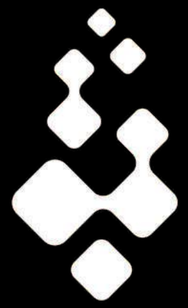


CLIP

Contrastive Language-Image Pre-training

Nancy Kalaj

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

⟶ DOG

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

⟶ DOG

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

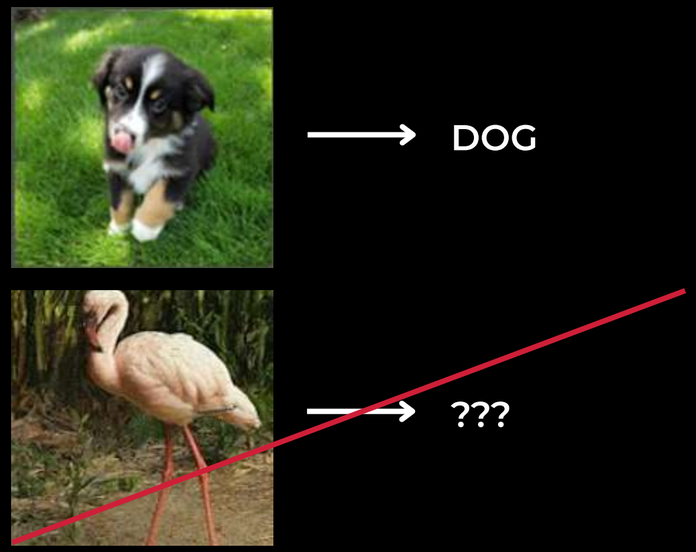
→ DOG

→ ???

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

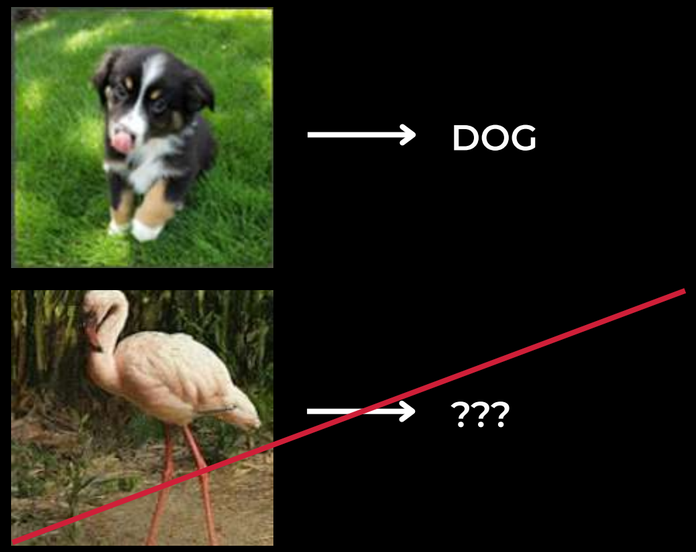


Retrain the model for new categories

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)



Retrain the model for new categories

Learn a different model for each task

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

→ DOG

→ ???

Retrain the model for new categories

Learn a different model for each task

NATURAL LANGUAGE SUPERVISION

Create a training set of pairs: (image, caption)

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

→ DOG

→ ???

Retrain the model for new categories

Learn a different model for each task

NATURAL LANGUAGE SUPERVISION

Create a training set of pairs: (image, caption)

→ “Pepper the aussie pup”

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

→ DOG

→ ???

Retrain the model for new categories

Learn a different model for each task

NATURAL LANGUAGE SUPERVISION

Create a training set of pairs: (image, caption)

→ “Pepper the aussie pup”

→ “A pink flamingo walking”

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

→ DOG

→ ???

Retrain the model for new categories

Learn a different model for each task

NATURAL LANGUAGE SUPERVISION

Create a training set of pairs: (image, caption)

→ “Pepper the aussie pup”

→ “A pink flamingo walking”

→ “An orange cat eating from a bowl”

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

→ DOG

→ ???

Retrain the model for new categories

Learn a different model for each task

NATURAL LANGUAGE SUPERVISION

Create a training set of pairs: (image, caption)

→ “Pepper the aussie pup”

→ “A pink flamingo walking”

→ “An orange cat eating from a bowl”

Paradigm Shift

TRADITIONAL LEARNING PARADIGM

Create a training set of pairs: (image, label)

→ **DOG**

→ **???**

Retrain the model for new categories

Learn a different model for each task

NATURAL LANGUAGE SUPERVISION

Create a training set of pairs: (image, caption)

→ “Pepper the aussie pup”

→ “A pink flamingo walking”

→ “An orange cat eating from a bowl”

→ **A PHOTO OF A { … }**

Non-trivial zero-shot transfer to downstream tasks

WebImageText (WIT)

NATURAL LANGUAGE SUPERVISION

Create a training set of pairs: (image, caption)

→ “Pepper the aussie pup”

→ “A pink flamingo walking”

→ “An orange cat eating from a bowl”

WebImageText (WIT)

NATURAL LANGUAGE SUPERVISION

Create a training set of pairs: (image, caption)

→ “Pepper the aussie pup”
→ “A pink flamingo walking”
→ “An orange cat eating from a bowl”

400M image-text pairings

WebImageText (WIT)

NATURAL LANGUAGE SUPERVISION

Create a training set of pairs: (image, caption)

→ “Pepper the aussie pup”
→ “A pink flamingo walking”
→ “An orange cat eating from a bowl”

400M image-text pairings
(minimally-curated, web-scraped data)

WebImageText (WIT)

NATURAL LANGUAGE SUPERVISION

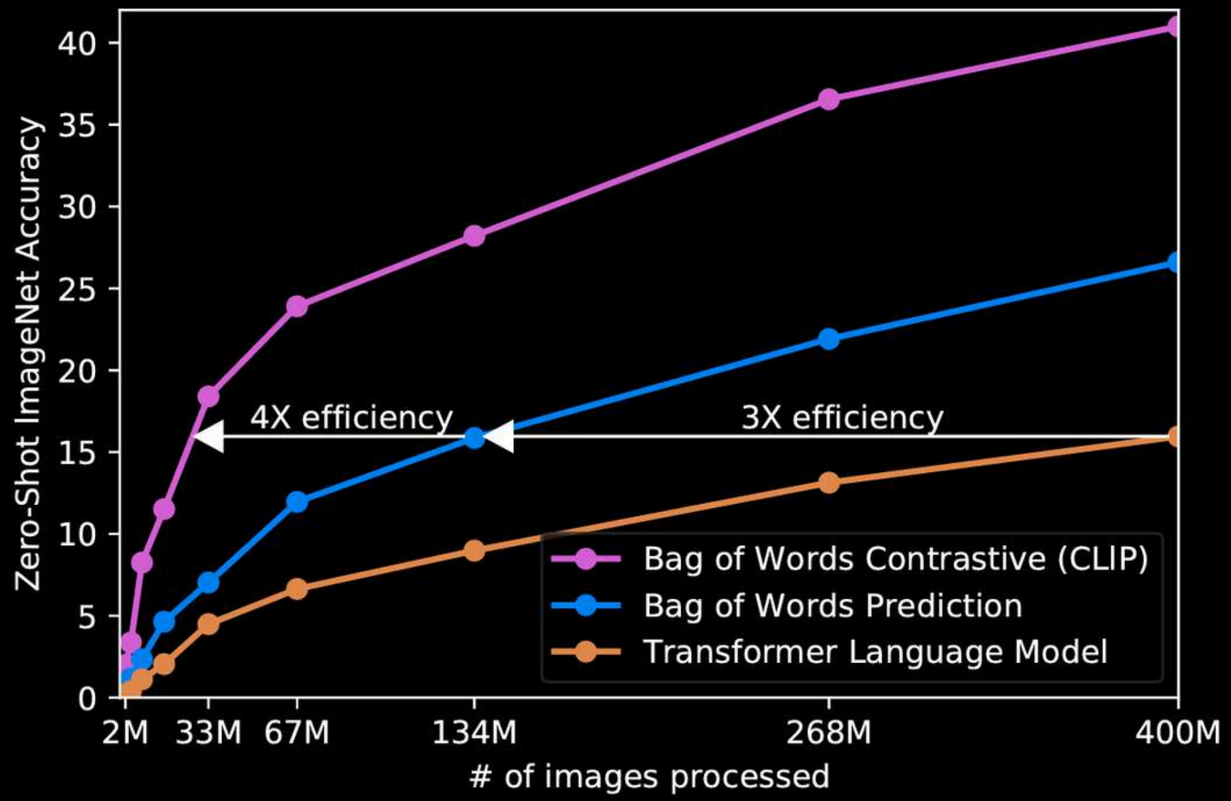
Create a training set of pairs: (image, caption)

→ “Pepper the aussie pup”
→ “A pink flamingo walking”
→ “An orange cat eating from a bowl”

400M image-text pairings
(minimally-curated, web-scraped data)

BIASES (?)

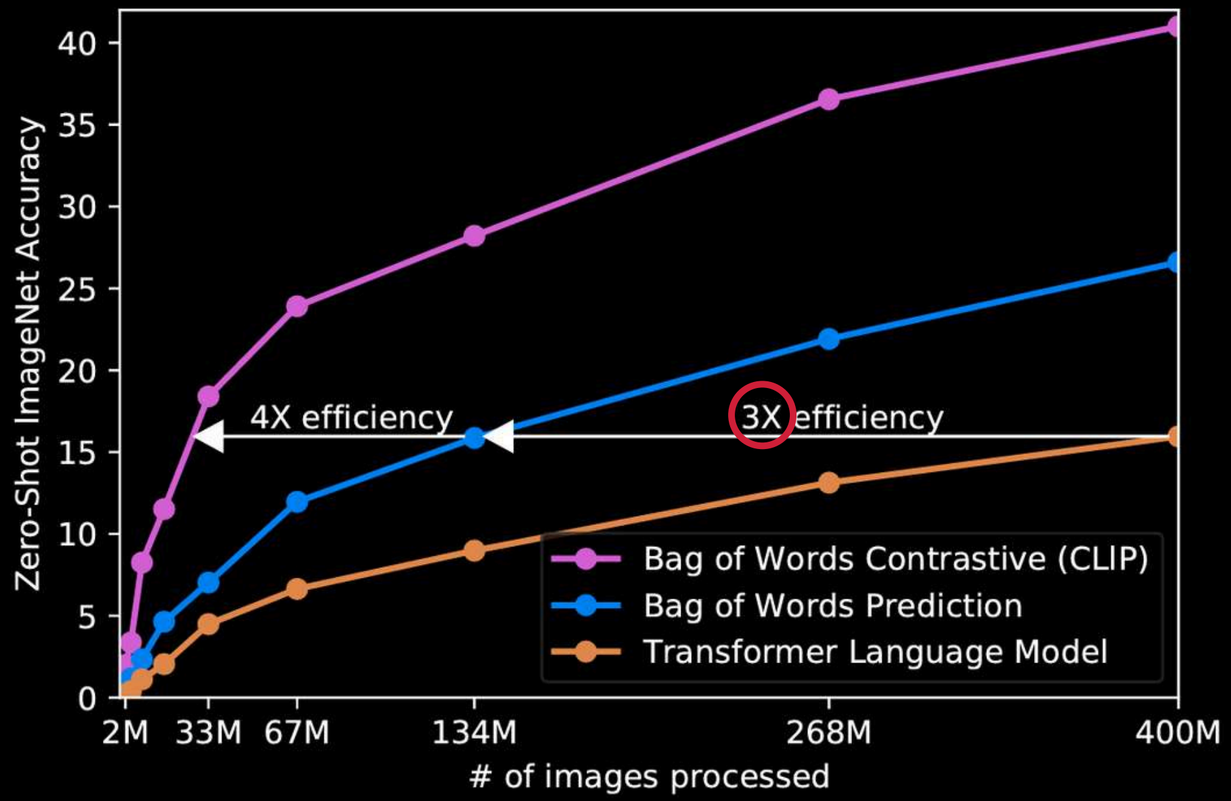
Efficient Pre-training



Transformer Language Model

ResNet-50 + text transformer to predict the caption of an image

Efficient Pre-training



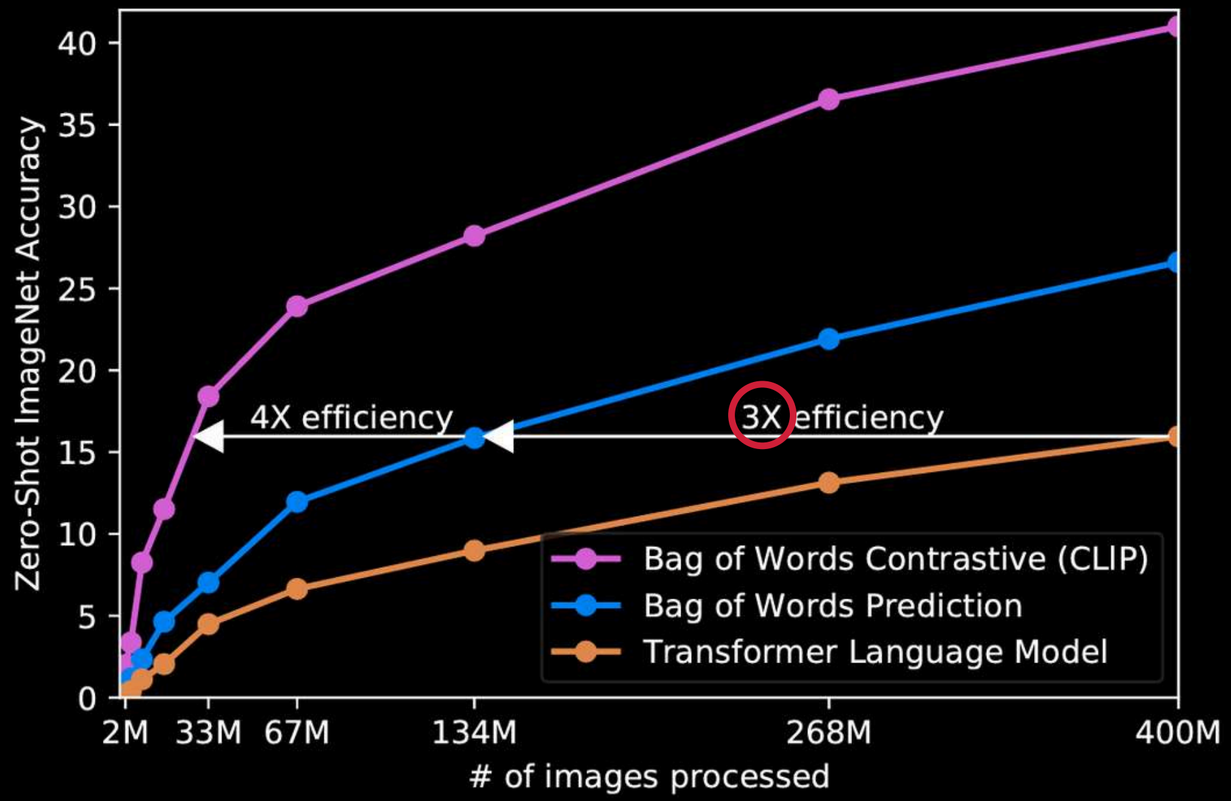
Bag of Words Prediction

ResNet-50 + text transformer to predict a BoW encoding of the same text

Transformer Language Model

ResNet-50 + text transformer to predict the caption of an image

Efficient Pre-training



Bag of Words Prediction

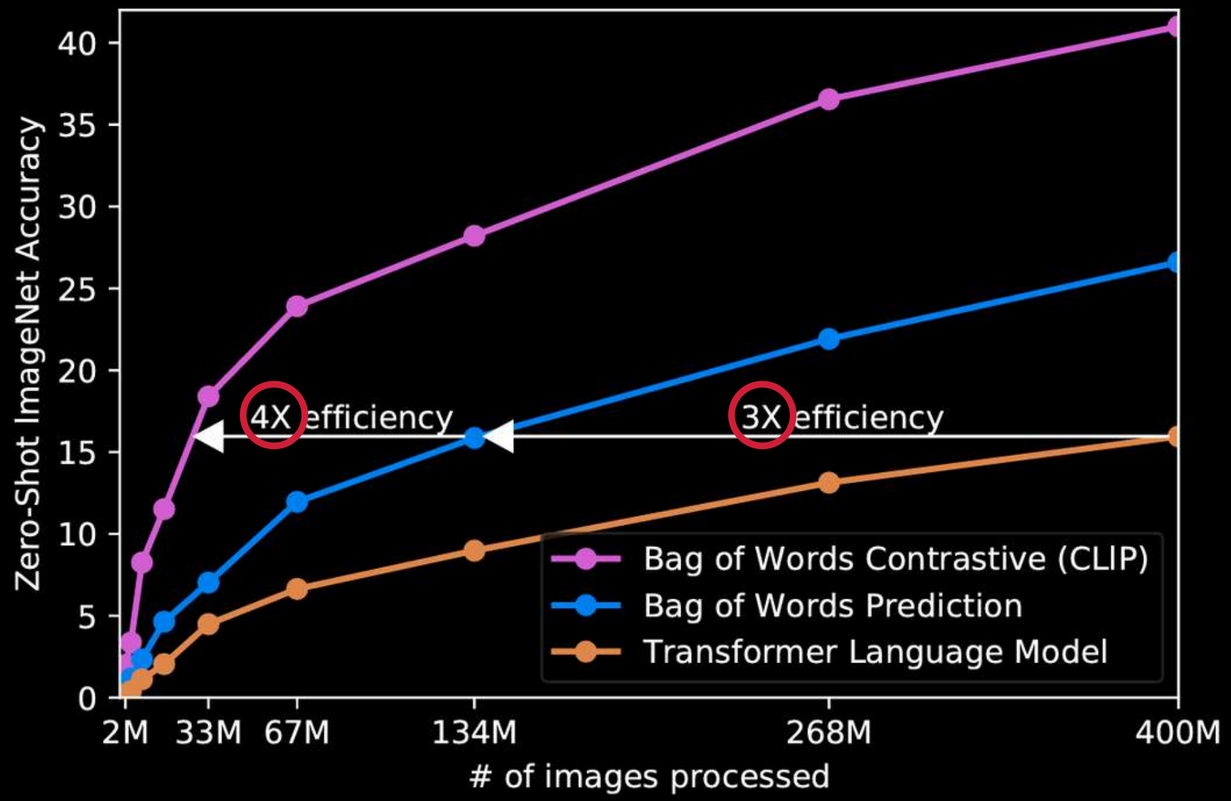
ResNet-50 + text transformer to predict a BoW encoding of the same text

Transformer Language Model

ResNet-50 + text transformer to predict the caption of an image

exact words

Efficient Pre-training



Bag of Words Contrastive (CLIP)

ResNet-50 + text transformer with a **contrastive** objective

Bag of Words Prediction

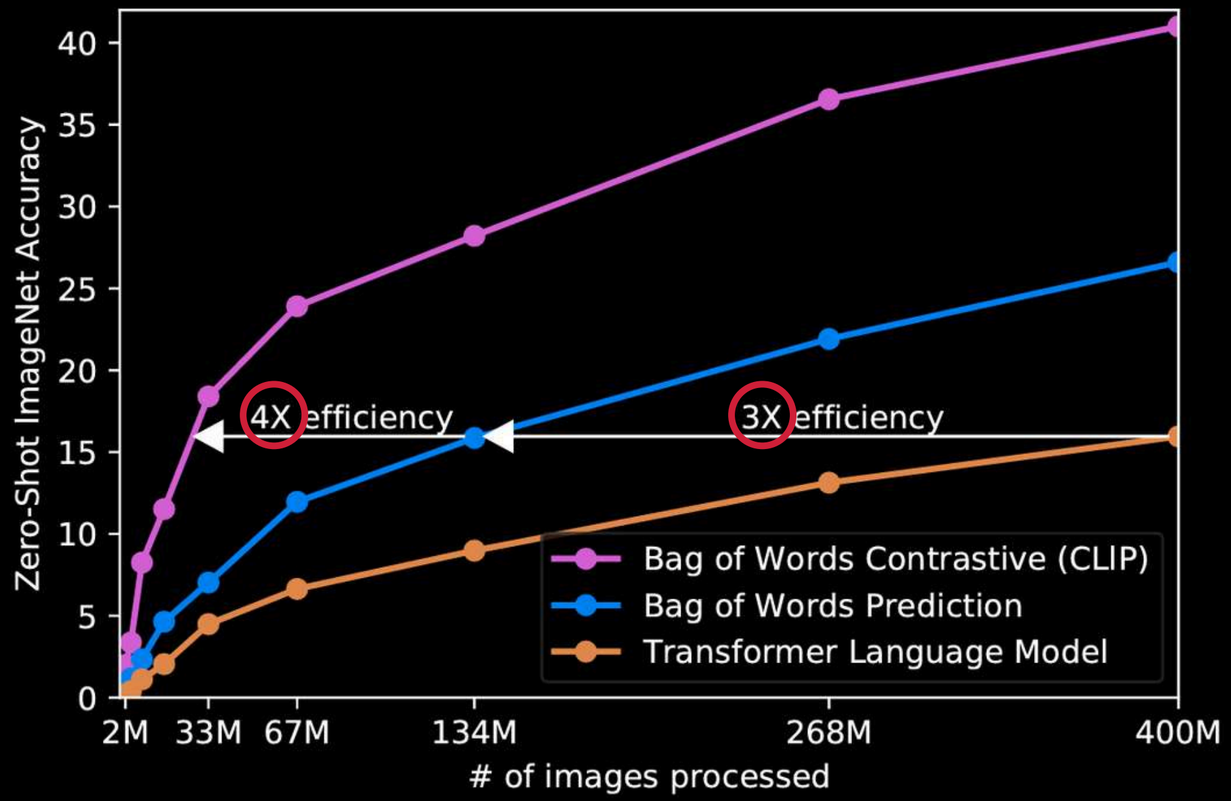
ResNet-50 + text transformer to predict a BoW encoding of the same text

Transformer Language Model

ResNet-50 + text transformer to predict the caption of an image

exact words

Efficient Pre-training



Bag of Words Contrastive (CLIP)

ResNet-50 + text transformer with a **contrastive** objective

Bag of Words Prediction

ResNet-50 + text transformer to predict a BoW encoding of the same text

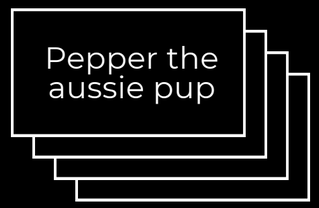
Transformer Language Model

ResNet-50 + text transformer to predict the caption of an image

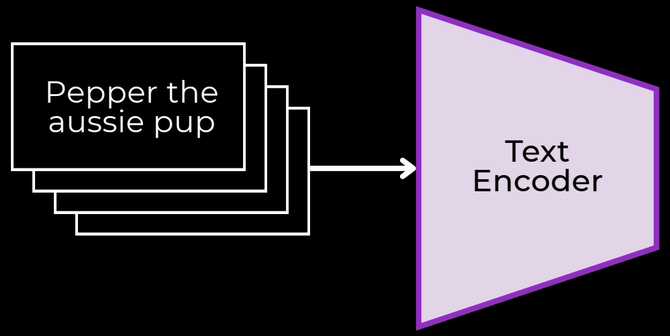
exact words

"Pepper the aussie pup"

