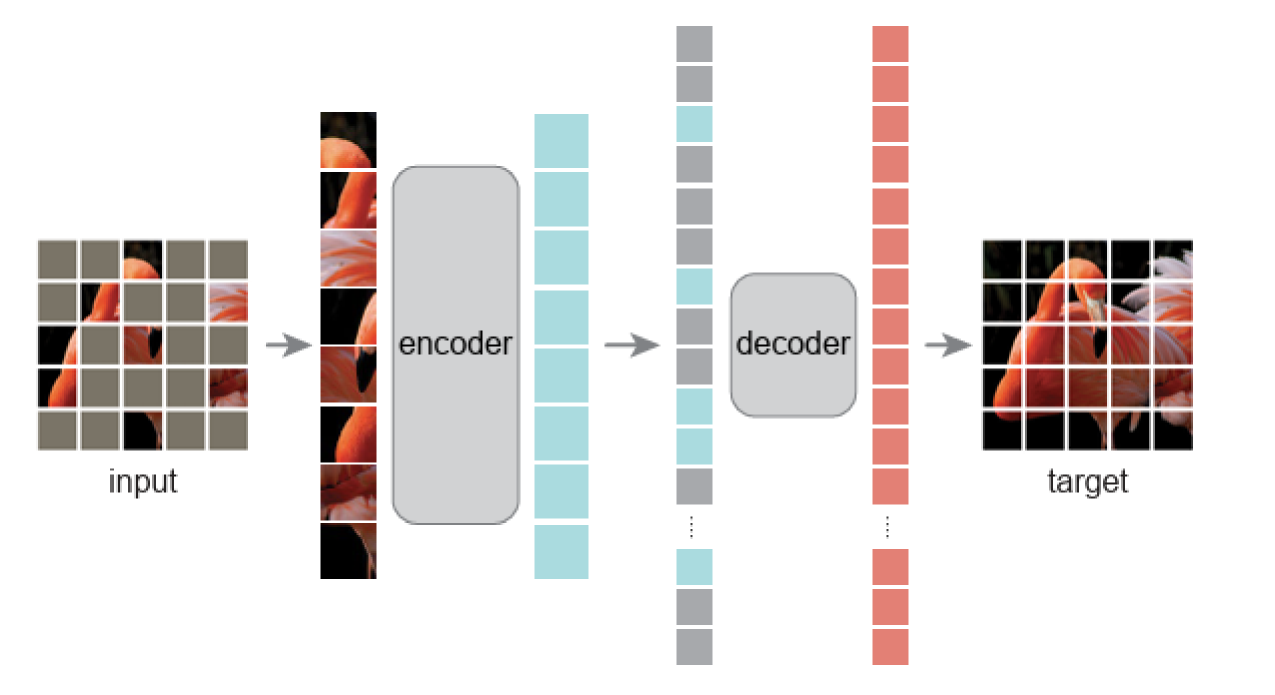
AIL 862

Lecture 18

MAE



MAE encoder

- Just like a standard ViT

- However, operates only on a small subset of the full set of the patches

Non-overlapping patches

- Why?

Non-overlapping patches

- Overlapping patches introduce redundancy.
- Non-overlapping patches enforce a stronger learning signal since the model must infer missing parts without redundant information.

MAE decoder

- The input to the MAE decoder is the full set of tokens consisting of encoded visible patches and mask tokens.
- Positional embeddings are added to all tokens in this full set.
- The decoder has another series of Transformer blocks.

Reconstruction target

Reconstructs the input by predicting the pixel values for each masked patch. Each element in the decoder's output is a vector of pixel values representing a patch.

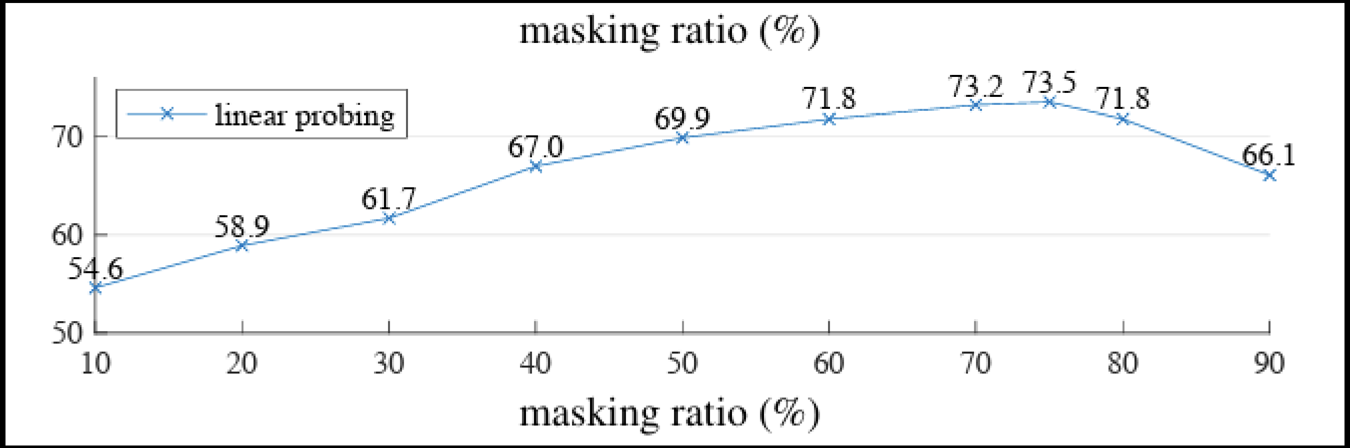
The last layer of the decoder is a linear projection whose number of output channels equals the number of pixel values in a patch.

