Masked Autoencoders Are Scalable Vision Learners

Han Lin

04/03/2024

Motivation

- **Self-supervised pertaining in NLP:**
  - Remove a portion of the data and learn to predict the removed content
  - This method enables training of generalizable NLP models with >100B parameters

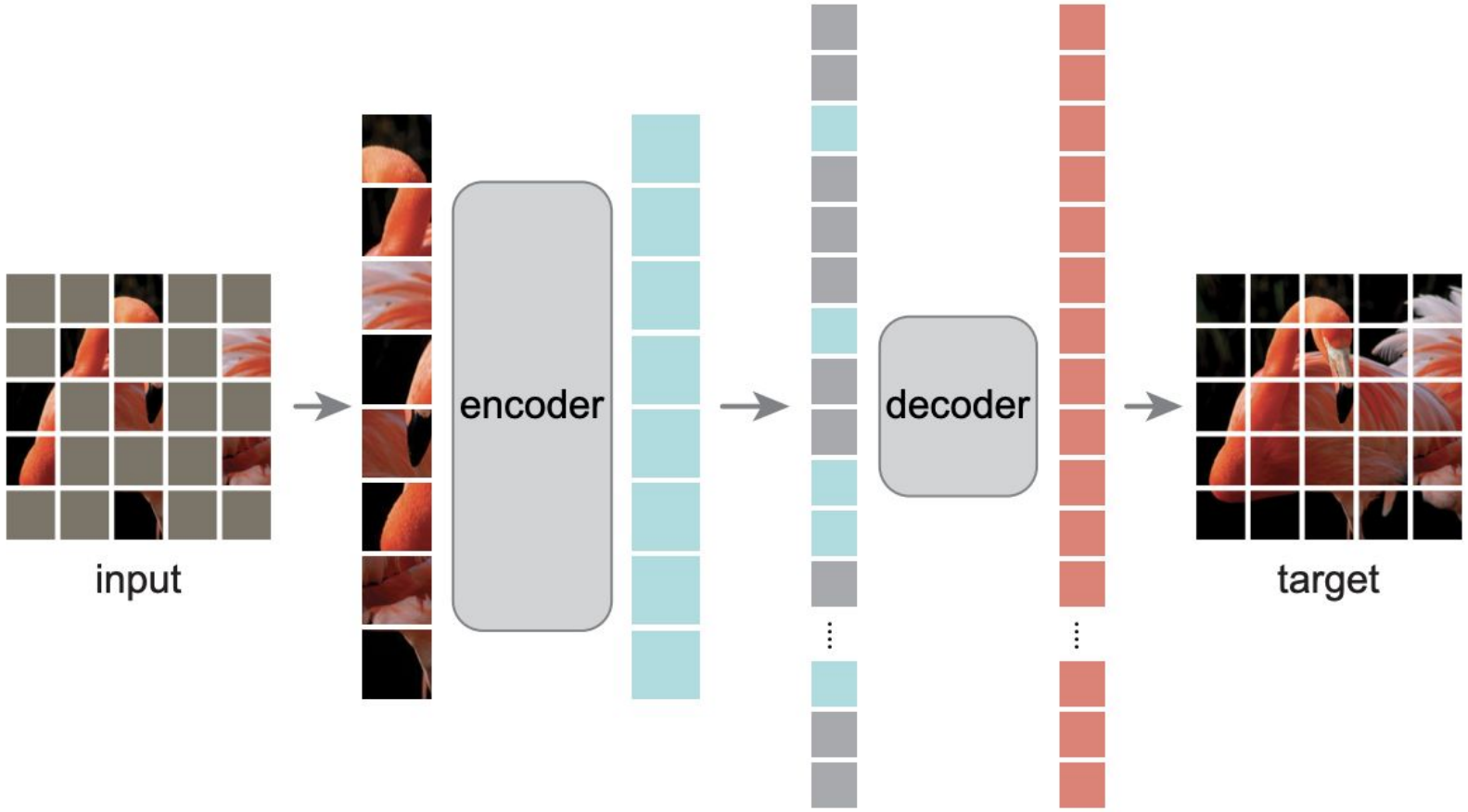
Masked autoencoding in vision and language

What makes masked autoencoding different between vision and language?

