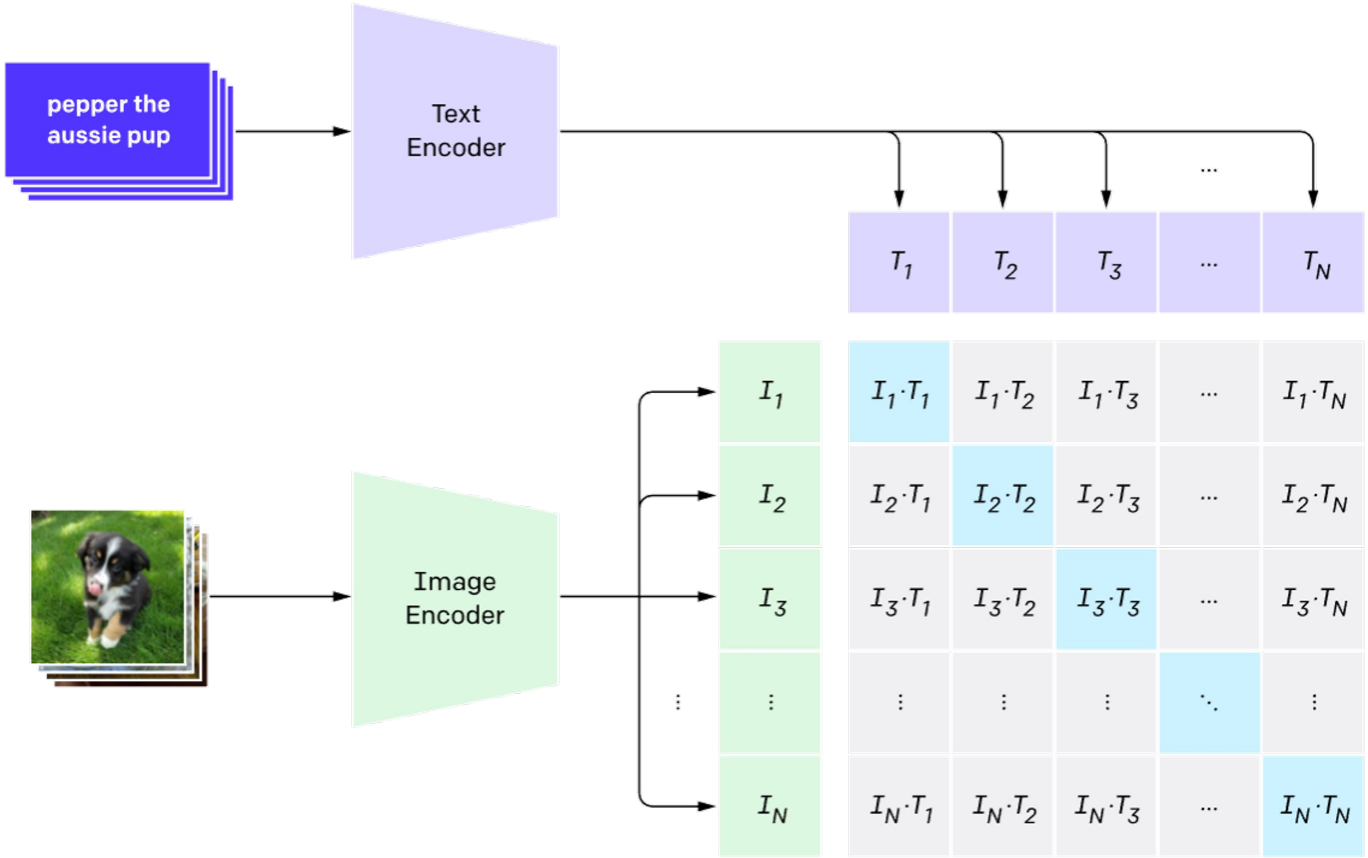
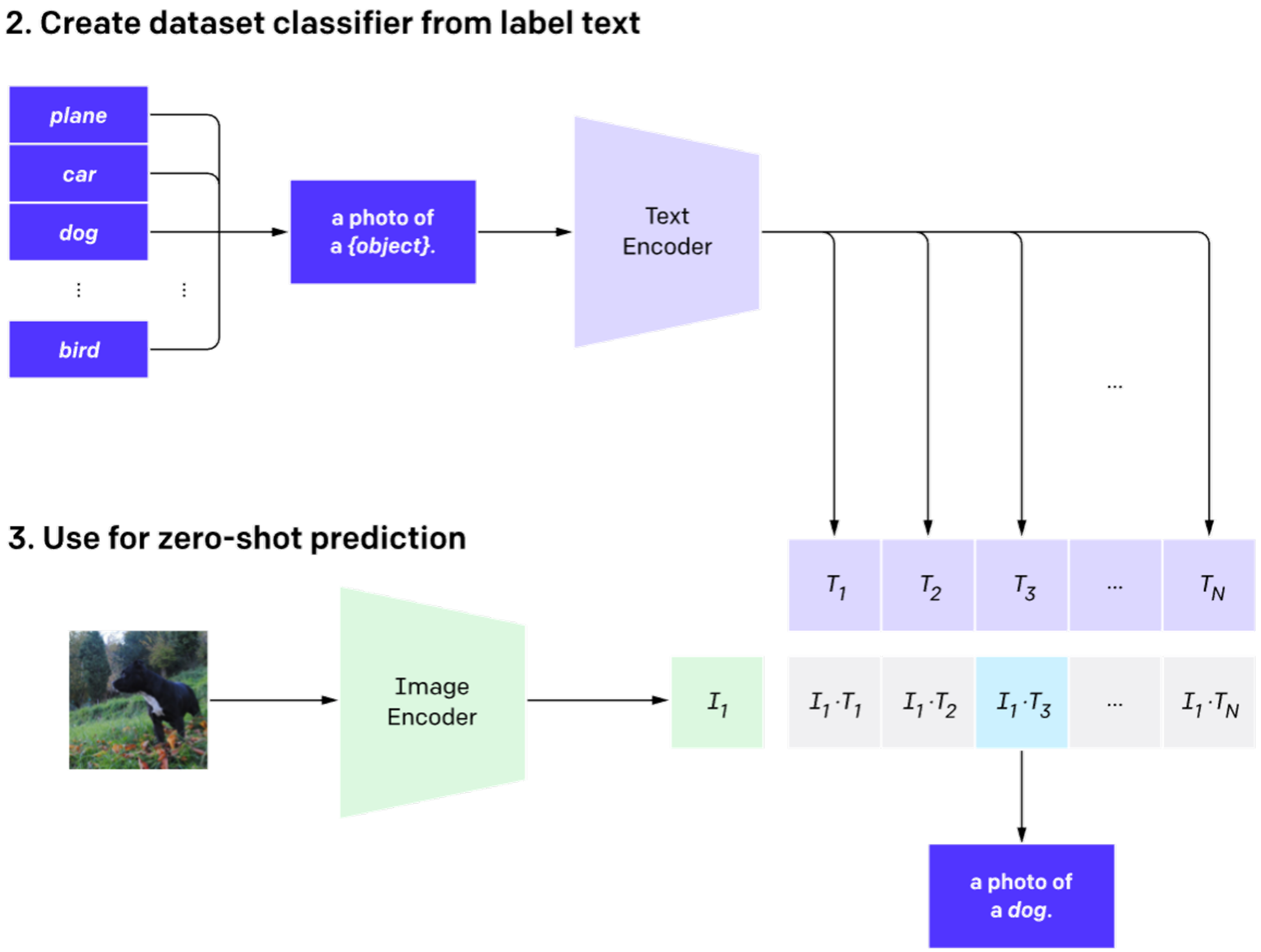

CLIP
One of the 10 most relevant works in the last decade in AI
Zero-shot adaptation to any visual classification task

1. Contrastive pre-training



Radford, Alec, et al. "Learning transferable visual models from natural language supervision."ICML , 2021.

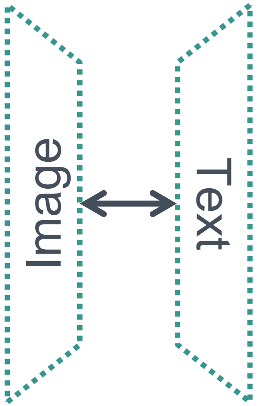


Radford, Alec, et al. "Learning transferable visual models from natural language supervision."ICML , 2021.

CLIP paradigm evolution

= Frozen pretrained model

= trained from scratch



CLIP

CLIP paradigm evolution

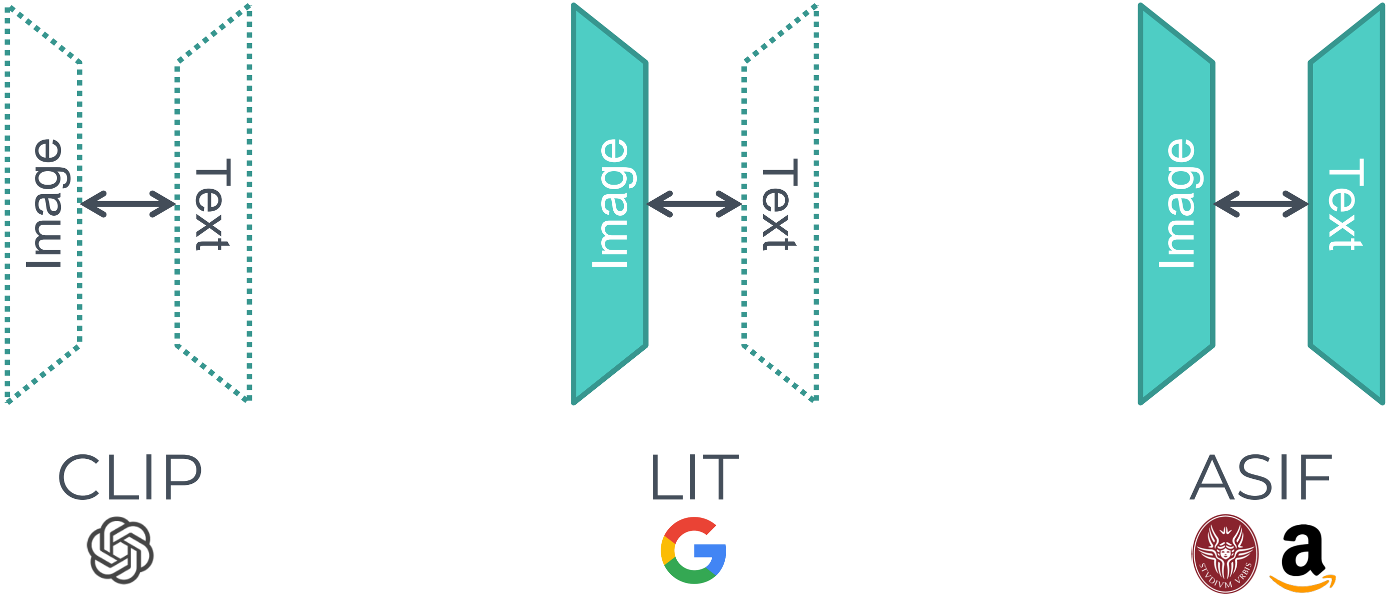
= Frozen pretrained model = trained from scratch