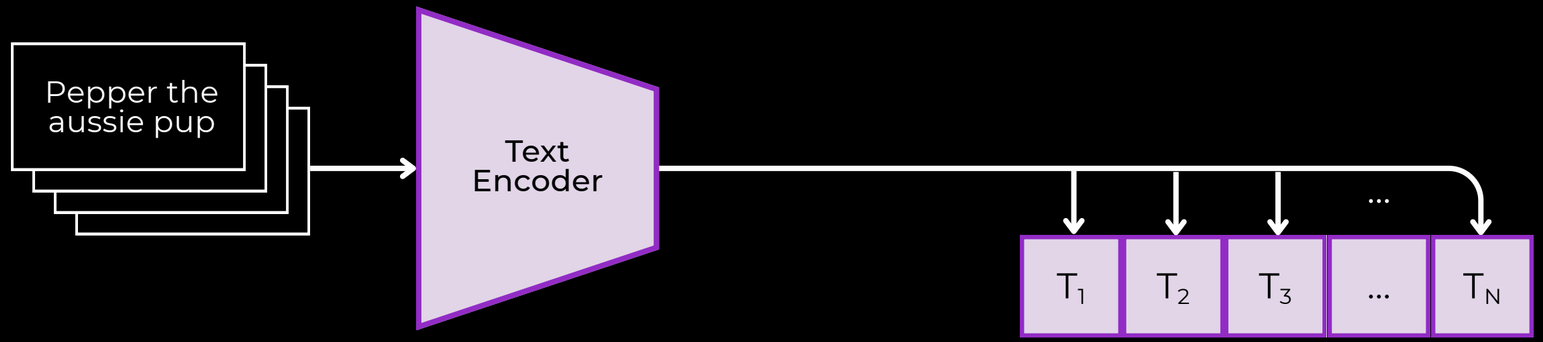
Contrastive Pre-training



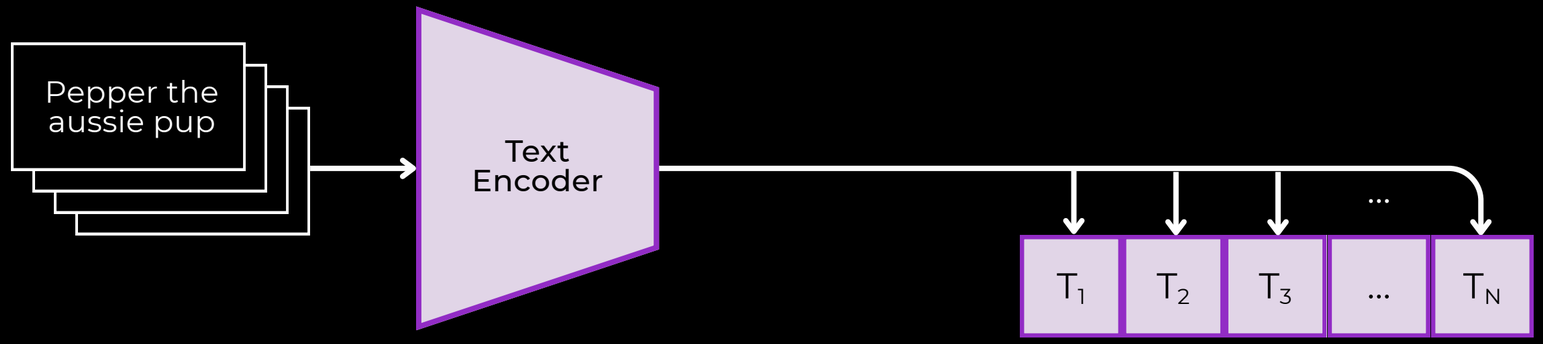
Contrastive Pre-training



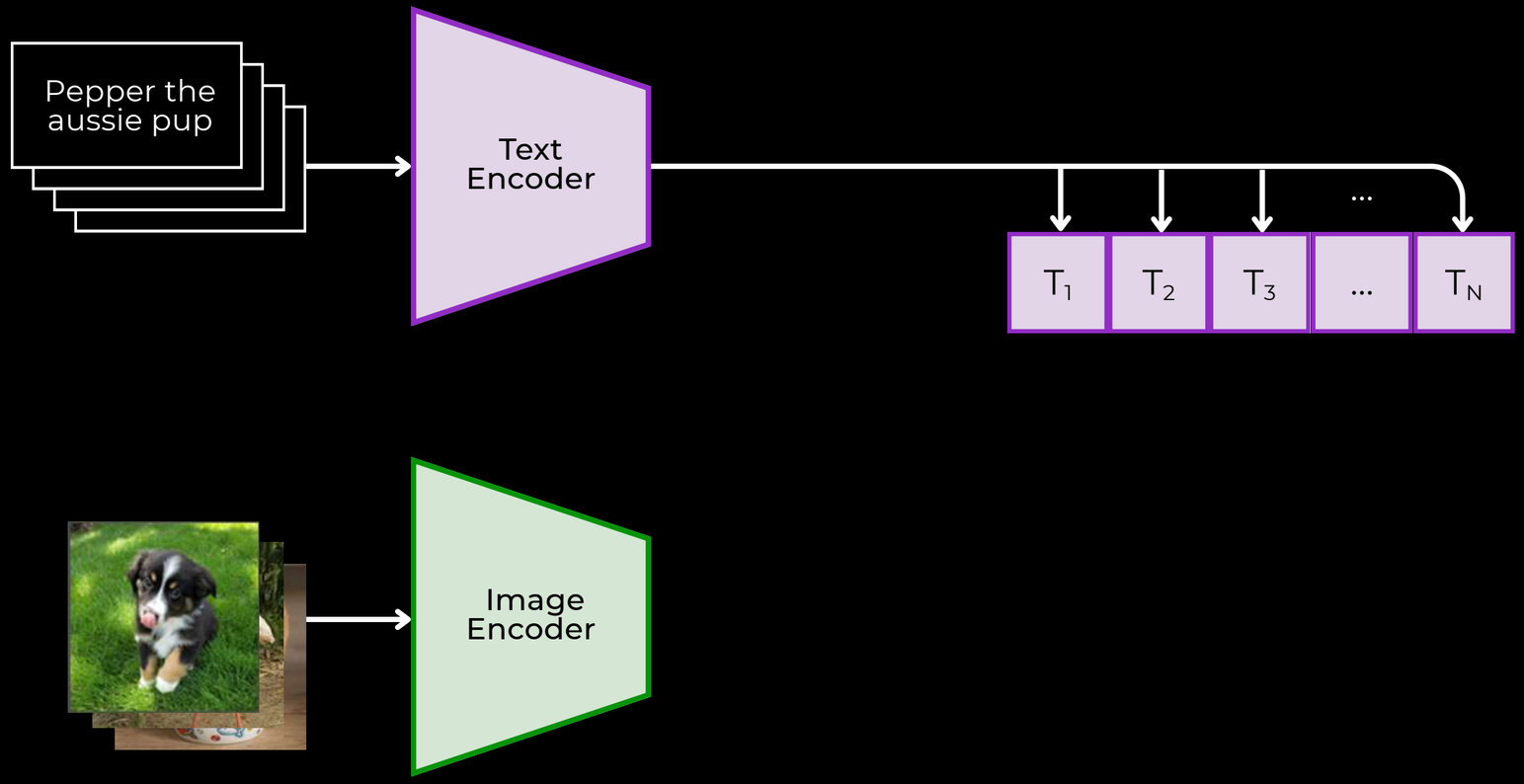
Contrastive Pre-training



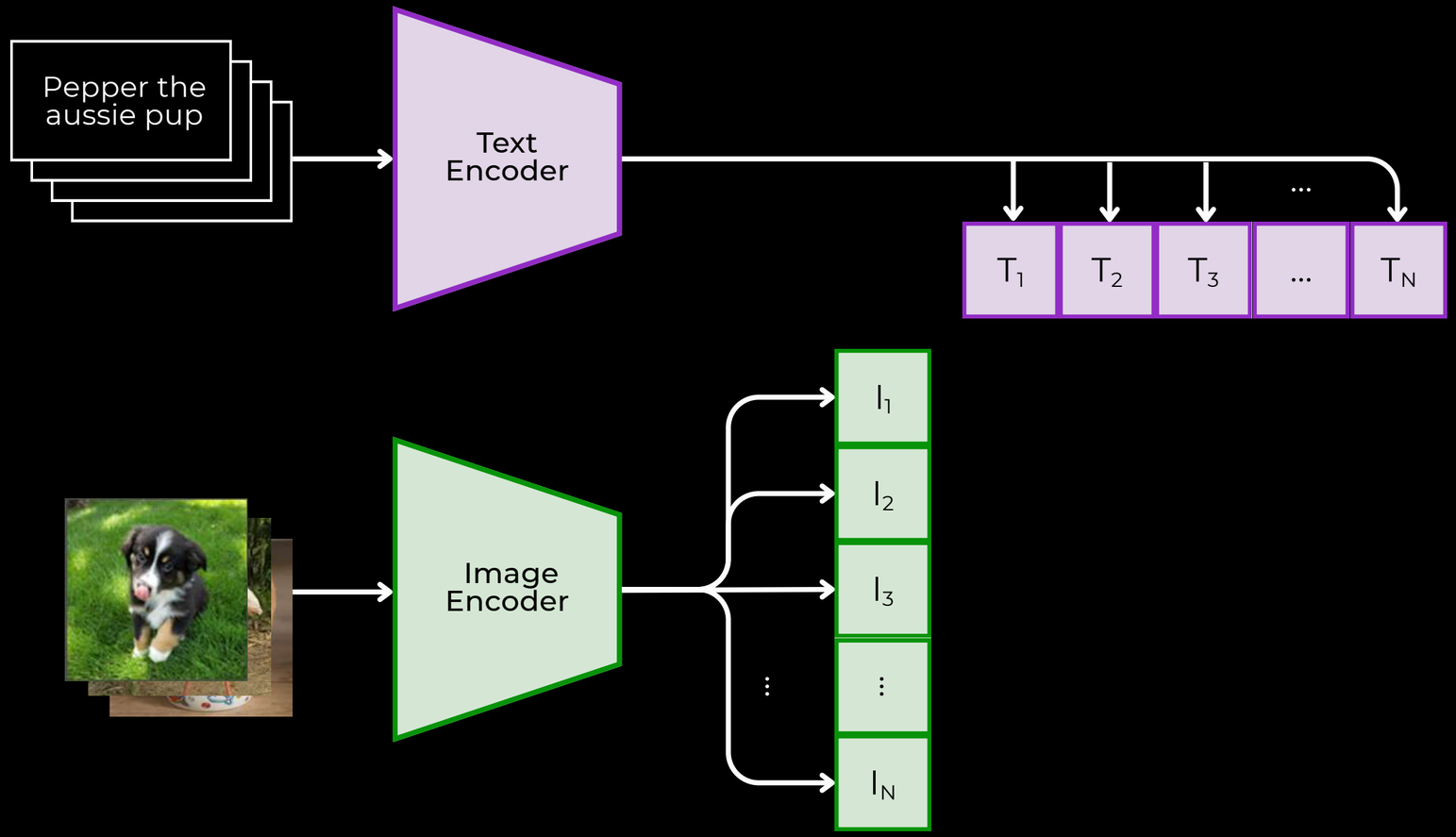
Contrastive Pre-training



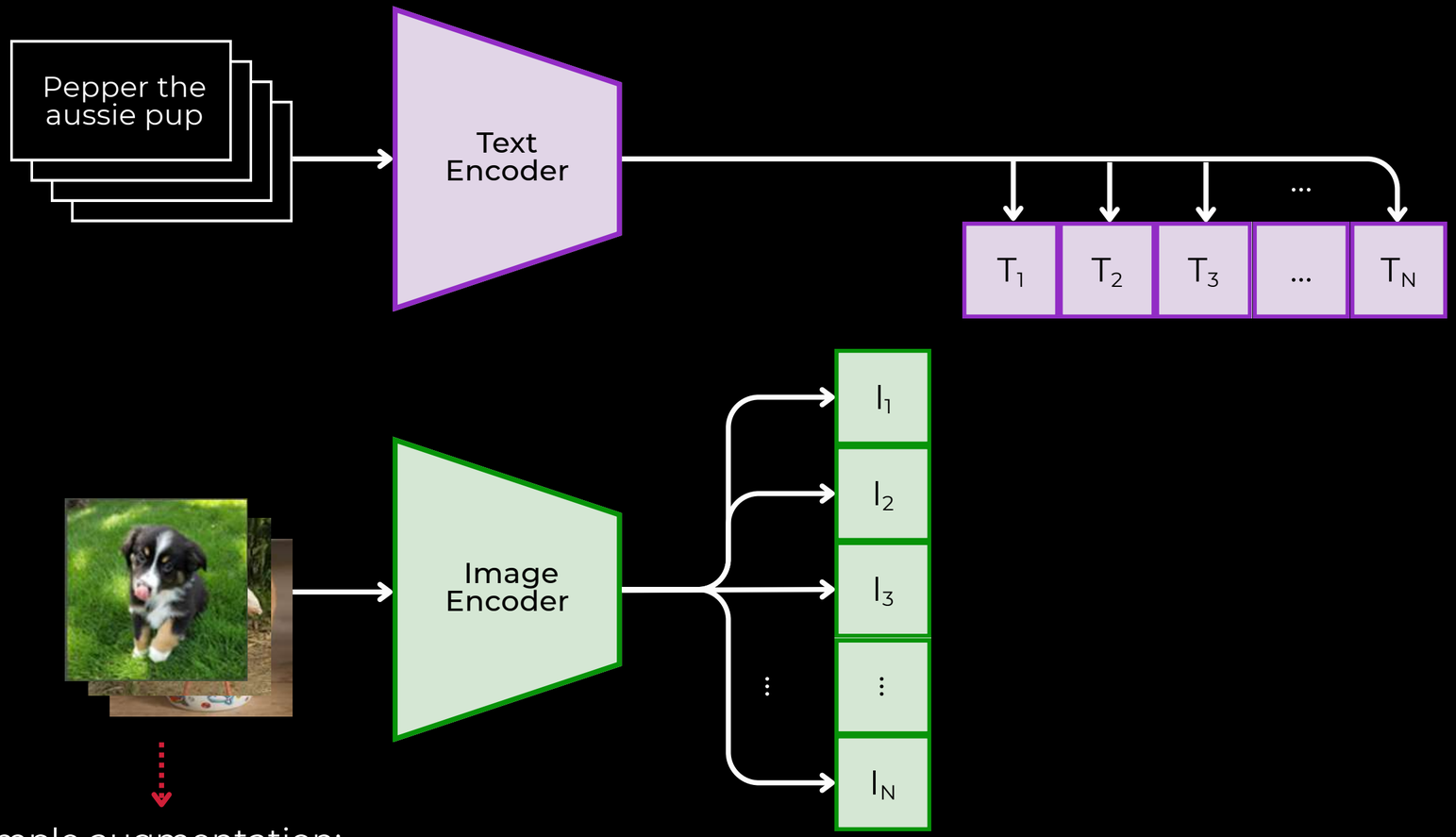
Contrastive Pre-training



Contrastive Pre-training

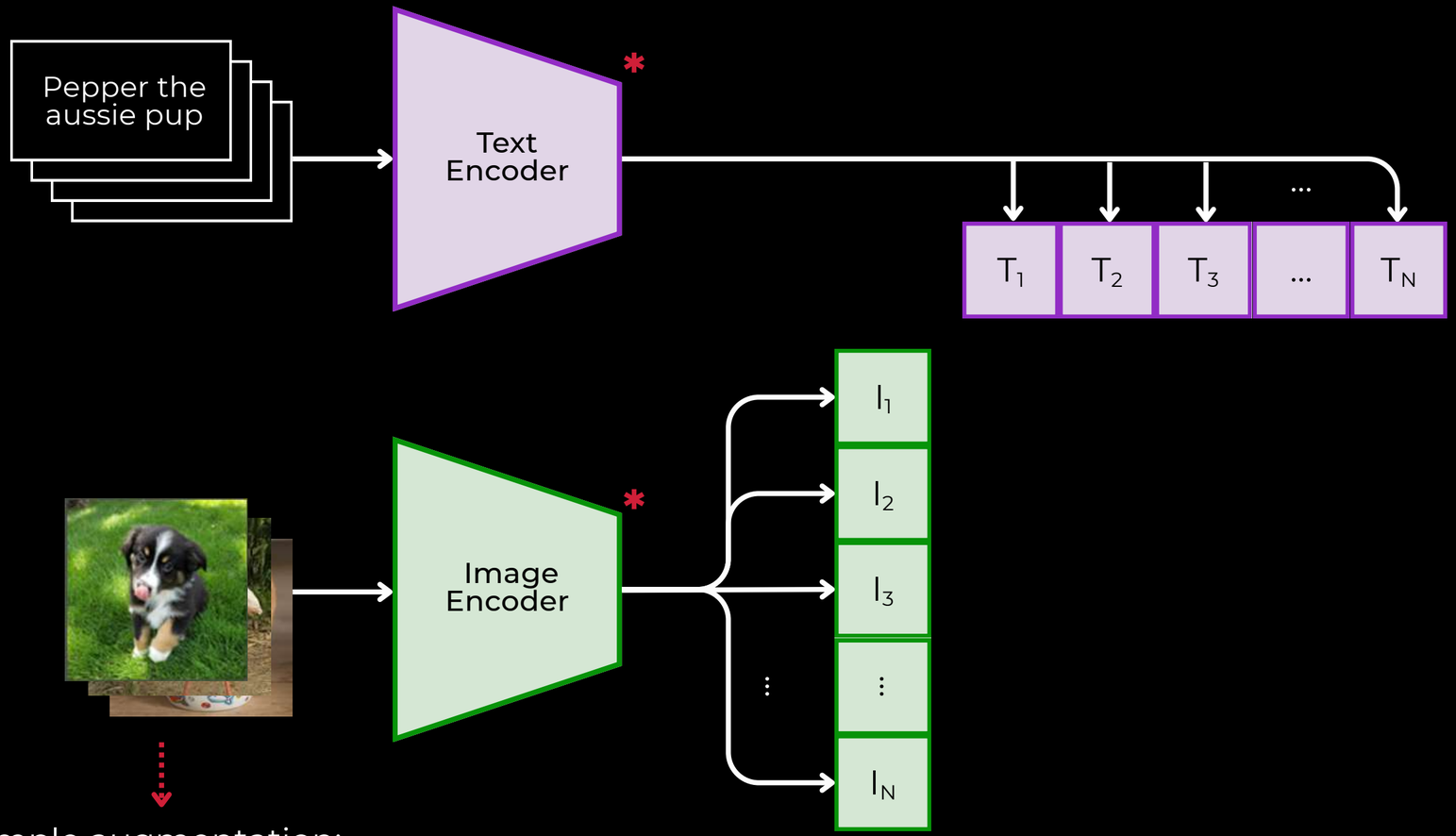


Contrastive Pre-training



simple augmentation:
random square crop
from resized images

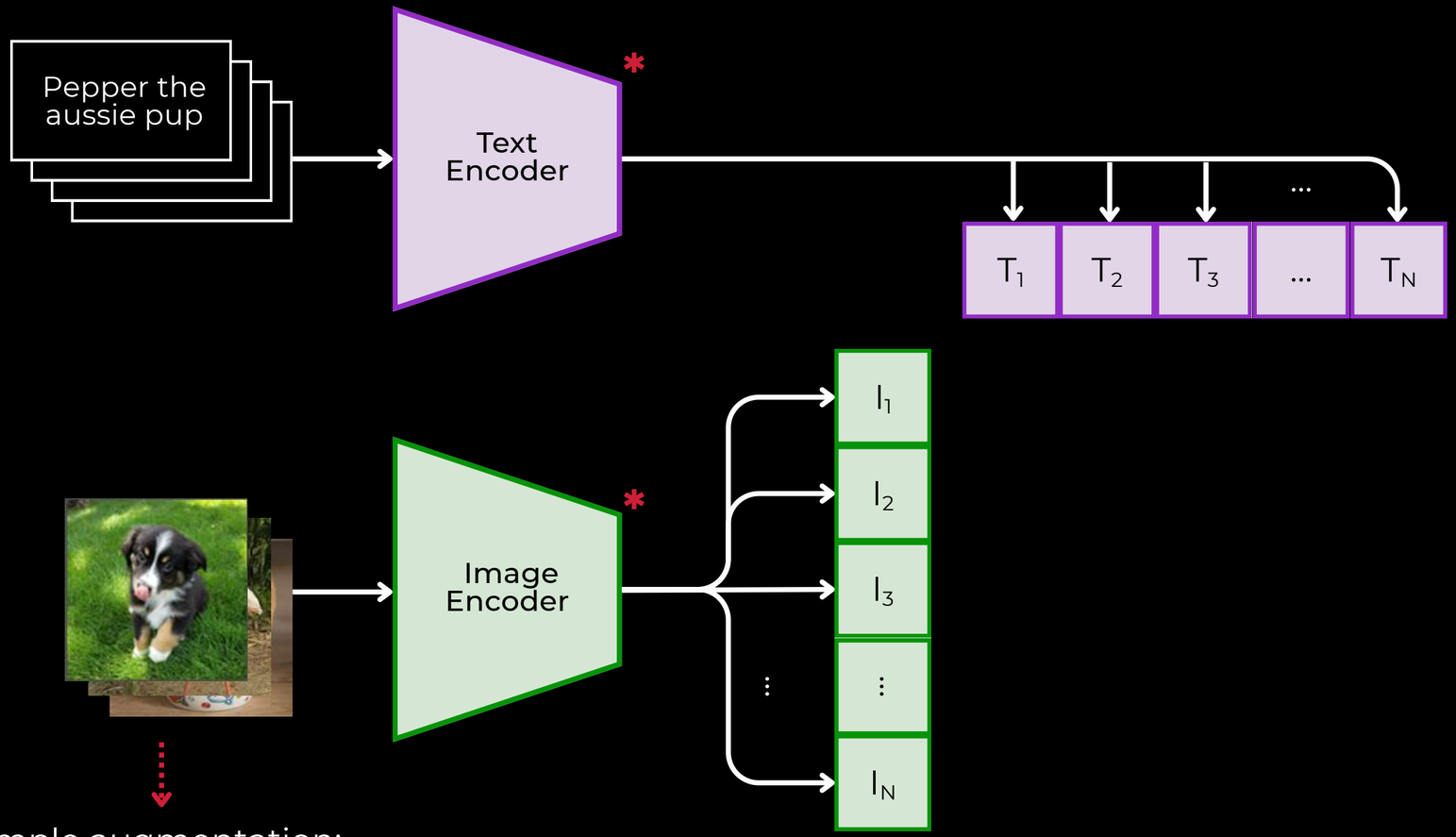
Contrastive Pre-training



simple augmentation:
random square crop
from resized images

* both encoders are trained from scratch since WIT is large

Contrastive Pre-training

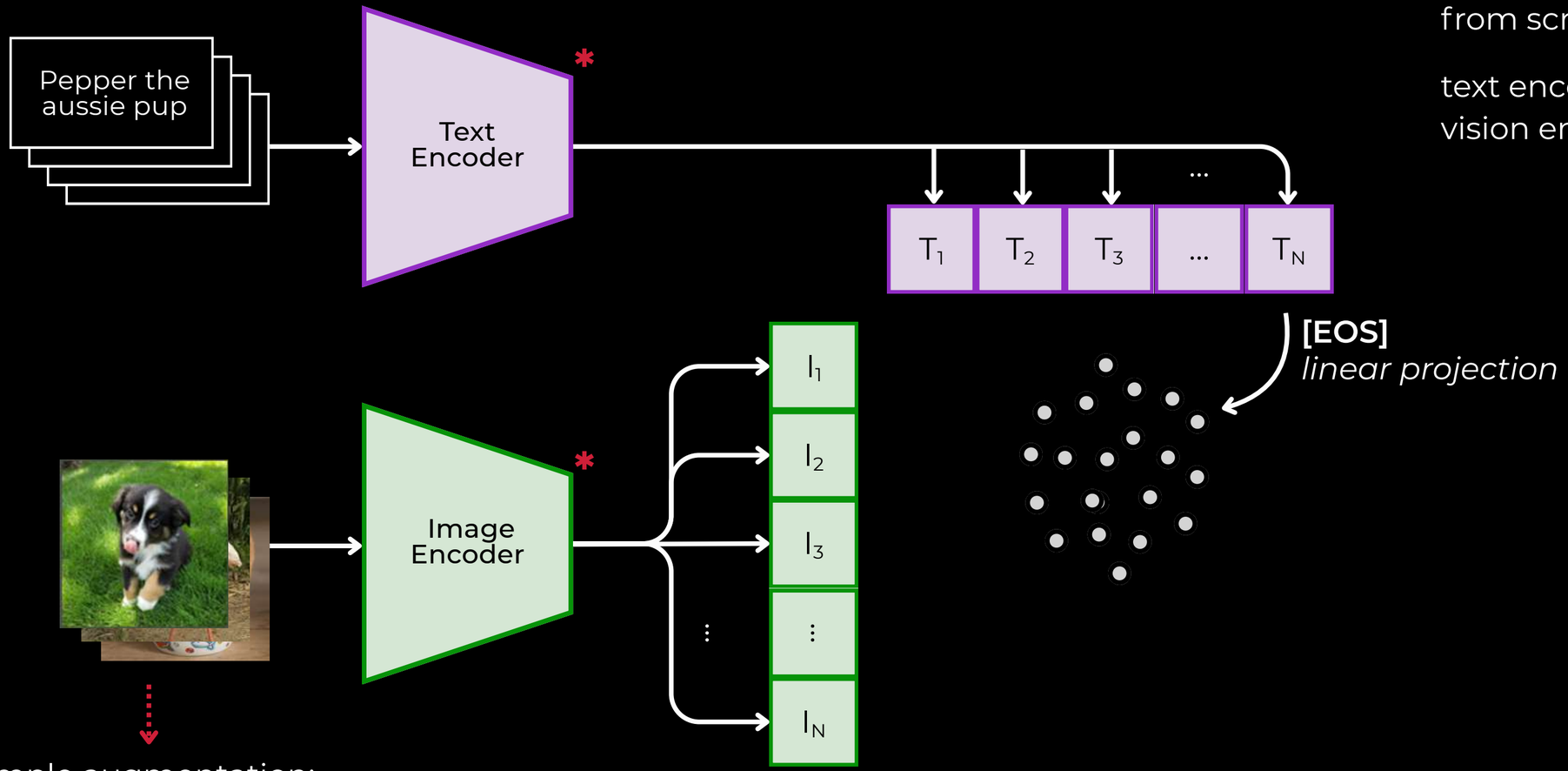


simple augmentation:
random square crop
from resized images

* both encoders are trained from scratch since WIT is large

text encoder → Transformer
vision encoder → ResNet-50 or ViT

Contrastive Pre-training

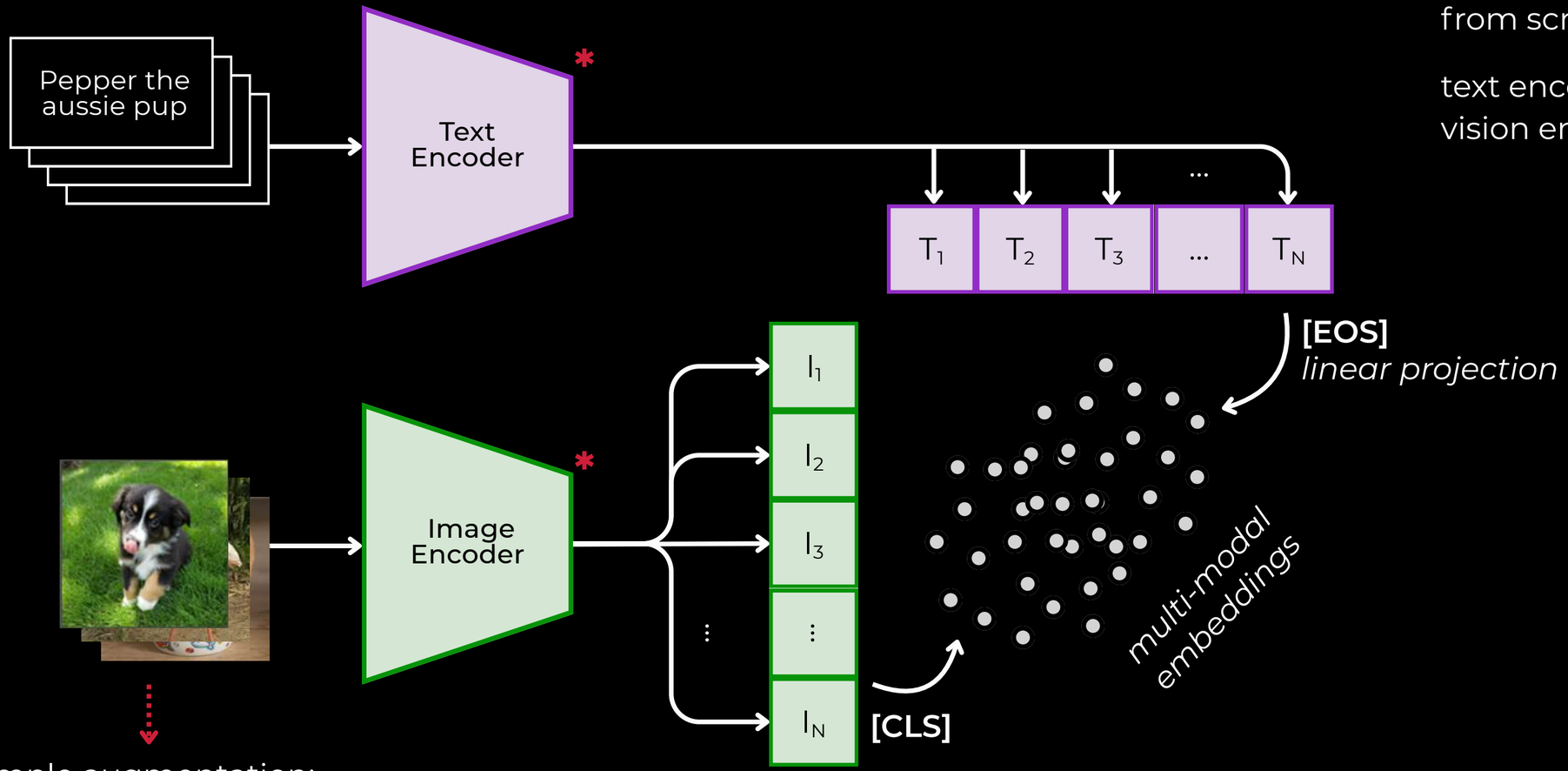


simple augmentation:
random square crop
from resized images

* both encoders are trained from scratch since WIT is large

text encoder → Transformer
vision encoder → ResNet-50 or ViT

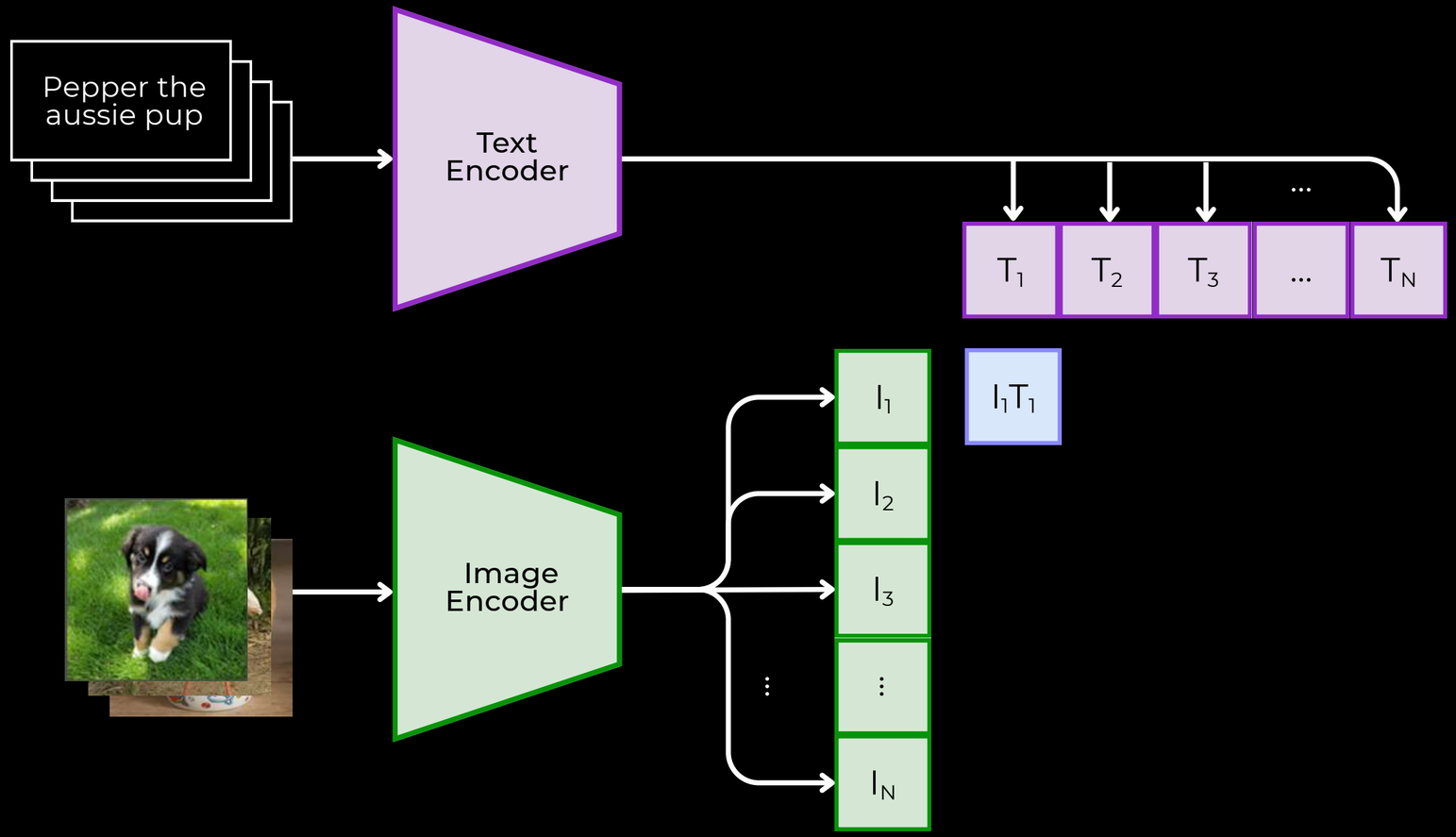
Contrastive Pre-training



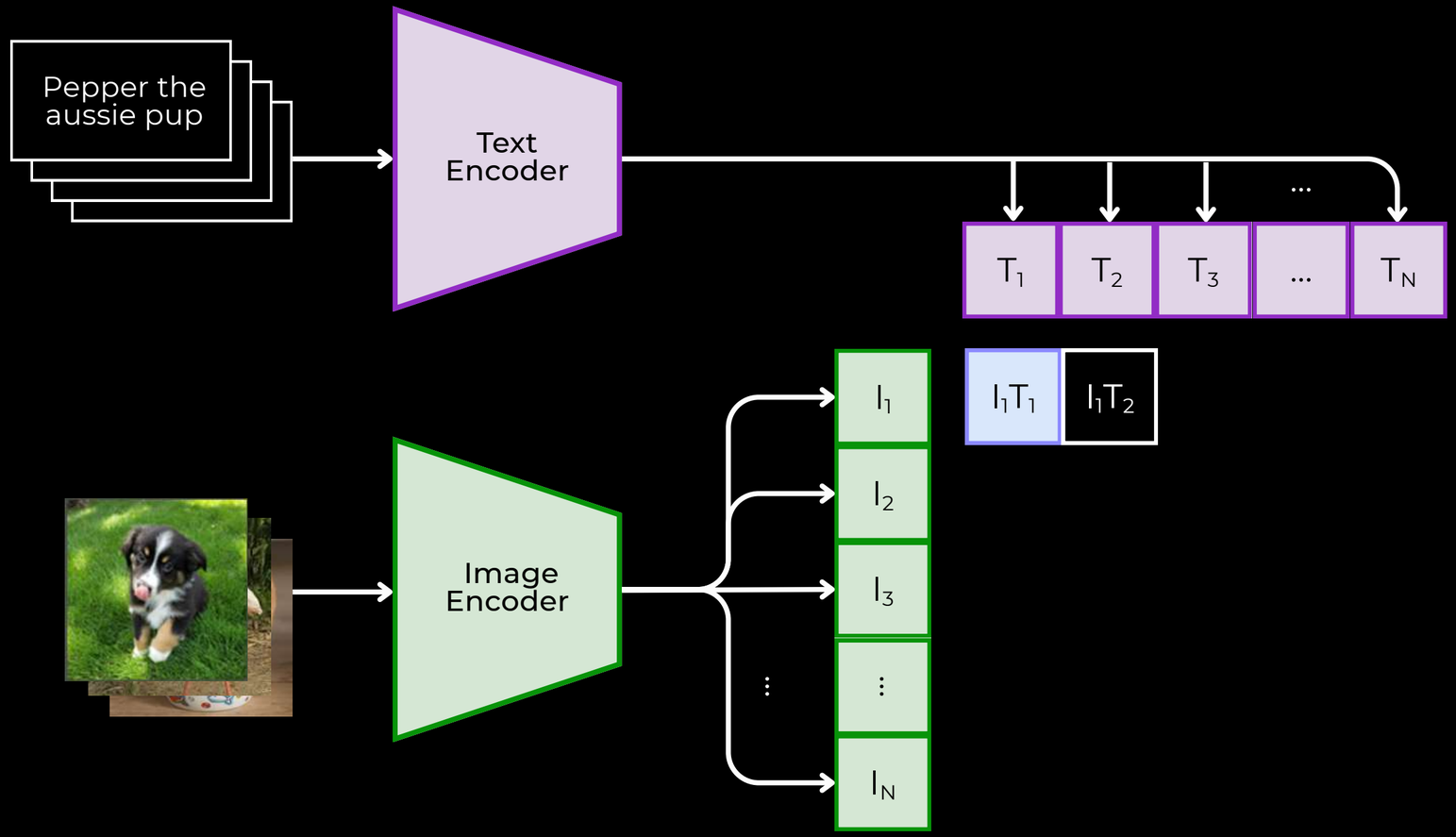
* both encoders are trained from scratch since WIT is large