- **Architecture difference:**
  - **CNN** operates on regular grids, and it is not straightforward to integrate mask tokens or positional embeddings into convolutional networks
- **Information density:**
  - Languages are highly semantic and information-dense.
    Missing words prediction needs **sophisticated language understanding**
  - Images are natural signals with heavy spatial redundancy.
    Masking **a very high portion** of random patches largely reduces redundancy
    Creates a challenging self-supervised task that requires **holistic understanding** beyond low-level image statistics.
- **Decoder:**
  - **Language:** predicts missing words that contain rich semantic information (e.g., in BERT the decoder is just a MLP)
  - **Vision:** reconstructs pixels, which is of **lower semantic level** than languages

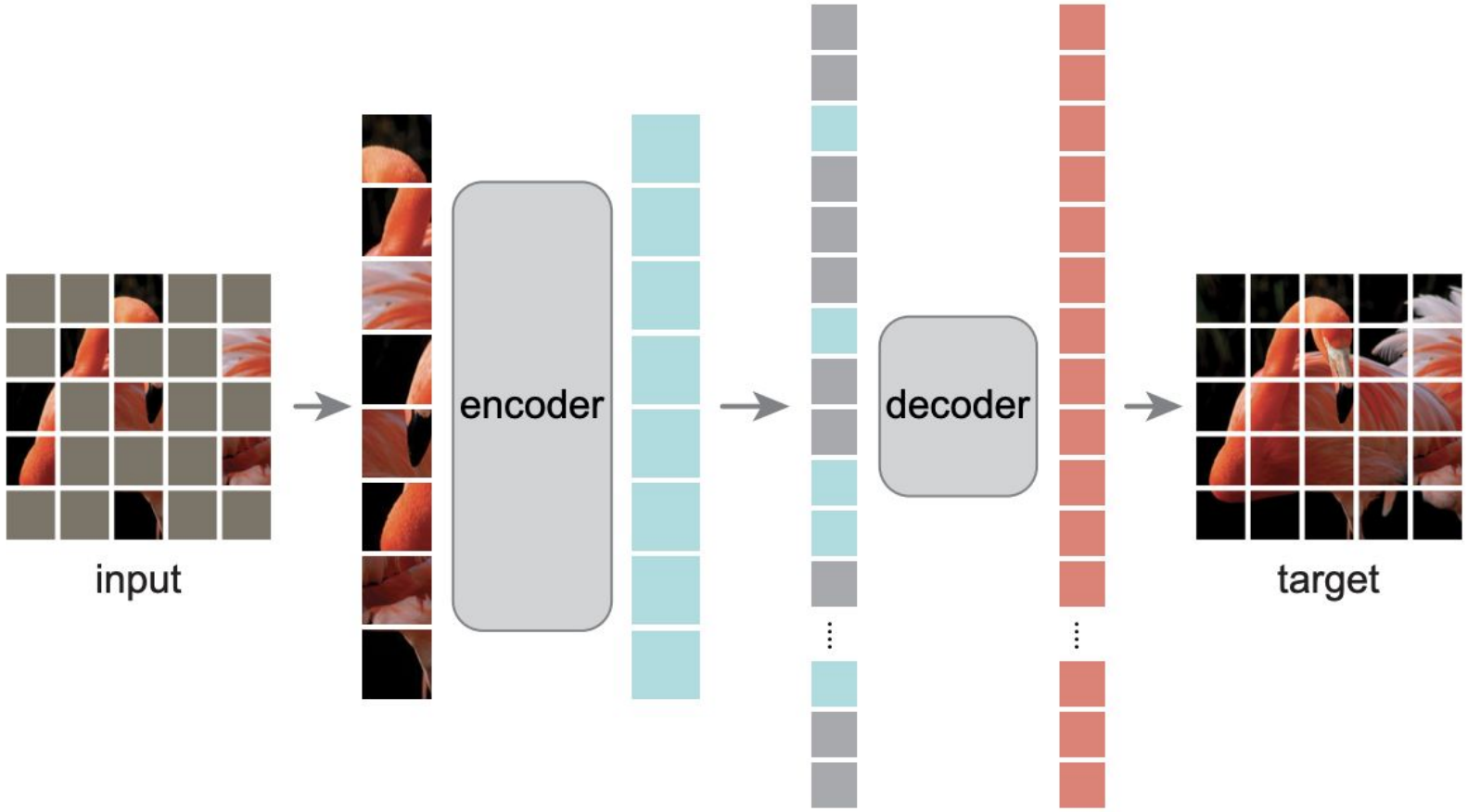
MAE Architecture

- **Encoder:**
  - Only operates on unmasked patches



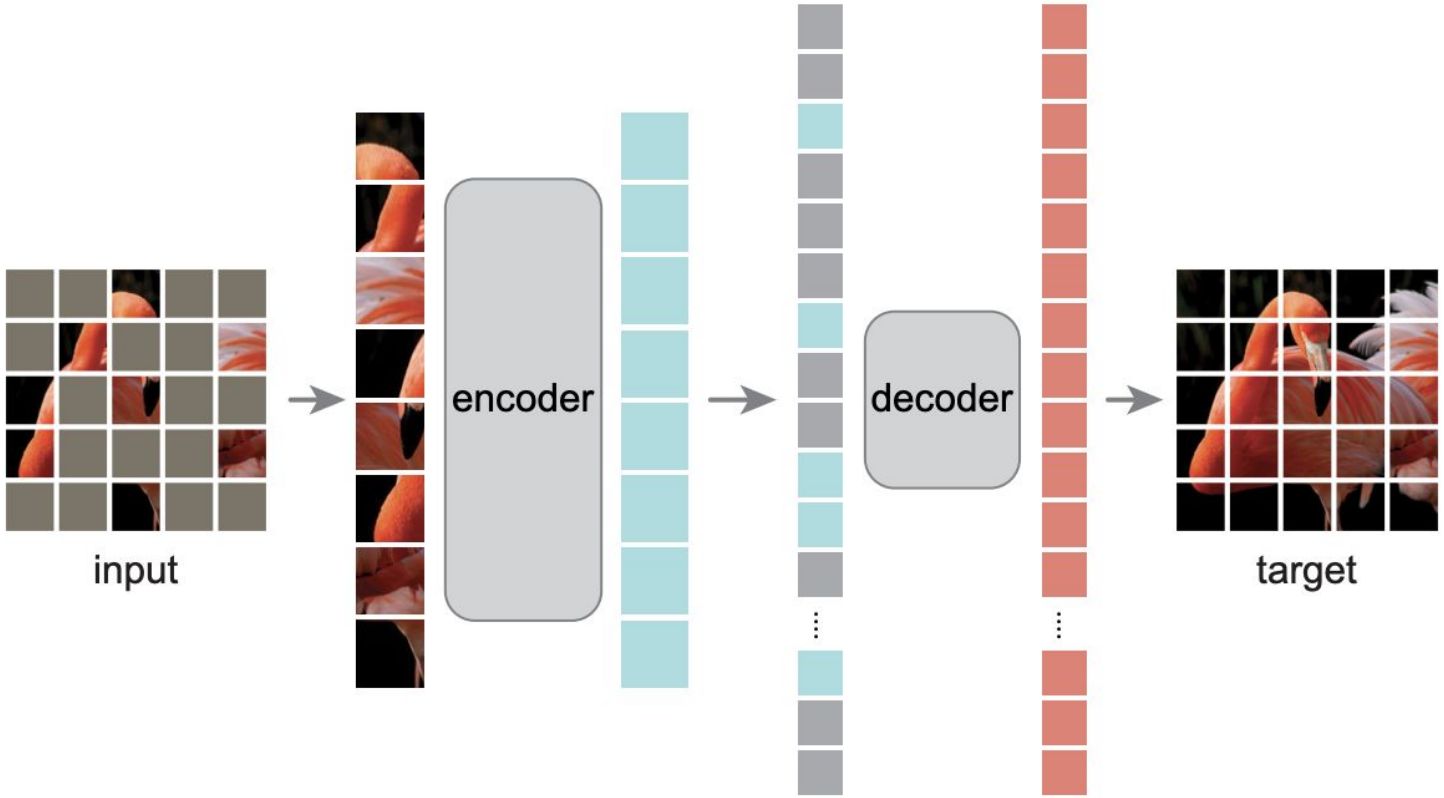
MAE Architecture

- **Decoder:**
  - Encoded visible patches
  - Mask token: shared, learned vector with positional embeddings



MAE Architecture

- **Reconstruction target:**
  - To predict the pixel values for each masked patch
  - Loss function computes the mean squared error (MSE) between the reconstructed and original images in the pixel space



