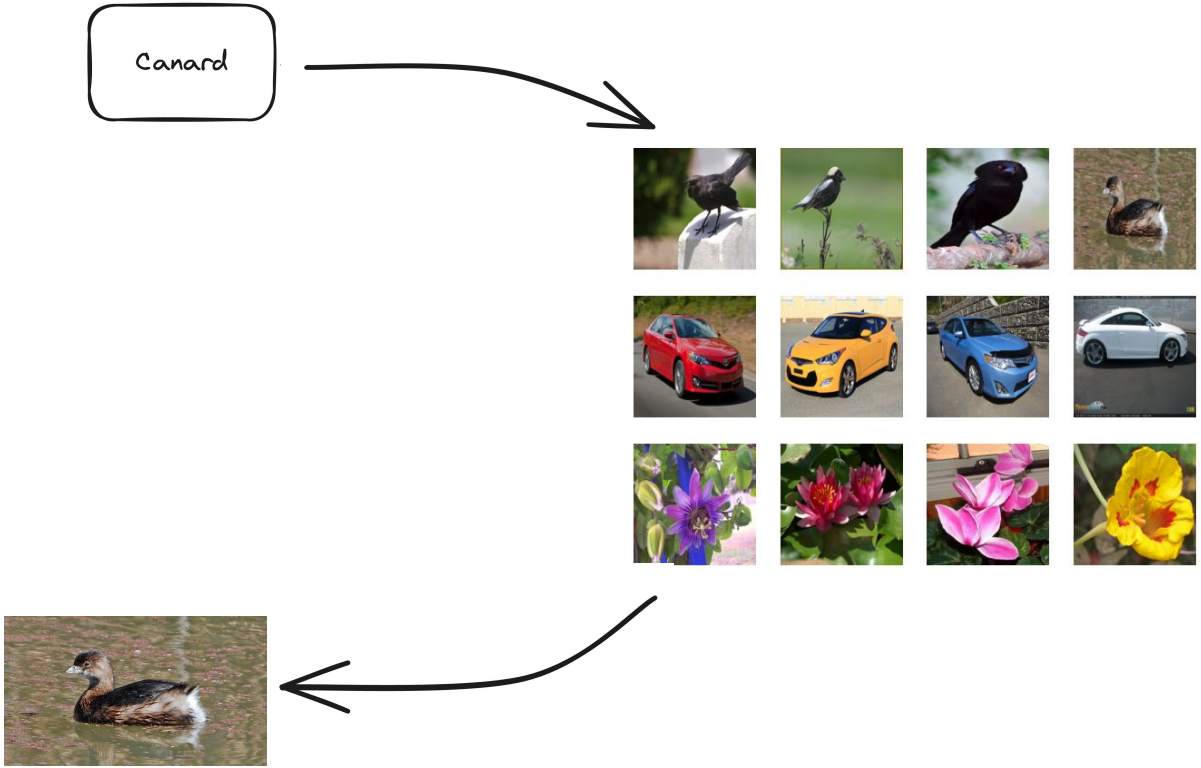
Free Zero-shot classification



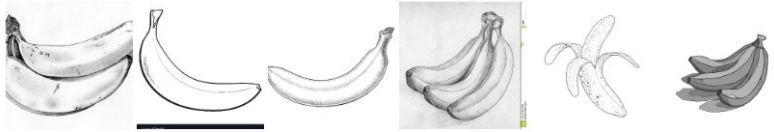
Free Zero-shot classification



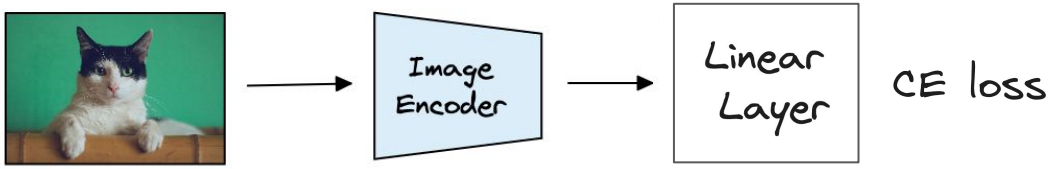
Perfect for Image retrieval



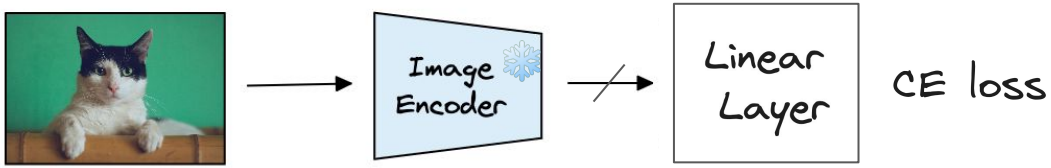
Performance

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%

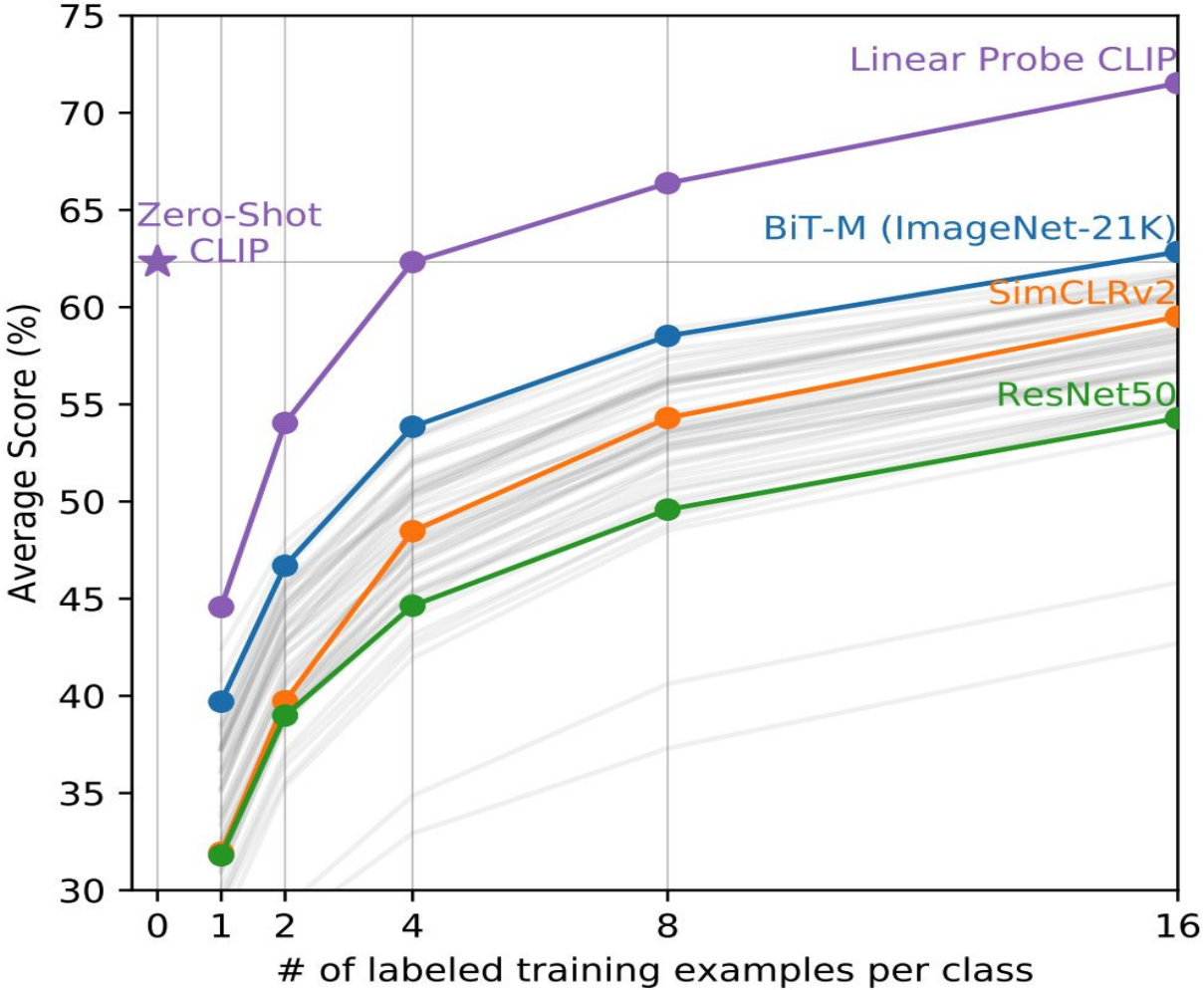
Linear Probing