text encoder → Transformer
vision encoder → ResNet-50 or ViT

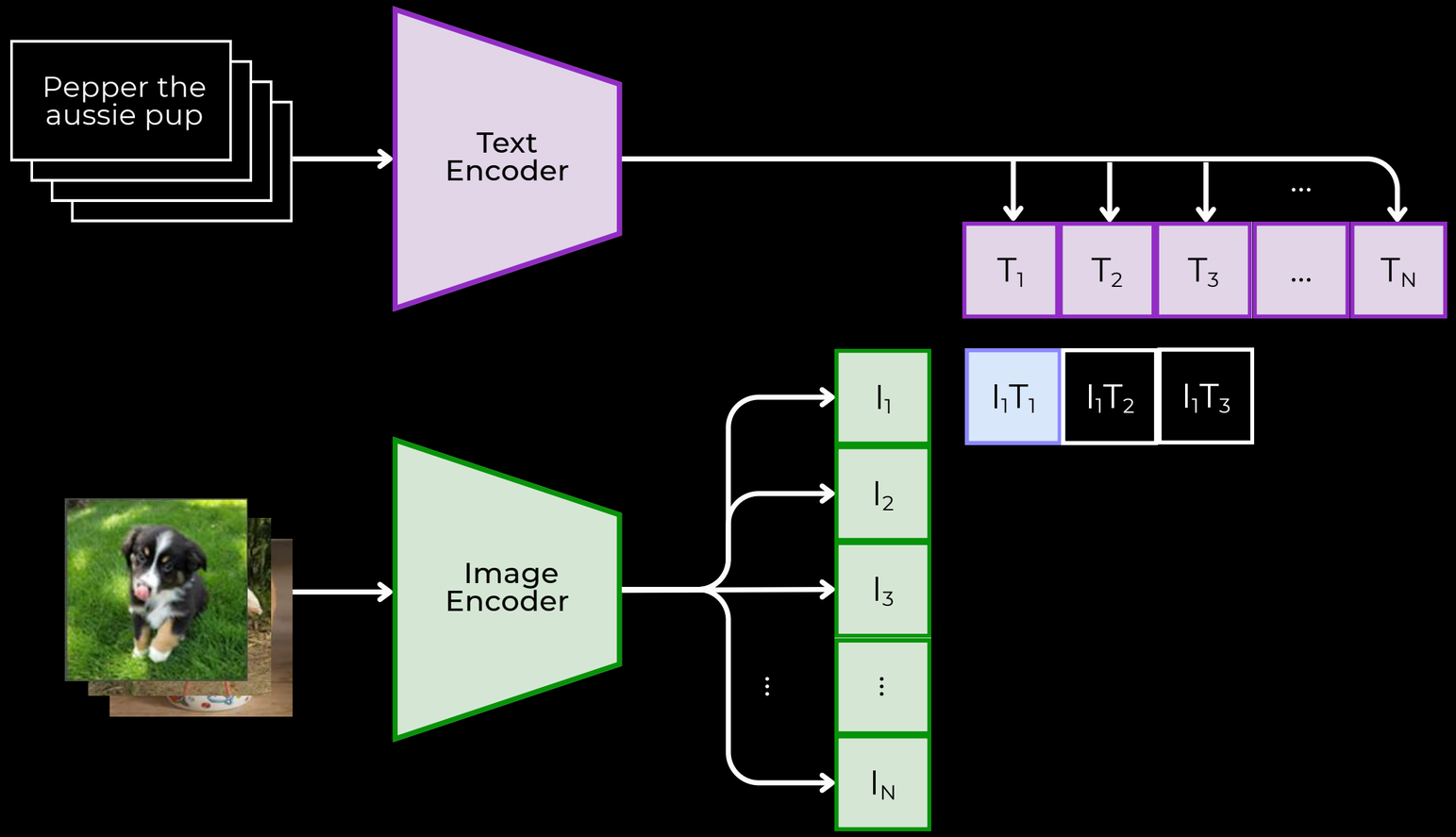
Contrastive Pre-training



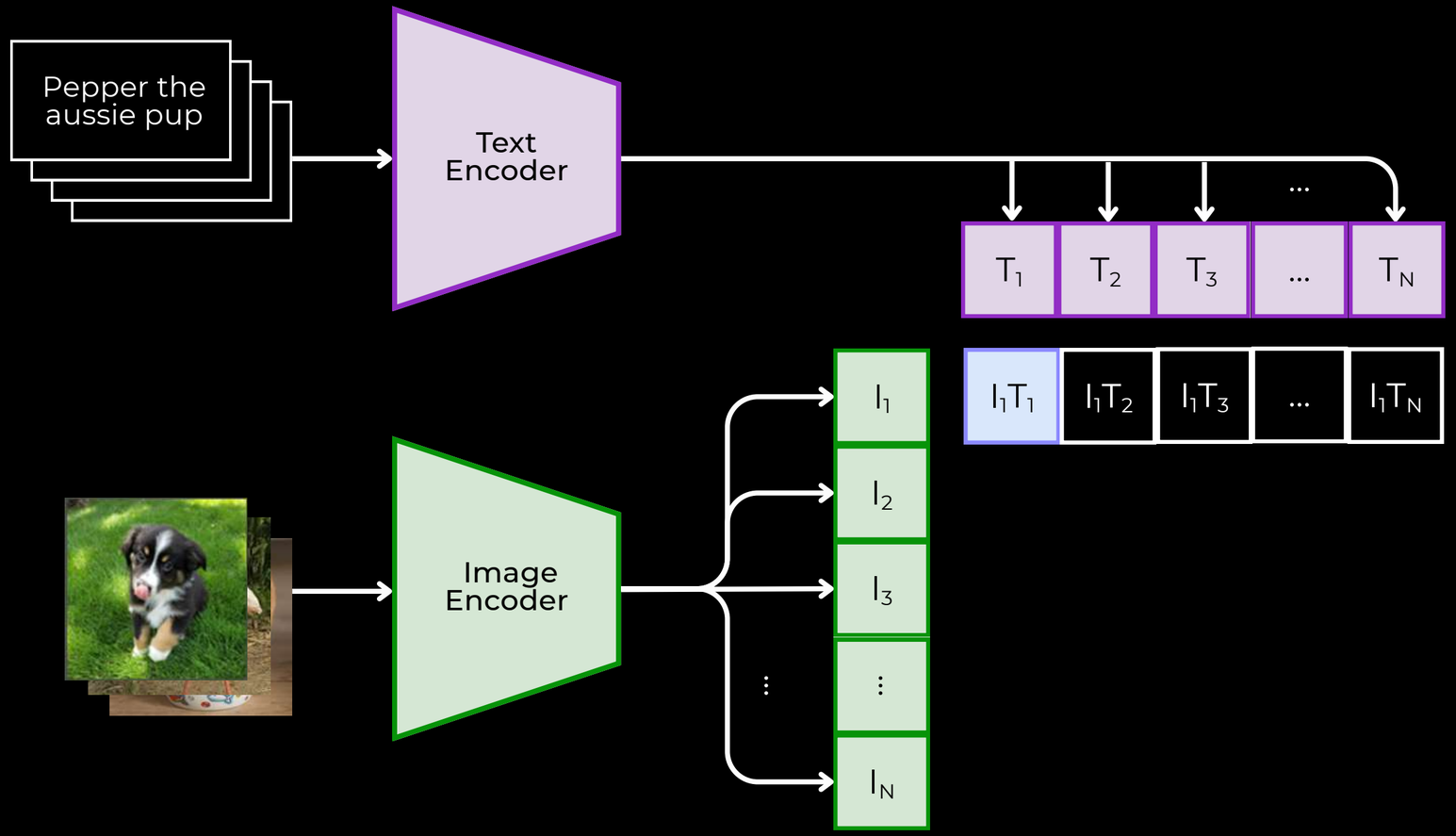
Contrastive Pre-training



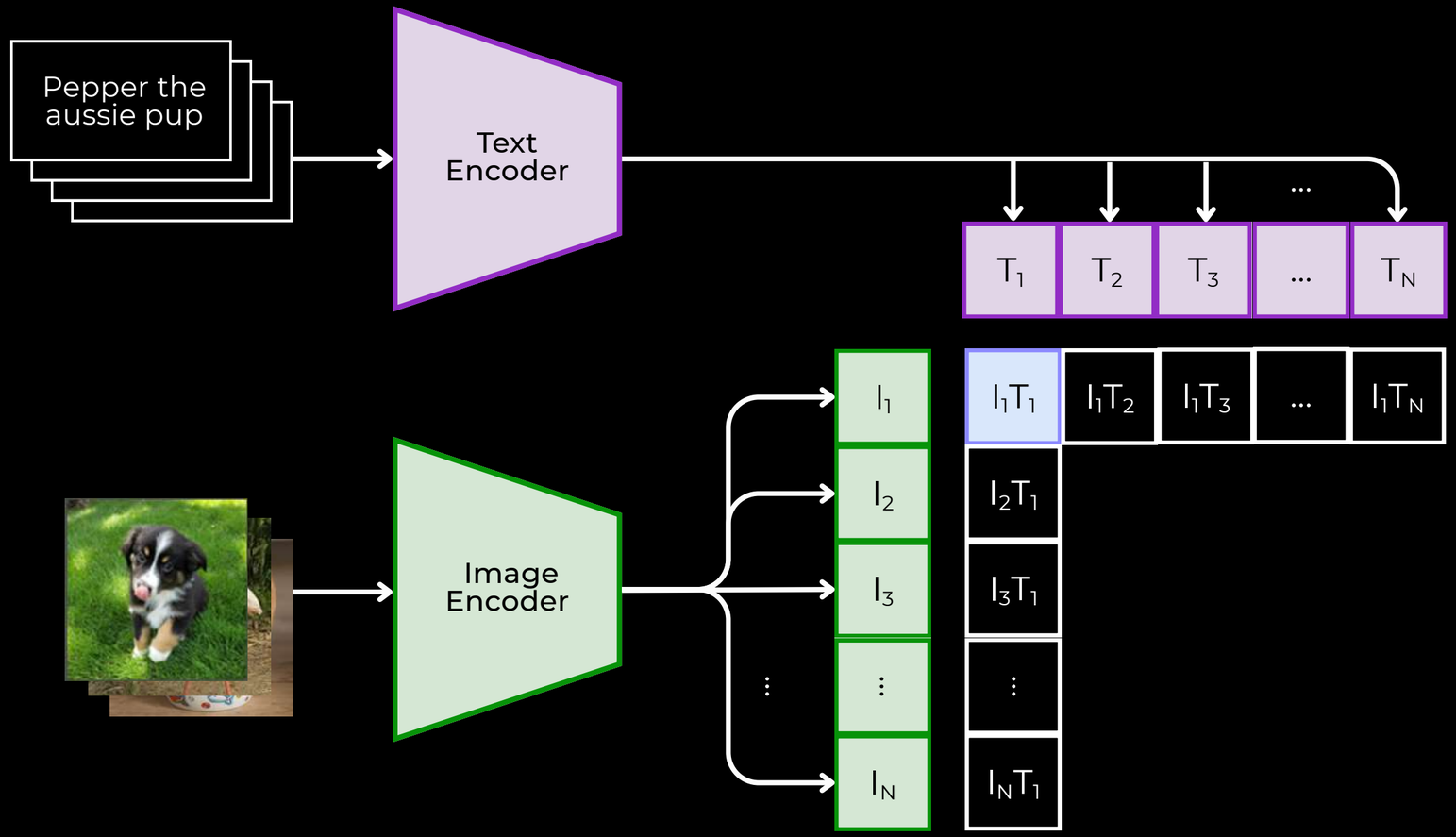
Contrastive Pre-training



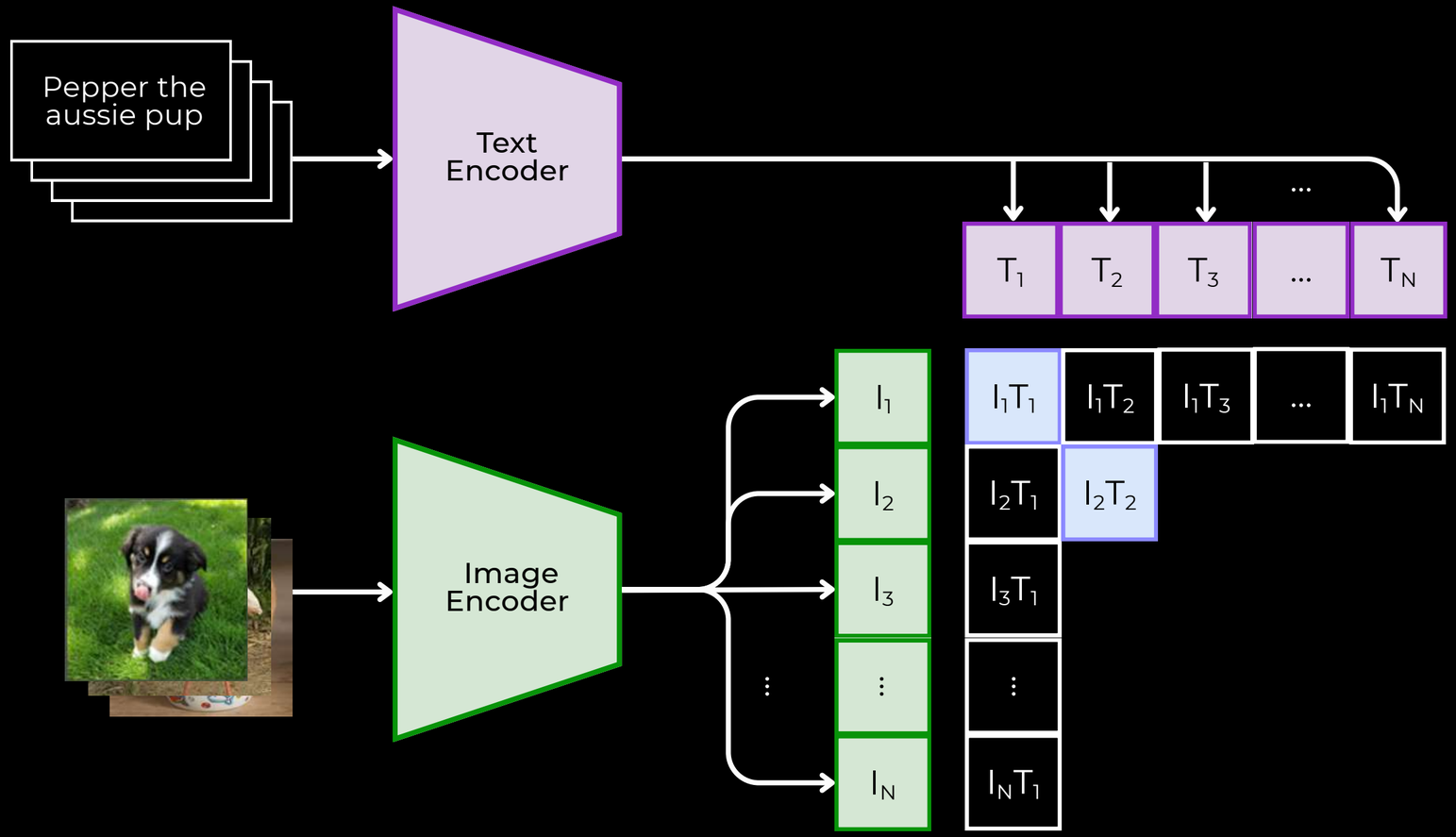
Contrastive Pre-training



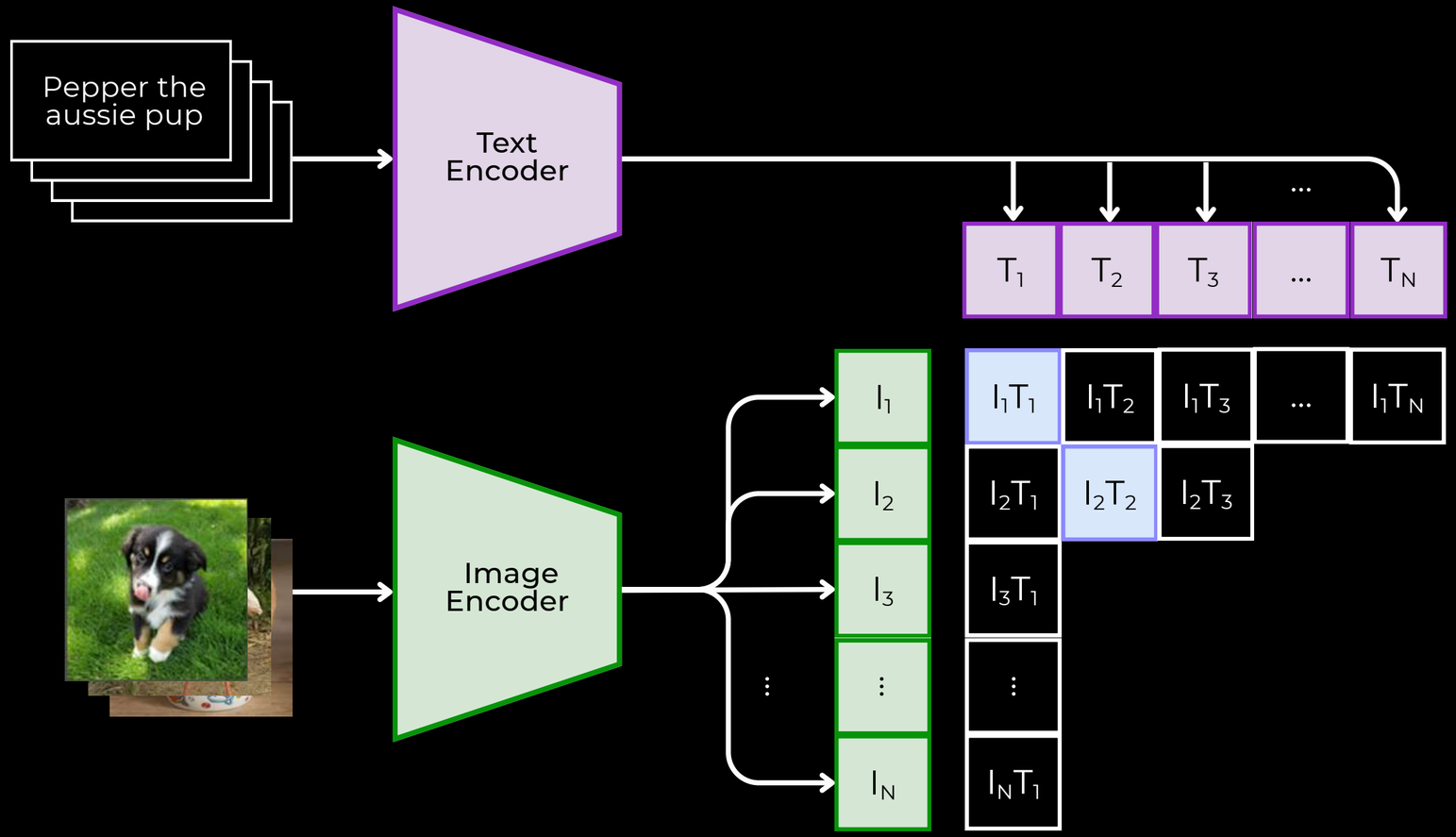
Contrastive Pre-training



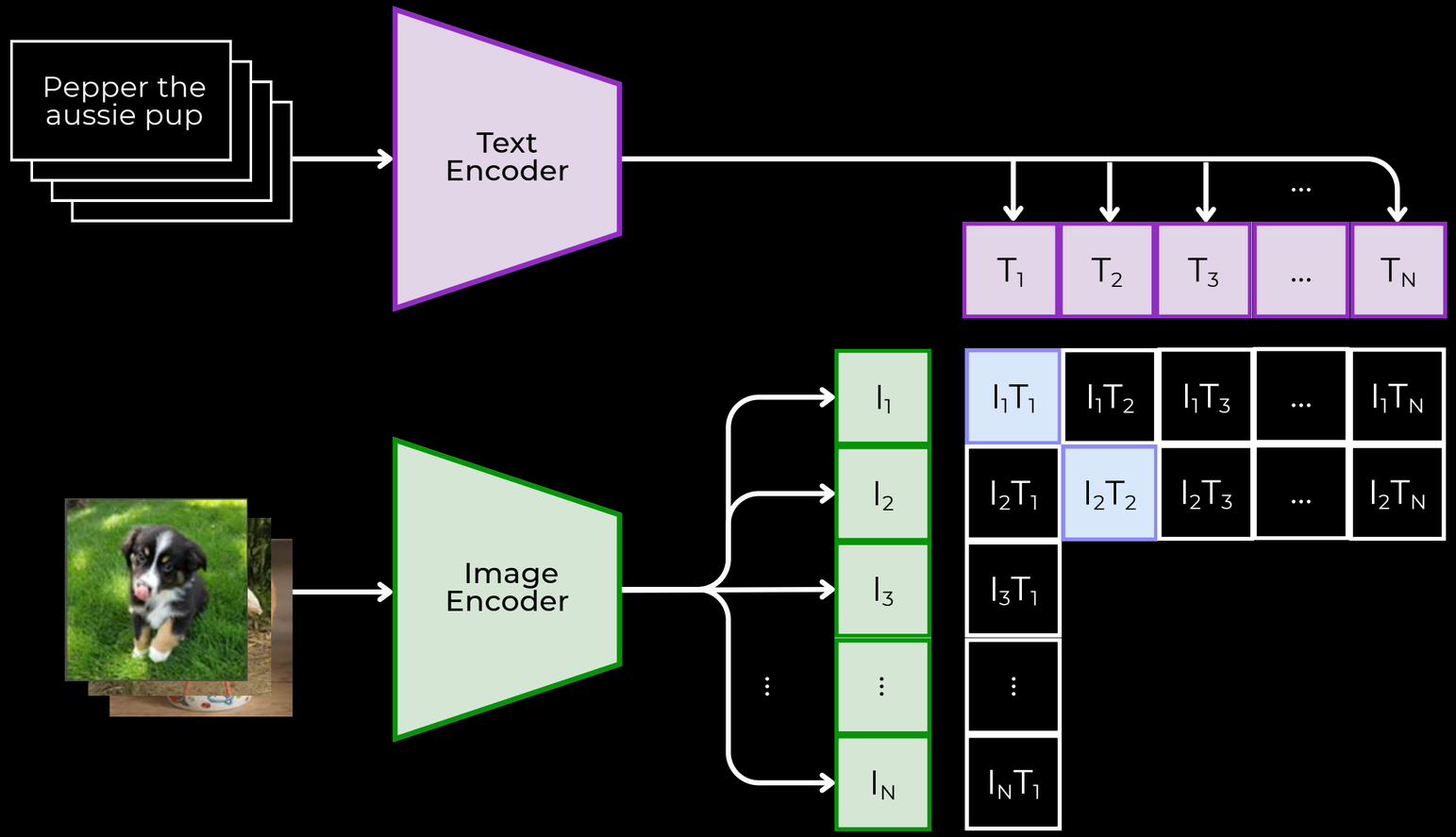
Contrastive Pre-training



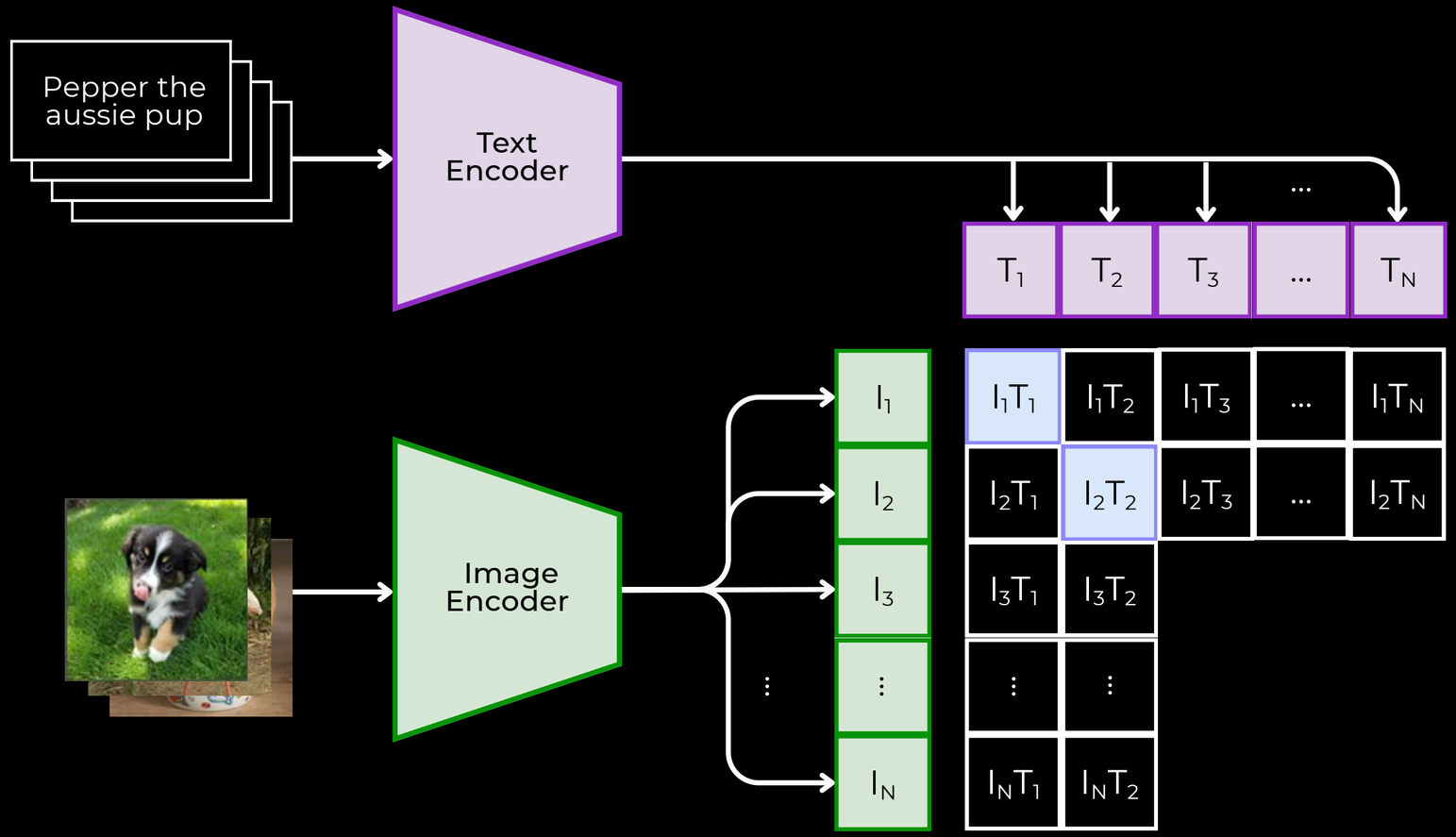
Contrastive Pre-training



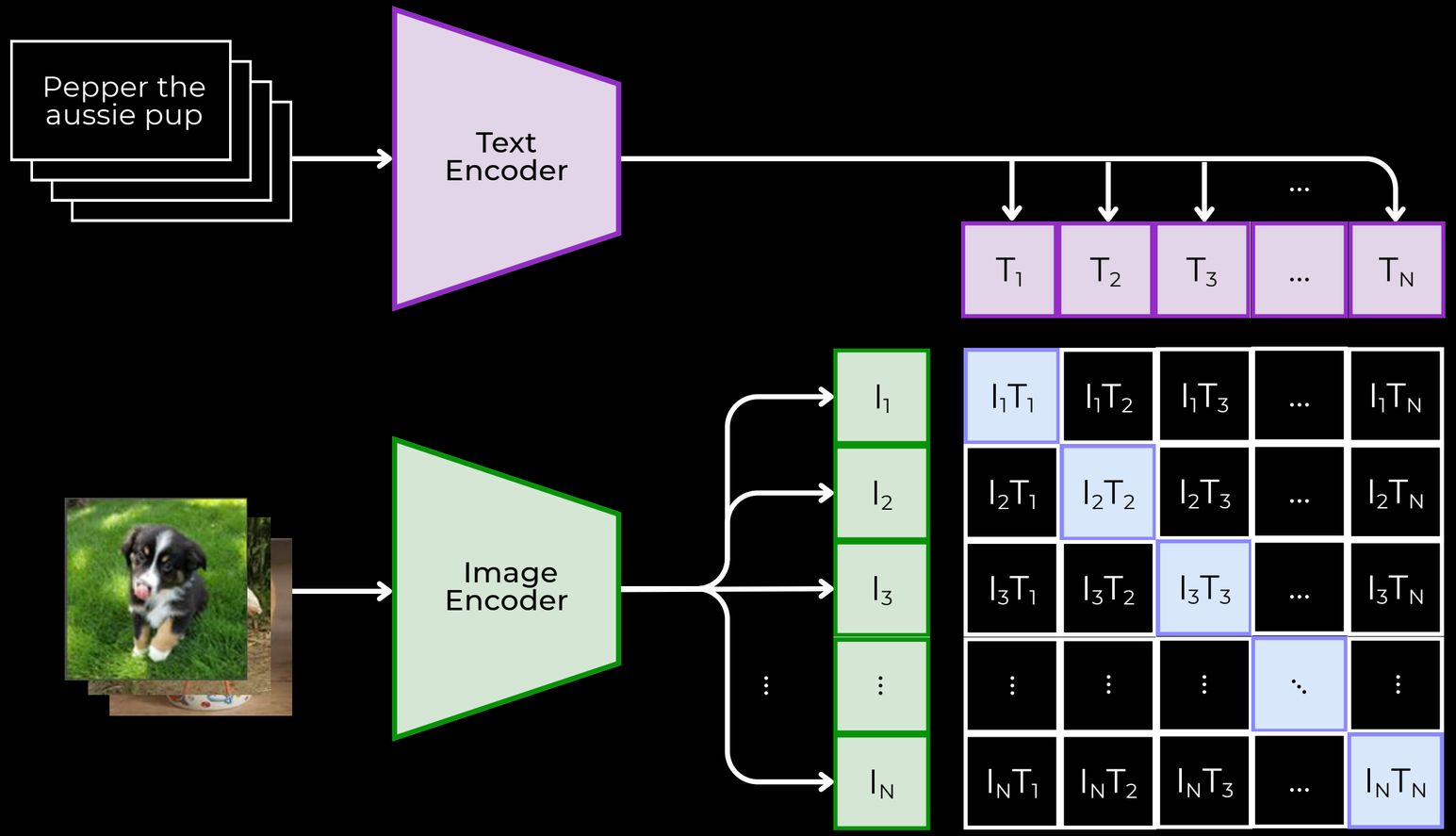
Contrastive Pre-training



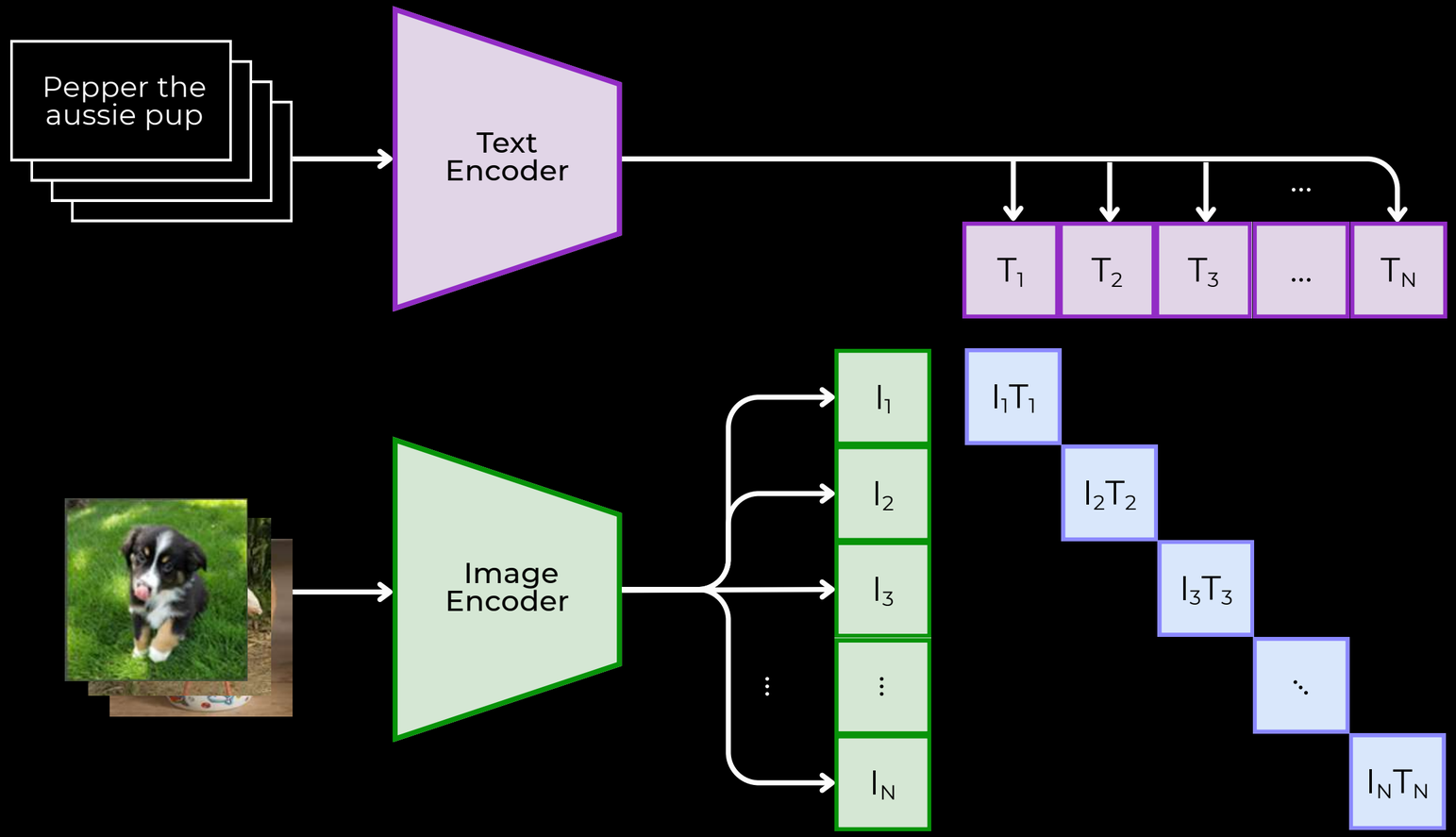
Contrastive Pre-training



Contrastive Pre-training



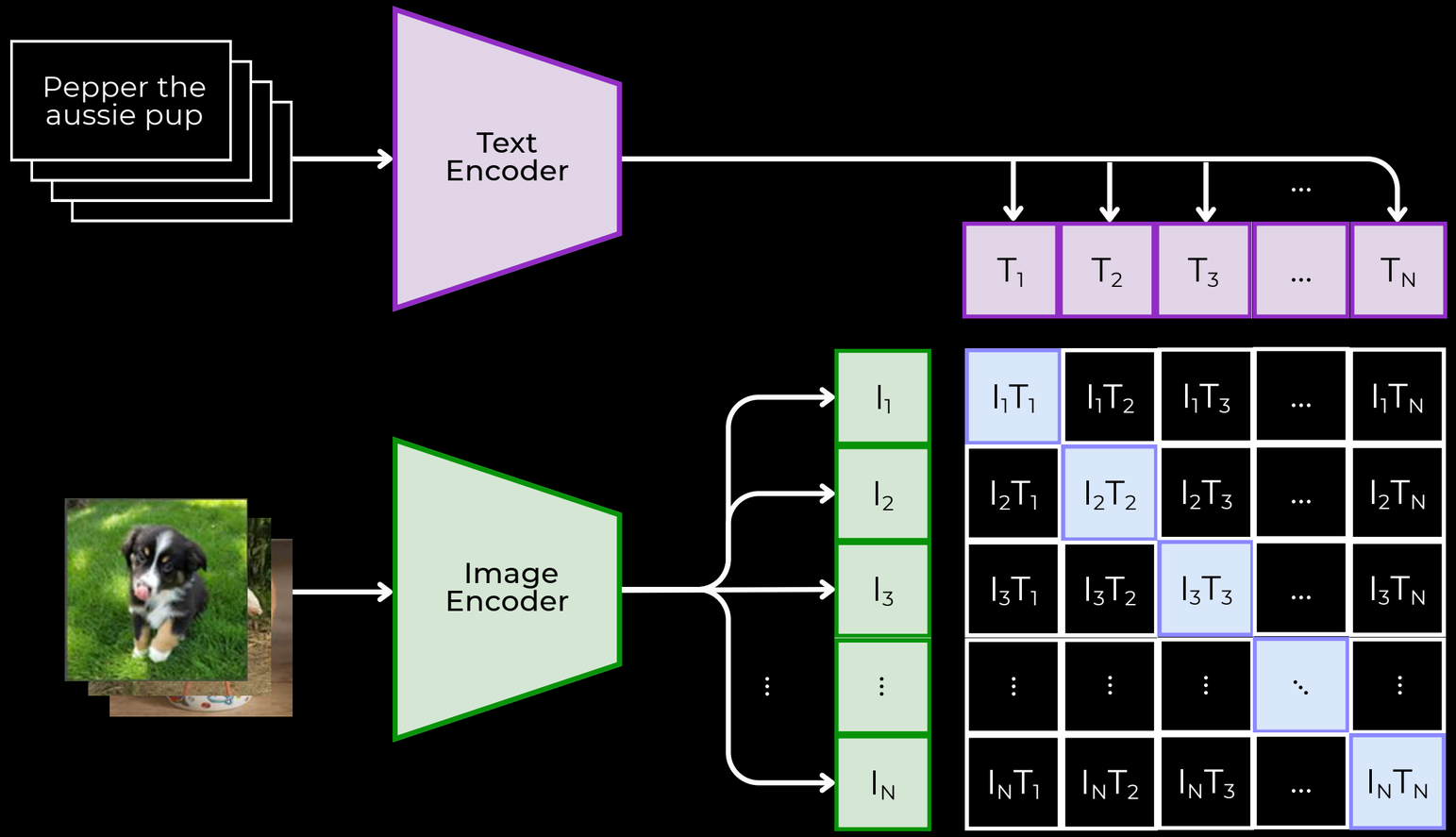
Contrastive Pre-training



LOSS

1. maximize the cosine similarity of the image and text embeddings of the *N* real pairs in the batch

Contrastive Pre-training

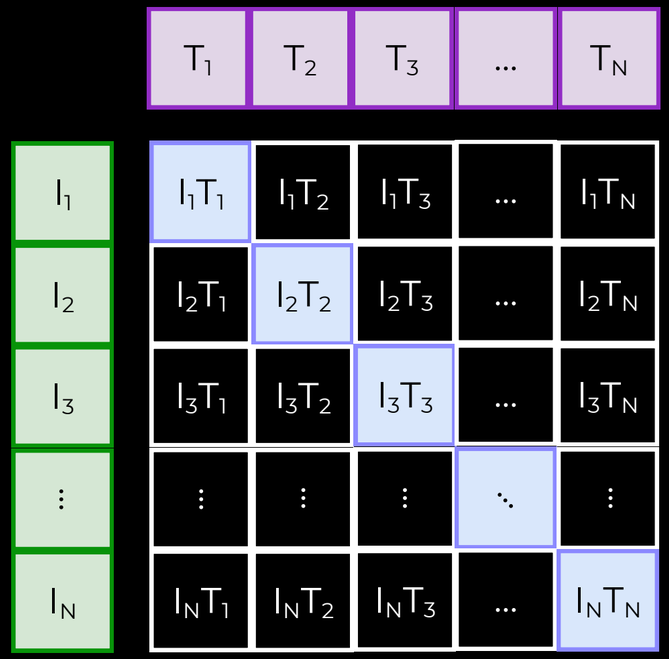