Experiments: Baseline

- **Baseline**
  - Uses ViT-L/16 as the backbone for ablation study

Experiment Results

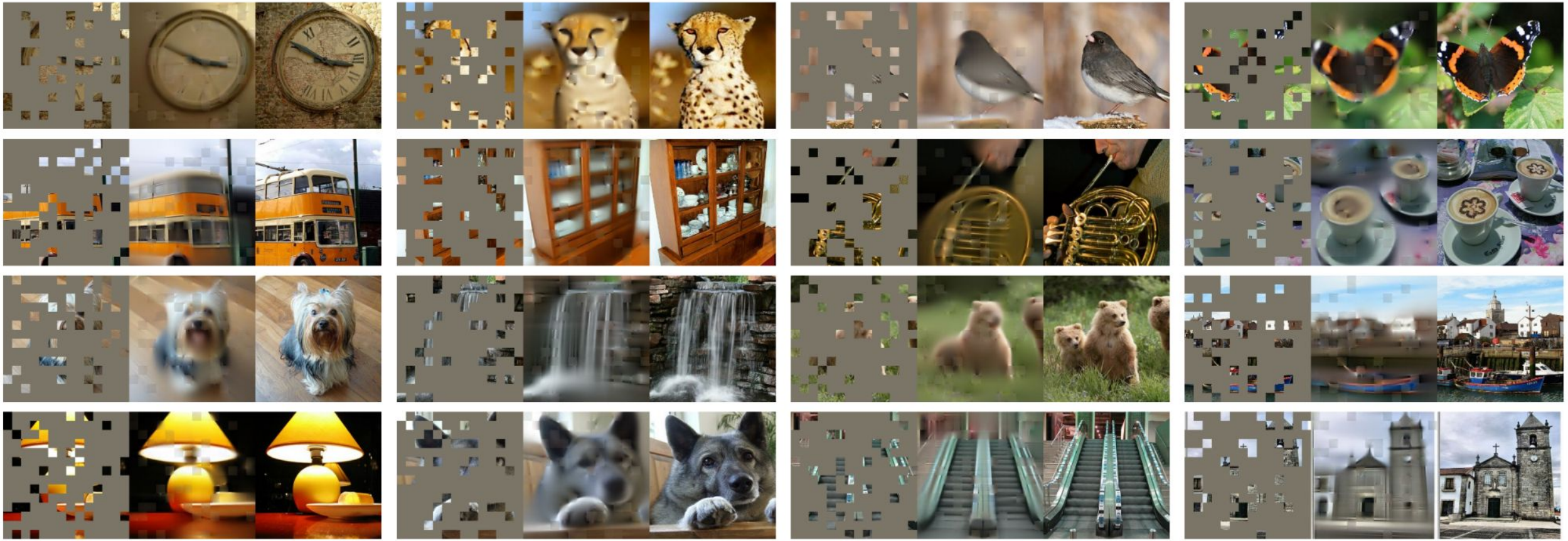


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.