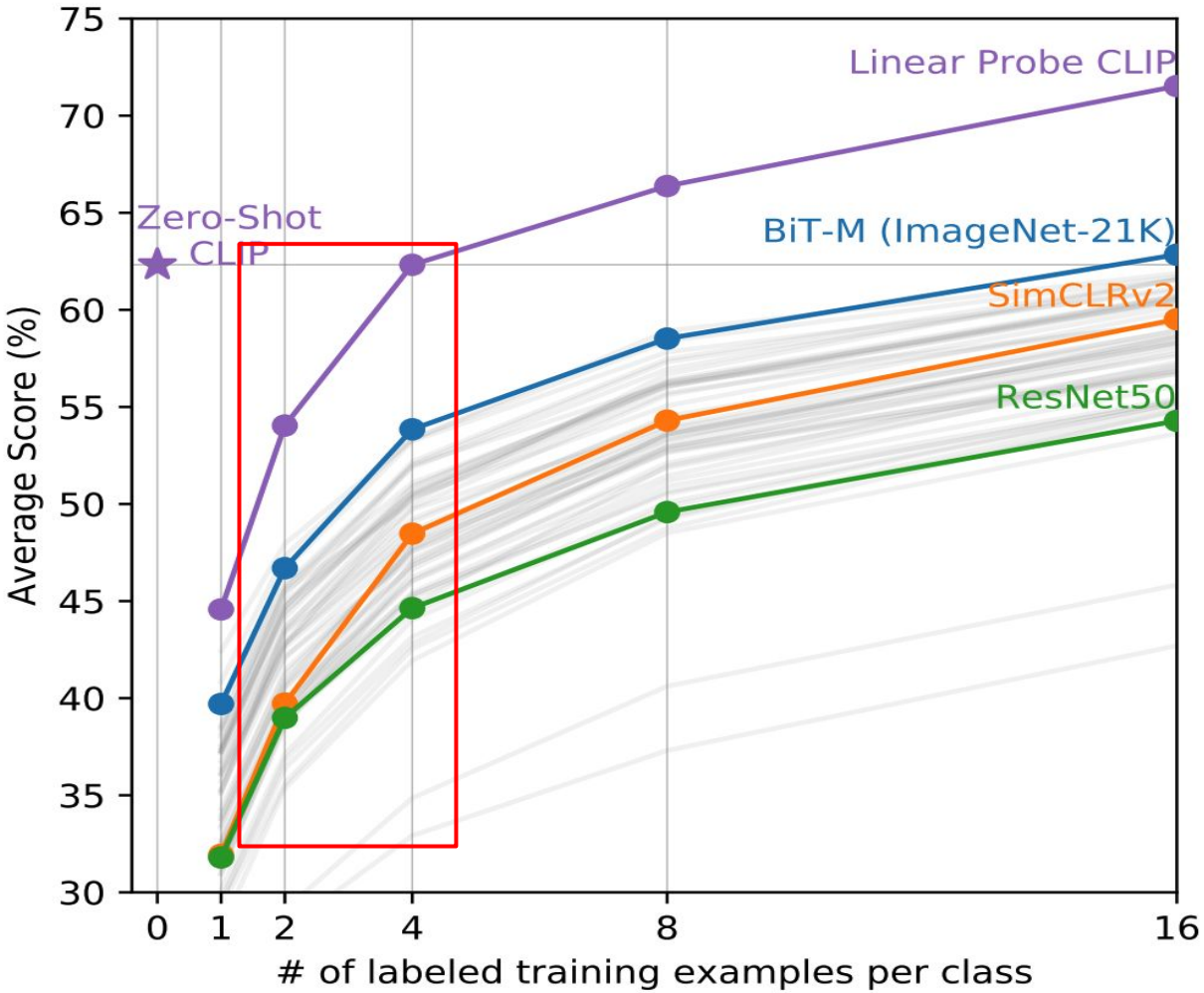
Linear Probing



Performance: Linear probing



Performance: Linear probing



Prompt tuning

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29

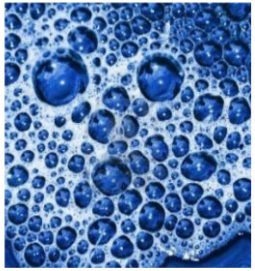
(a)

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

(b)

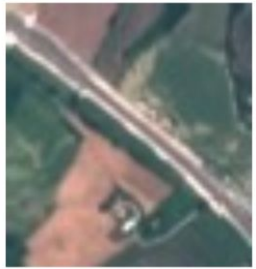
Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32

(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

(d)

Prompt tuning

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**

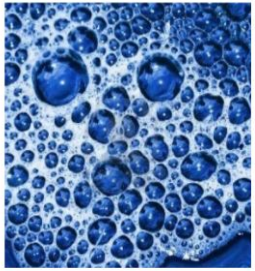
(a)

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)

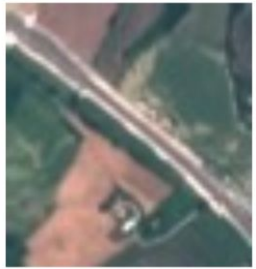
Describable Textures (DTD)



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)