CLIP

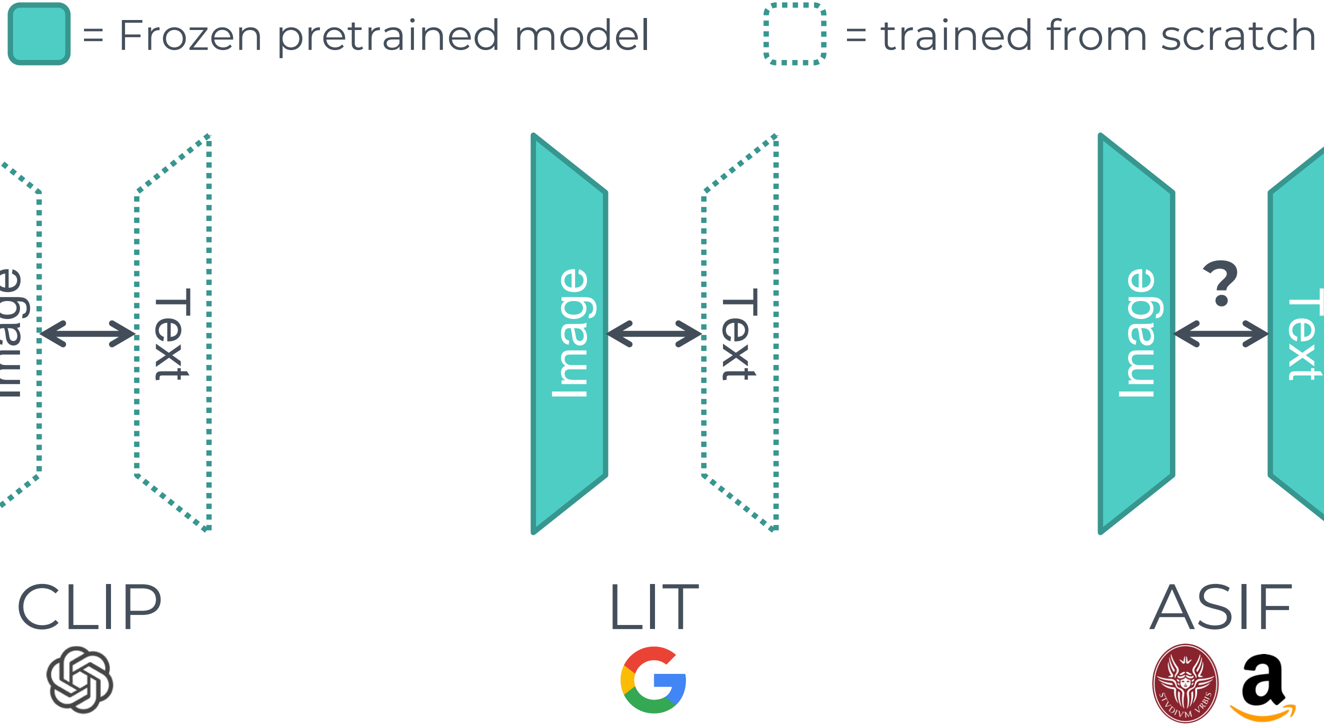
LIT

CLIP paradigm evolution

= Frozen pretrained model   = trained from scratch

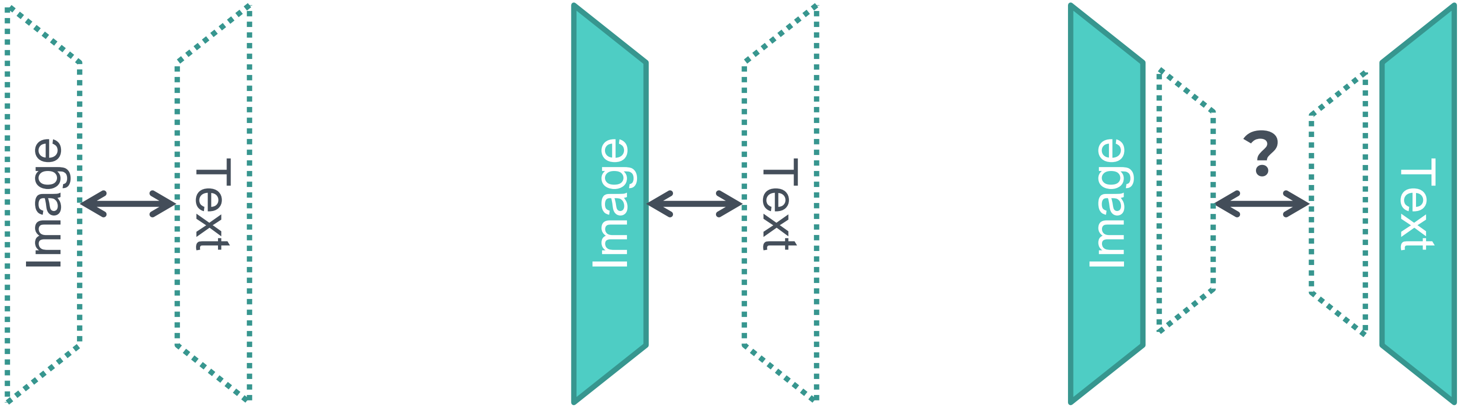


CLIP paradigm evolution



CLIP paradigm evolution

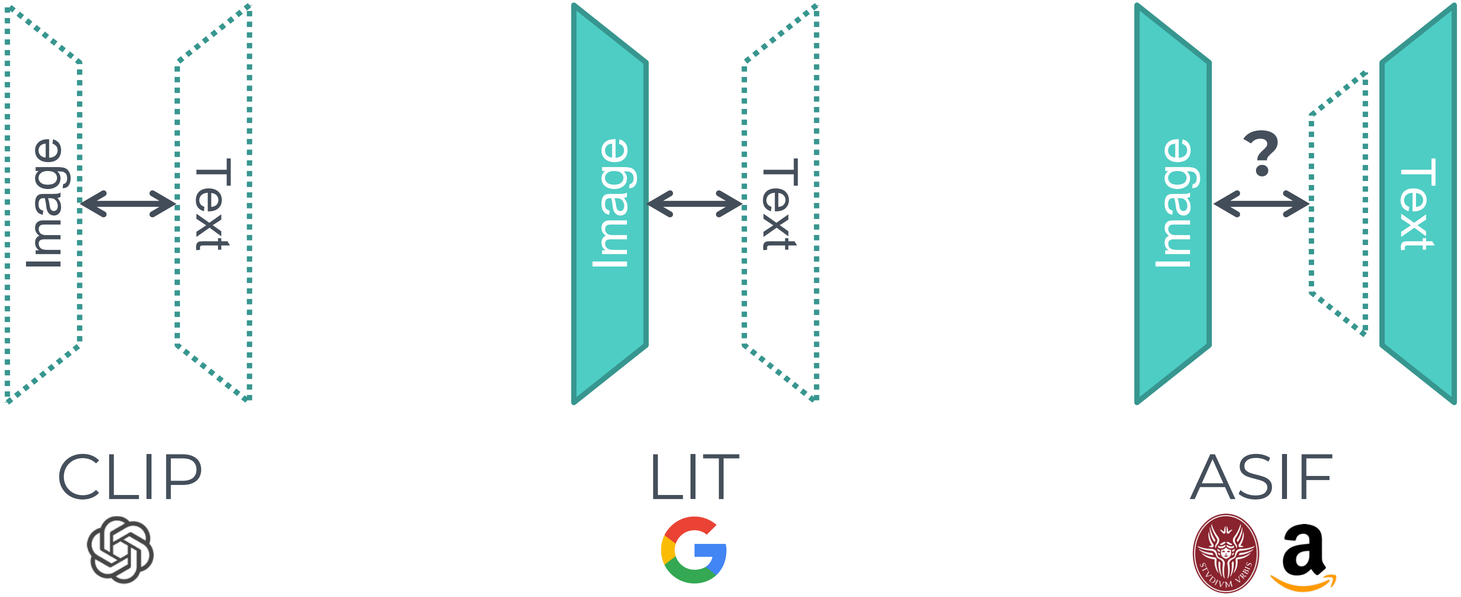
= Frozen pretrained model   = trained from scratch