†*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

Ablation: Masking Ratio

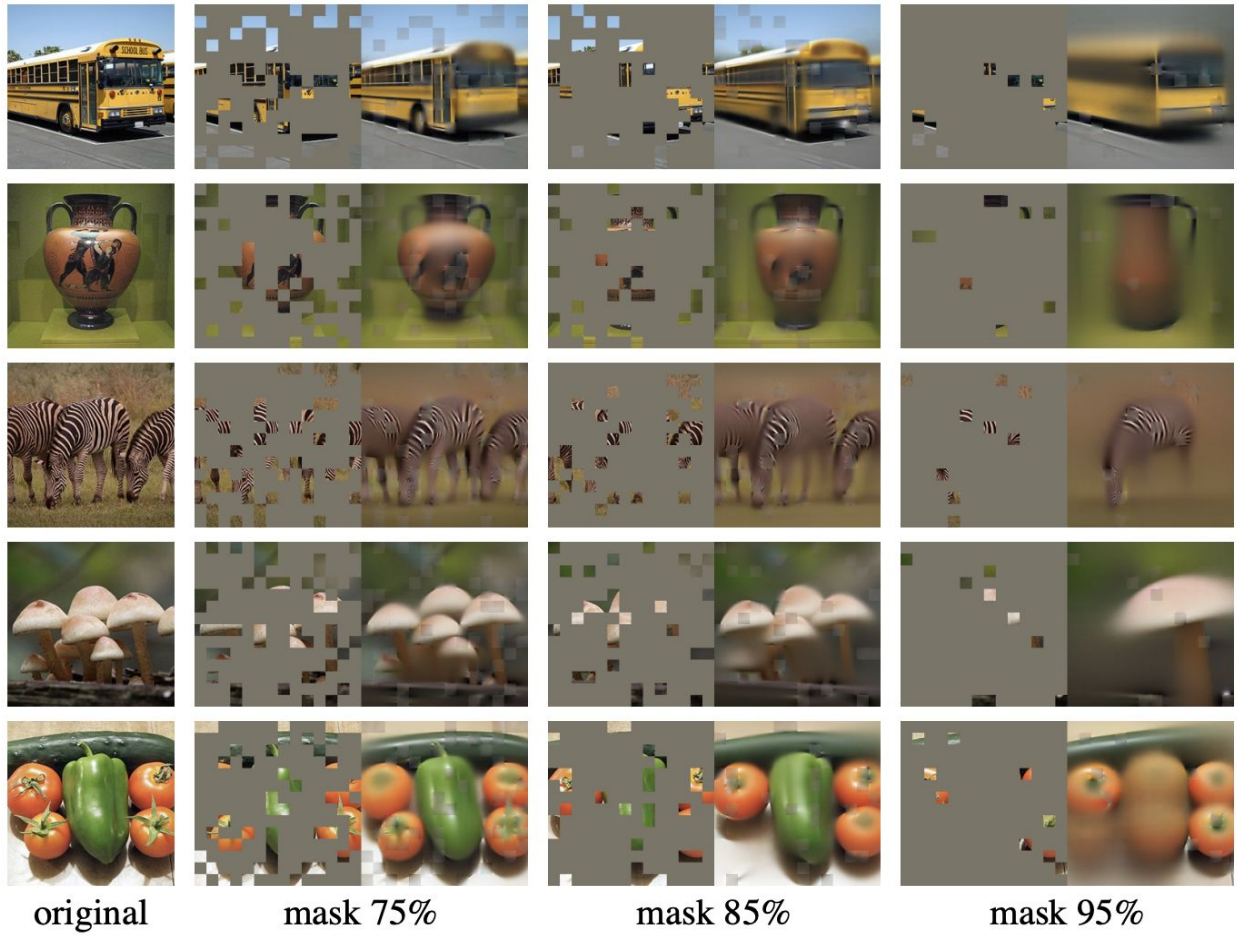


Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.

We hypothesize that this reasoning-like behavior is linked to the learning of useful representations

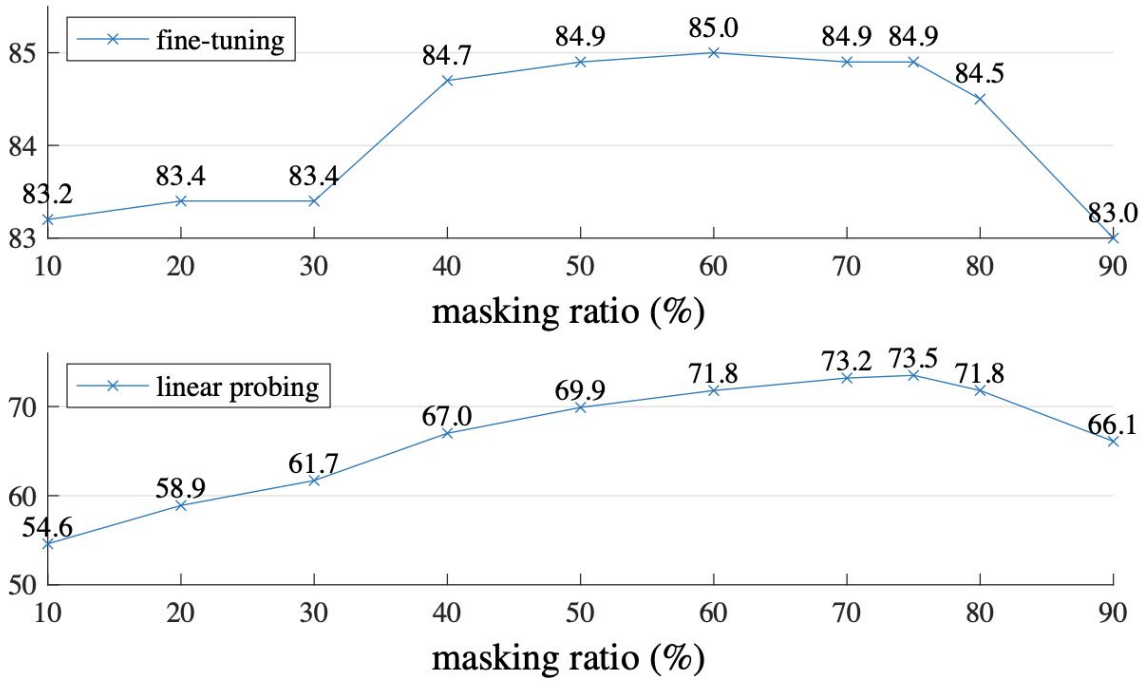


Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

Ablation: Decoder Depth

- **A sufficiently deep decoder is important**
  - The layer several layers of decoder are more specialized for reconstruction, but are less relevant for recognition.
  - A reasonably deep decoder can leave the latent representations at a more abstract level

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

Ablation: Decoder Width

- **A sufficiently deep decoder is important**
  - We use 512-d by default, which performs well under fine-tuning and linear probing

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

Ablation: Mask Token

- If the encoder uses mask tokens, it performs worse
- By skipping the mask token in the encoder, we increase training FLIPs by 3.3×

case	ft	lin	FLOPs
encoder w/ [M]	84.2	59.6	3.3×
encoder w/o [M]	**84.9**	**73.5**	**1**×

Ablation: Data Augmentation

- MAE works well even without data augmentation
- In contrast, using cropping-only augmentation in BYOL and SimCLR reduces accuracy by 13% and 28%

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.