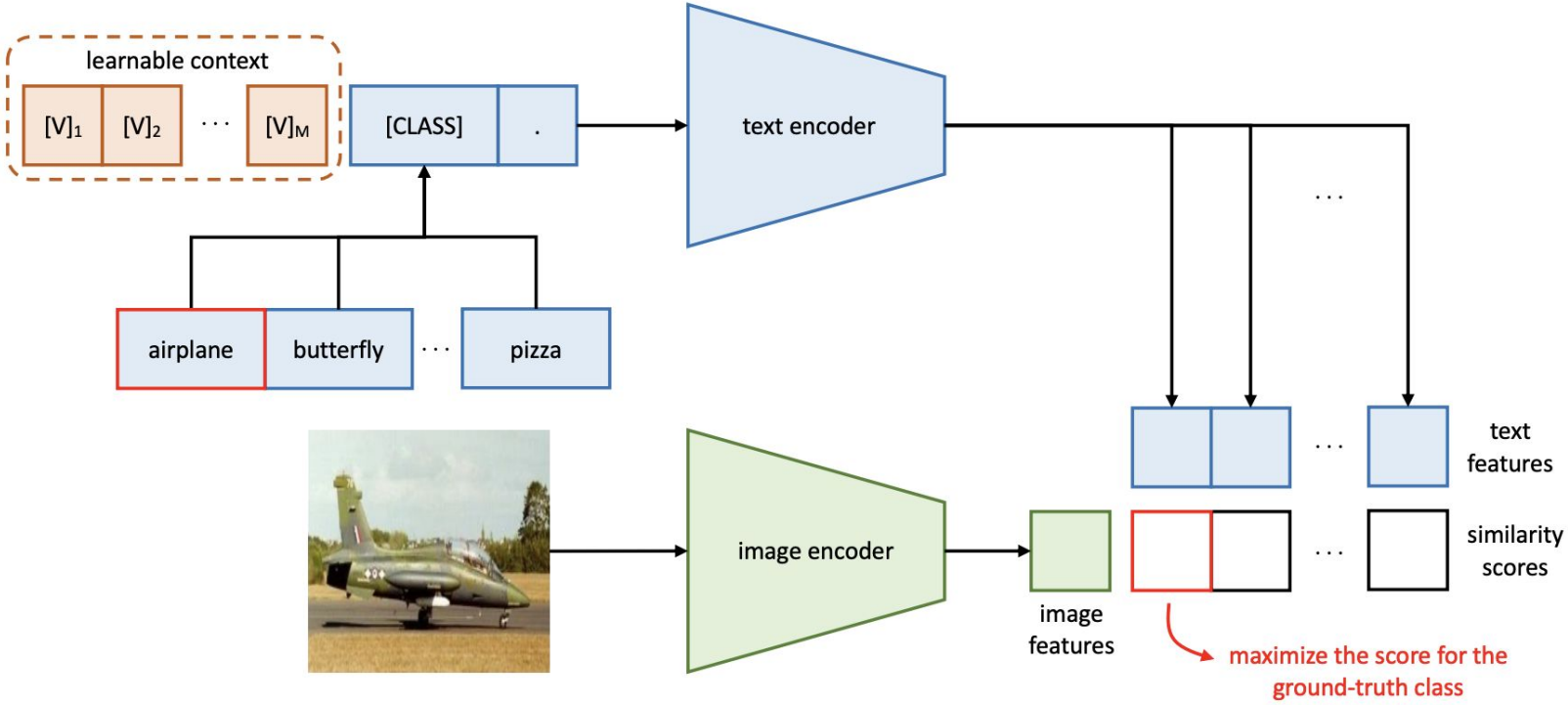
EuroSAT



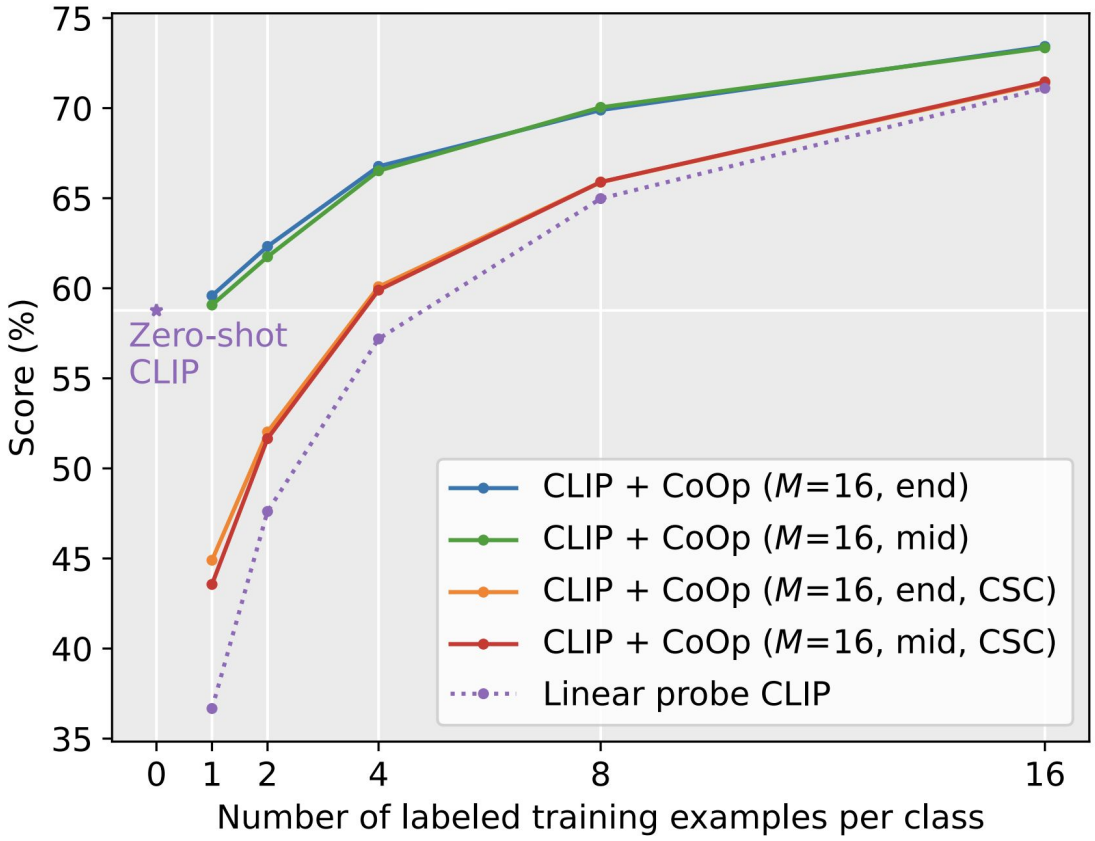
Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