CLIP   LIT   ASIF

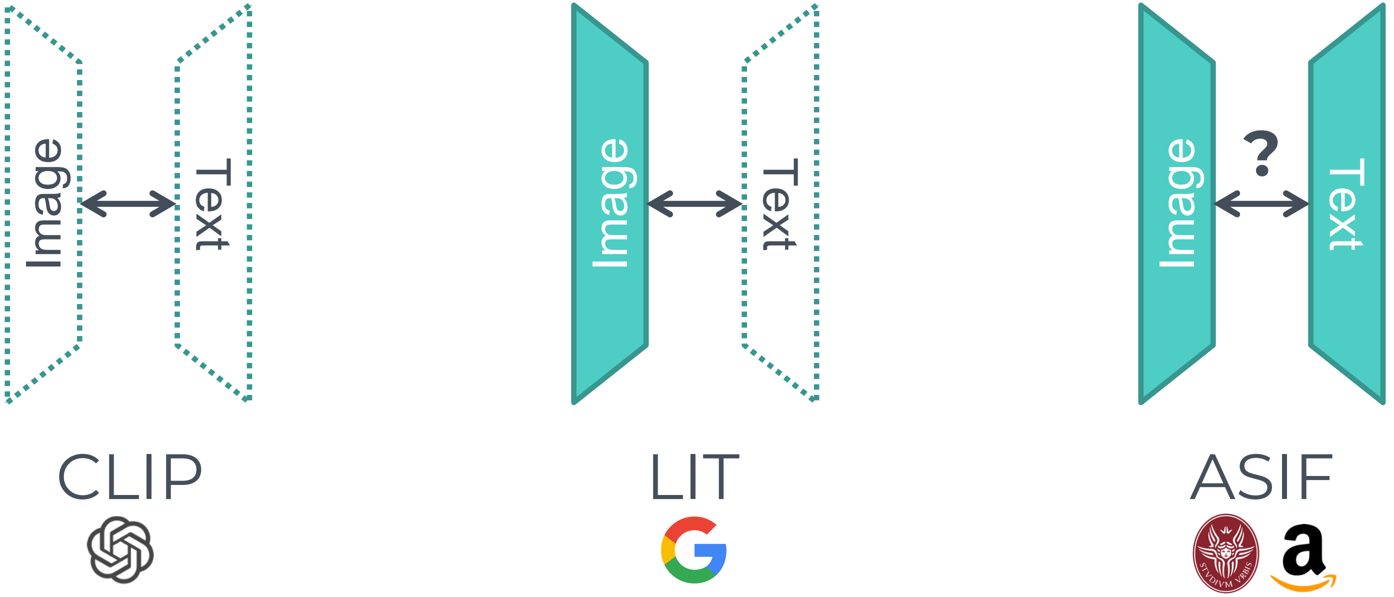
CLIP paradigm evolution

= Frozen pretrained model   = trained from scratch



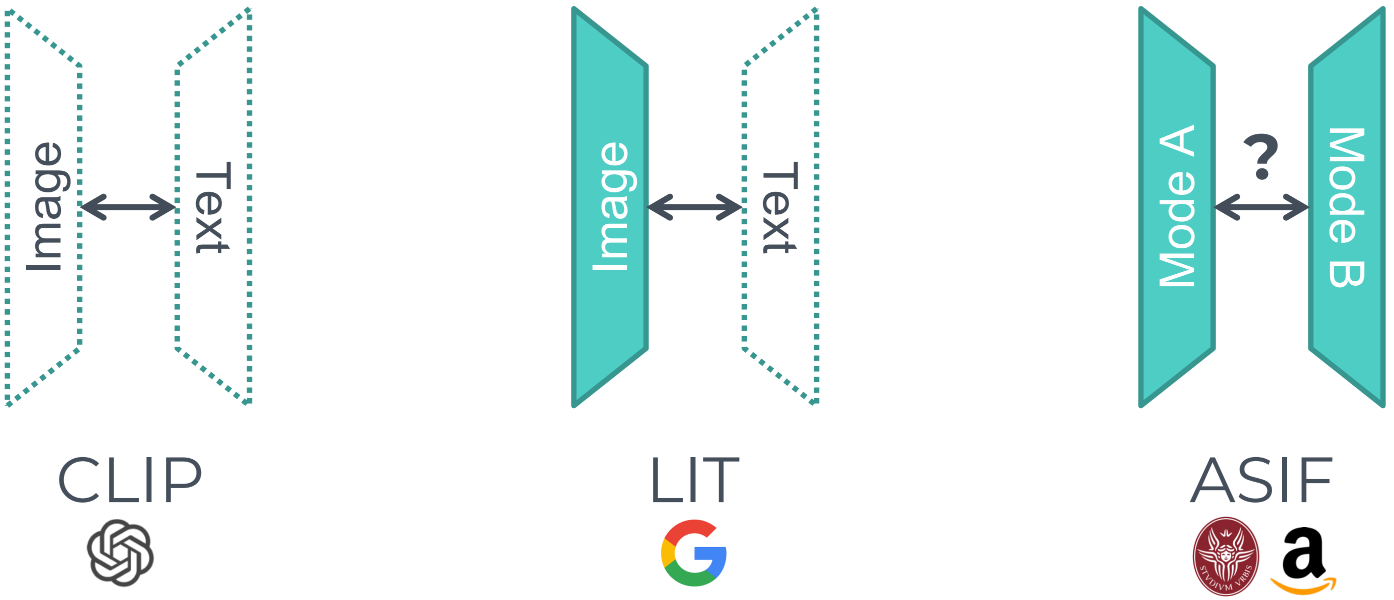
CLIP paradigm evolution

= Frozen pretrained model   = trained from scratch



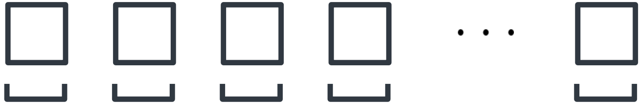
CLIP paradigm evolution

= Frozen pretrained model = trained from scratch



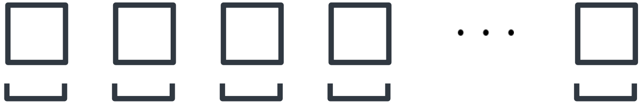
ASIF

a green car in the forest



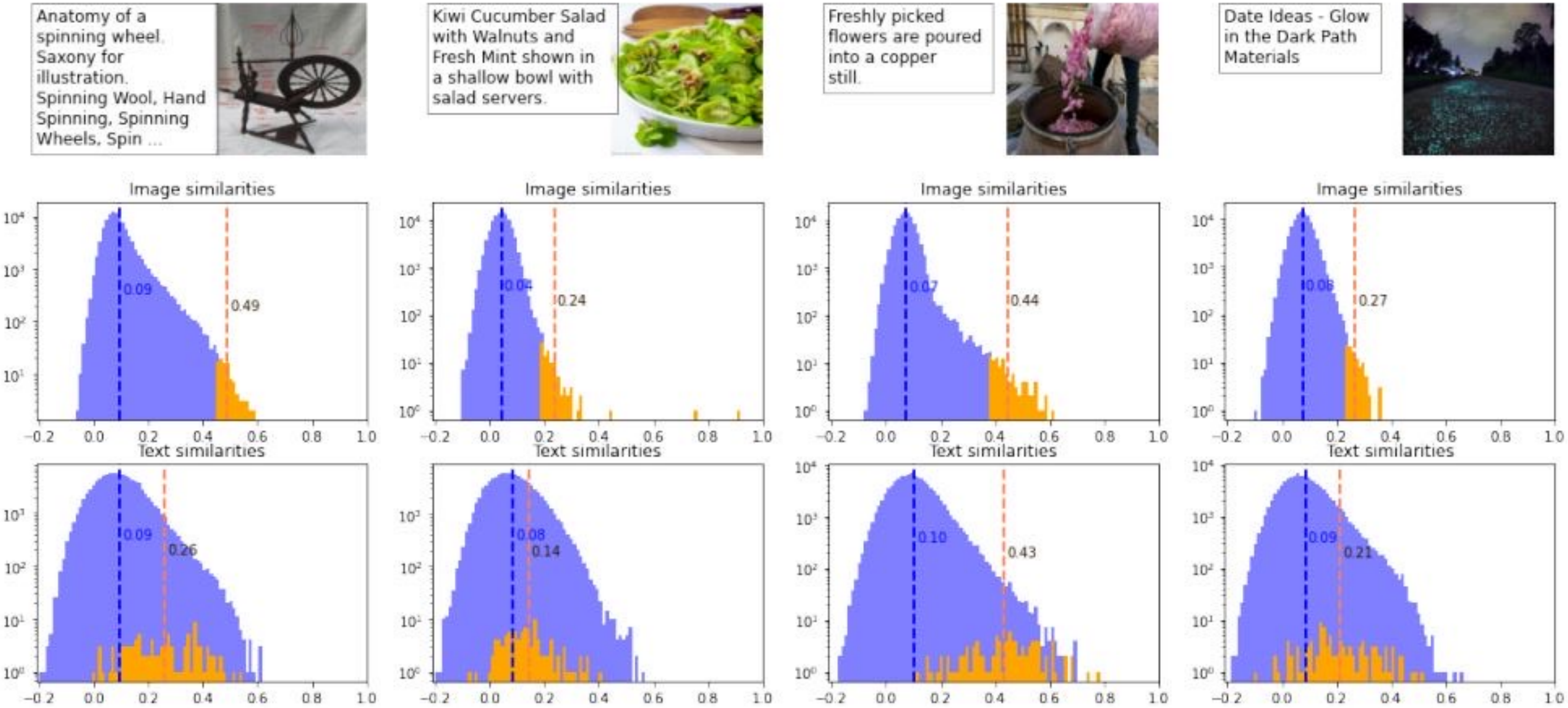
Multimodal dataset

Captions of similar images are themselves similar



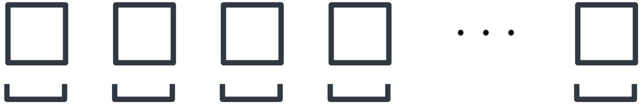
Multimodal dataset

a green car in the forest



ASIF

Problem: best caption for a given image



Multimodal dataset

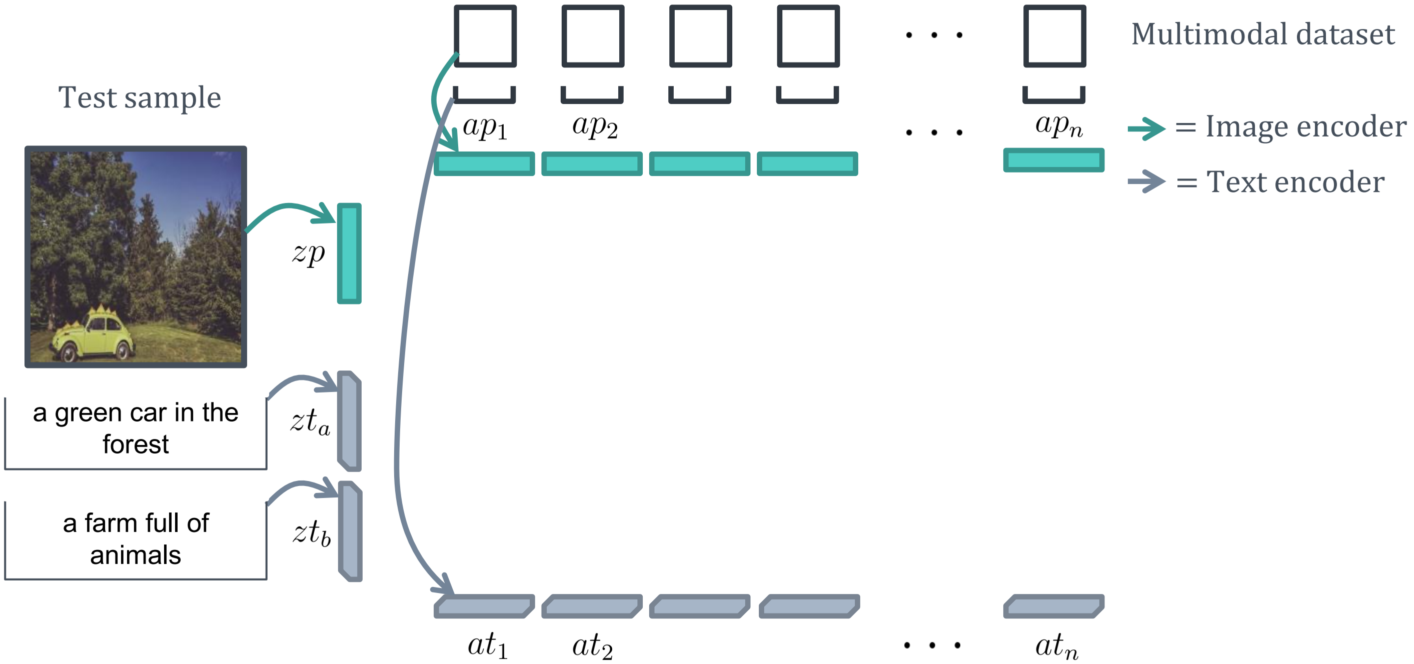
Test sample

a green car in the forest

a farm full of animals

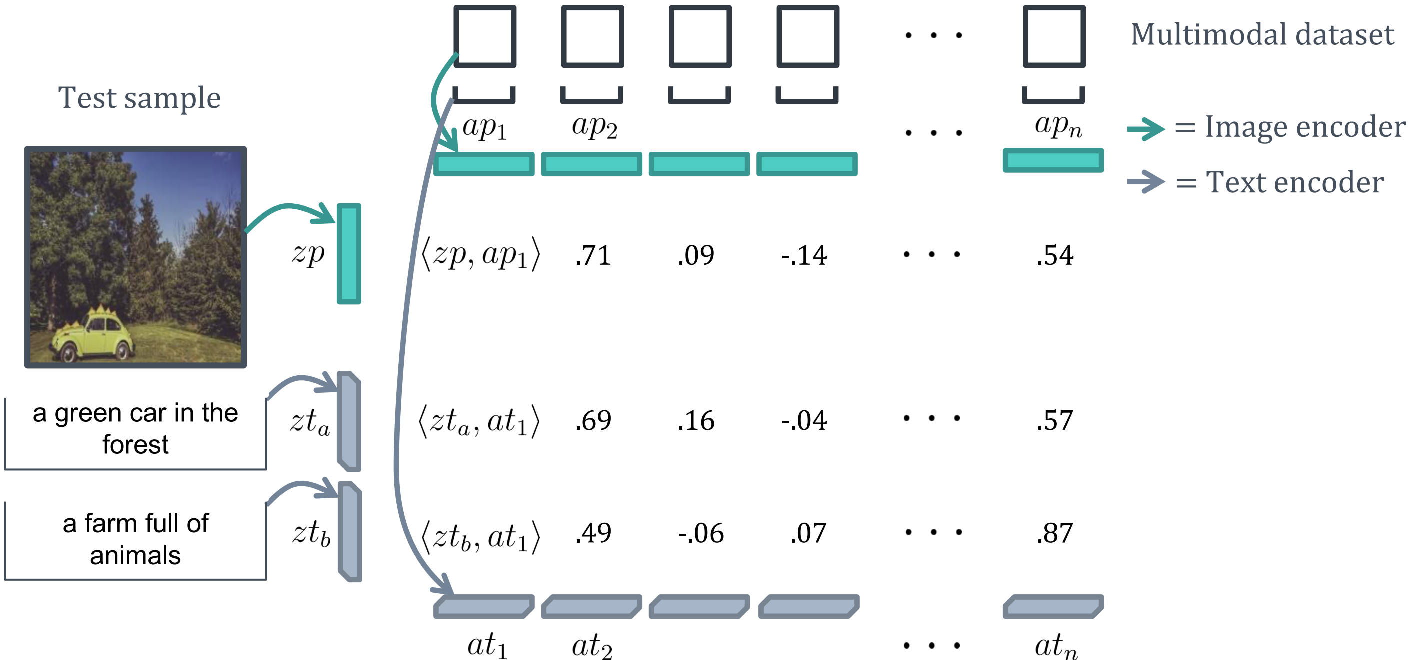
ASIF

Problem: best caption for a given image



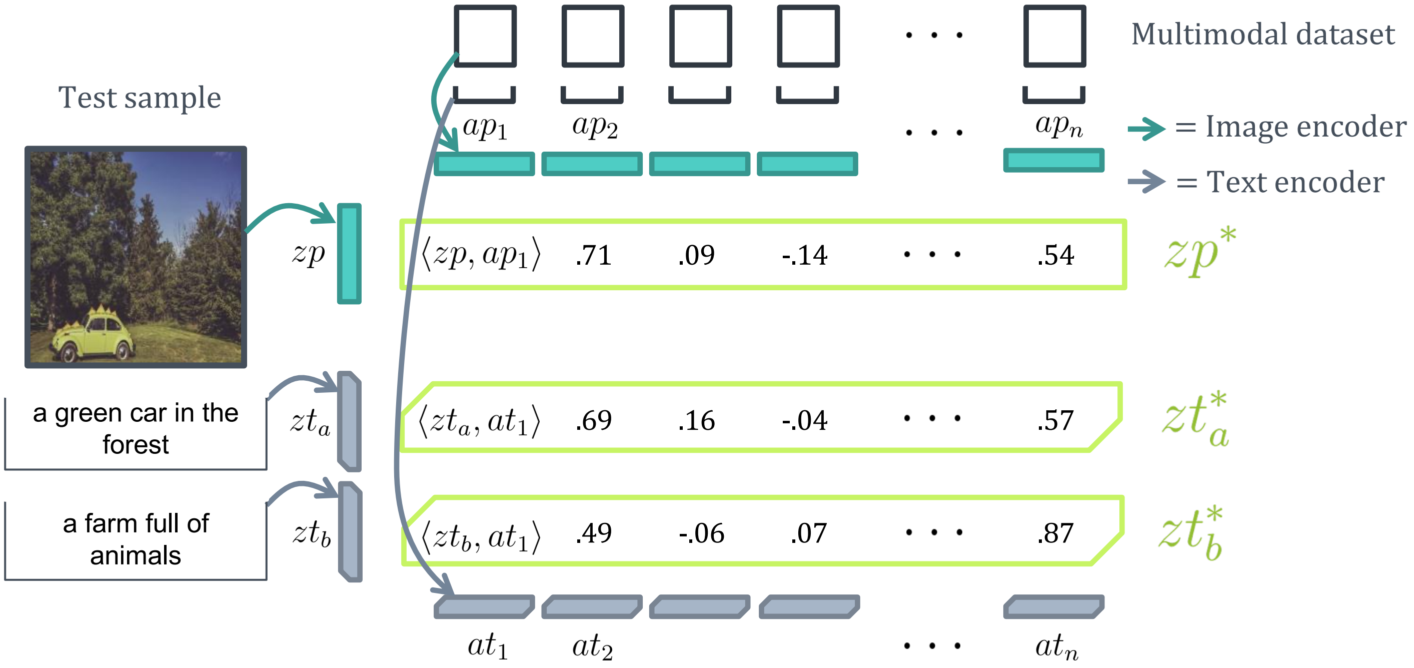
ASIF

Problem: best caption for a given image



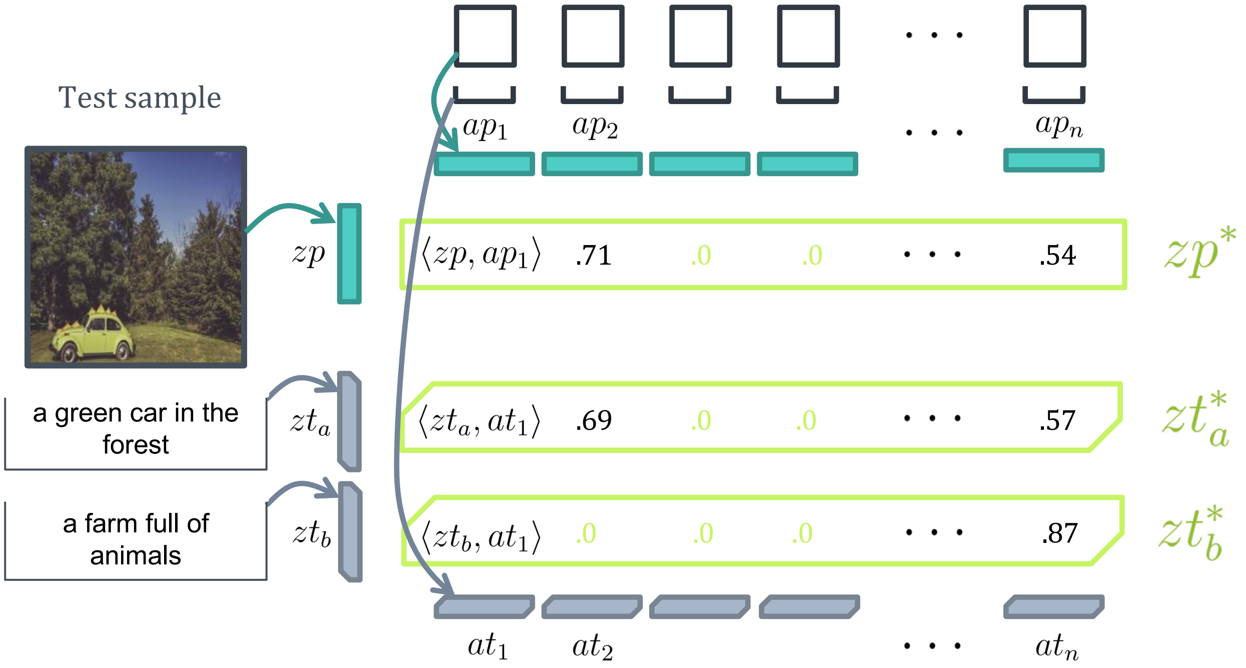
ASIF

Problem: best caption for a given image



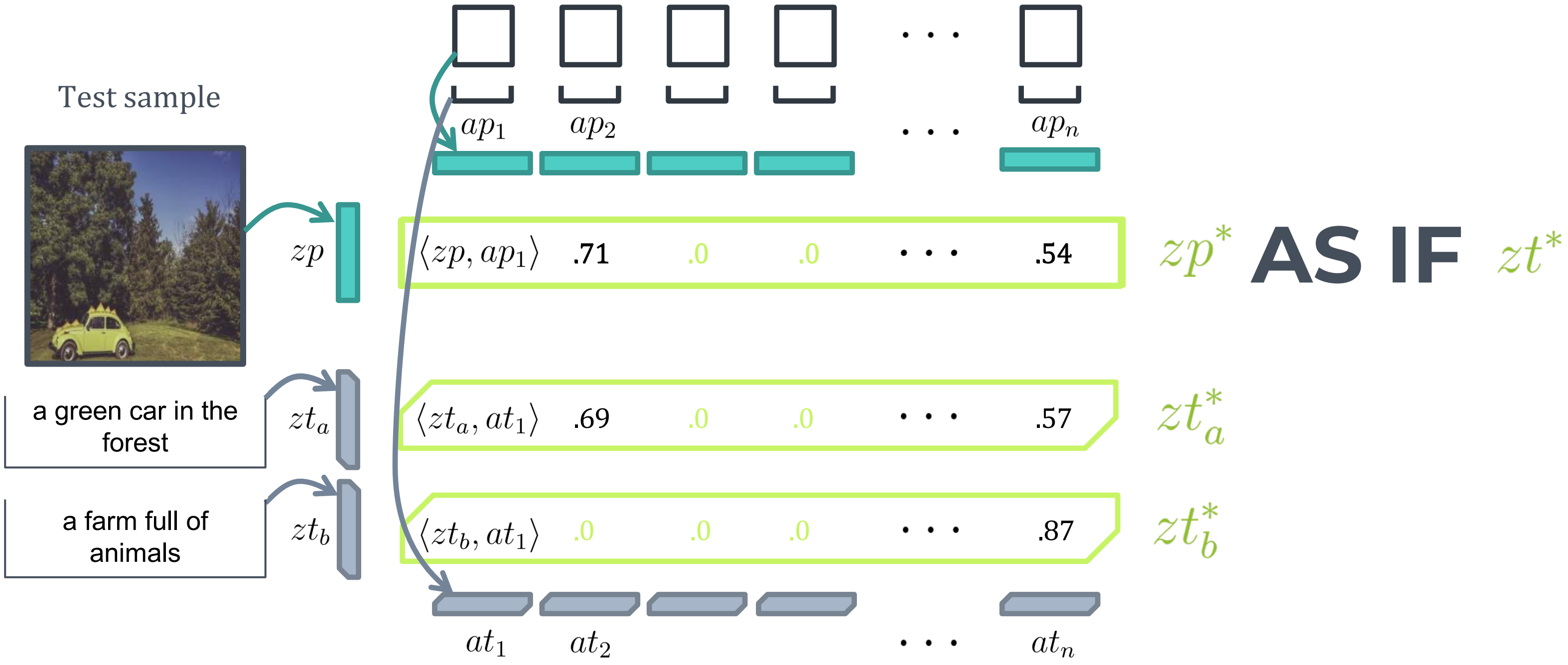
ASIF

- Sparse representations

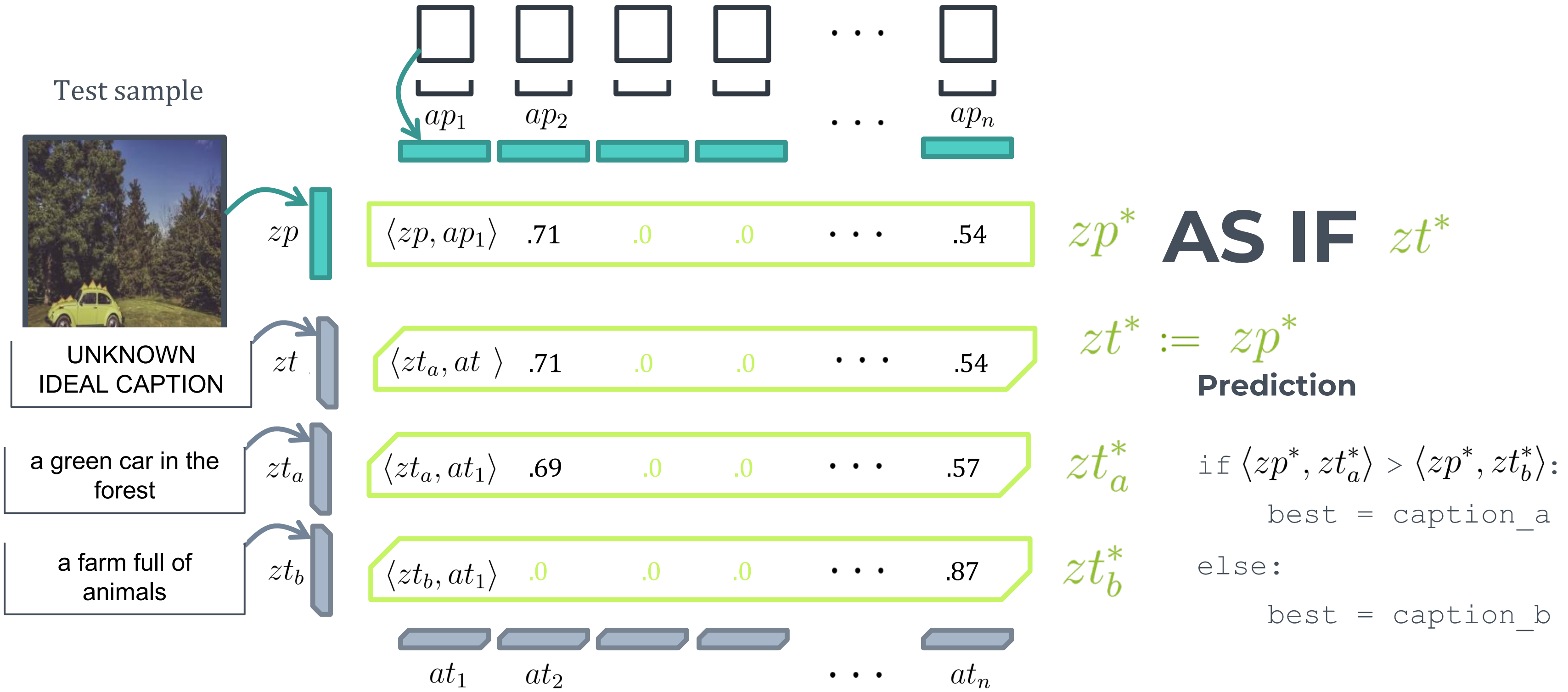


ASIF

- Sparse representations

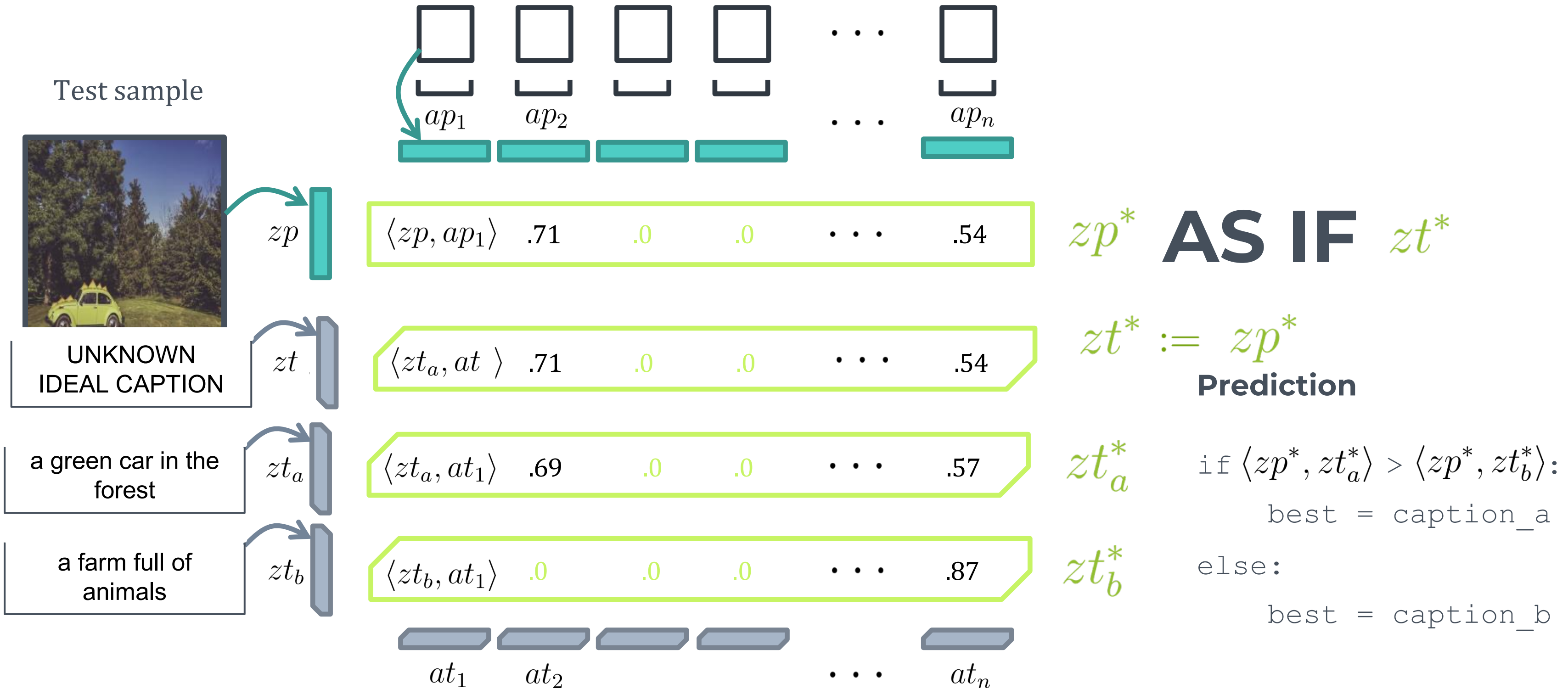