Ablation: Training Epochs

- The accuracy improves steadily with longer training
- In contrast, MoCo v3 saturates at 300 epochs for ViT-L

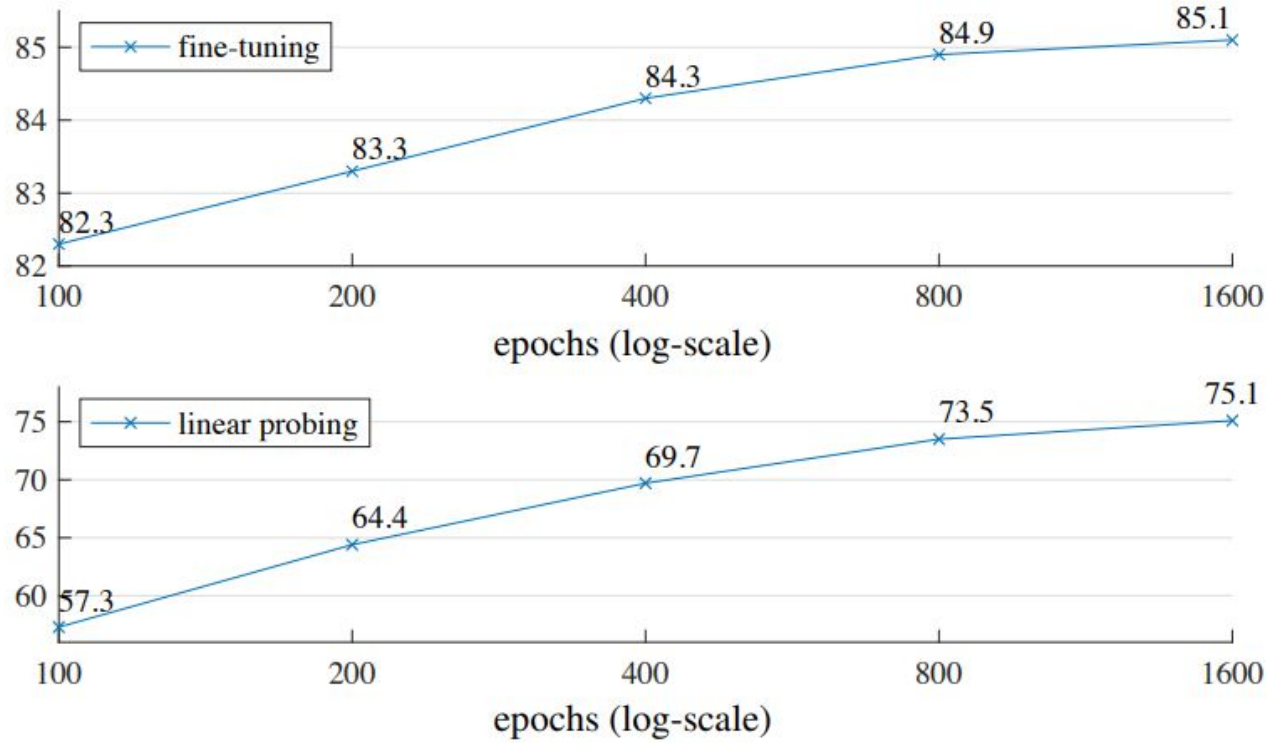
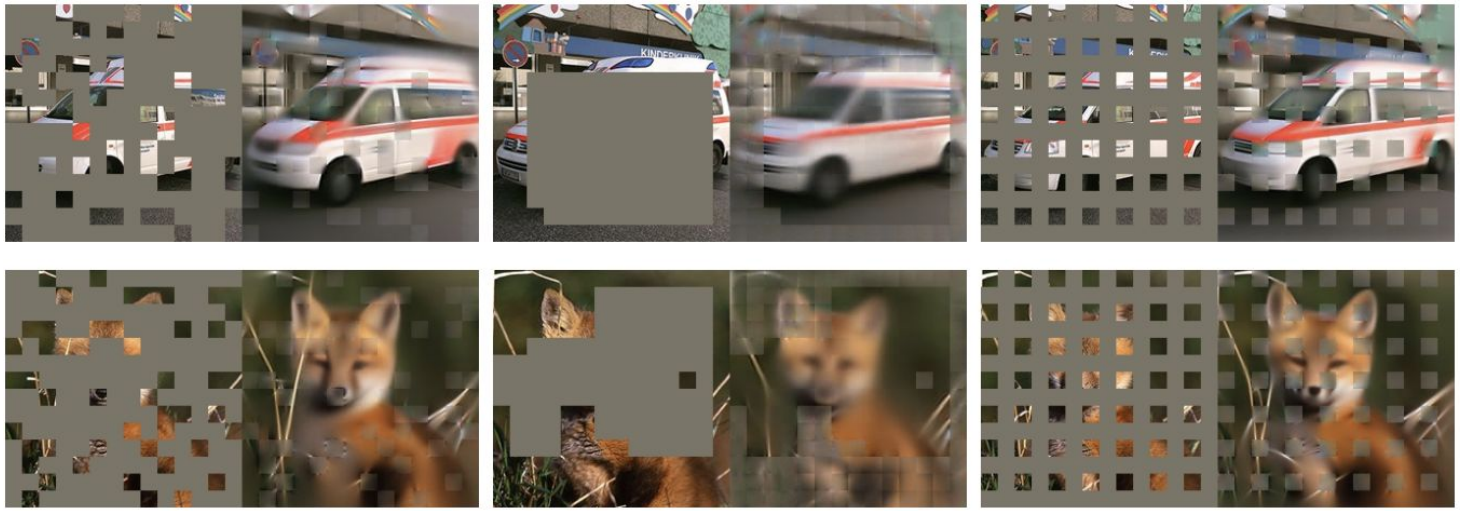