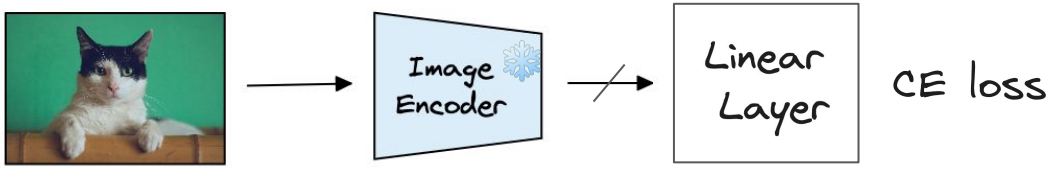
Prompt tuning



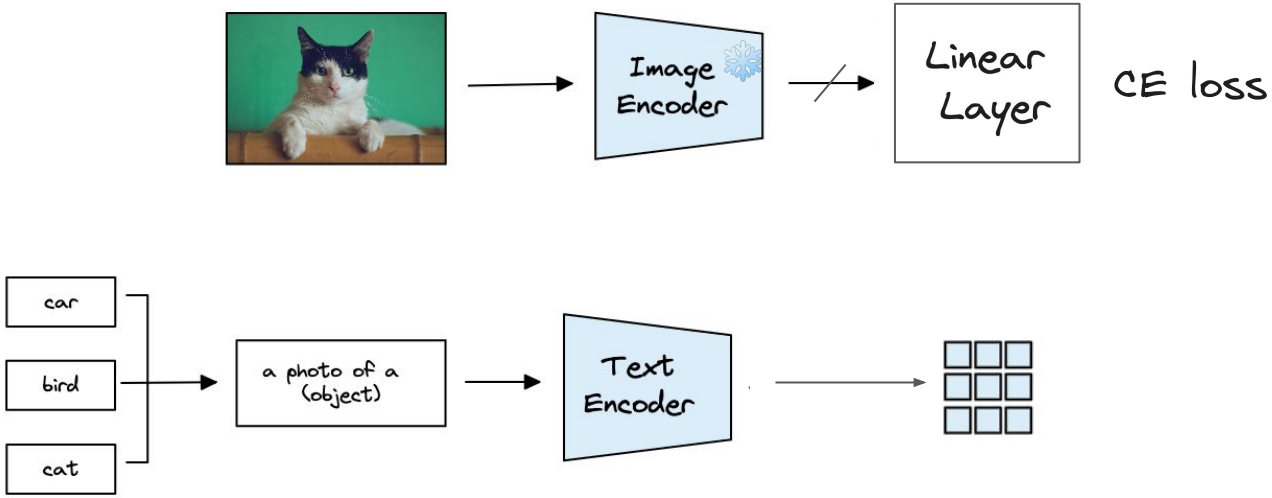
Prompt tuning



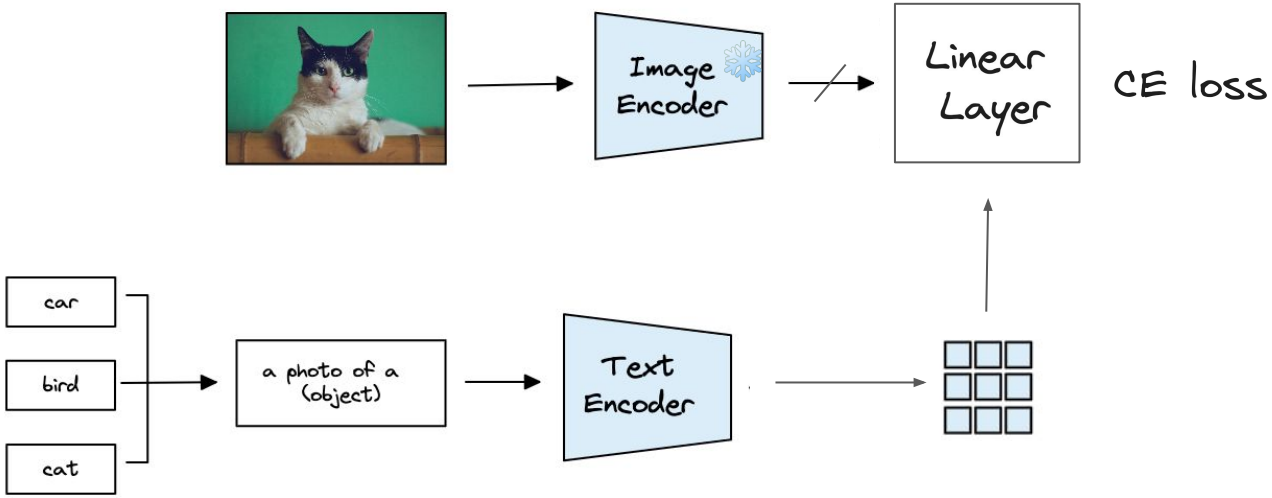
Few-shot Adaptation



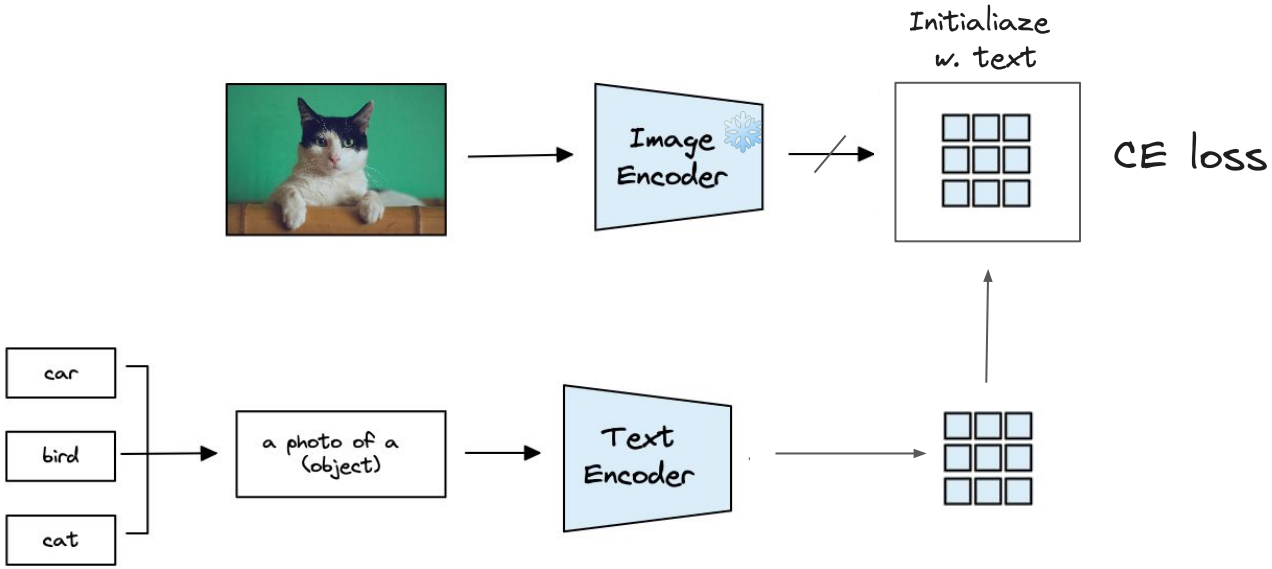
Few-shot Adaptation



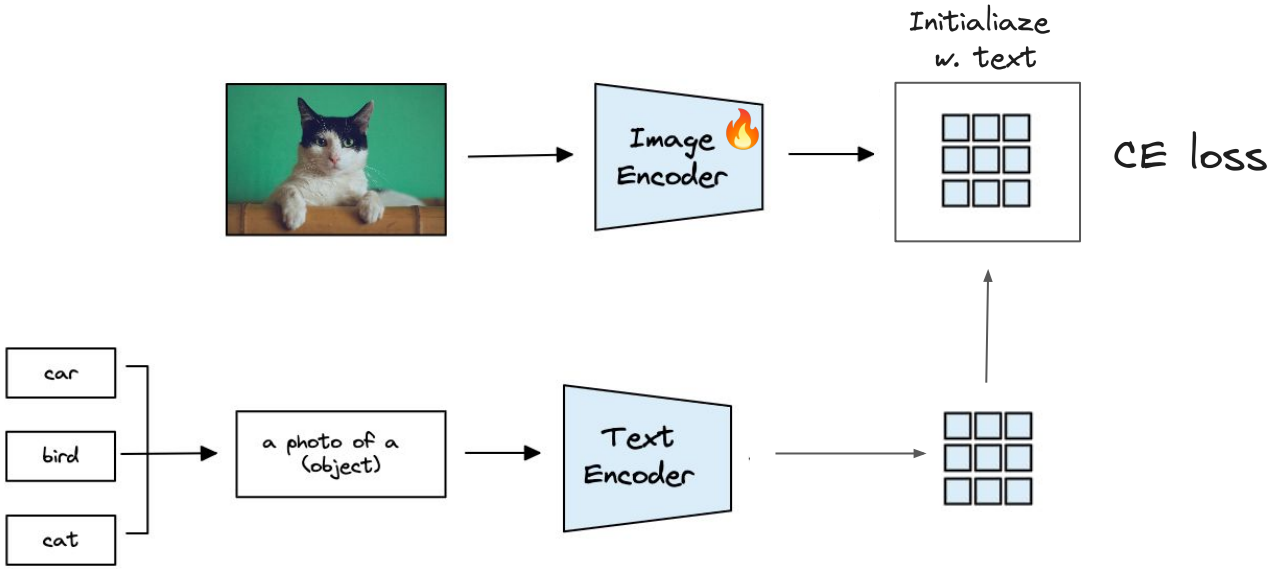
Few-shot Adaptation



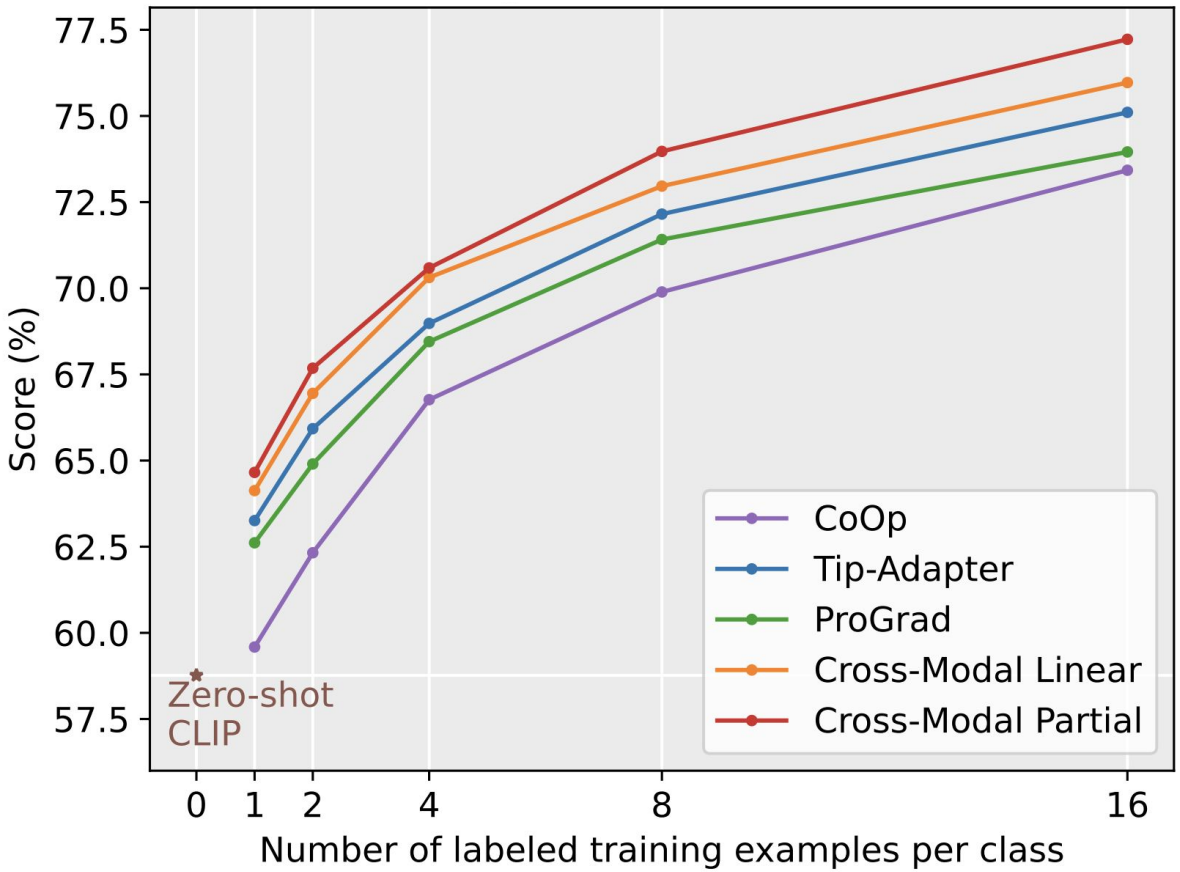
Few-shot Adaptation



Few-shot Adaptation



Multi-modal Few-shot



Better CLIPs

Open-CLIP