The decoder's output is reshaped to form a reconstructed image.

Loss function computes the (MSE between the reconstructed and original images in the pixel space.

Masking ratio

Masking ratio



Comparison to supervision

to overfit. The following is a comparison between ViT-L trained from scratch *vs*. fine-tuned from our baseline MAE:

scratch, original [16]	scratch, our impl.	baseline MAE
76.5	82.5	84.9

Data augmentation

- Is it needed here?

Data augmentation

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

Mask sampling - random

- Random: sample random patches without replacement

- Follow uniform distribution – avoids bias

- High masking ratio eliminates redundancy – thus creating a task that cannot be easily solved by extrapolating from neighboring patches

Mask sampling - block

- Remove large random blocks

Mask sampling - grid

- To remove 75% patches, remove one of every four patches

Mask sampling

case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(f) **Mask sampling**. Random sampling works the best. See Figure 6 for visualizations.

Transfer to other tasks

method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementa-

MAE applications in different domains

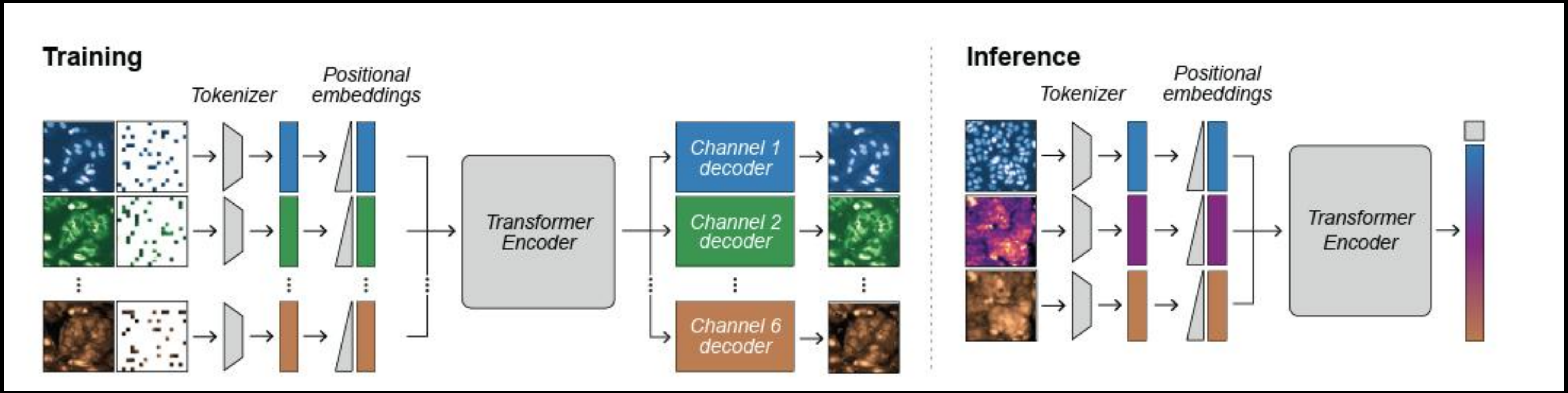
- Have been used in several domains such as medical, EO

MAE for microscopy

- Channel-agnostic MAE

Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology, 2024

MAE for microscopy



Masked Autoencoders for Microscopy are Scalable Learners of Cellular Biology, 2024

For EO with multi-sensor reconstruction

i-MAE

- Basic idea

i-MAE

- Mixed representation fed to encoder

i-MAE

- Two linear layers acting on the mixed representation to obtain two latent representations

i-MAE

- Shared decoder to reconstruct each input

i-MAE

MAE with pixel reconstruction (mix ratio:0.3, mask ratio:0.5)