LOSS

1. maximize the cosine similarity of the image and text embeddings of the *N* real pairs in the batch

2. minimize the cosine dimilarity of the embeddings of the *N(N-1)* incorrect pairings

Contrastive Pre-training

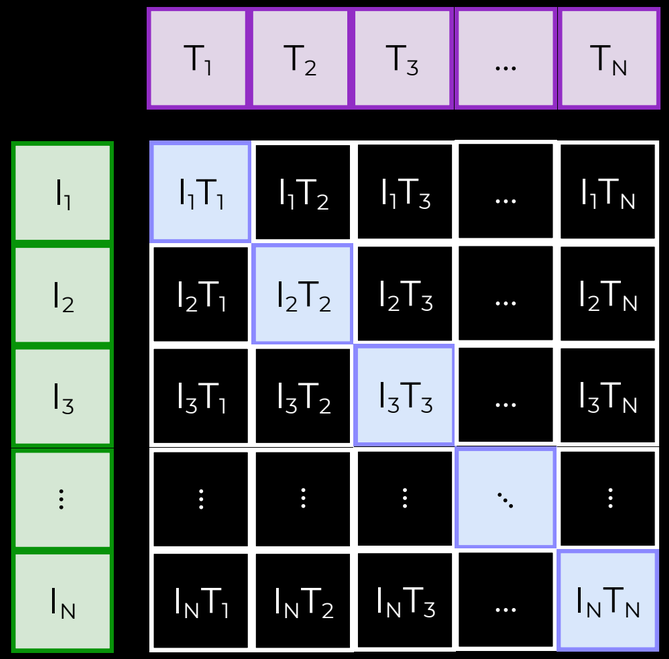


LOSS

1. maximize the cosine similarity of the image and text embeddings of the *N* real pairs in the batch

2. minimize the cosine dimilarity of the embeddings of the *N(N-1)* incorrect pairings

Contrastive Pre-training



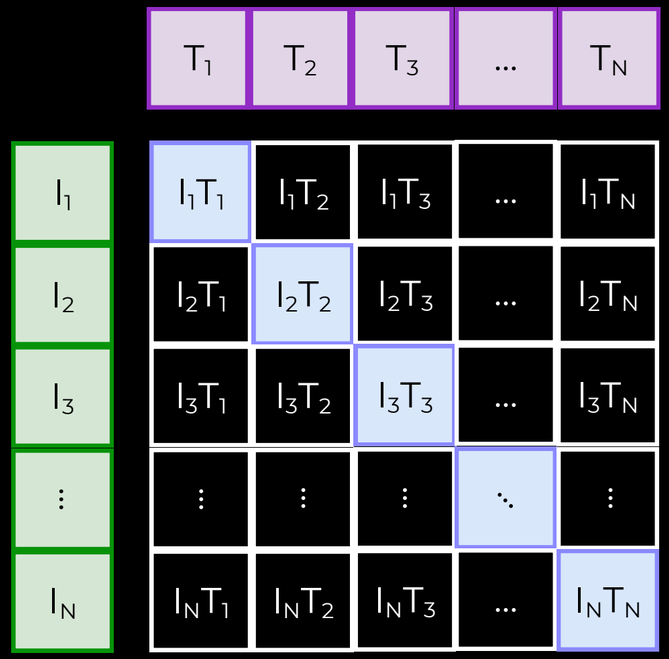
LOSS

1. maximize the cosine similarity of the image and text embeddings of the *N* real pairs in the batch

2. minimize the cosine dimilarity of the embeddings of the *N(N-1)* incorrect pairings

$$[\mathcal{L} = -\frac{1}{2|B|}\sum_{i=1}^{|B|}\left(\log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_i^\top y_j)} + \log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_j^\top y_i)}\right)]$$