ASIF



Test sample
ap_1
ap_2
ap_n
zp
$\langle zp, ap_1 \rangle$ .71 .0 .0 .54
zp^* AS IF zt^*
UNKNOWN IDEAL CAPTION
zt
$\langle zt_a, at \rangle$ .71 .0 .0 .54
$zt^* := zp^*$
Prediction
a green car in the forest
zt_a
$\langle zt_a, at_1 \rangle$ .69 .0 .0 .57
zt_a^*
if $\langle zp^*, zt_a^* \rangle > \langle zp^*, zt_b^* \rangle$:
best = caption_a
else:
best = caption_b
a farm full of animals
zt_b
$\langle zt_b, at_1 \rangle$ .0 .0 .0 .87
zt_b^*
at_1
at_2
at_n

ASIF: coupled data turns unimodal models to multimodal without training



Our implementation

- Pretrained image encoder: VITb8; DINO VITs16

  -Training dataset: labeled Imagenet 22k; same unlabeled.
  -Learning task: supervised classification; unsupervised self-distillation
  -Embedding size: 768; 384

Antonio Norelli et al. "ASIF: coupled data turns unimodal models to multimodal without training" under review, 2022.

Our implementation

- Pretrained image encoder: VITb8; DINO VITs16

  -Training dataset: labeled Imagenet 22k; same unlabeled.
  -Learning task: supervised classification; unsupervised self-distillation
  -Embedding size: 768; 384

- Pretrained text encoder: SentenceT

  -Training dataset: >1B sentences scraped from the internet (Reddit, Wiki, SO, ...).