Figure 7. **Training schedules**. A longer training schedule gives a noticeable improvement. Here each point is a full training schedule. The model is ViT-L with the default setting in Table 1.

Ablation: Masking Strategies



random 75% block 50% grid 75%

Figure 6. **Mask sampling strategies** determine the pretext task difficulty, influencing reconstruction quality and representations (Table 1f). Here each output is from an MAE trained with the specified masking strategy. Left: random sampling (our default). Middle: block-wise sampling [2] that removes large random blocks. Right: grid-wise sampling that keeps one of every four patches. Images are from the validation set.

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.

Comparison with Self-Supervised Methods

- Scalability: MAE can scale up easily with steady improvement from bigger models

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

Table 3. **Comparisons with previous results on ImageNet-1K**. The pre-training data is the ImageNet-1K training set (except the tokenizer in BEiT was pre-trained on 250M DALLE data [50]). All self-supervised methods are evaluated by end-to-end fine-tuning. The ViT models are B/16, L/16, H/14 [16]. The best for each column is underlined. All results are on an image size of 224, except for ViT-H with an extra result on 448. Here our MAE reconstructs normalized pixels and is pre-trained for 1600 epochs.

Comparison with Supervised Methods

- Scalability: MAE follows a trend similar to the JFT-300M supervised pre-training

