Masked Autoencoders Are Scalable Vision Learners

Kaiming He, Xinlei Chen Saining Xie Yanghao Li Piotr Doll´ar Ross Girshick

Presented by Jiang Li and Shreyash Malhotra

Agenda

Methods

- Motivation
- Overall architecture
- Related work
- Design details

Experiments

- Setup
- ImageNet-1K
- Transfer Learning

Summary

Motivations

Masked autoencoding based method achieves great success in NLP domain, but the same idea doesn't work as well in vision. Why the same success has not been replicated in vision?

The author tries to find reasons from three aspects

- Architecture
- Information density
- Decoder design

Motivations - Backbone Architecture

CNN(**C**onvolutional **N**eural **N**etwork) vs ViT(**Vi**sual **T**ransformer)

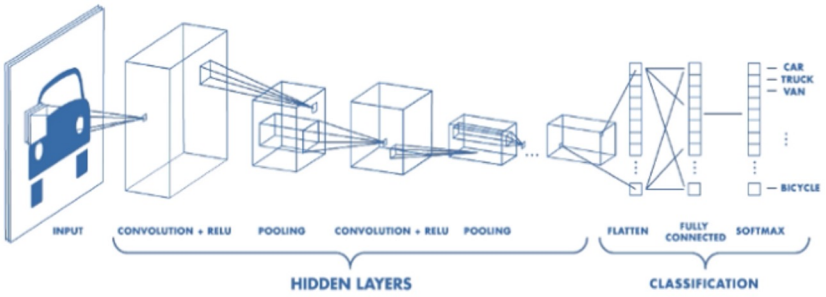


Image source: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

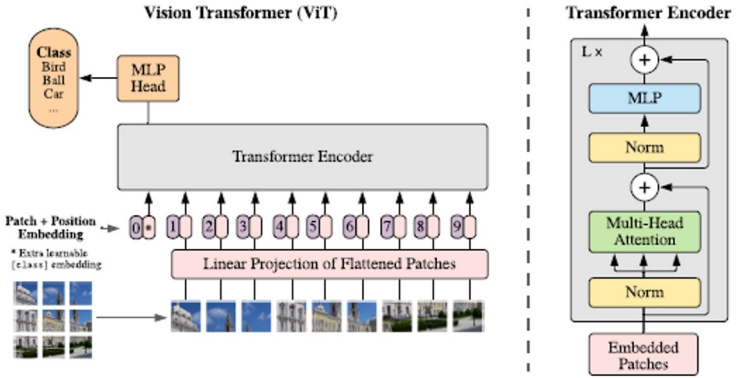


Image source: https://arxiv.org/pdf/2010.11929v2.pdf

Model	ResNet-50	ResNet-101	ResNet-152	ViT-Base	ViT-Large	ViT-Huge
# parameters	25.6M	44.5M	60.2M	86M	307M	632M

CNN architecture may also work well.
ConvMAE:
https://arxiv.org/pdf/2205.03892.pdf

Motivations - Information Density

Difference between language and Image

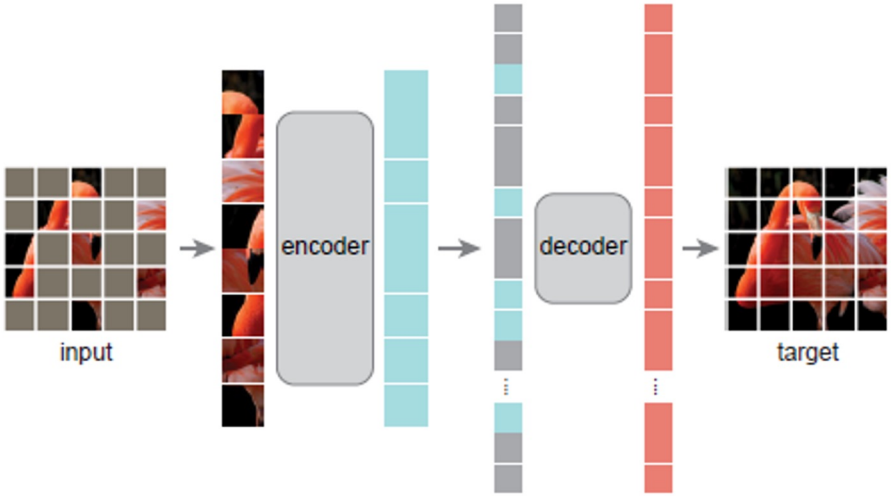
- Language
  - Human generated signal, discrete
  - Strong semantic structure
  - Abstract, compact and highly information-dense.
- Image
  - Nature signal, continuous
  - Heavy spatial redundancy, low information density.