  -Learning task: BERT-like then contrastive with couples of sentences.
  -Embedding size: 768

Antonio Norelli et al. "ASIF: coupled data turns unimodal models to multimodal without training" under review, 2022.

Our implementation

- Pretrained image encoder: VITb8; DINO VITs16
  -Training dataset: labeled Imagenet 22k; same unlabeled.
  -Learning task: supervised classification; unsupervised self-distillation
  -Embedding size: 768; 384
- Pretrained text encoder: SentenceT
  -Training dataset: >1B sentences scraped from the internet (Reddit, Wiki, SO,
  -Learning task: BERT-like then contrastive with couples of sentences.
  -Embedding size: 768
- Analogy collection: subset of CC12M
  Images and alttexts scraped from the internet. CC12M size is 10M, we used 1.5M analogies

<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo)

Hand holding a fresh mangosteen

#jellyfish #blue #ocean #pretty Sea Turtle Wallpaper, Aquarius Aesthetic, Blue Aesthetic Pastel, The Adventure Zone, Capricorn And <PERSON>, Life Aquatic, Ocean Life, Jellyfish, Marine Life

Memory impact?

We have to keep all the embeddings of the analogy collection in memory

Memory impact?

We have to keep all the embeddings of the analogy collection in memory, but:

- We can **compress embeddings**, e.g. by quantization.

Memory impact?