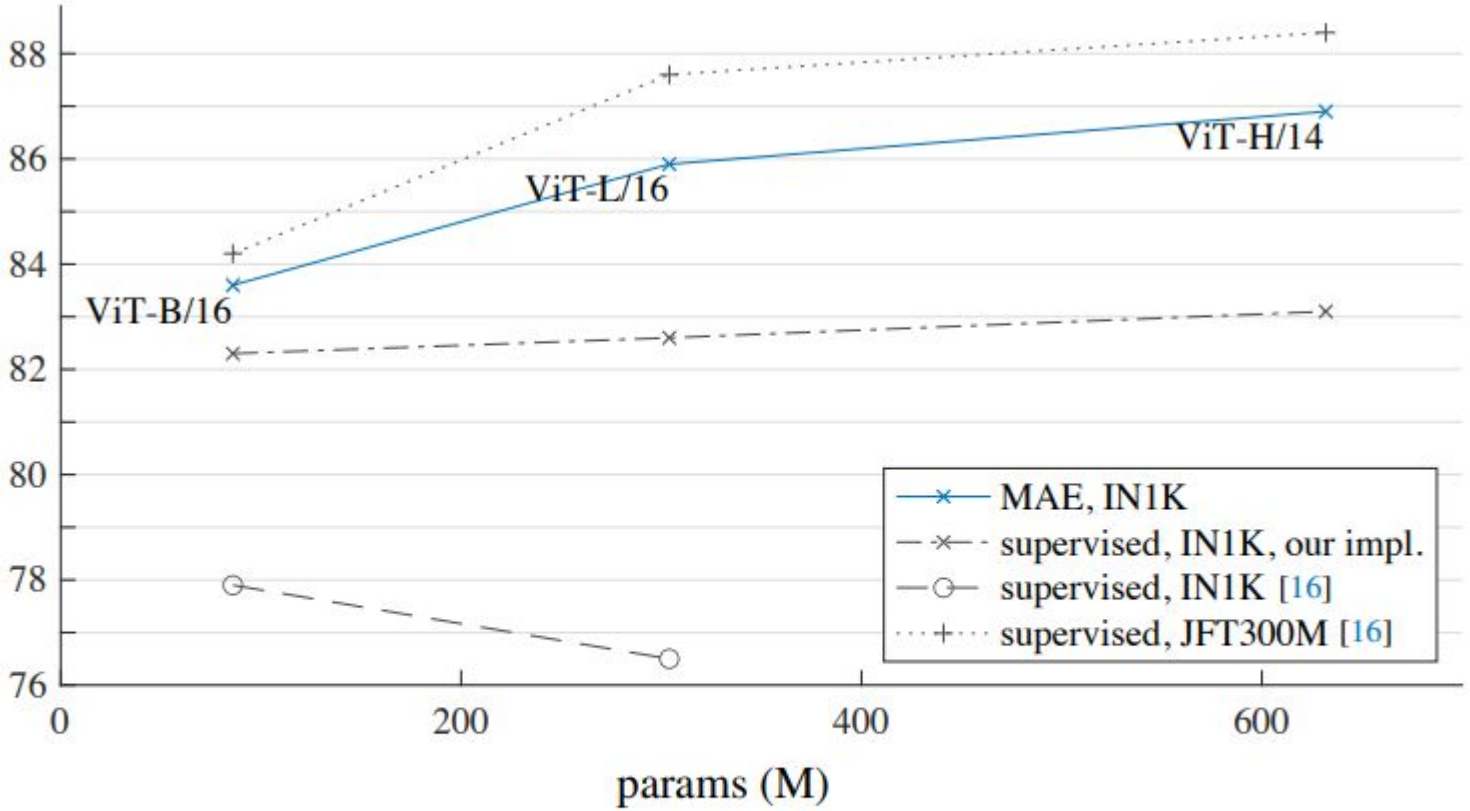


Figure 8. **MAE pre-training *vs*. supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

Partial Fine-Tuning

- MoCo v3:
  - Higher linear probing accuracy
  - However, its partial fine-tuning results are worse than MAE
- MAE:
  - Stronger non-linear features and perform well when a non-linear head is tuned

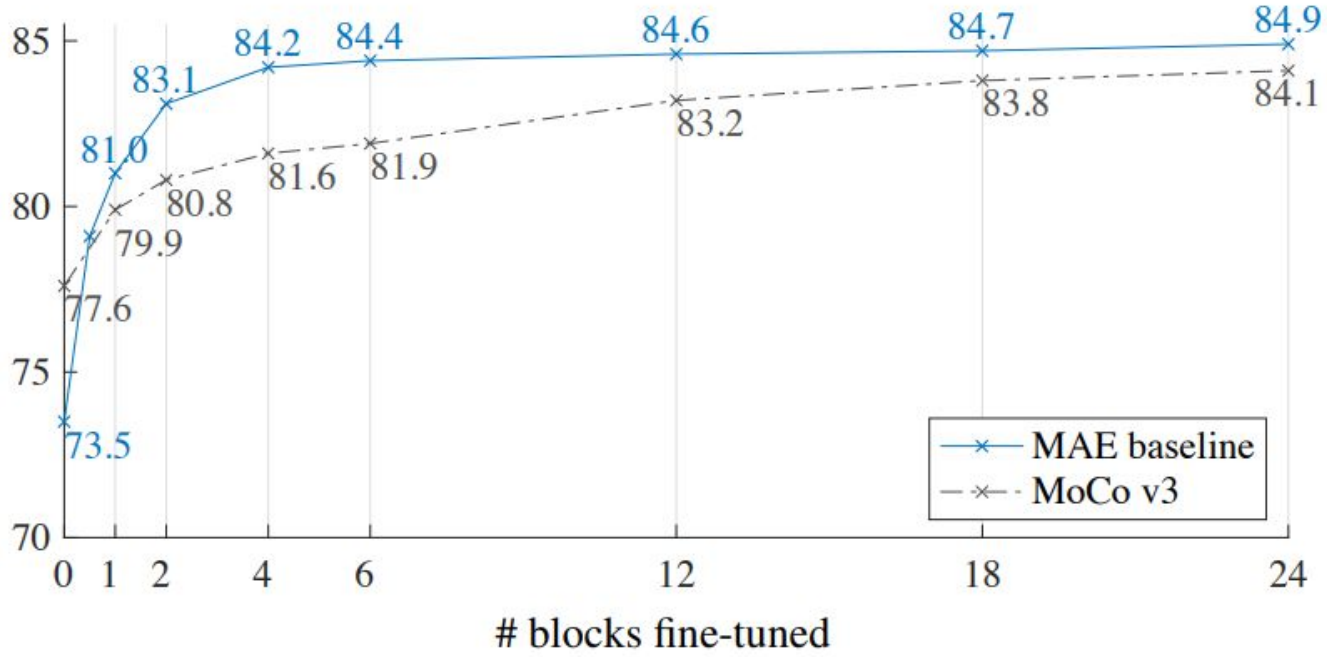


Figure 9. **Partial fine-tuning** results of ViT-L w.r.t. the number of fine-tuned Transformer blocks under the default settings from Table 1. Tuning 0 blocks is linear probing; 24 is full fine-tuning. Our MAE representations are less linearly separable, but are consistently better than MoCo v3 if one or more blocks are tuned.

Conclusion

- Images and languages are signals of a different nature
- MAE infers complex, holistic reconstructions, suggesting it has learned good semantics concepts
- This behavior occurs by way of a rich hidden representation inside the MAE

Thanks for your attention!