Language is more information-dense than image. So, masking 15% of words in a sentence is not equivalent to making 15% of patches in an image. We may need to mask more patches in an image to reach the same level of information loss in a sentence.

Motivations - Decoder Design

- Decoder design plays a key role in determining the semantic level of the learned latent representations.
- If the decoder is too simple, the pixel reconstruction quality will be bad and the encoder is difficult to learn meaningful representations.
- If the decoder is too complicated, the pixel reconstruction quality will be good but the more semantic representation may shift to the decoder side and leave the learned features from the encoder less semantic
- We need to make a good trade-off in the complexity of the decoder so that the encoder can learn more semantic/high-level features.

Masked Autoencoder - Overall Architecture



- Asymmetric encoder and decoder architecture
- Encoder only operates on un-maksed patches
- Decoder operates on encoded visible patches and masked tokens

MAE - Design Details

Masking

- Randomly sample a subset of patches and mask (i.e., remove) the remaining ones.
- High masking ratio

Encoder

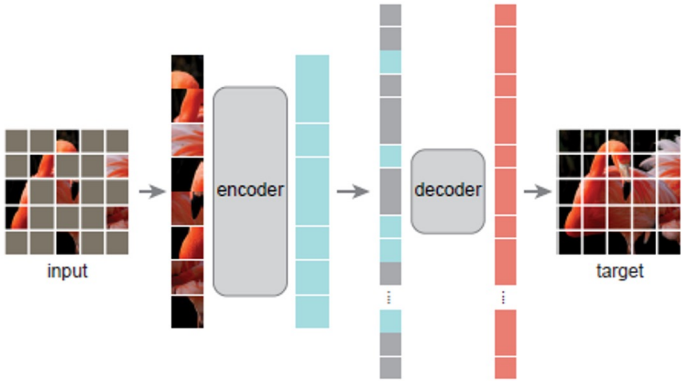
- Based on ViT
- Only operates on visible patch.

Decoder

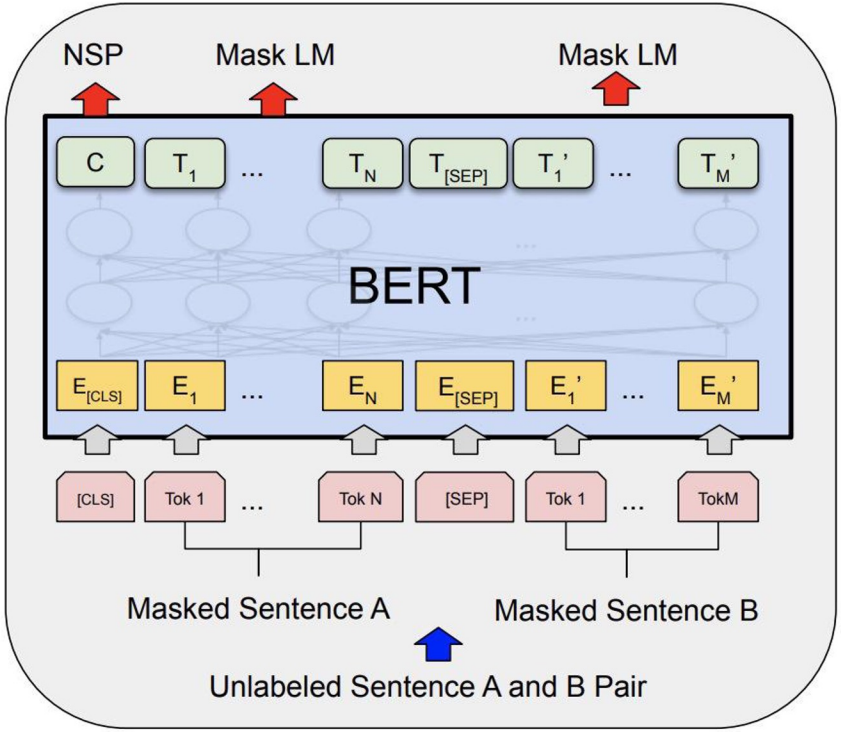
- Operates on both encoded visible patches and mask tokens
- Only used during pre-training.