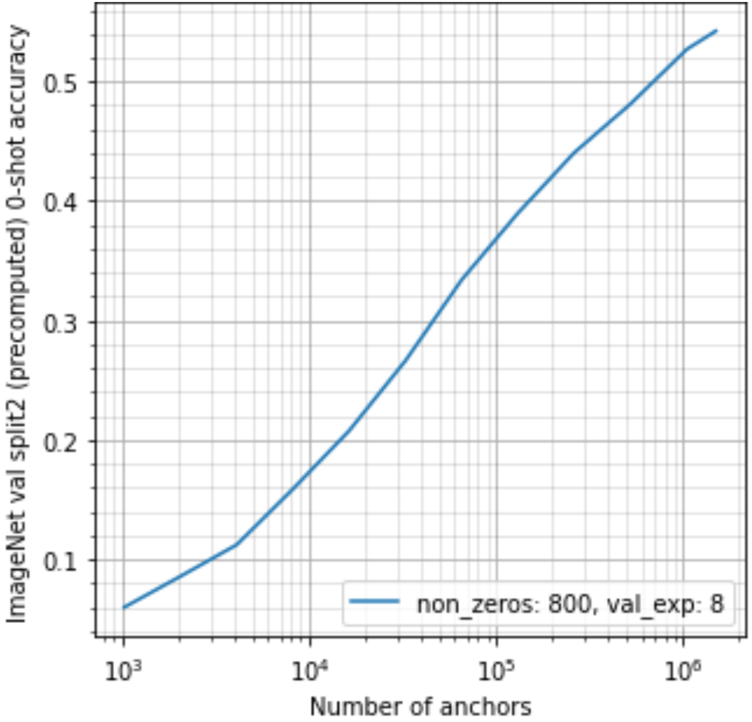
We have to keep all the embeddings of the analogy collection in memory, but:

- We can **compress embeddings**, e.g. by quantization.
- If we want a specialized model, we can perform **fine pruning**

1. How can we do this?

2. Benefits of this approach

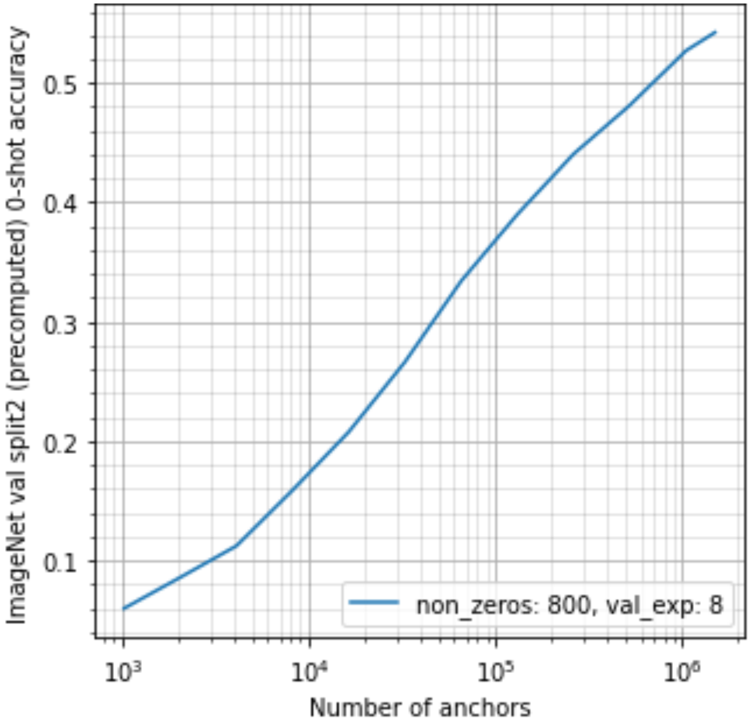
Zero-shot capabilities emerge early with small multimodal datasets



Antonio Norelli et al. "ASIF: coupled data turns unimodal models to multimodal without training" under review, 2022.

Zero-shot capabilities emerge early with small multimodal datasets

Method	Dataset size	ImageNet
CLIP (Radford et al., 2021)	400M (private)	68.6
CLIP (Radford et al., 2021)	15M (public)	31.3
LIT (Zhai et al., 2022)	10M (public)	66.9
CLIP (Zhai et al., 2022, `uu`)	901M (private)	50.6
LIT (Zhai et al., 2022)	901M (private)	70.1
ASIF (sup vis. encoder)	1.6M (public)	55.4*



Antonio Norelli et al. "ASIF: coupled data turns unimodal models to multimodal without training" under review, 2022.

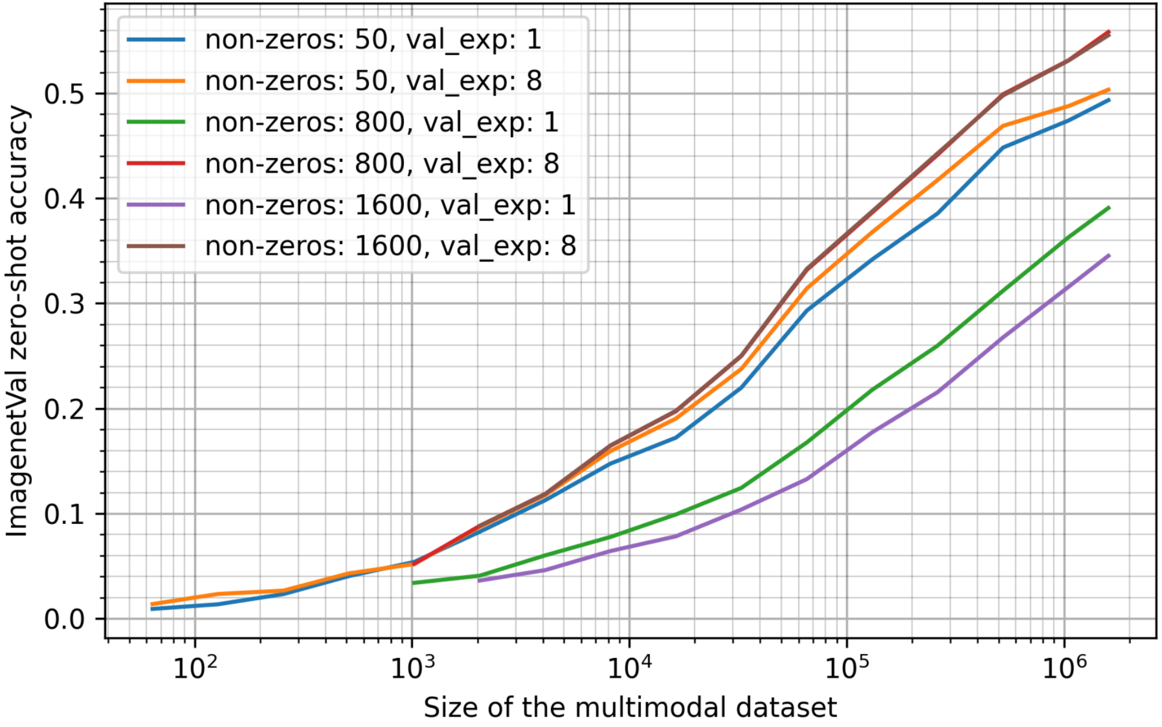
Zero-shot capabilities emerge early with small multimodal datasets

Method	Dataset size	ImageNet	CIFAR100	Pets	ImageNet v2
CLIP (Radford et al., 2021)	400M (private)	68.6	68.7	88.9	-
CLIP (Radford et al., 2021)	15M (public)	31.3	-	-	-
LIT (Zhai et al., 2022)	10M (public)	66.9	-	-	-
CLIP (Zhai et al., 2022, `uu`)	901M (private)	50.6	47.9	70.3	43.3
LIT (Zhai et al., 2022)	901M (private)	70.1	70.9	88.1	61.7
ASIF (sup vis. encoder)	1.6M (public)	55.4*	63.3	71.5	45.6

Antonio Norelli et al. "ASIF: coupled data turns unimodal models to multimodal without training" under review, 2022.

Zero-shot capabilities emerge early with small multimodal datasets

Method	Dataset size	ImageNet
CLIP (Radford et al., 2021)	400M (private)	68.6
CLIP (Radford et al., 2021)	15M (public)	31.3
LIT (Zhai et al., 2022)	10M (public)	66.9
CLIP (Zhai et al., 2022, uu)	901M (private)	50.6
LIT (Zhai et al., 2022)	901M (private)	70.1
ASIF (sup vis. encoder)	1.6M (public)	55.4*
ASIF (unsup vis. encoder)	1.6M (public)	53.0*

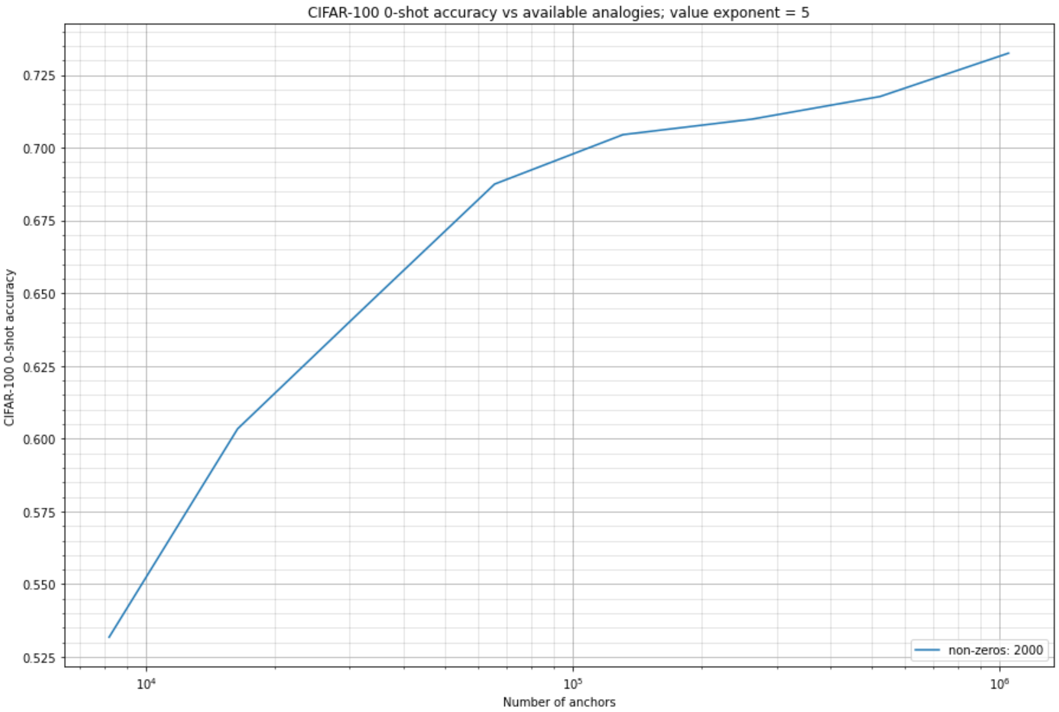


Antonio Norelli et al. "ASIF: coupled data turns unimodal models to multimodal without training" under review, 2022.

Zero-shot capabilities emerge early with small multimodal datasets

Test dataset:
CIFAR 100

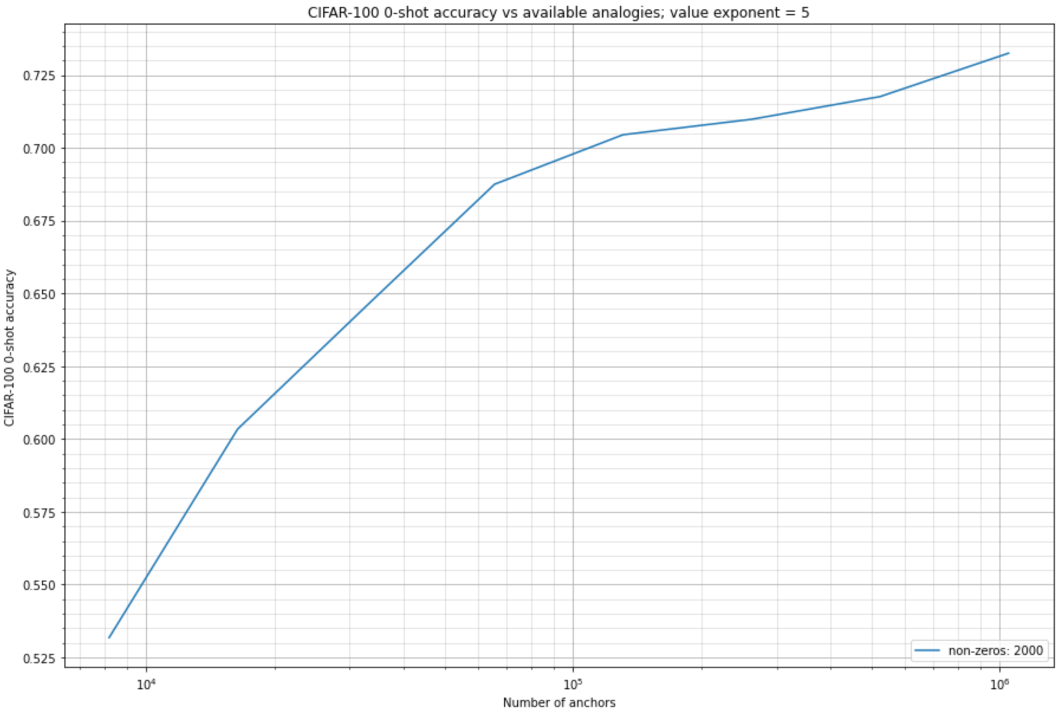
Model	Accuracy	Image-text couples seen
CLIP (VITb16)	68.7	400M
LIT (VITb16)	70.9	900M
ASIF (VITb16)	73.3	1.5M