Contrastive Pre-training



LOSS

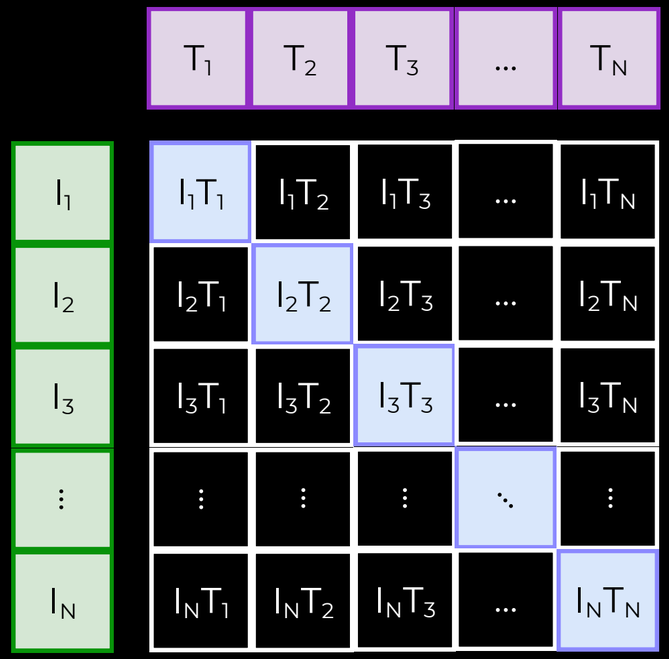
1. maximize the cosine similarity of the image and text embeddings of the *N* real pairs in the batch

2. minimize the cosine dimilarity of the embeddings of the *N(N-1)* incorrect pairings

$$[\mathcal{L} = -\frac{1}{2|B|}\sum_{i=1}^{|B|}\left(\log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_i^\top y_j)} + \log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_j^\top y_i)}\right)]$$

"minibatch"
of size 32,768

Contrastive Pre-training



LOSS

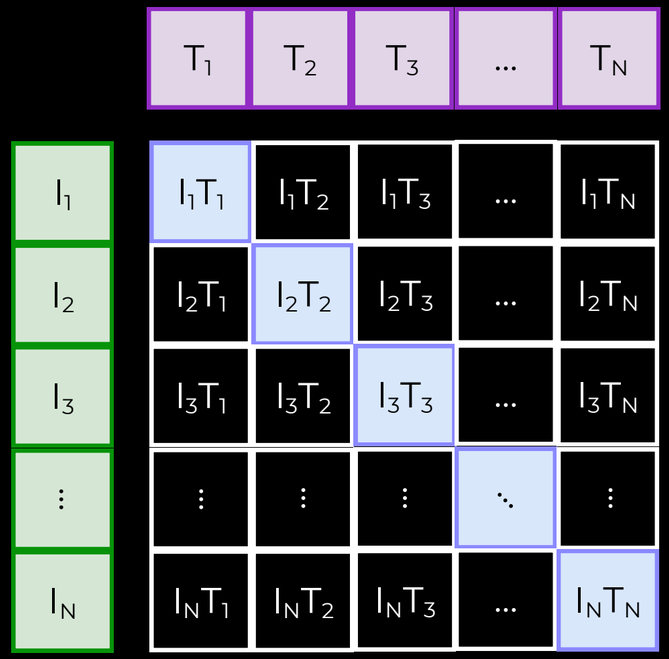
1. maximize the cosine similarity of the image and text embeddings of the *N* real pairs in the batch

2. minimize the cosine dimilarity of the embeddings of the *N(N-1)* incorrect pairings

$$[\mathcal{L} = -\frac{1}{2|B|}\sum_{i=1}^{|B|}\left(\log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_i^\top y_j)} + \log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_j^\top y_i)}\right)]$$

image → text softmax

Contrastive Pre-training



LOSS

1. maximize the cosine similarity of the image and text embeddings of the *N* real pairs in the batch

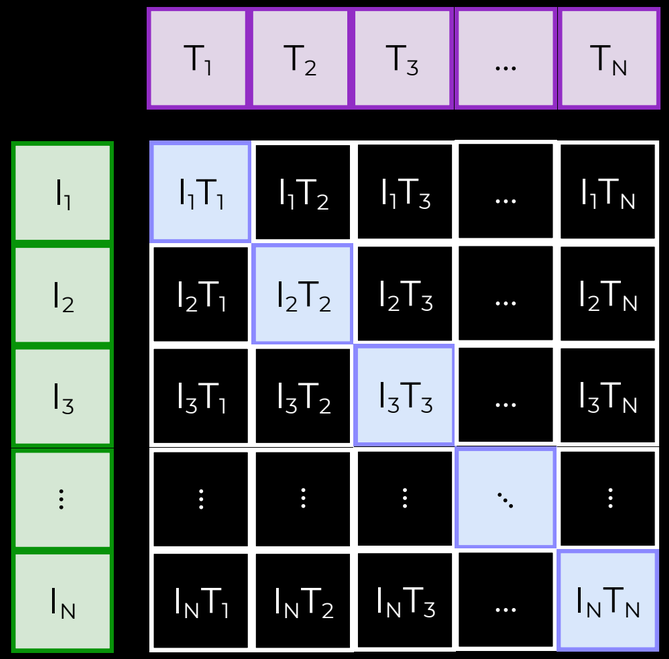
2. minimize the cosine dimilarity of the embeddings of the *N(N-1)* incorrect pairings

$$[\mathcal{L} = -\frac{1}{2|B|}\sum_{i=1}^{|B|}\left(\log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_i^\top y_j)} + \log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_j^\top y_i)}\right)]$$

image → text softmax

text → image softmax

Contrastive Pre-training



LOSS

1. maximize the cosine similarity of the image and text embeddings of the *N* real pairs in the batch

2. minimize the cosine dimilarity of the embeddings of the *N(N-1)* incorrect pairings

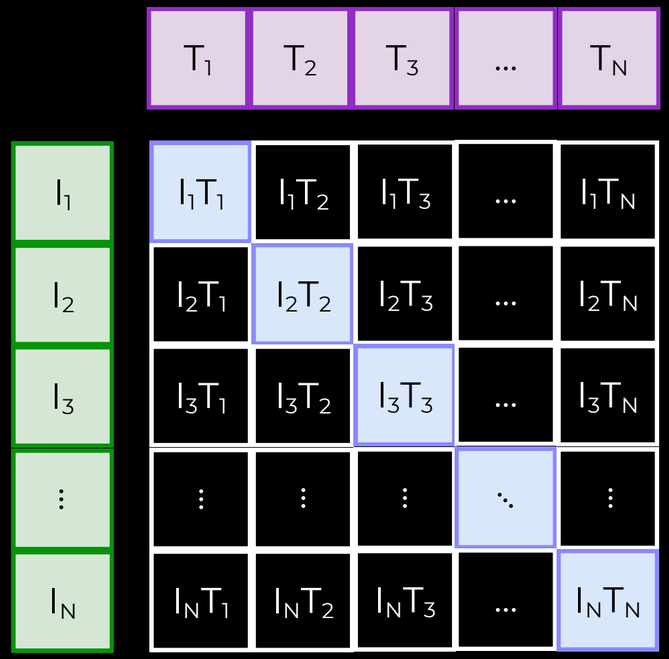
$$[\mathcal{L} = -\frac{1}{2|B|}\sum_{i=1}^{|B|}\left(\log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_i^\top y_j)} + \log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_j^\top y_i)}\right)]$$

image → text softmax

text → image softmax

where $[x_i = \frac{f(I_i)}{\|f(I_i)\|_2}, \quad y_i = \frac{g(T_i)}{\|g(T_i)\|_2}]$

Contrastive Pre-training



LOSS

1. maximize the cosine similarity of the image and text embeddings of the *N* real pairs in the batch

2. minimize the cosine dimilarity of the embeddings of the *N(N-1)* incorrect pairings

$$[\mathcal{L} = -\frac{1}{2|B|}\sum_{i=1}^{|B|}\left(\log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_i^\top y_j)} + \log\frac{\exp(x_i^\top y_i)}{\sum_{j=1}^{|B|}\exp(x_j^\top y_i)}\right)]$$

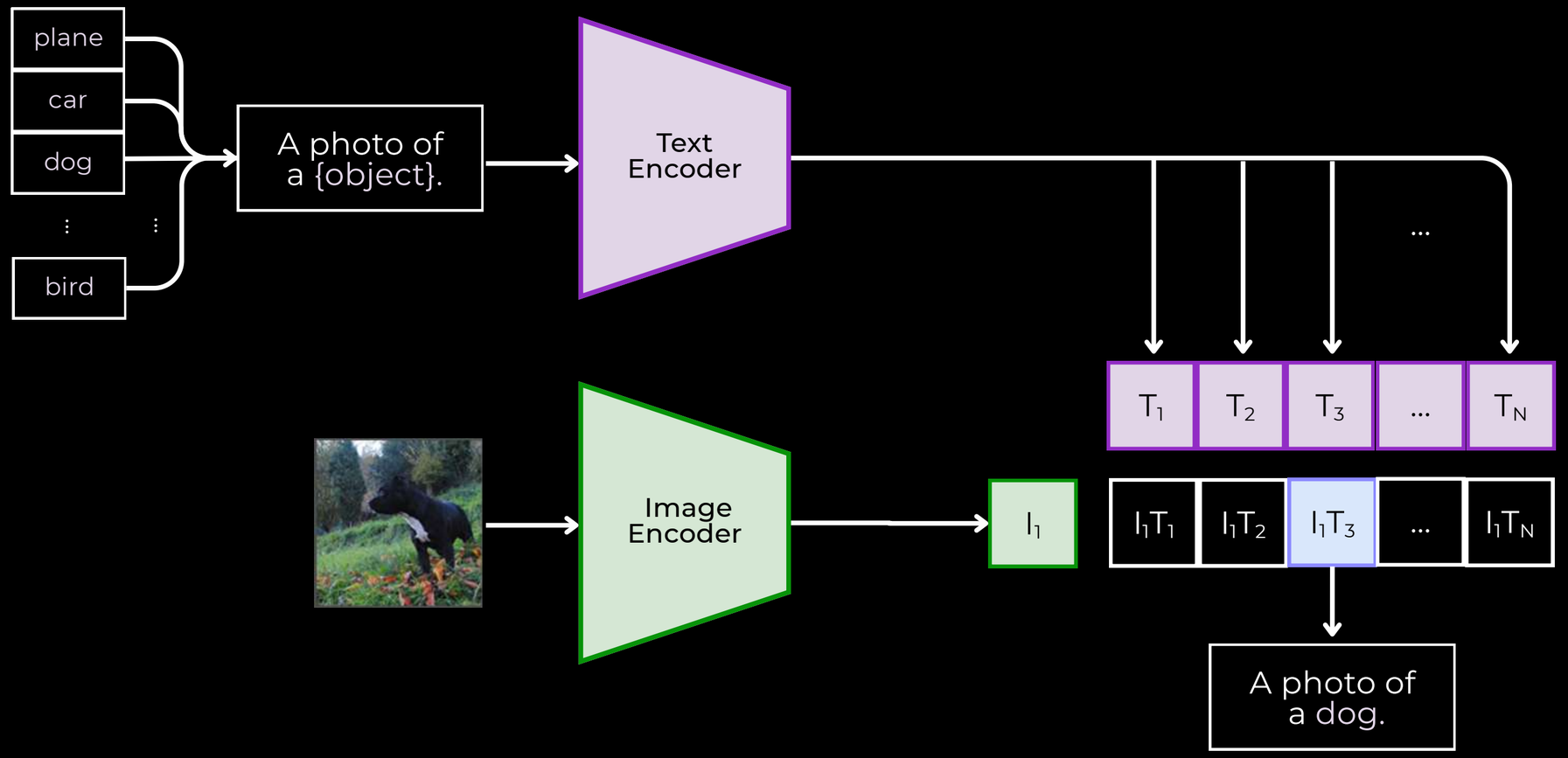
image → text softmax

text → image softmax

where $[x_i = \frac{f(I_i)}{\|f(I_i)\|_2}, \quad y_i = \frac{g(T_i)}{\|g(T_i)\|_2}]$

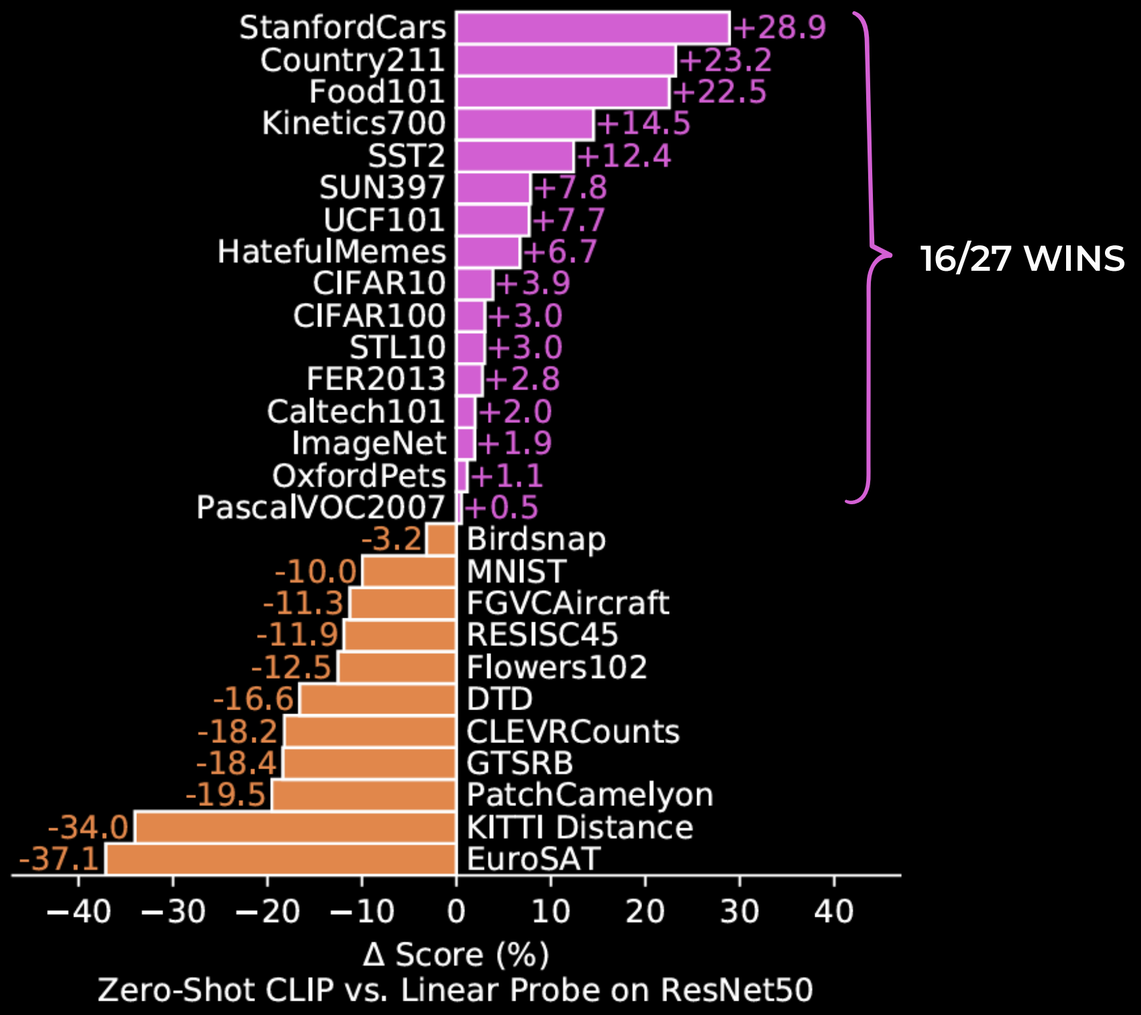
symmetric **cross-entropy** *loss over similarity scores (InfoNCE)*

Inference


plane
car
dog
bird
A photo of a {object}.
Text Encoder
...
T_1
T_2
T_3
...
T_N
Image Encoder
I_1
I_1T_1
I_1T_2
I_1T_3
...
I_1T_N
A photo of a dog.

Zero-shot learning

ZS CLIP is compared to a fully supervised linear classifier fitted on ResNet-50 features across 27 datasets



Zero-shot learning

ZS CLIP is compared to a fully supervised linear classifier fitted on ResNet-50 features across 27 datasets

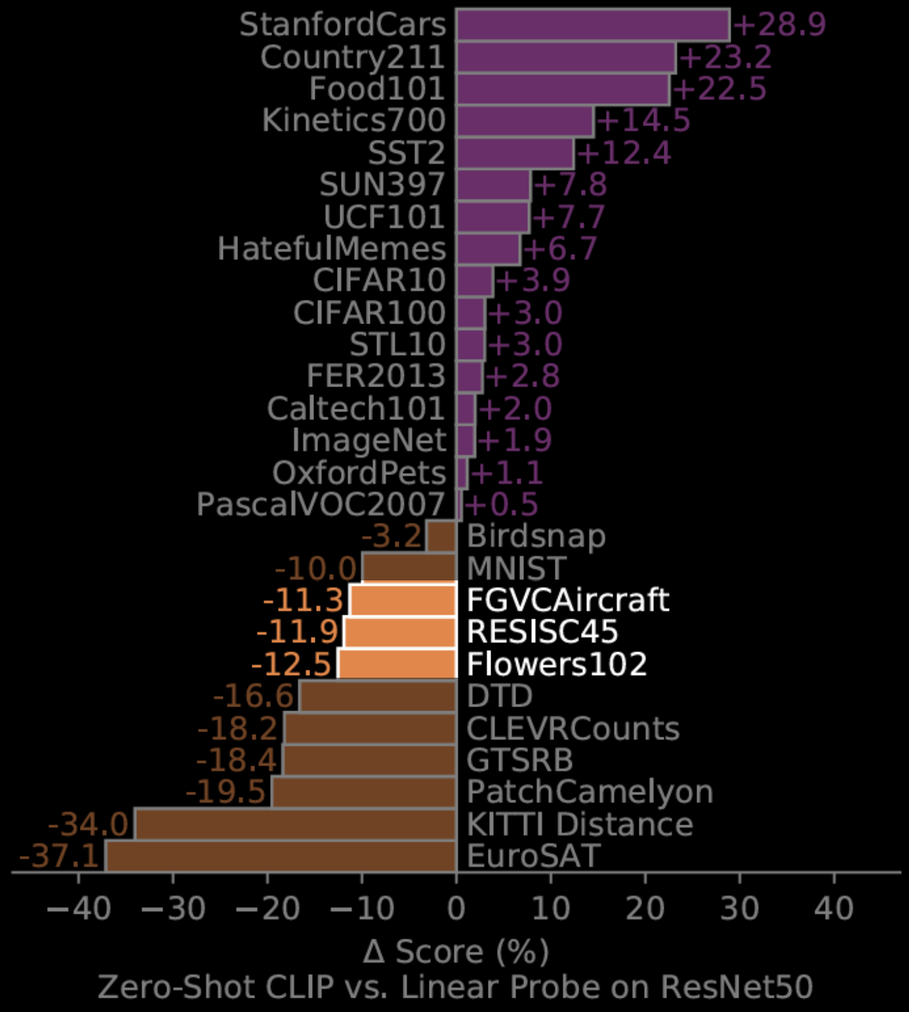
ZS CLIP outperforms logistic regression by over 20%

Zero-shot learning

ZS CLIP is compared to a fully supervised linear classifier fitted on ResNet-50 features across 27 datasets

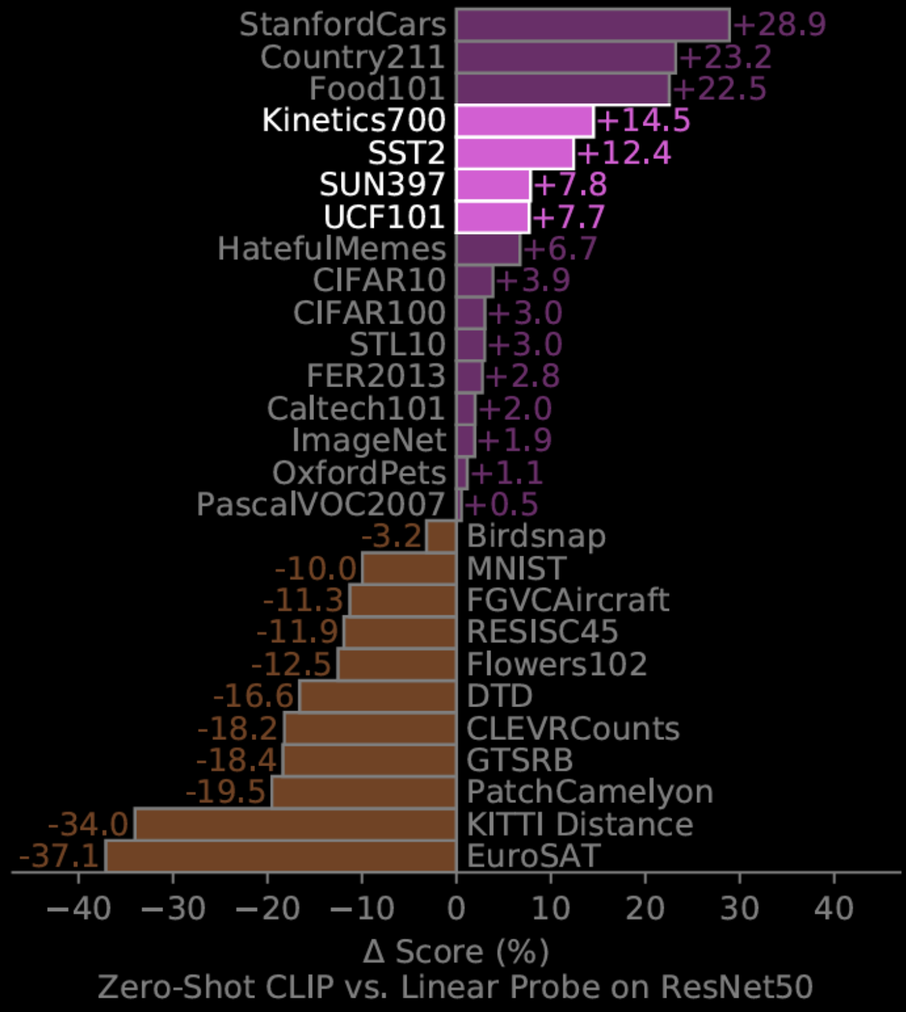
ZS CLIP underperforms logistic regression by over 10%

per-task supervision in WIT vs ImageNet (?)



Zero-shot learning

ZS CLIP is compared to a fully supervised linear classifier fitted on ResNet-50 features across 27 datasets

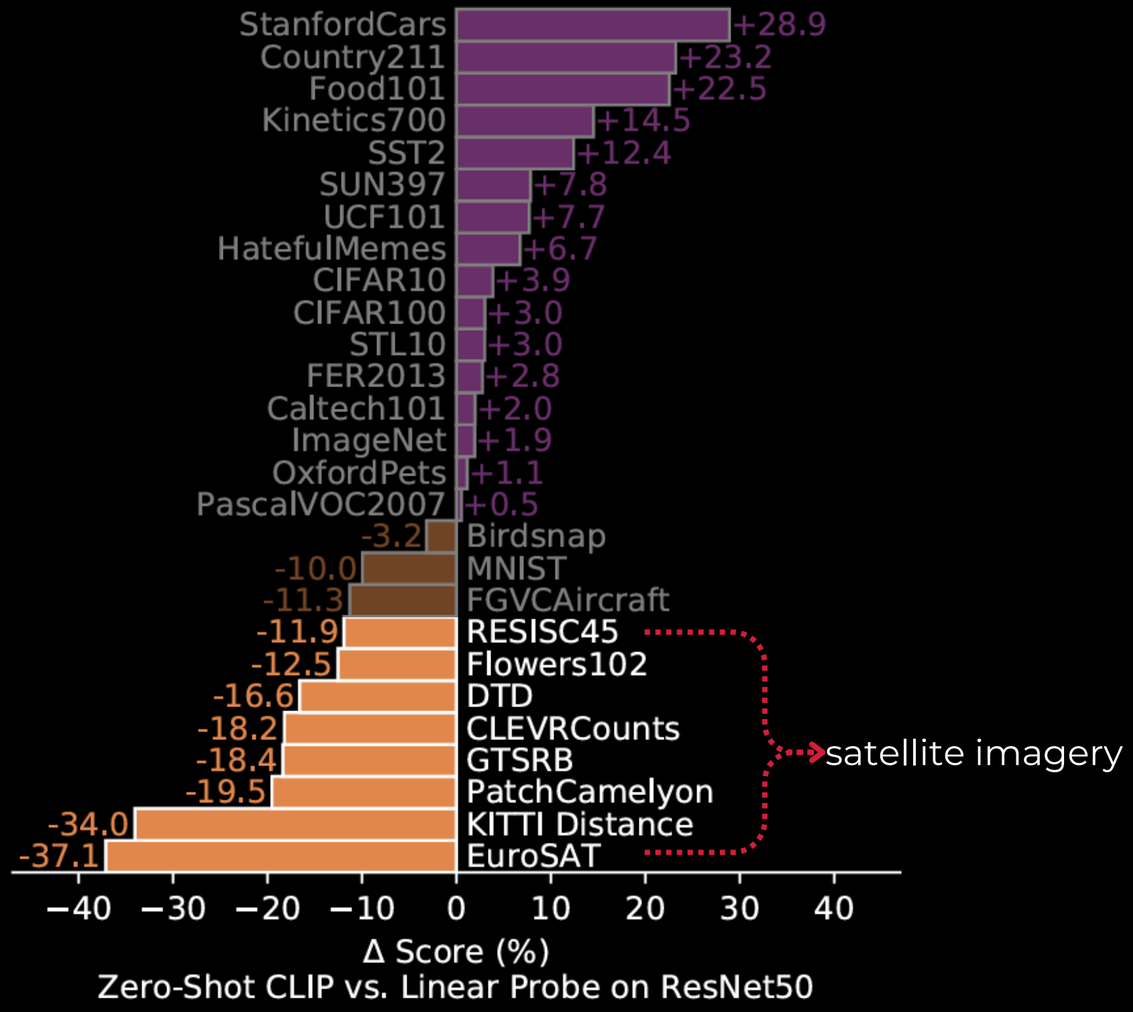


ZS CLIP overperforms logistic regression on both Kinetics700 and UCF101 (action recognition in videos)

- ImageNet
  noun-centric object supervision
- natural language
  wider supervision for visual concepts involving verbs