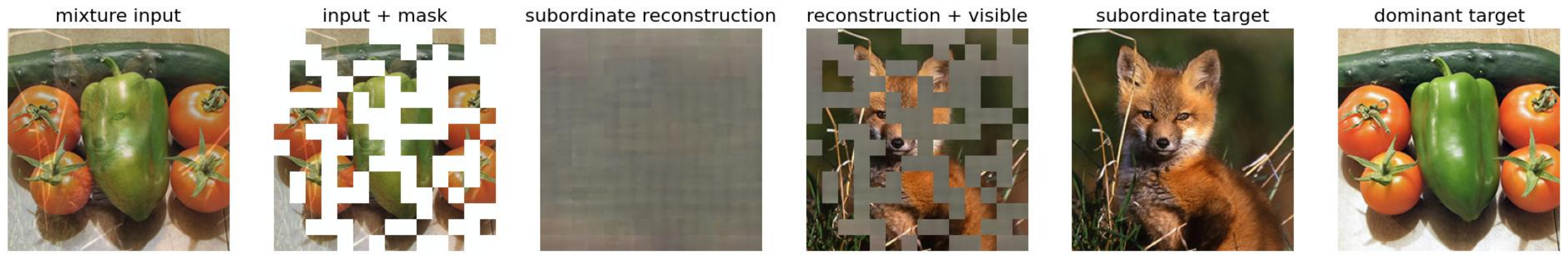


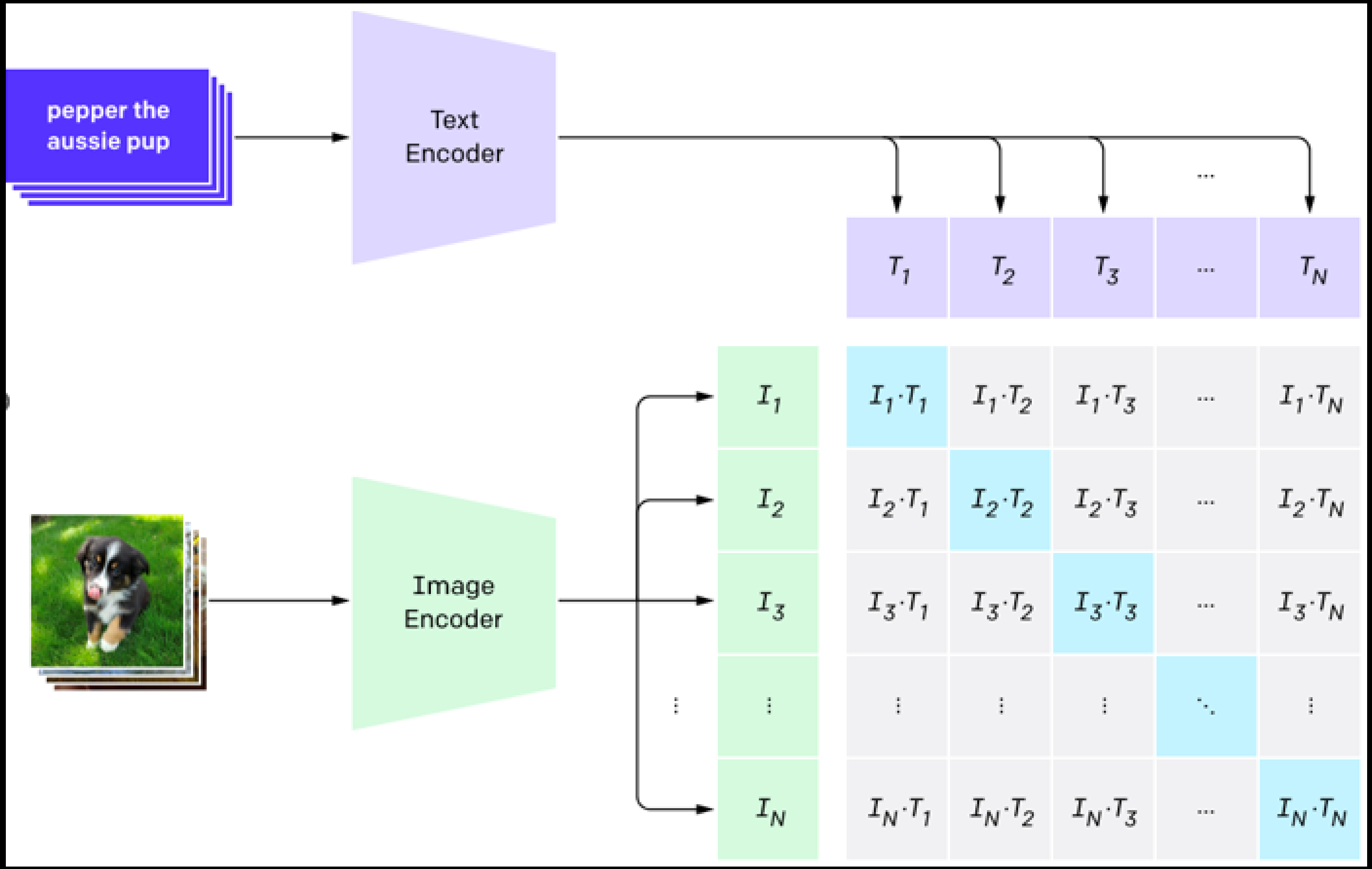
Image + Text

- History – image captioning etc. tasks

CLIP

Dataset

CLIP



pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

CLIP

Image encoder

CLIP

Text encoder