Zero-shot capabilities emerge early with small multimodal datasets

Test dataset:
CIFAR 100

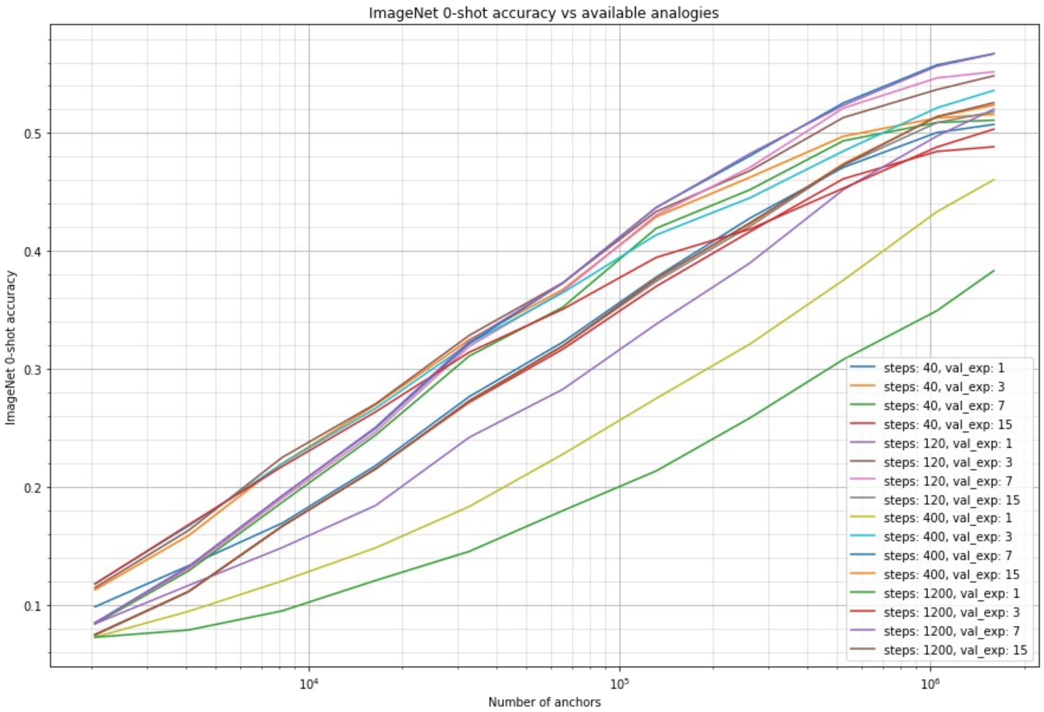
- 55% accuracy with just 10,000 image-text couples



Zero-shot capabilities emerge early with small multimodal datasets

Test dataset:
ImageNet

Model	Accuracy	Image-text couples seen
CLIP (VITb16)	68.6	400M
LIT (VITb16)	70.1	900M
ASIF (VITb16)	57.0	1.5M



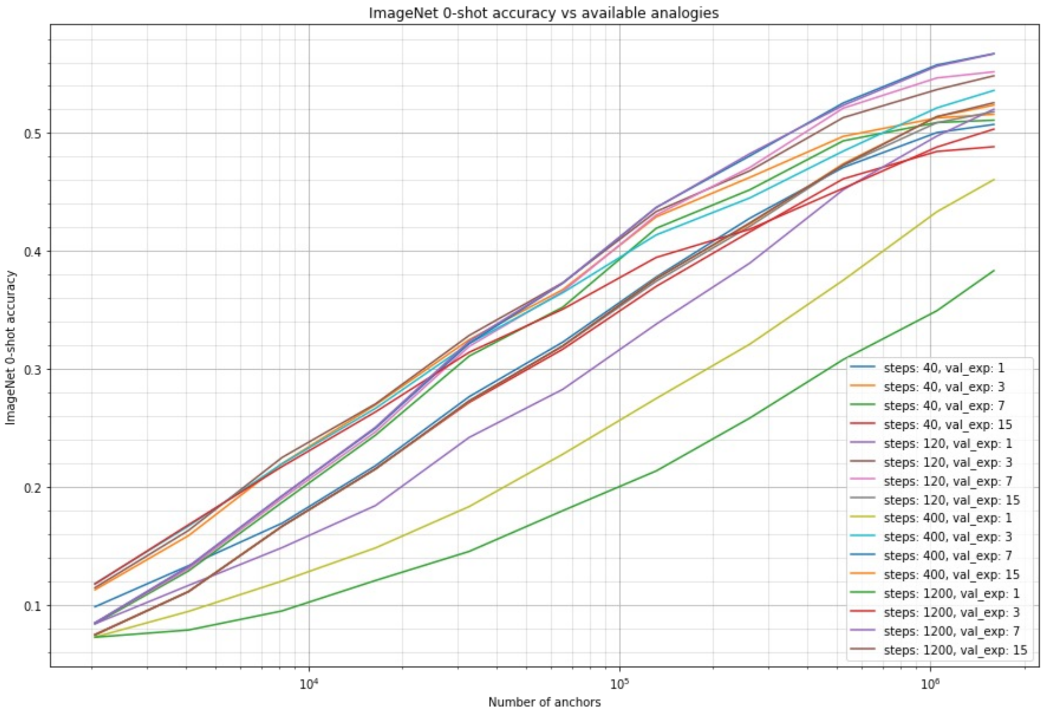
Zero-shot capabilities emerge early with small multimodal datasets

Test dataset:

ImageNet

Hyperparameters relevance:

- # nonzeros (steps)
- In. product exponent



Zero-shot capabilities emerge early with small multimodal datasets

Test dataset:
ImageNetv2

Model	Accuracy	Image-text couples seen
CLIP (VITb16)	43.3*	900M*
LIT (VITb16)	61.7	900M
ASIF (VITb16)	47.1	1.5M

Test dataset:
PETS

Model	Accuracy	Image-text couples seen
CLIP (VITb16)	70.3*	900M*
LIT (VITb16)	88.1	900M
ASIF (VITb16)	72.9	1.5M

**tested by the LIT authors*

- Zero-shot capabilities emerge early with small multimodal datasets

Encoders can be pretrained in a completely unsupervised way

- ASIF with DINO visual encoder remains effective.

Method	Dataset size	ImageNet	CIFAR100	Pets	ImageNet v2
CLIP (Radford et al., 2021)	400M (private)	68.6	68.7	88.9	-
CLIP (Radford et al., 2021)	15M (public)	31.3	-	-	-
LIT (Zhai et al., 2022)	10M (public)	66.9	-	-	-
CLIP (Zhai et al., 2022, uu)	901M (private)	50.6	47.9	70.3	43.3
LIT (Zhai et al., 2022)	901M (private)	70.1	70.9	88.1	61.7
ASIF (sup vis. encoder)	1.6M (public)	55.4*	63.3	71.5	45.6

Antonio Norelli et al. "ASIF: coupled data turns unimodal models to multimodal without training" under review, 2022.

- Zero-shot capabilities emerge early with small multimodal datasets

Encoders can be pretrained in a completely unsupervised way

- ASIF with DINO visual encoder remains effective.

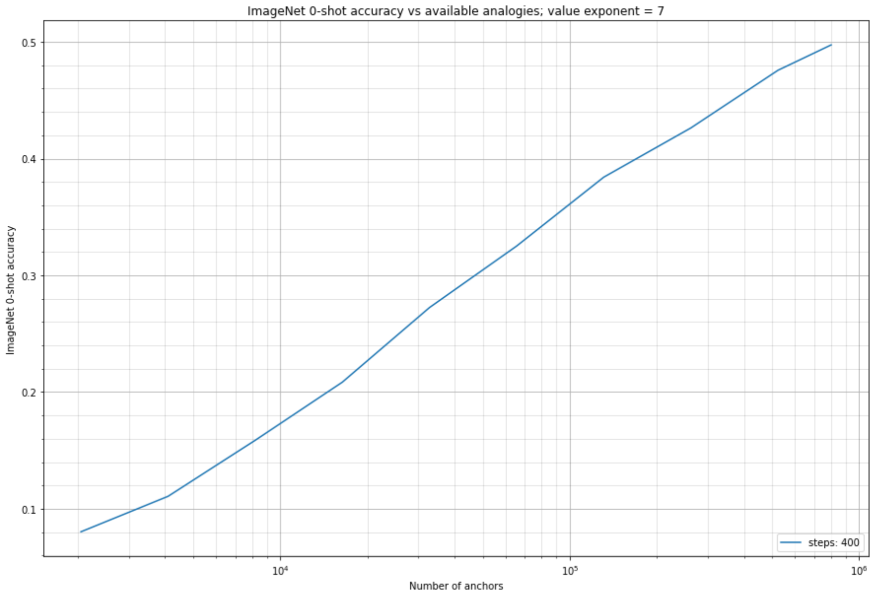
Method	Dataset size	ImageNet	CIFAR100	Pets	ImageNet v2
CLIP (Radford et al., 2021)	400M (private)	68.6	68.7	88.9	-
CLIP (Radford et al., 2021)	15M (public)	31.3	-	-	-
LIT (Zhai et al., 2022)	10M (public)	66.9	-	-	-
CLIP (Zhai et al., 2022, `uu`)	901M (private)	50.6	47.9	70.3	43.3
LIT (Zhai et al., 2022)	901M (private)	70.1	70.9	88.1	61.7
ASIF (sup vis. encoder)	1.6M (public)	55.4*	63.3	71.5	45.6
ASIF (unsup vis. encoder)	1.6M (public)	53.0*	46.5	74.7	45.9

Antonio Norelli et al. “ASIF: coupled data turns unimodal models to multimodal without training” under review, 2022.