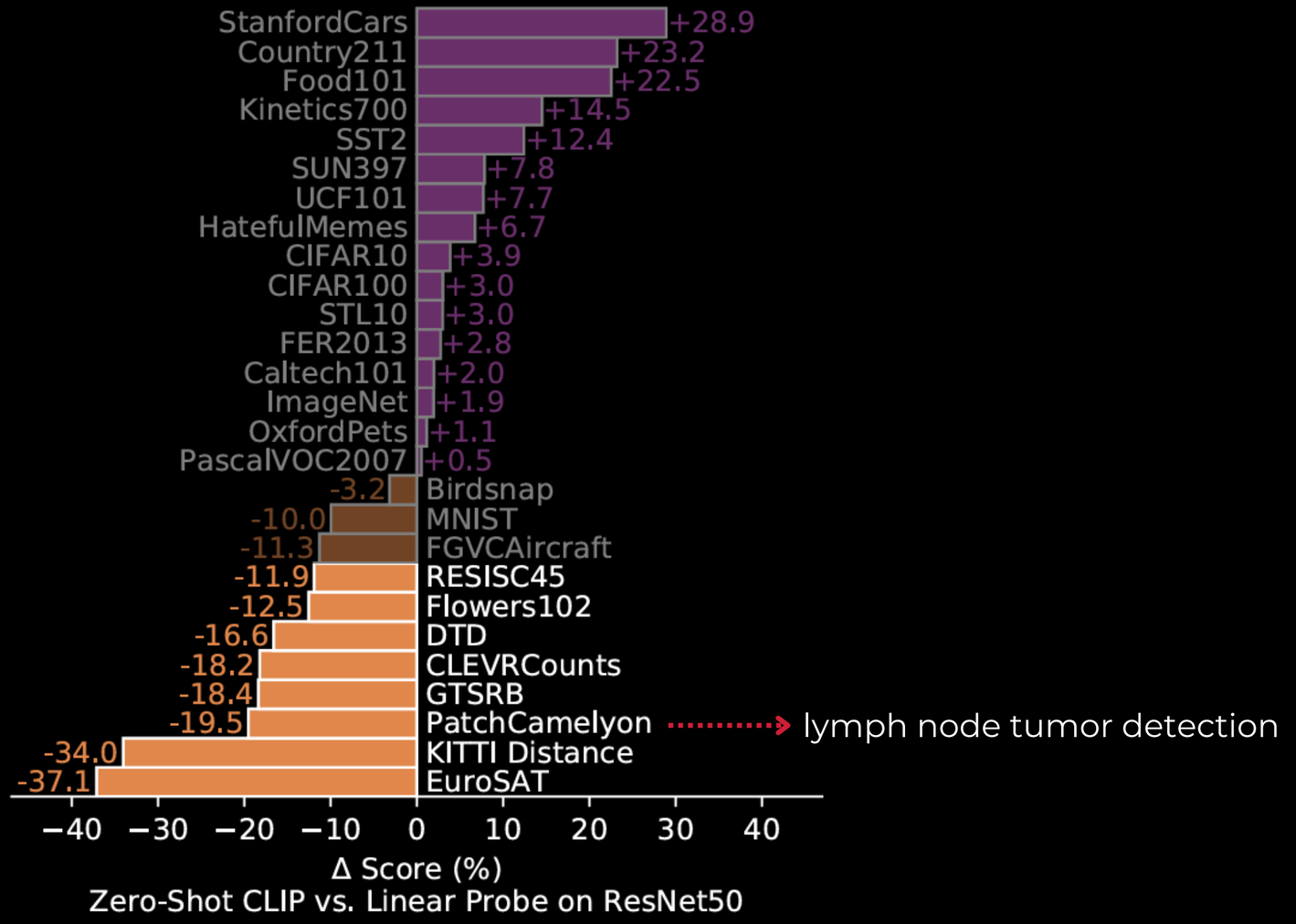
Zero-shot learning

ZS CLIP is compared to a fully supervised linear classifier fitted on ResNet-50 features across 27 datasets



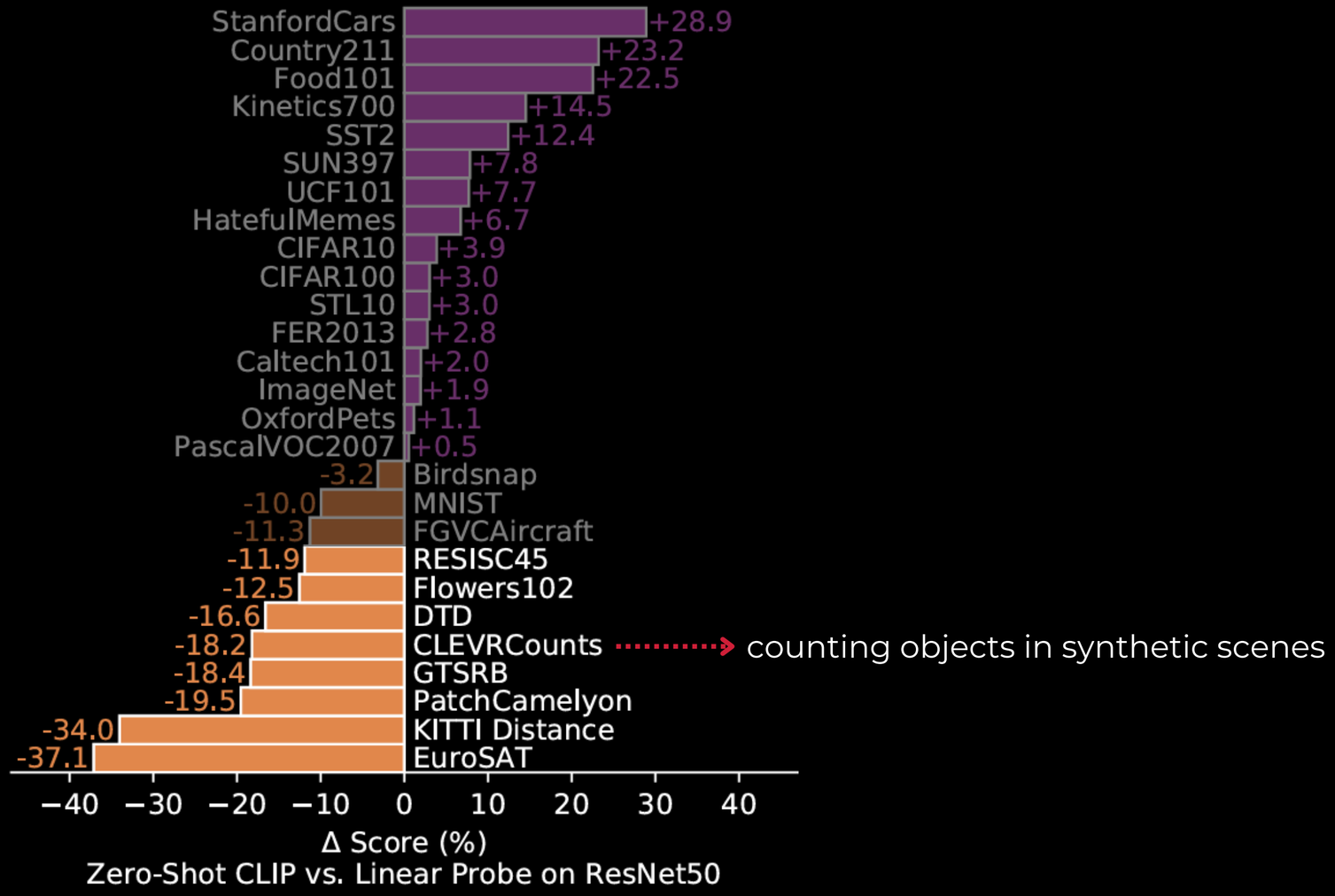
Zero-shot learning

ZS CLIP is compared to a fully supervised linear classifier fitted on ResNet-50 features across 27 datasets



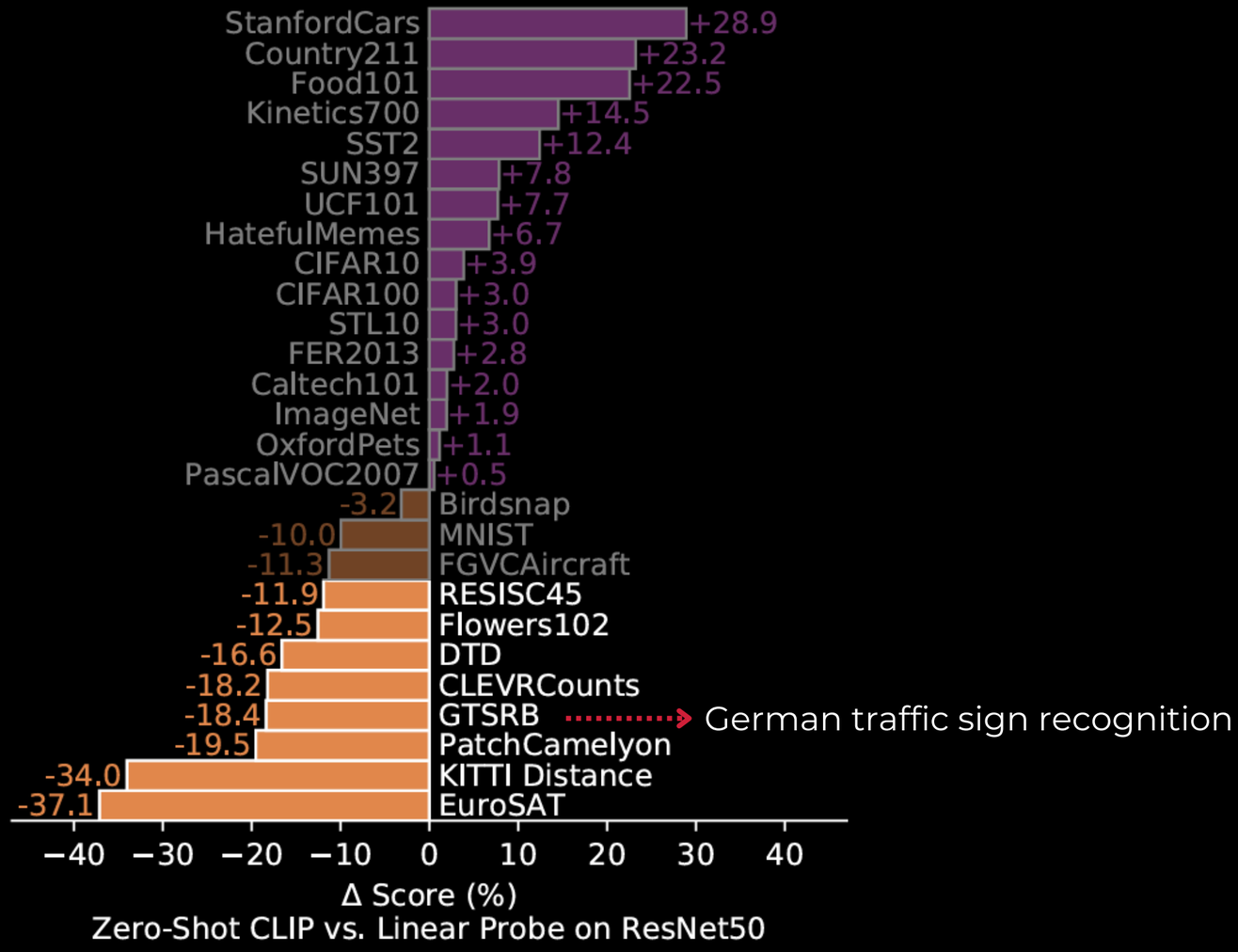
Zero-shot learning

ZS CLIP is compared to a fully supervised linear classifier fitted on ResNet-50 features across 27 datasets



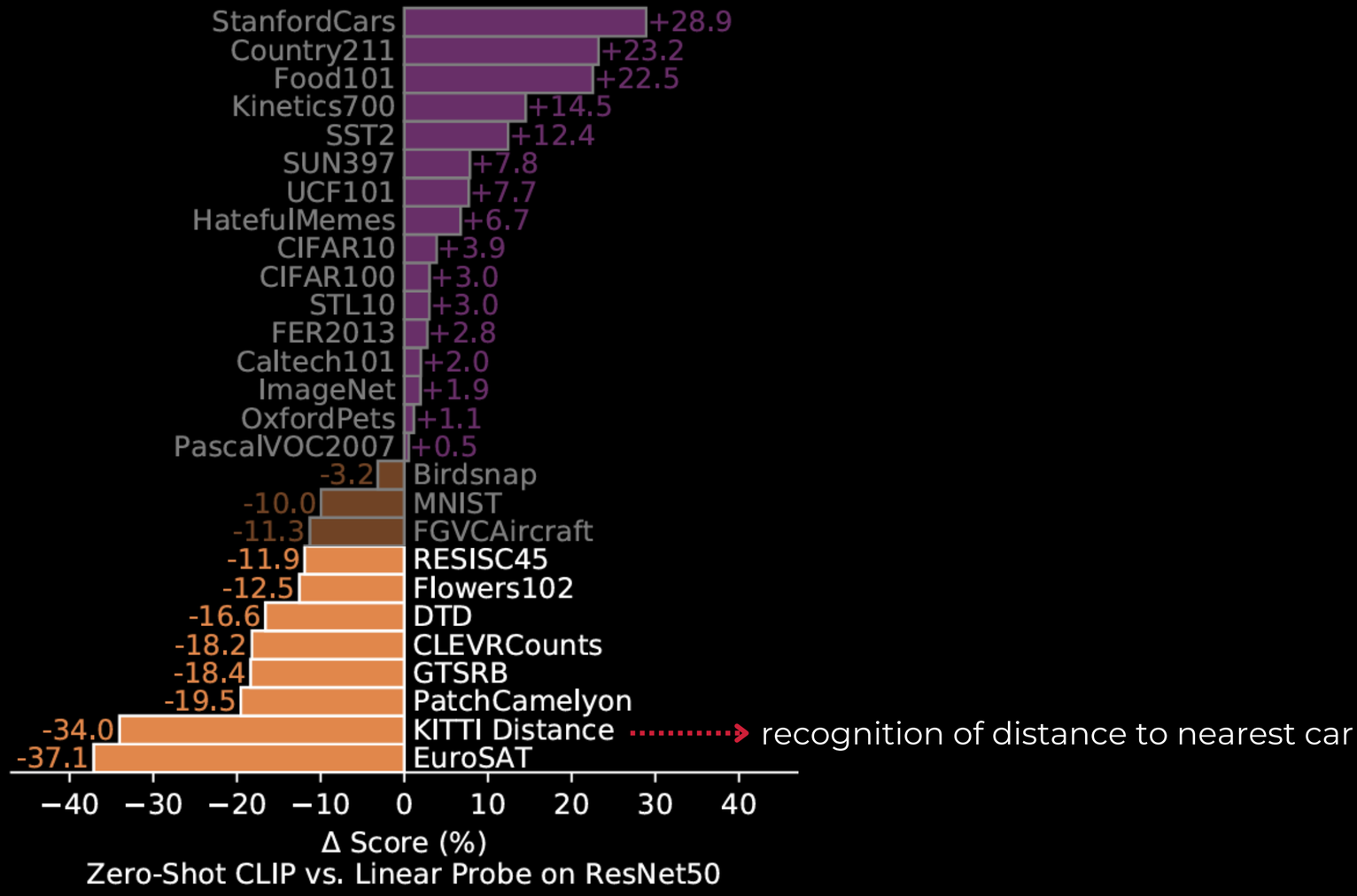
Zero-shot learning

ZS CLIP is compared to a fully supervised linear classifier fitted on ResNet-50 features across 27 datasets



Zero-shot learning

ZS CLIP is compared to a fully supervised linear classifier fitted on ResNet-50 features across 27 datasets

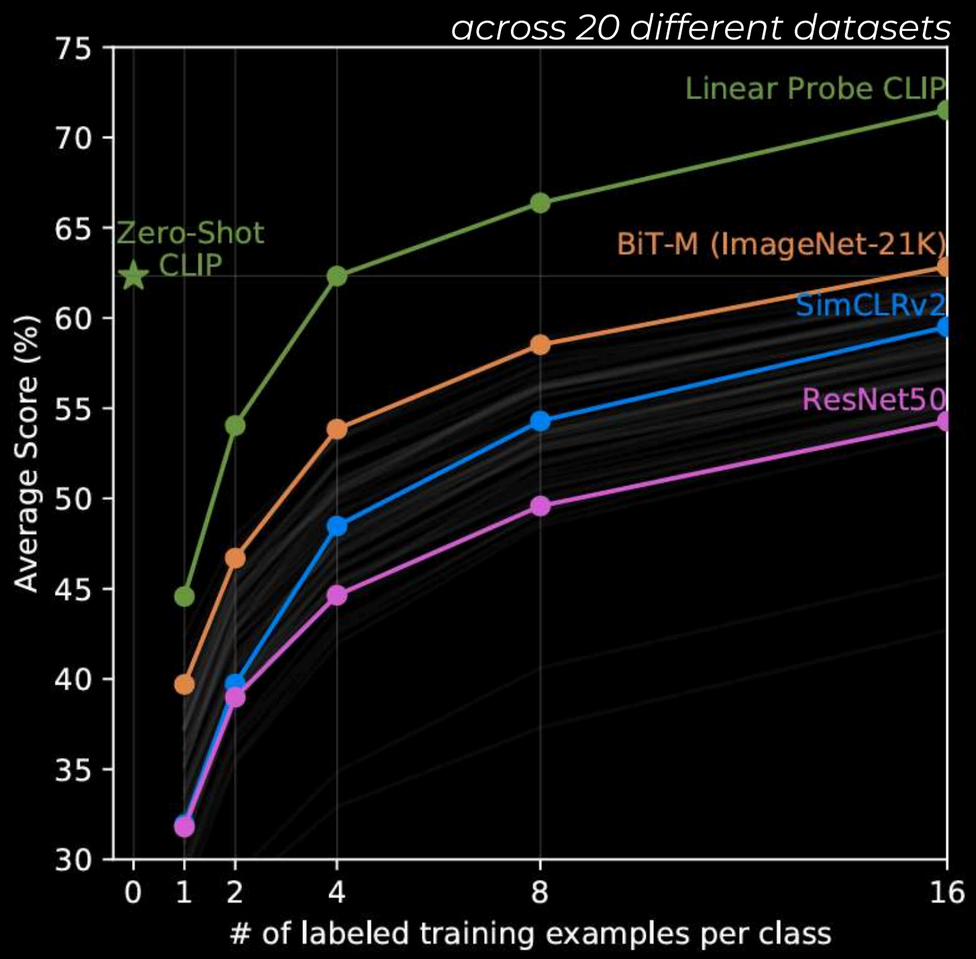


Zero-shot learning

ZS CLIP is compared to few-shot logistic regression on the features of many image models including CLIP itself

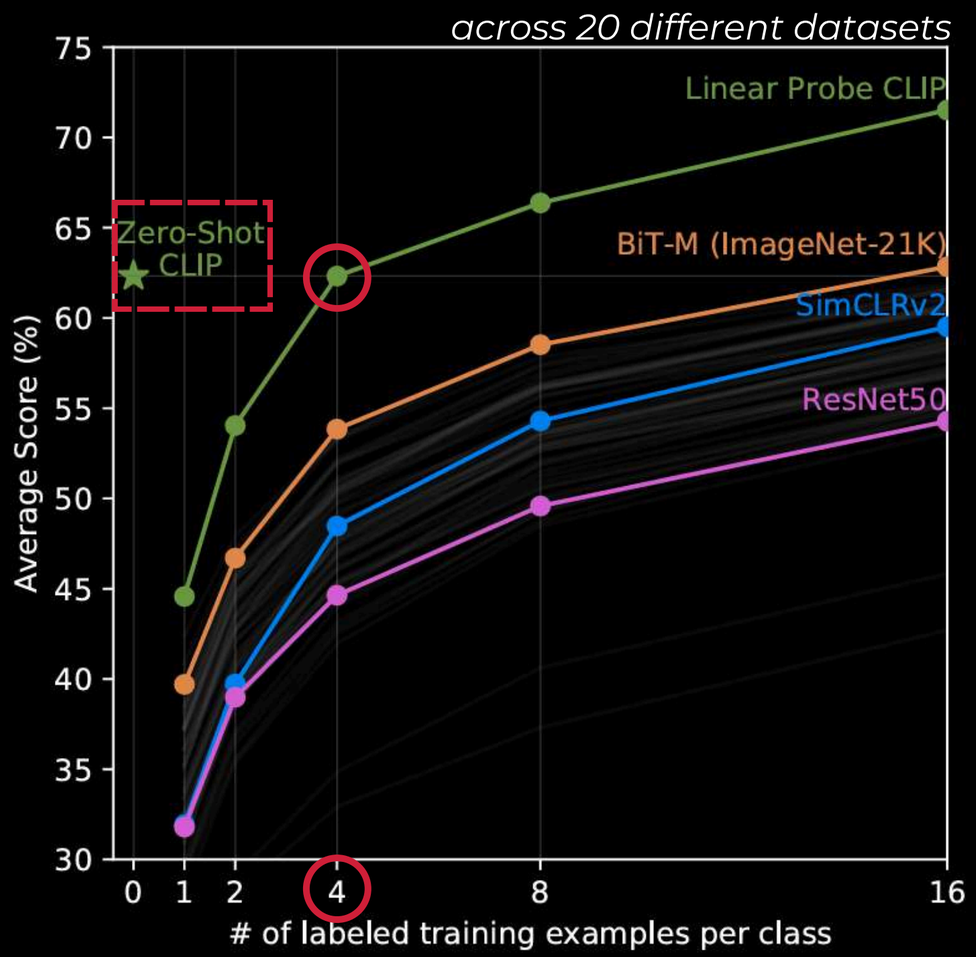
Zero-shot learning

ZS CLIP is compared to few-shot logistic regression on the features of many image models including CLIP itself



Zero-shot learning

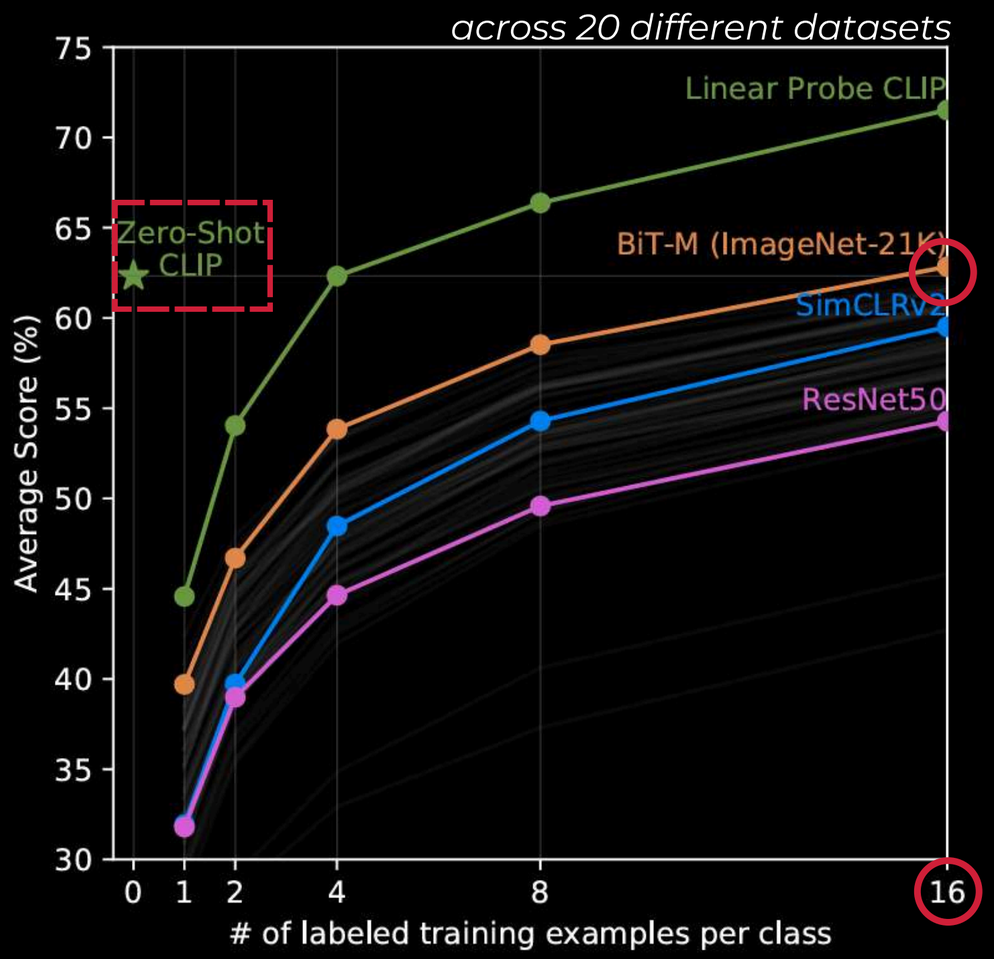
ZS CLIP is compared to few-shot logistic regression on the features of many image models including CLIP itself



ZS CLIP matches the performance of 4-shot logistic regression on the same feature space

Zero-shot learning

ZS CLIP is compared to few-shot logistic regression on the features of many image models including CLIP itself

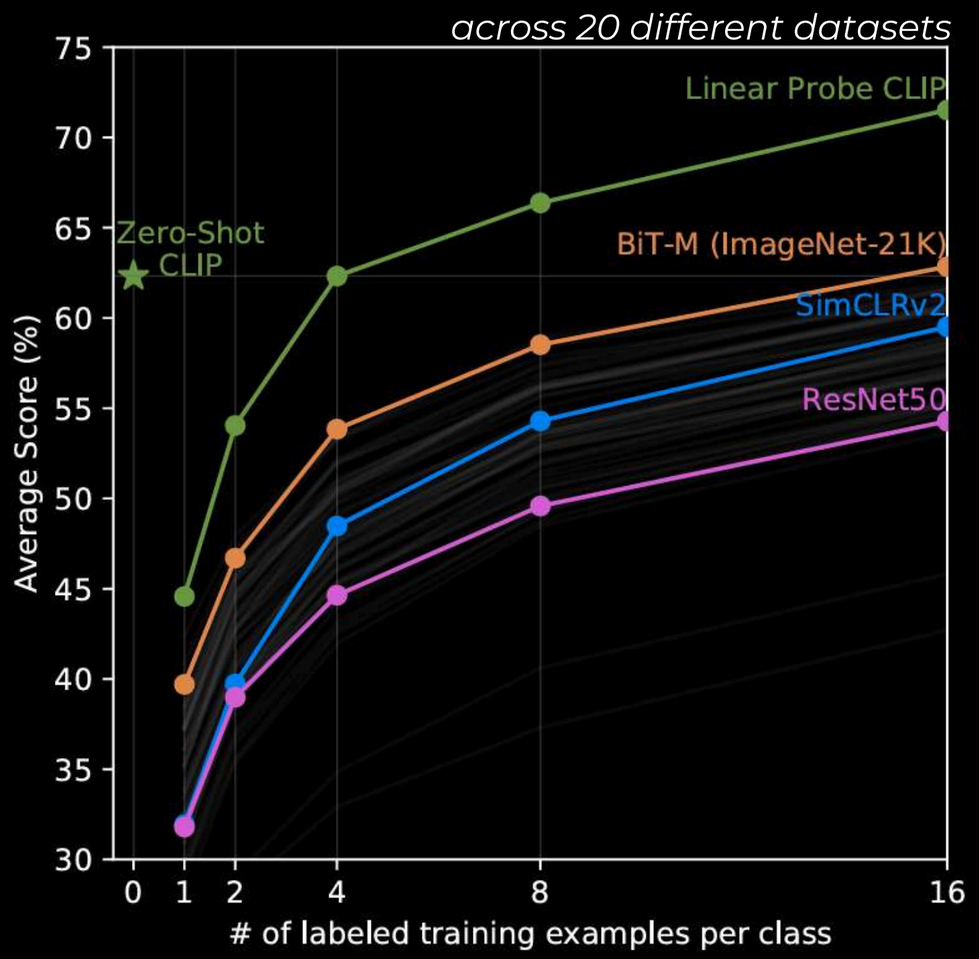


ZS CLIP matches the performance of 4-shot logistic regression on the same feature space

ZS CLIP almost matches the performance of the best performing 16-shot classifier (BiT-M ResNet-152x2)

Zero-shot learning

ZS CLIP is compared to few-shot logistic regression on the features of many image models including CLIP itself



ZS CLIP matches the performance of 4-shot logistic regression on the same feature space

ZS CLIP almost matches the performance of the best performing 16-shot classifier (BiT-M ResNet-152x2)

DATASET-SPECIFIC PERFORMANCE?