Reproducible scaling laws for contrastive language-image learning

Mehdi Cherti[1,5] §§ **Romain Beaumont**[1] §§ **Ross Wightman**[1,3] §§
Mitchell Wortsman[4] §§ **Gabriel Ilharco**[4] §§ **Cade Gordon**[2]
Christoph Schuhmann[1] **Ludwig Schmidt**[4] °° **Jenia Jitsev**[1,5] §§°°
LAION[1] UC Berkeley[2] HuggingFace[3] University of Washington[4]
Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)[5]
contact@laion.ai, {m.cherti,j.jitsev}@fz-juelich.de
§§ Equal first contributions, °° Equal senior contributions

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS

COMMON CRAWL

Similar performance to OpenAI

OpenAI

Model	Training data	Resolution	ImageNet zero-shot acc.
ViT-B/32	DataComp-1B	256px	72.8%
ViT-B/16	DataComp-1B	224px	73.5%
ViT-L/14	LAION-2B	224px	75.3%
ViT-H/14	LAION-2B	224px	78.0%
ViT-L/14	DataComp-1B	224px	79.2%
ViT-G/14	LAION-2B	224px	80.1%
ViT-L/14	OpenAI's WIT	224px	75.5%

ImageBind

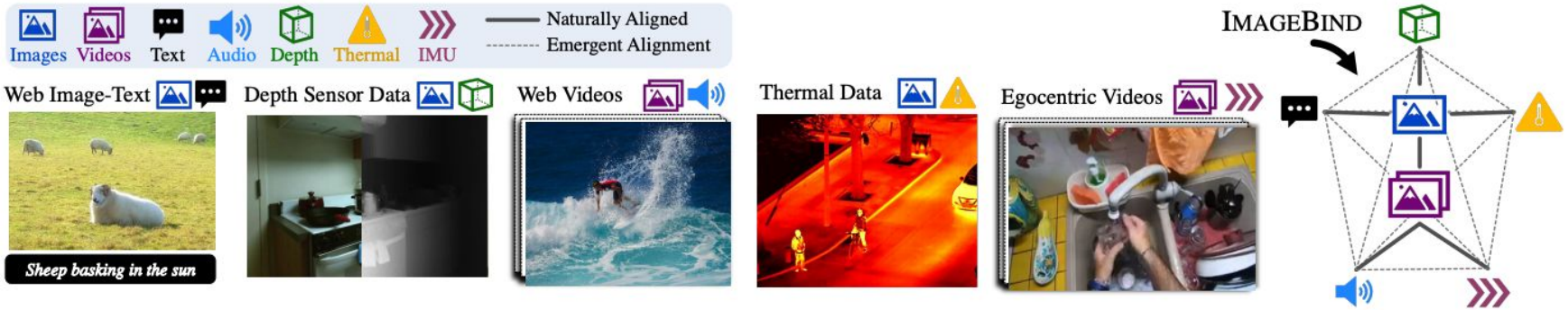
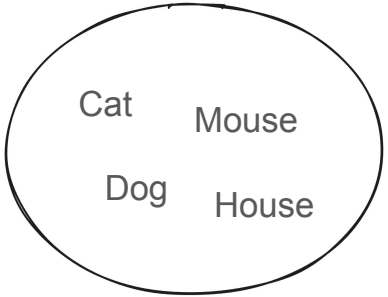


Figure 2. IMAGEBIND overview. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, *etc*. IMAGEBIND links all these modalities in a common embedding space, enabling new emergent alignments and capabilities.

Open-vocabulary models using Text

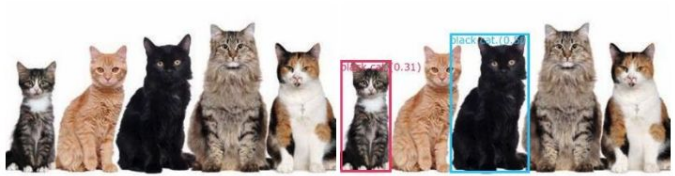
1. Classification



Closed set of classes

All vocabulary

2. Object Detection



Detection Prompt: `black cat`

3. Image Segmentation

Visual Question Answering (VQA)