- Zero-shot capabilities emerge early with small multimodal datasets

Encoders can be pretrained in a completely unsupervised way

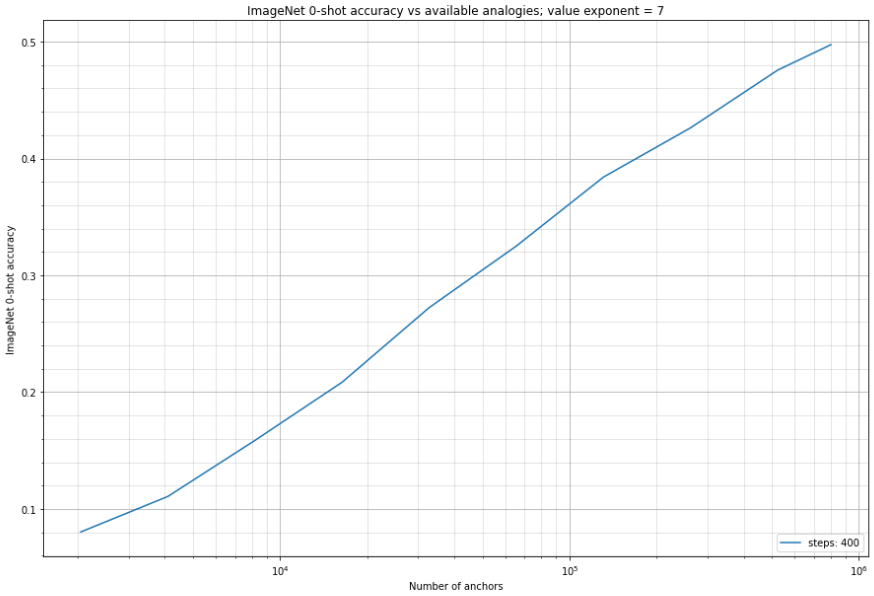
- Performance barely deteriorates with DINO encoder.



- Zero-shot capabilities emerge early with small multimodal datasets

Encoders can be pretrained in a completely unsupervised way

- Performance barely deteriorates with DINO encoder.
- 50% accuracy on ImageNet with 800k couples.



Highly interpretable representations

Query image
x*

x*

A photo of a triumphal arch

A photo of a mosque

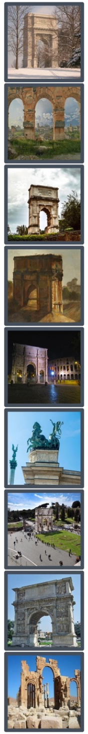
The Arch of Titus, Rome

AC Milan's players celebrate on a bus after winning the championship

Photo looking up at the Arch of Titus

The faded triumph of <PERSON>'s Arch in Benevento

The triumphal arch of Volubilis glowing in the sun



The arch at Valley Forge

Neoclassical Architecture Painting - A View through Three Arches of the …

The Arch of Titus in Rome, Italy. Rome landmark.

The Arch of Septimus Severus, Rome

The Arch of Constantine and the <PERSON> in Rome.

<PERSON> and warrior on a chariot. <PERSON> bronze statue atop …

Arch of Constantine near the Colosseum in Rome, Italy stock photography

The Arch of Trajan: all what you need to know

The Arch of Triumph or Arch of <PERSON>, Palmyra, Syria, 2005

Large stone triumphal arch with trees on the other side in Dougga

Ruins of the Roman triumphal arch at Palmyra as photographed in 2006.

A watercolor sketch or illustration of the Brandenburg gate

Model of the Arch of Constantine Probably by <PERSON>

The Arch of <PERSON> and the aqueduct

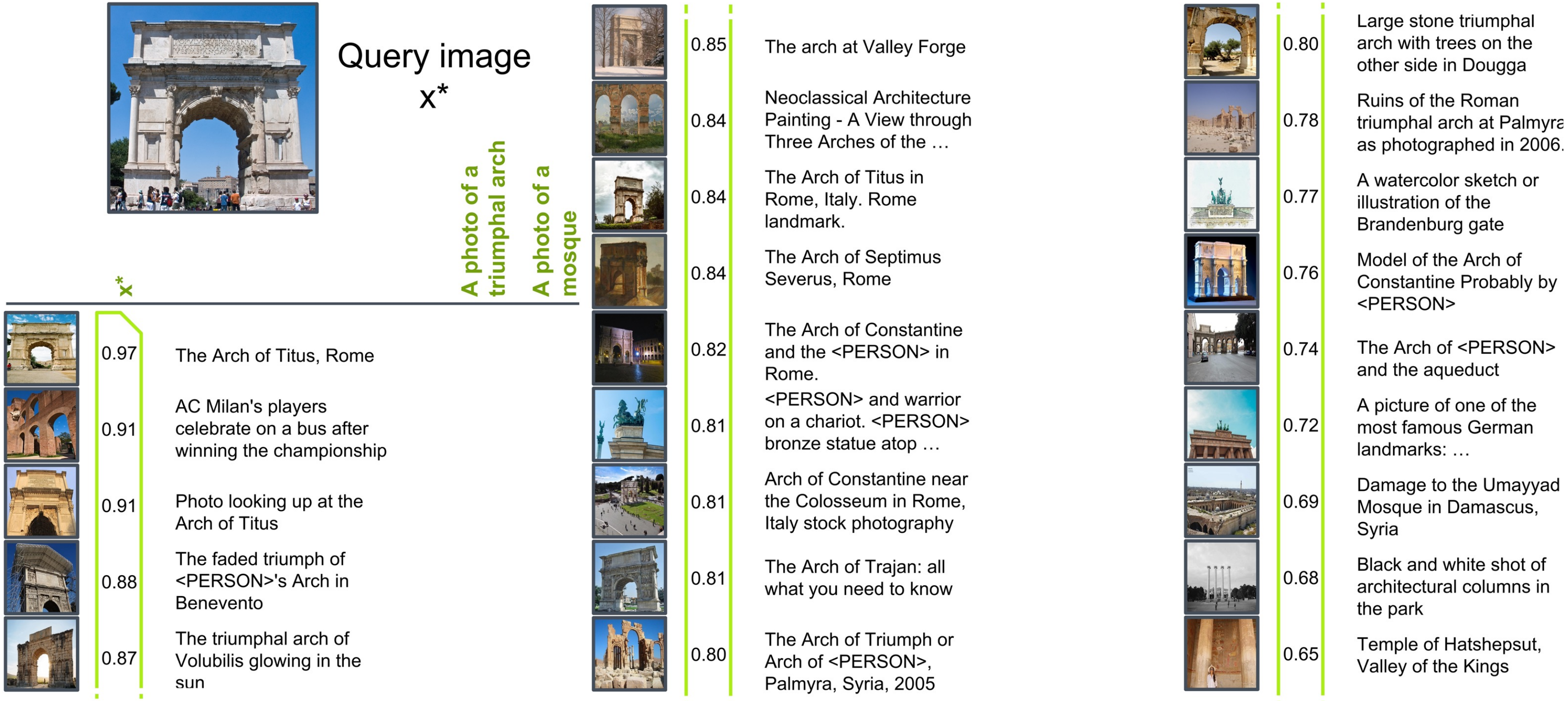
A picture of one of the most famous German landmarks: …

Damage to the Umayyad Mosque in Damascus, Syria

Black and white shot of architectural columns in the park

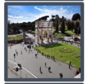
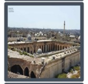
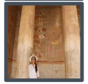
Temple of Hatshepsut, Valley of the Kings

Highly interpretable representations



Highly interpretable representations

Query image
x*

	x*		A photo of a triumphal arch	A photo of a mosque
	0.97	The Arch of Titus, Rome	0.88	0
	0.91	AC Milan's players celebrate on a bus after winning the championship	0	0
	0.91	Photo looking up at the Arch of Titus	0.83	0
	0.88	The faded triumph of <PERSON>'s Arch in Benevento	0.93	0
	0.87	The triumphal arch of Volubilis glowing in the sun	0.95	0
	0.85	The arch at Valley Forge	0.85	0
	0.84	Neoclassical Architecture Painting - A View through Three Arches of the …	0.81	0
	0.84	The Arch of Titus in Rome, Italy. Rome landmark.	0.89	0
	0.84	The Arch of Septimus Severus, Rome	0.84	0
	0.82	The Arch of Constantine and the <PERSON> in Rome.	0.87	0
	0.81	<PERSON> and warrior on a chariot. <PERSON> bronze statue atop …	0	0
	0.81	Arch of Constantine near the Colosseum in Rome, Italy stock photography	0.87	0
	0.81	The Arch of Trajan: all what you need to know	0.86	0
	0.80	The Arch of Triumph or Arch of <PERSON>, Palmyra, Syria, 2005	0.95	0.66
	0.80	Large stone triumphal arch with trees on the other side in Dougga	0.95	0
	0.78	Ruins of the Roman triumphal arch at Palmyra as photographed in 2006.	0.96	0
	0.77	A watercolor sketch or illustration of the Brandenburg gate	0	0
	0.76	Model of the Arch of Constantine Probably by <PERSON>	0.84	0
	0.74	The Arch of <PERSON> and the aqueduct	0.83	0
	0.72	A picture of one of the most famous German landmarks: …	0	0
	0.69	Damage to the Umayyad Mosque in Damascus, Syria	0	0.87
	0.68	Black and white shot of architectural columns in the park	0	0
	0.65	Temple of Hatshepsut, Valley of the Kings	0	0

- Zero-shot capabilities emerge early with small multimodal datasets
- Encoders can be pretrained in a completely unsupervised way

Highly interpretable representations

- Each feature comes from a single data-point.

- Zero-shot capabilities emerge early with small multimodal datasets
- Encoders can be pretrained in a completely unsupervised way

Highly interpretable representations

- Each feature comes from a single data-point.
- Each classification traces back to a small set of training data

- Zero-shot capabilities emerge early with small multimodal datasets

- Encoders can be pretrained in a completely unsupervised way

- Highly interpretable representations

We can add/remove training samples and update the model in seconds

King Charles gave his first Christmas speech

<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo)

Hand holding a fresh mangosteen

- Zero-shot capabilities emerge early with small multimodal datasets
- Encoders can be pretrained in a completely unsupervised way
- Highly interpretable representations

We can add/remove training samples and update the model in seconds

- Zero-shot capabilities emerge early with small multimodal datasets
- Encoders can be pretrained in a completely unsupervised way
- Highly interpretable representations

We can add/remove training samples and update the model in seconds

- Imagine using analogy collections built with movies and tv-series of different countries

- Zero-shot capabilities emerge early with small multimodal datasets
- Encoders can be pretrained in a completely unsupervised way
- Highly interpretable representations

We can add/remove training samples and update the model in seconds

- Imagine using analogy collections built with movies and tv-series of different countries

- Zero-shot capabilities emerge early with small multimodal datasets
- Encoders can be pretrained in a completely unsupervised way
- Highly interpretable representations
- We can add/remove training samples and update the model in seconds

ASIF knows what it does not know

If all inner products are ~0 we can output an unknown token

- Zero-shot capabilities emerge early with small multimodal datasets
- Encoders can be pretrained in a completely unsupervised way
- Highly interpretable representations
- We can add/remove training samples and update the model in seconds

ASIF knows what it does not know

If all inner products are ~0 we can output an unknown token

On ImageNet
threshold: 0.0039
accuracy: 0.495, unknown: 0.297, wrong: 0.208

- Zero-shot capabilities emerge early with small multimodal datasets

- Encoders can be pretrained in a completely unsupervised way

- Highly interpretable representations

- We can add/remove training samples and update the model in seconds

- ASIF knows what it does not know

Fine pruning

If we specialize the model, we can keep few couples

- **What is the difference between learning and retrieval?**
- **Are neural encoders just sensors?**

Thanks!

Any questions?

ASIF: Coupled Data Turns Unimodal Models to Multimodal Without Training

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, Francesco Locatello

Check the paper on arXiv!