Robustness to Distribution Shift

Many of the correlations and patterns learned by DL models do not hold for distributions other than the one of the training set

Robustness to Distribution Shift

Many of the correlations and patterns learned by DL models do not hold for distributions other than the one of the training set

TAORI ET AL. (2020)

Evaluated ImageNet models on *natural distribution shifts*
→ differ from *synthetic ditribution shifts*
(i.g. adversarial attacks)

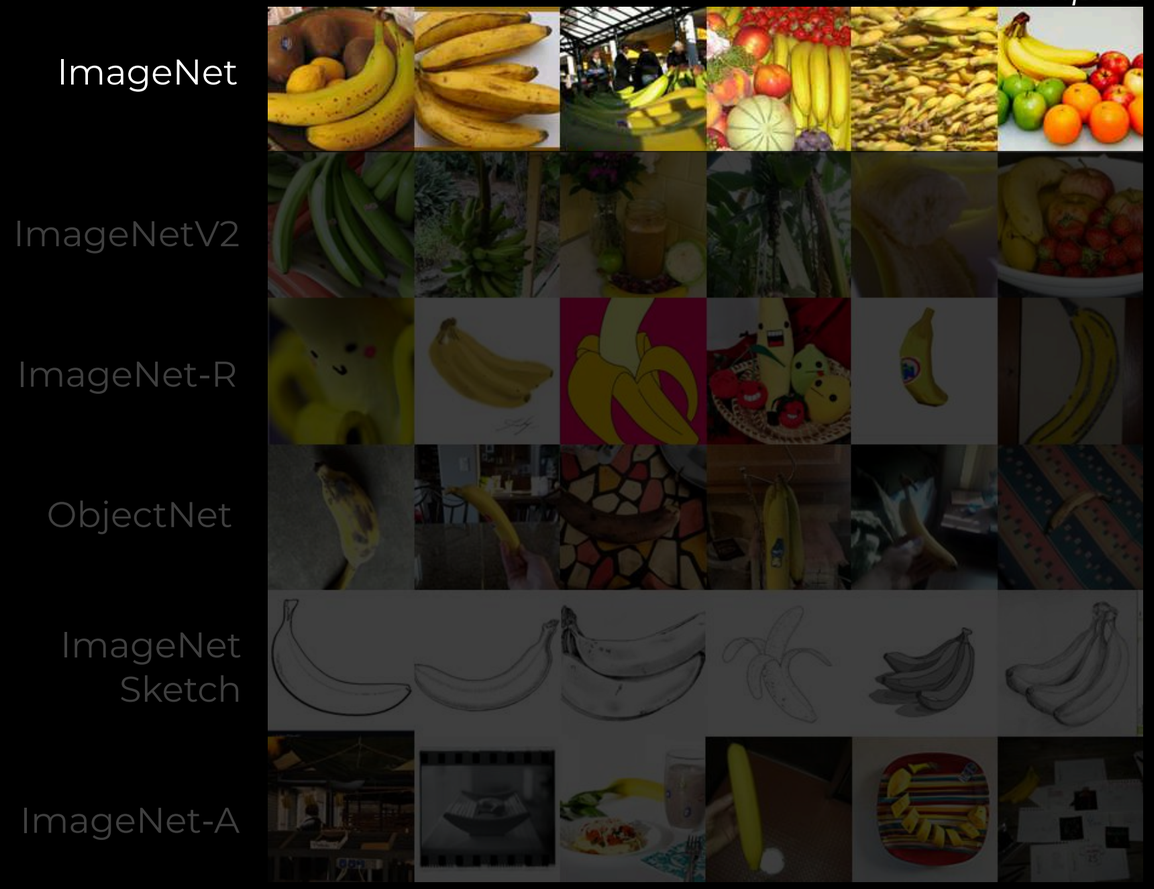
Robustness to Distribution Shift

Many of the correlations and patterns learned by DL models do not hold for distributions other than the one of the training set

TAORI ET AL. (2020)

Evaluated ImageNet models on *natural distribution shifts*
→ differ from *synthetic ditribution shifts*
(i.g. adversarial attacks)

dataset examples



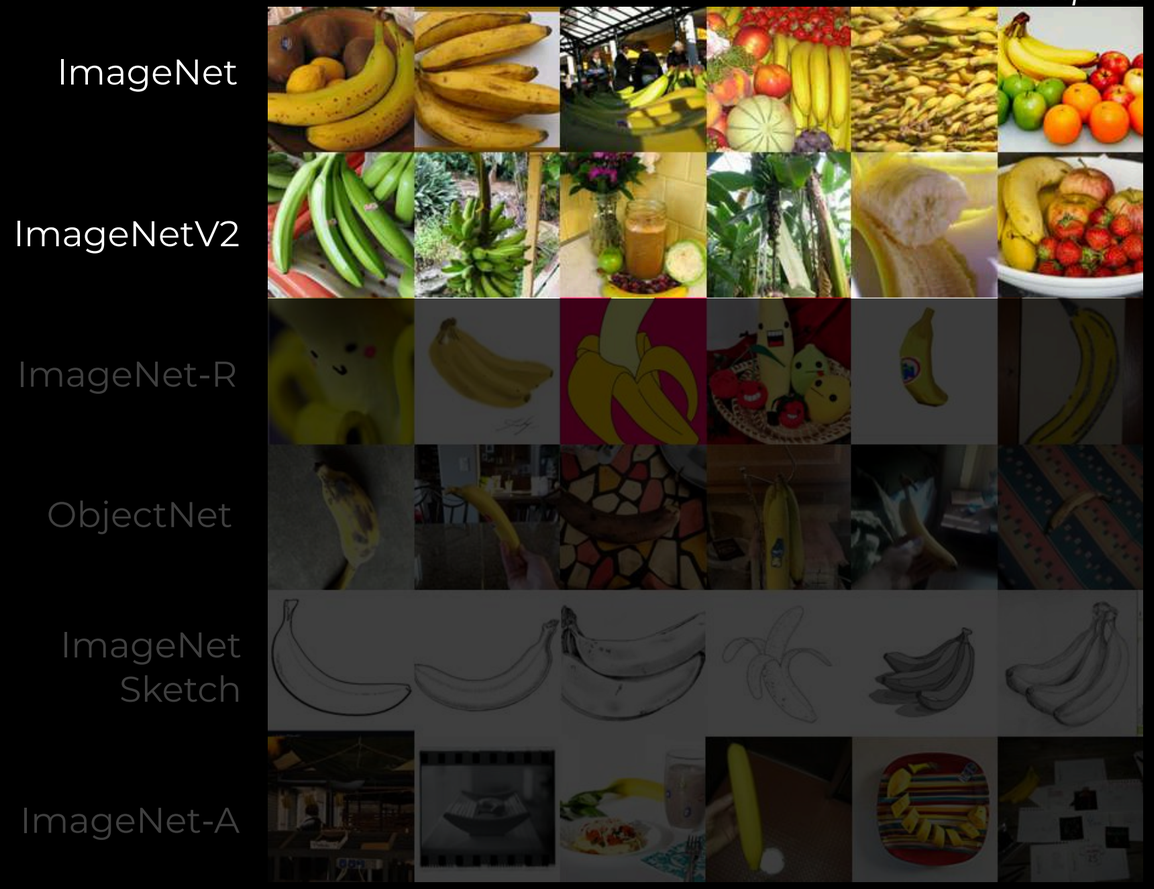
Robustness to Distribution Shift

Many of the correlations and patterns learned by DL models do not hold for distributions other than the one of the training set

TAORI ET AL. (2020)

Evaluated ImageNet models on *natural distribution shifts*
→ differ from *synthetic ditribution shifts*
(i.g. adversarial attacks)

dataset examples



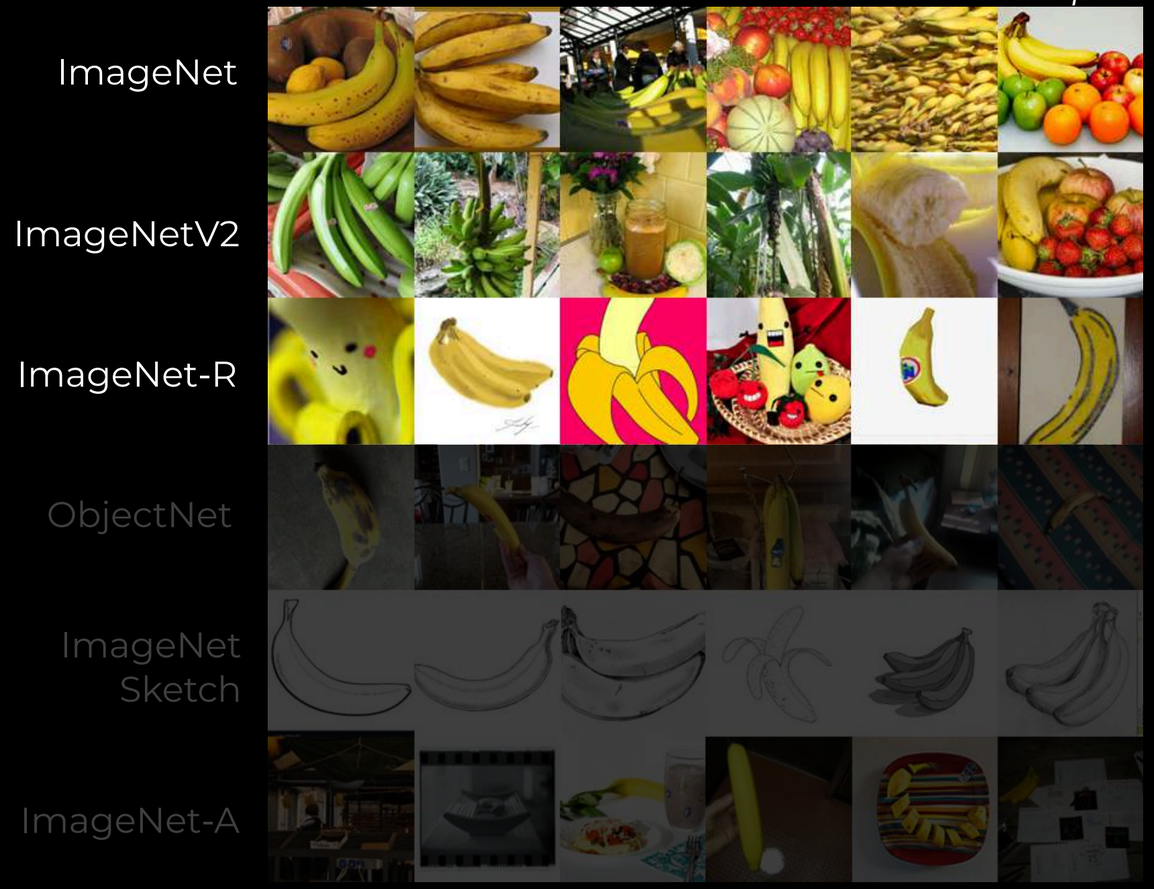
Robustness to Distribution Shift

Many of the correlations and patterns learned by DL models do not hold for distributions other than the one of the training set

TAORI ET AL. (2020)

Evaluated ImageNet models on *natural distribution shifts*
→ differ from *synthetic ditribution shifts*
(i.g. adversarial attacks)

dataset examples



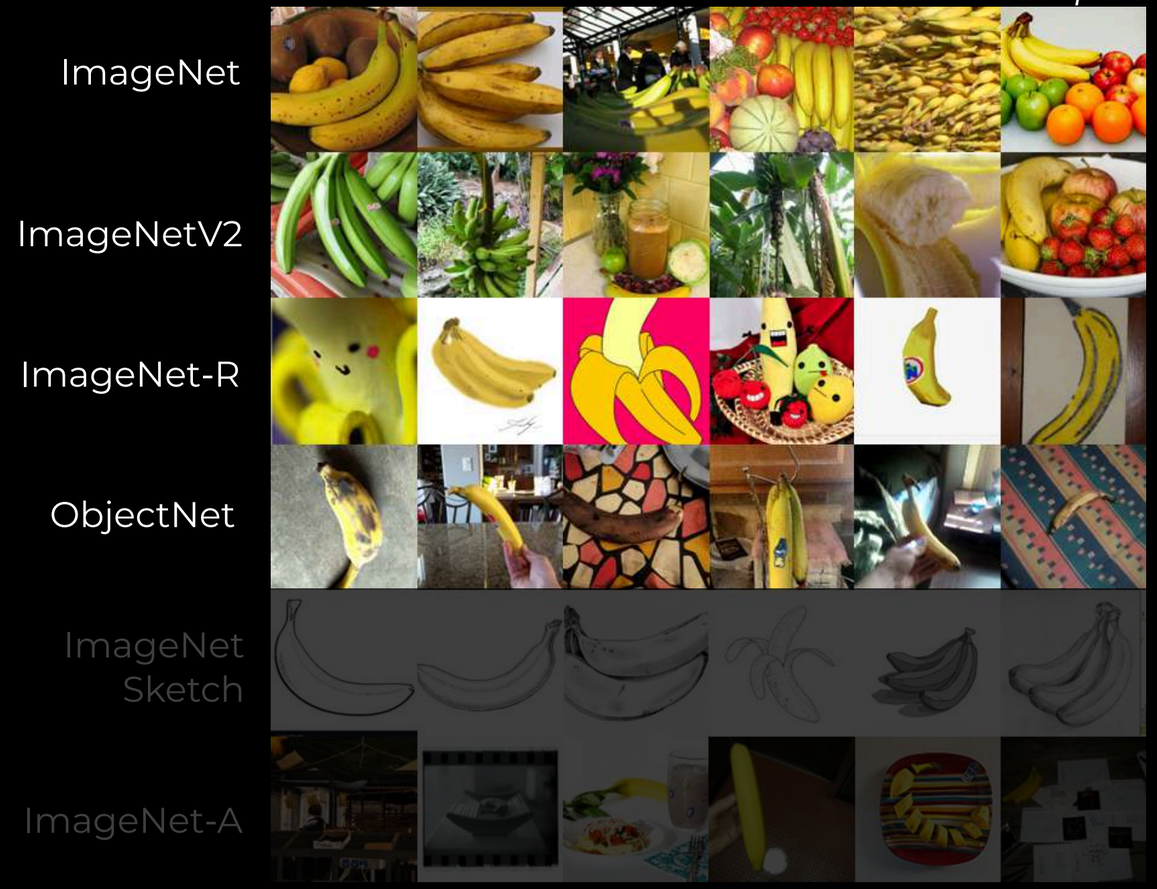
Robustness to Distribution Shift

Many of the correlations and patterns learned by DL models do not hold for distributions other than the one of the training set

TAORI ET AL. (2020)

Evaluated ImageNet models on *natural distribution shifts*
→ differ from *synthetic ditribution shifts*
(i.g. adversarial attacks)

dataset examples



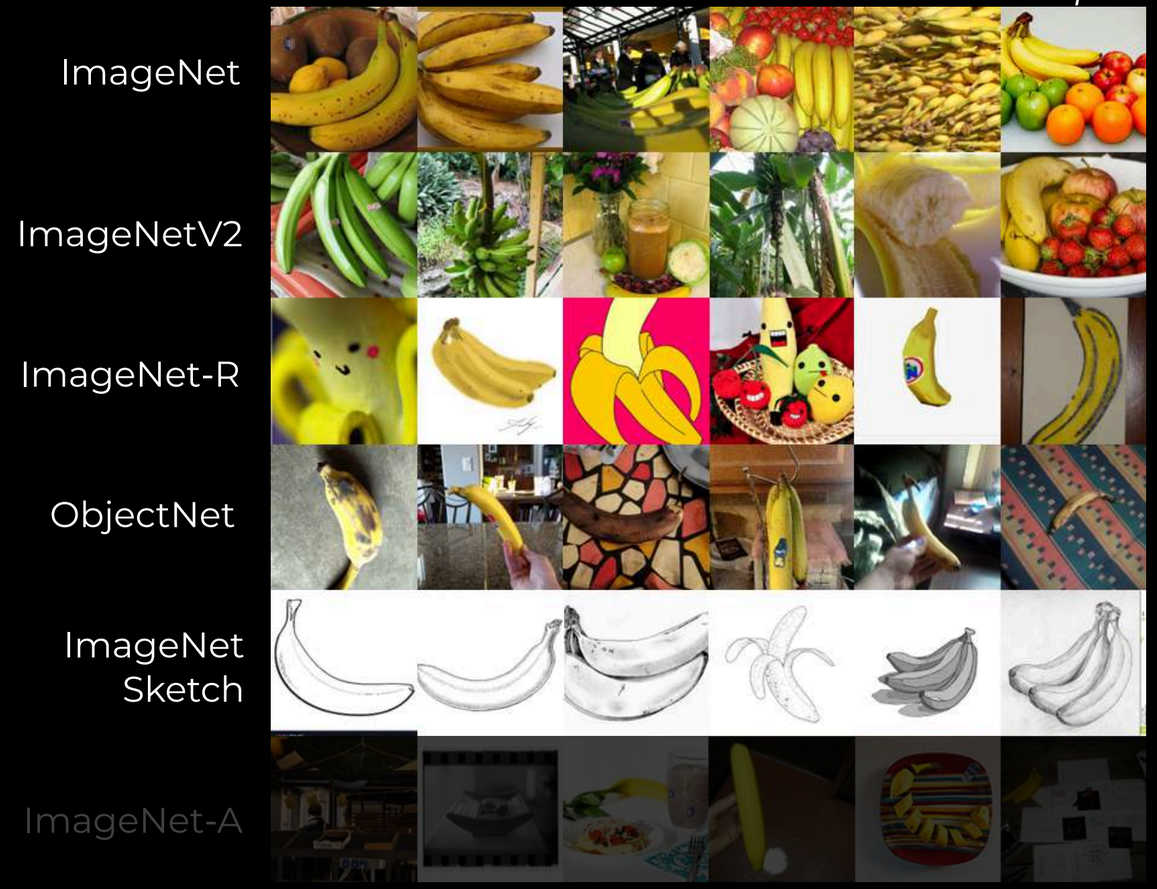
Robustness to Distribution Shift

Many of the correlations and patterns learned by DL models do not hold for distributions other than the one of the training set

TAORI ET AL. (2020)

Evaluated ImageNet models on *natural distribution shifts*
→ differ from *synthetic ditribution shifts*
(i.g. adversarial attacks)

dataset examples



Robustness to Distribution Shift

Many of the correlations and patterns learned by DL models do not hold for distributions other than the one of the training set

TAORI ET AL. (2020)

Evaluated ImageNet models on *natural distribution shifts*
→ differ from *synthetic ditribution shifts*
(i.g. adversarial attacks)

dataset examples

Robustness to Distribution Shift

Many of the correlations and patterns learned by DL models do not hold for distributions other than the one of the training set

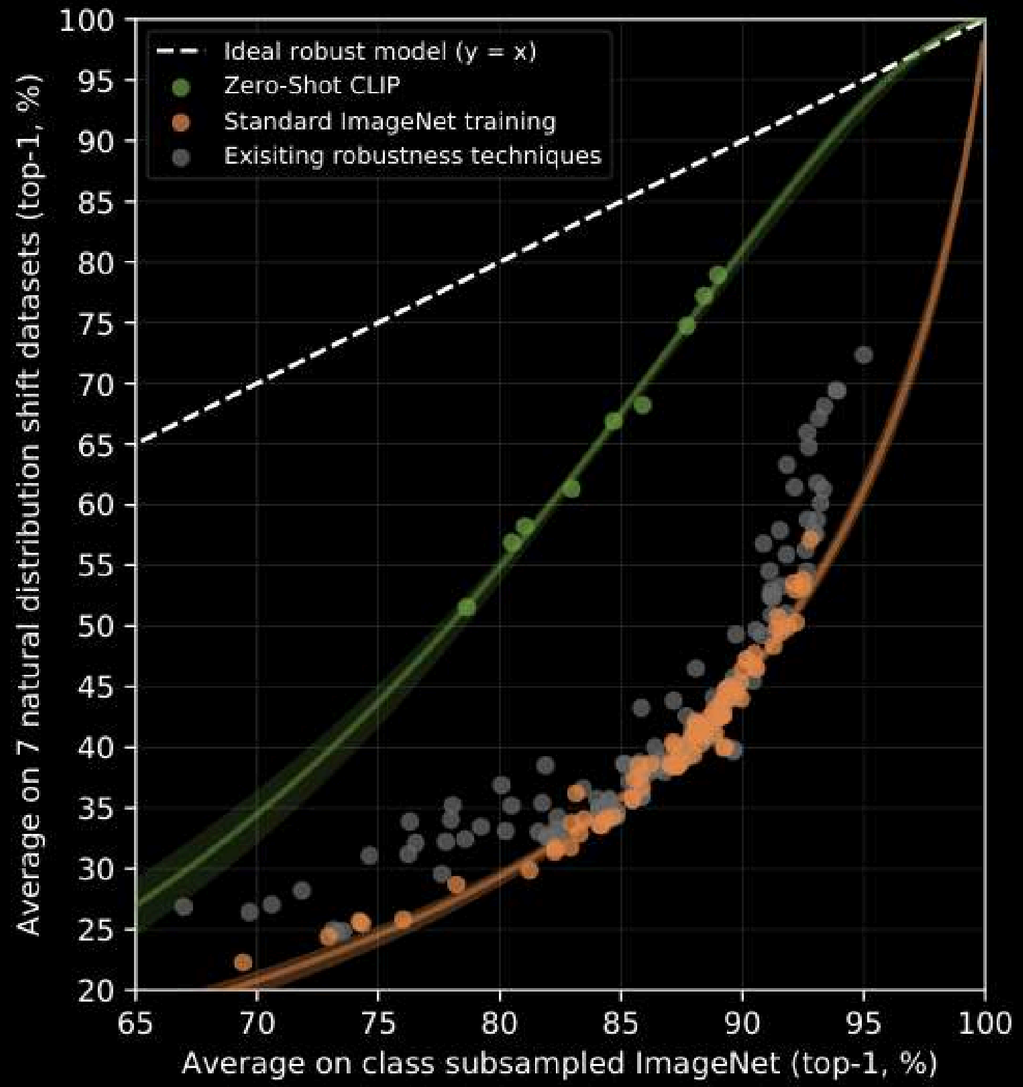
TAORI ET AL. (2020)

Evaluated ImageNet models on *natural distribution shifts*
→ differ from *synthetic ditribution shifts*
(i.g. adversarial attacks)

Found that accuracy under distribution shift increases predictably with ImageNet accuracy

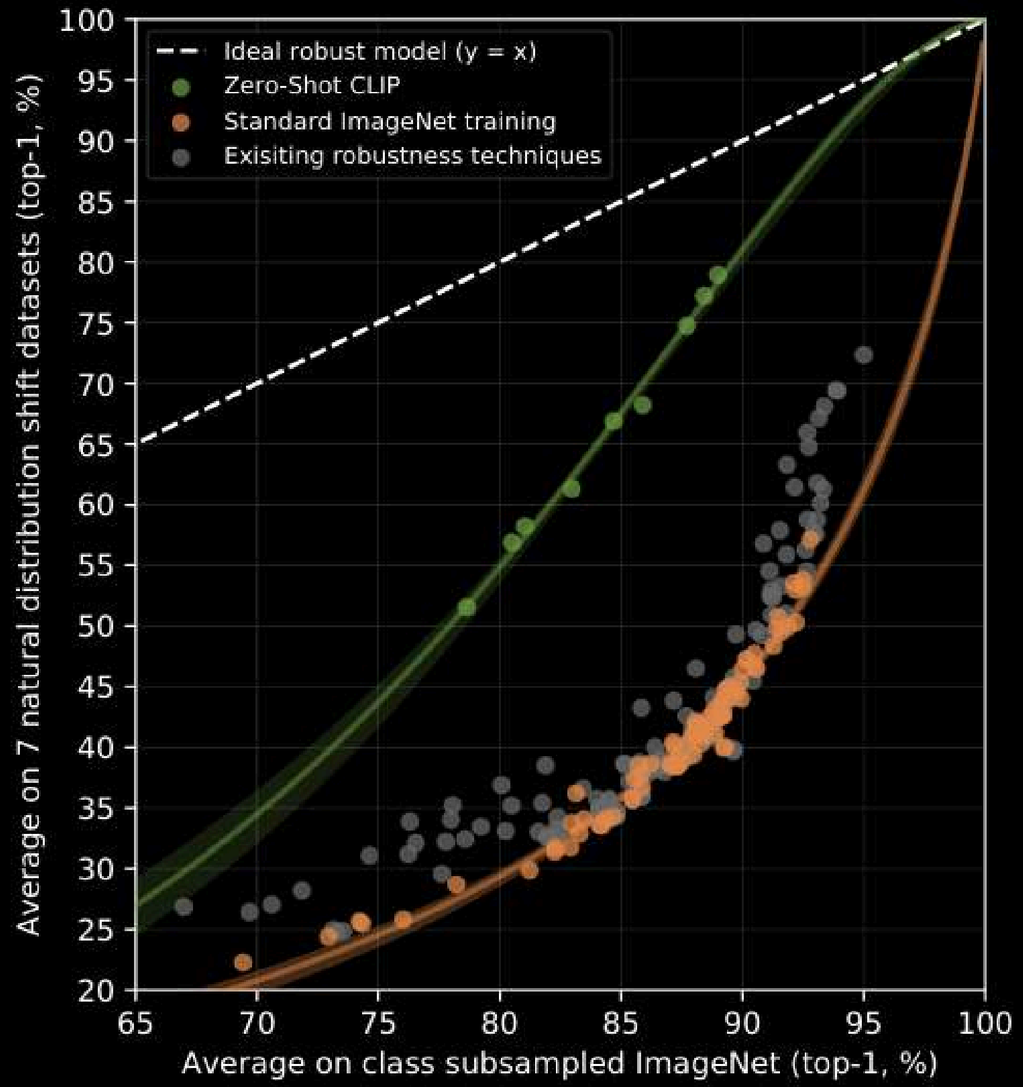
dataset examples

Robustness to Distribution Shift



All ZS CLIP models improve *effective* robustness and reduce the gap between ImageNet accuracy and accuracy under distribution shift by up to 75%

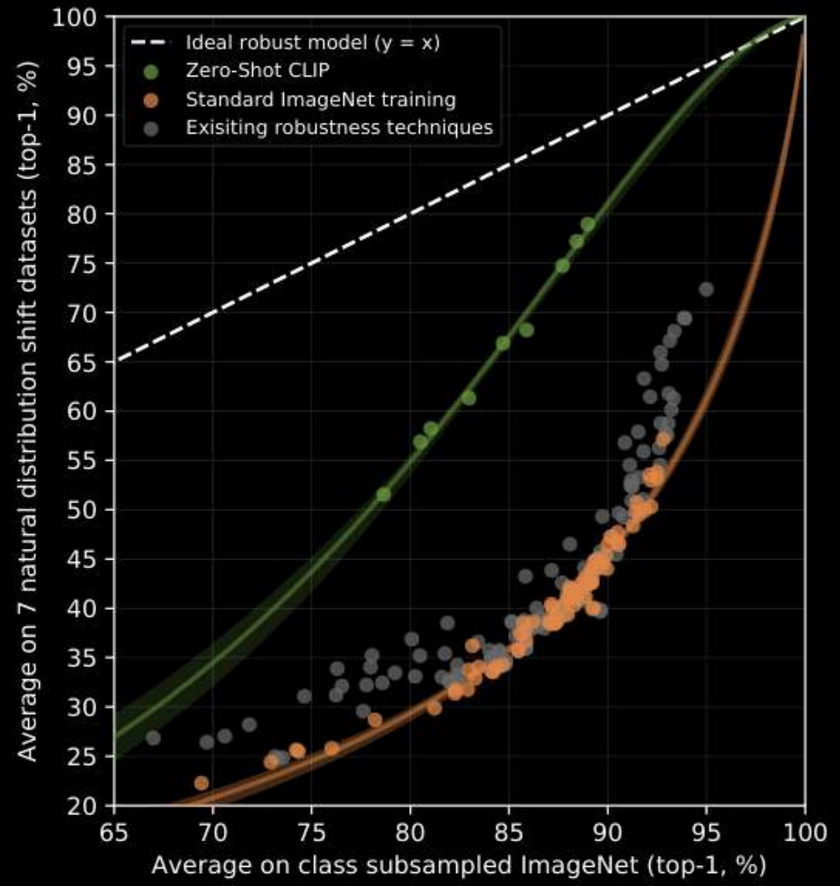
Robustness to Distribution Shift

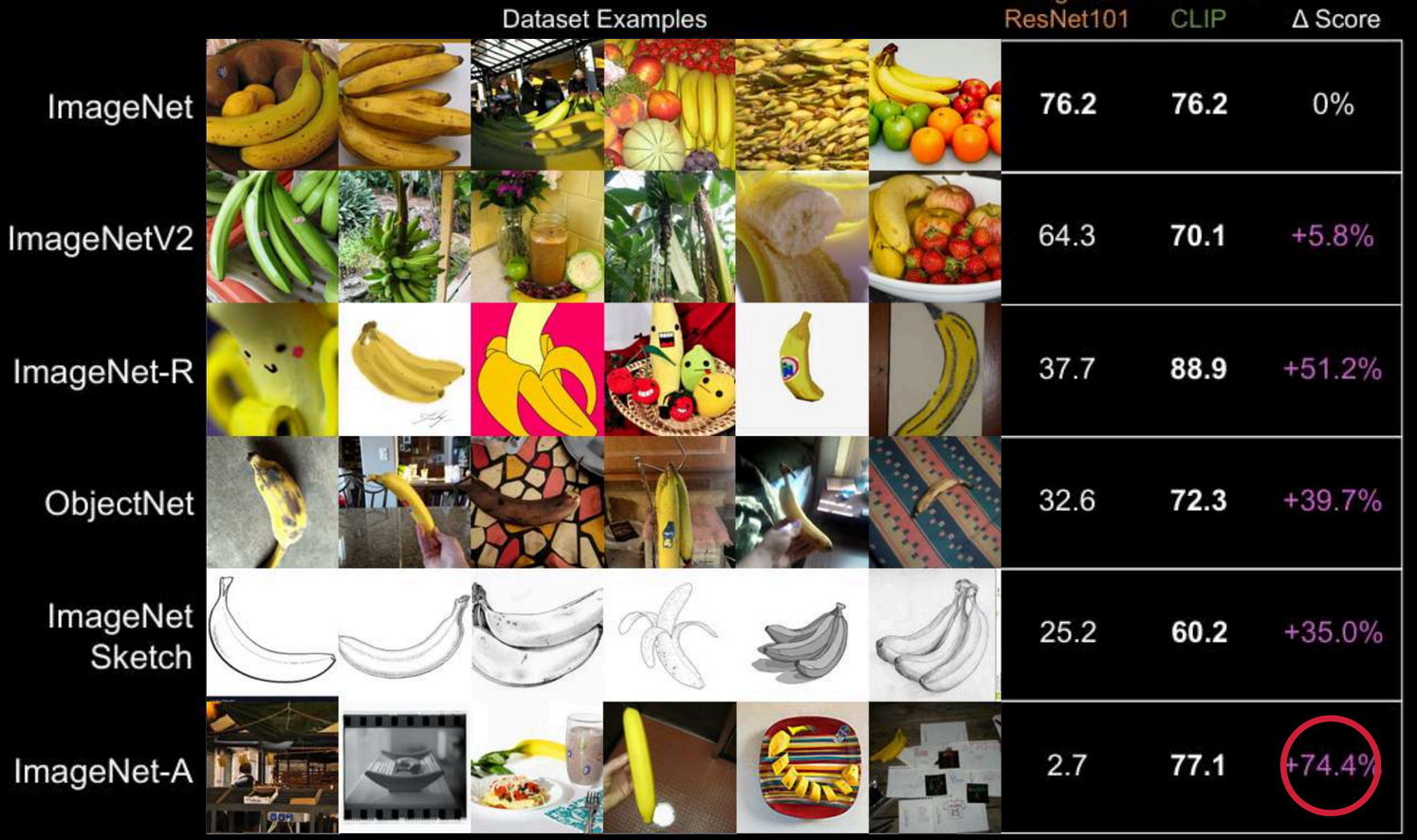


All ZS CLIP models improve *effective* robustness and reduce the gap between ImageNet accuracy and accuracy under distribution shift by up to 75%

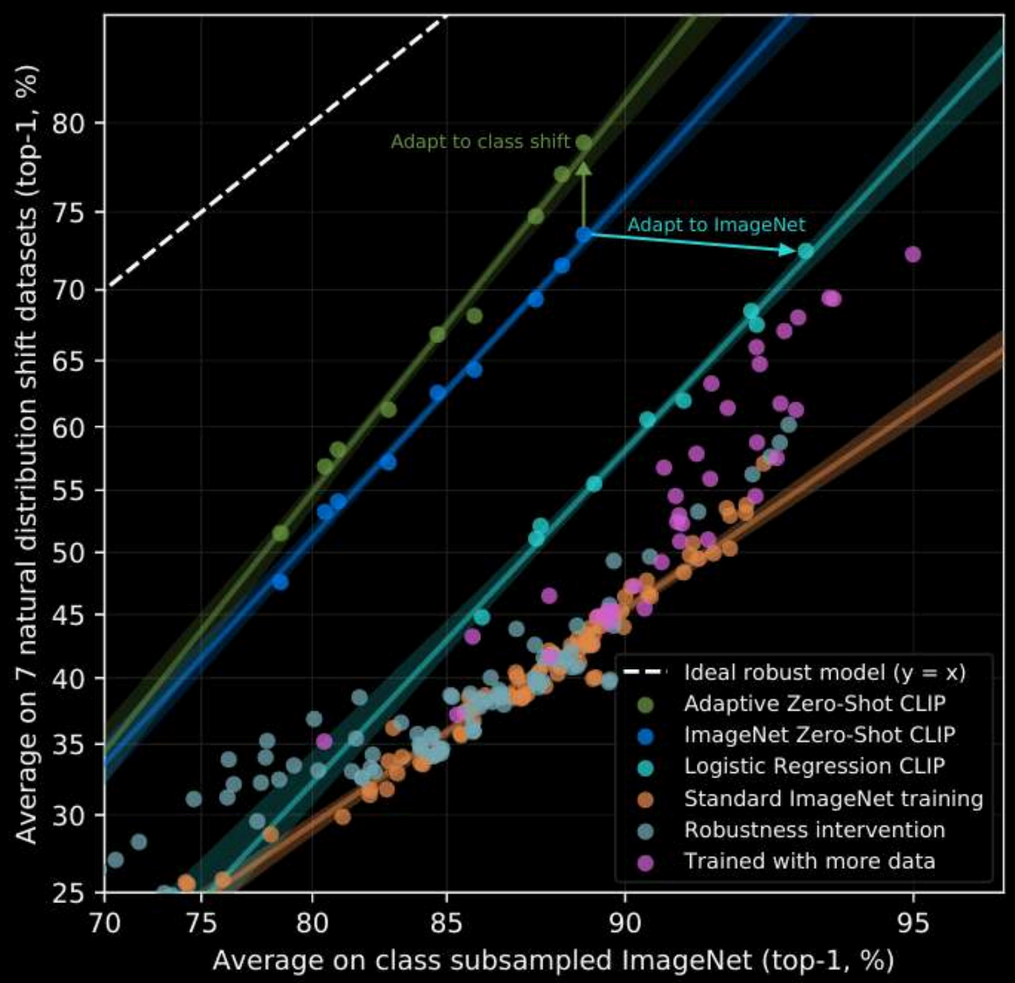
does supervised learning on ImageNet cause a robustness gap (?)

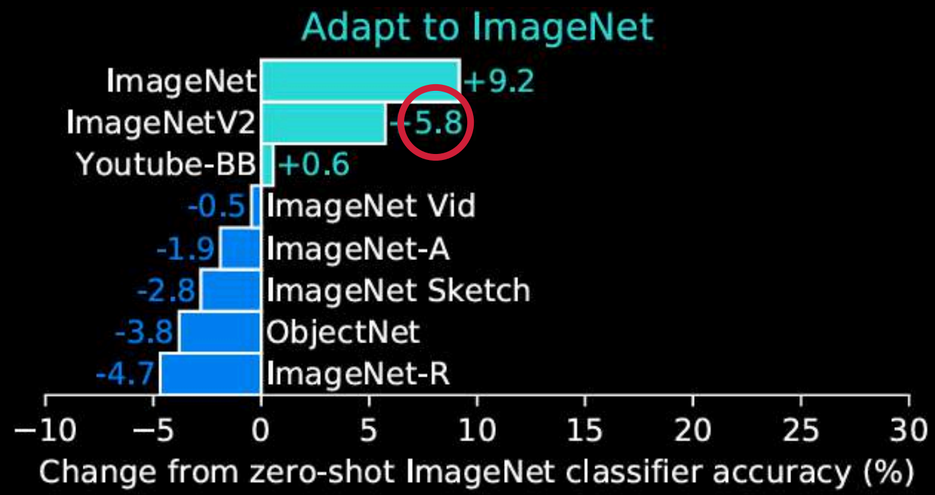
Robustness to Distribution Shift





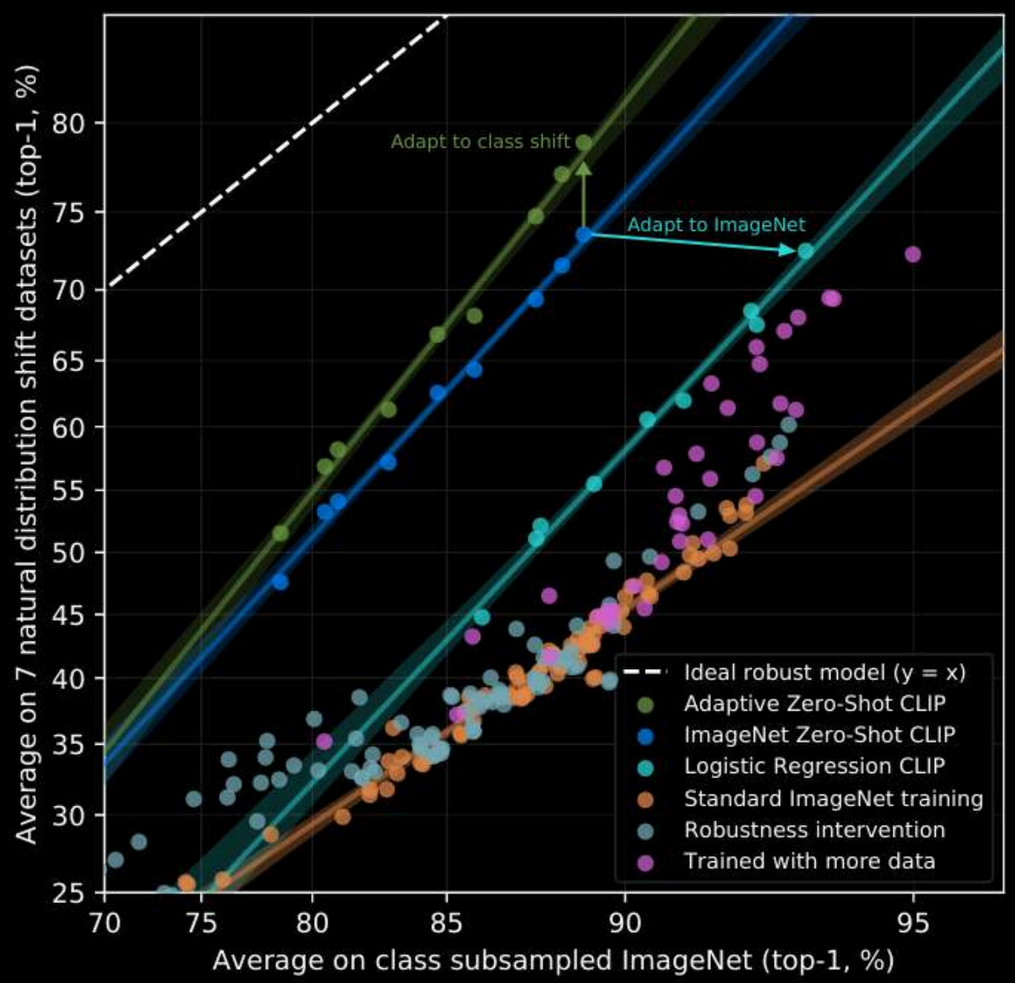
Robustness to Distribution Shift

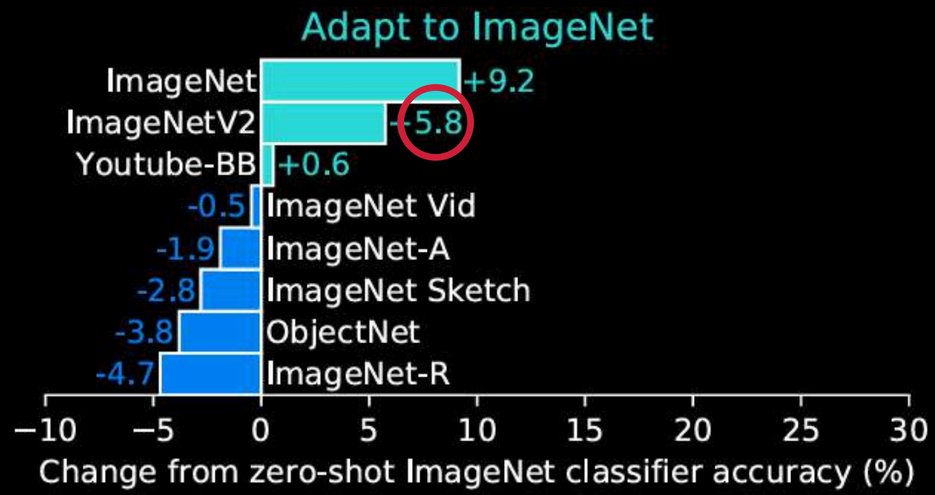




the 9.2% increase in accuracy fails to translate into any improvement in avg performance under distribution shift → sus

Robustness to Distribution Shift

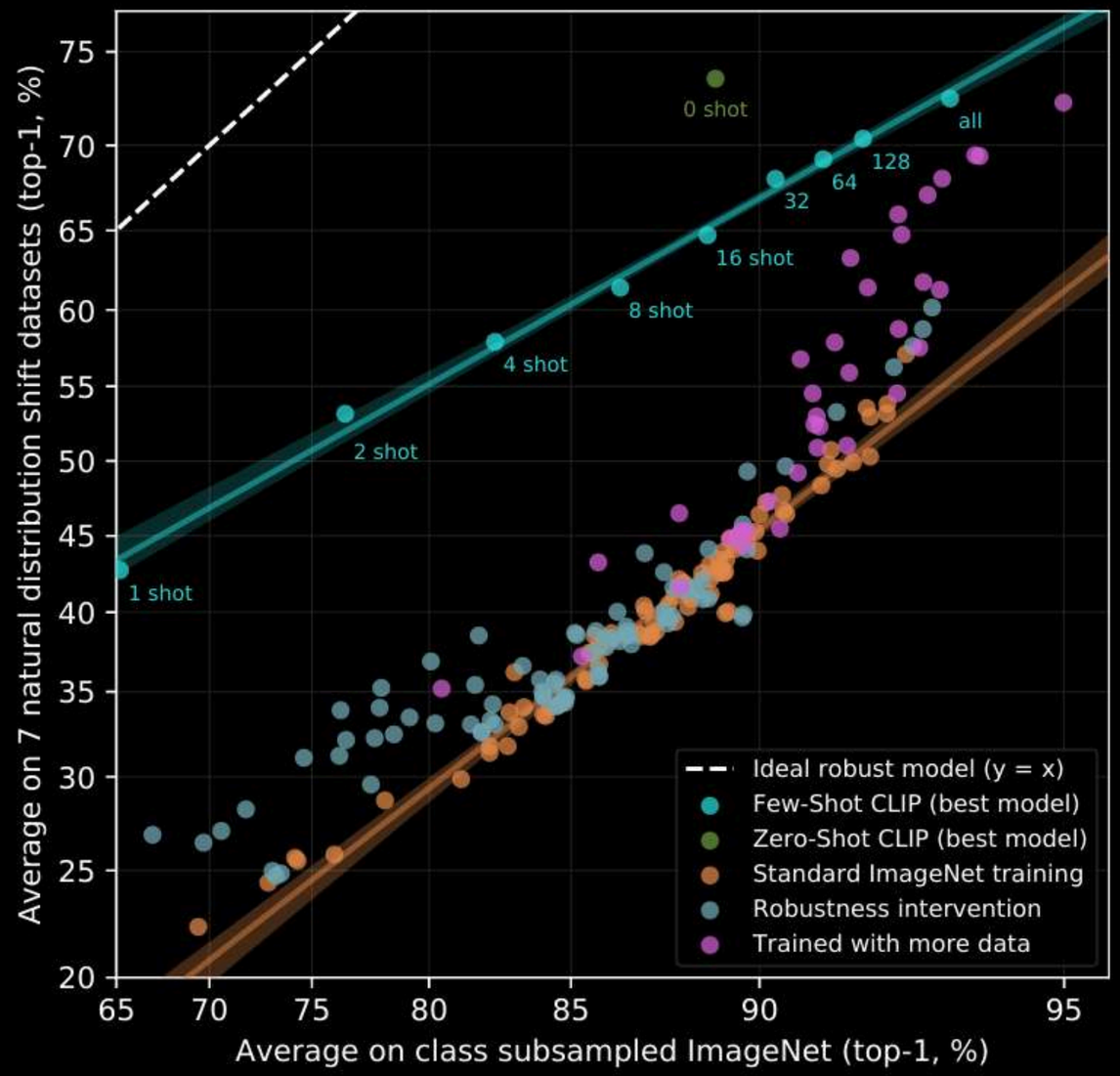




the 9.2% increase in accuracy fails to translate into any improvement in avg performance under distribution shift → sus

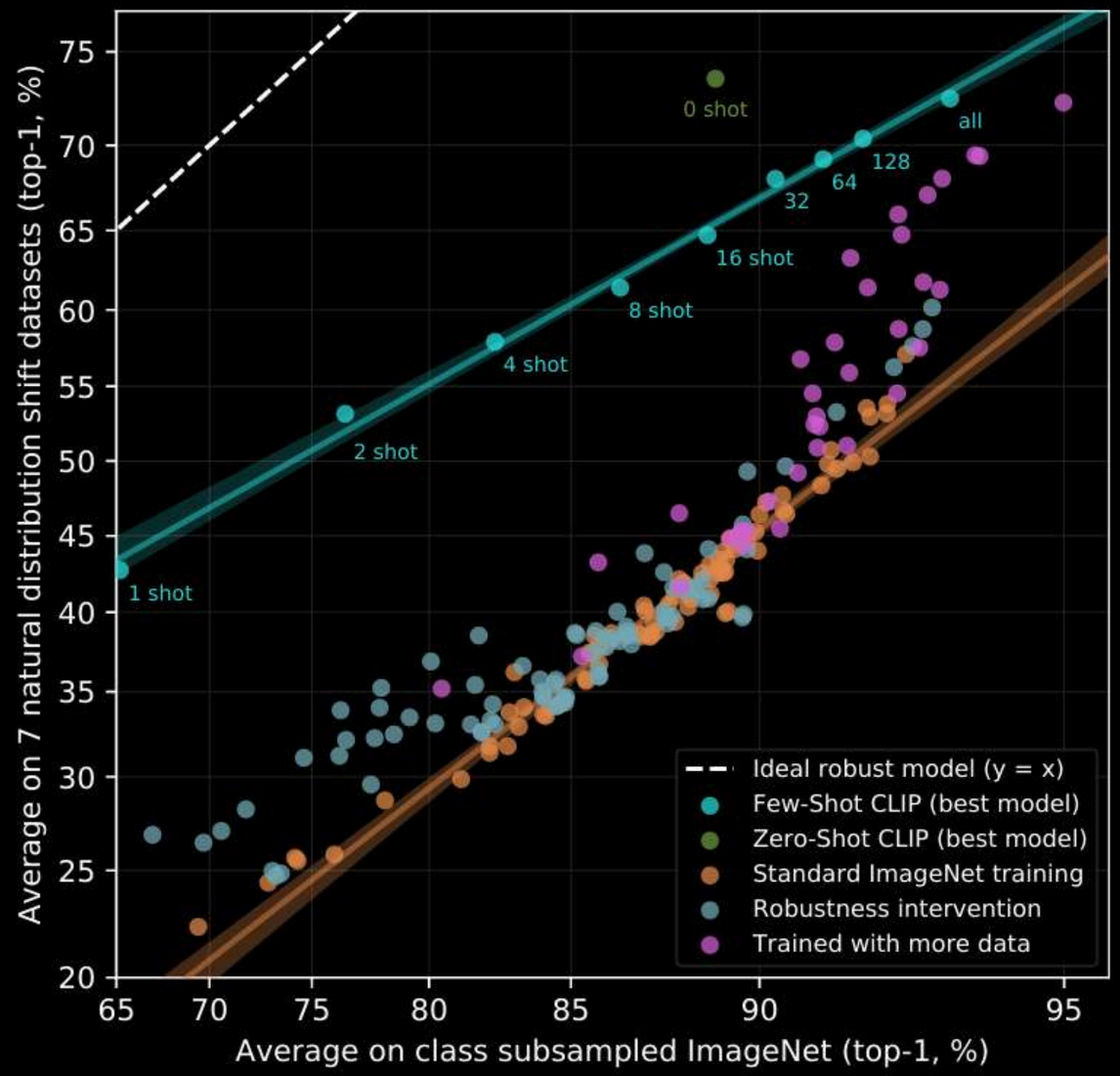
the robustness benefit seems entirely gone in a fully supervised setting

Robustness to Distribution Shift



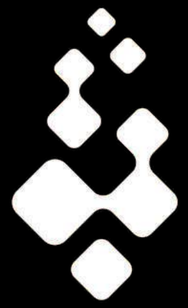
The robustness benefit fades as in-distribution performance increases with more training data and is mostly gone for the fully supervised model

Robustness to Distribution Shift



The robustness benefit fades as in-distribution performance increases with more training data and is mostly gone for the fully supervised model

minimize the amount of distribution specific data a model has access to



THANK YOU FOR THE ATTENTION