Reconstruction loss

- L2(MSE) loss, computed only on masked patches

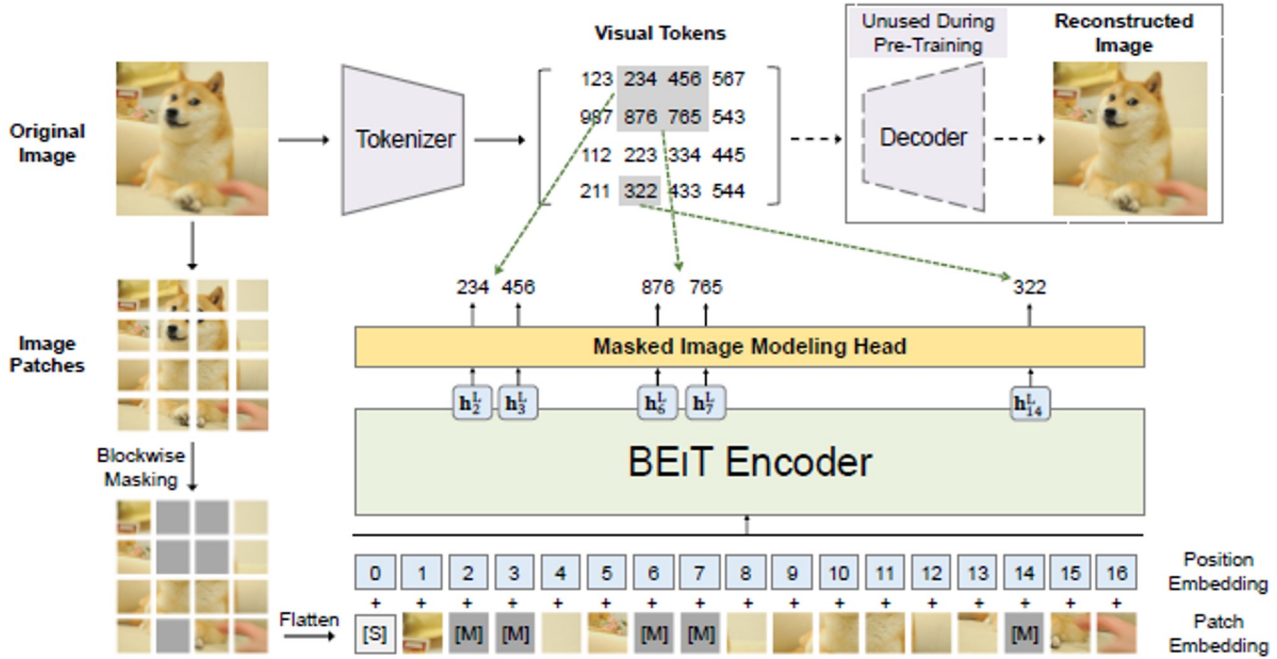


BERT(**B**idirectional **E**ncoder **R**epresentation from **T**ransformers)



Mask ratio:15%
Decoder/prediction: MLP
BERT-B (L=12, H=768, A=12, 110M)
BERT-L (L=24, H=1024, A=16, 340M)

BEiT(**B**idirectional **E**ncoder representation from **I**mage **T**ransformers)



Tokenize the image to discrete visual tokens, by using the latent codes of discrete VAE(VQ-VAE)

The model learns to recover the **visual tokens** of the original image, instead of the raw pixels of masked patches.

Experiment Setup

- Encoder: ViT-B, ViT-L, ViT- H, ViT-H448

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

- Decoder
  - Depth:1-8 blocks(of transformer)
  - Width: 128-1024 dim
- Need extra projection layer to match the encode and decode width
- Training VIT-L/H from scratch on ImageNet-1K can be very tricky, need strong regularization.
- Pre-training, fine-tune, linear probe, partial fine-tune

ImageNet-1K Results

Use ViT-L(307M parameters), pre-trained on ImageNet-1K.

Top-1 validation accuracy

Scratch, original	Scratch, w. strong reg.	MAE + fine-tuning
76.5	82.5	84.9

MAE pre-training outperforms supervised pre-training on ImageNet-1K

ImageNet-1K Results

Masking ratio

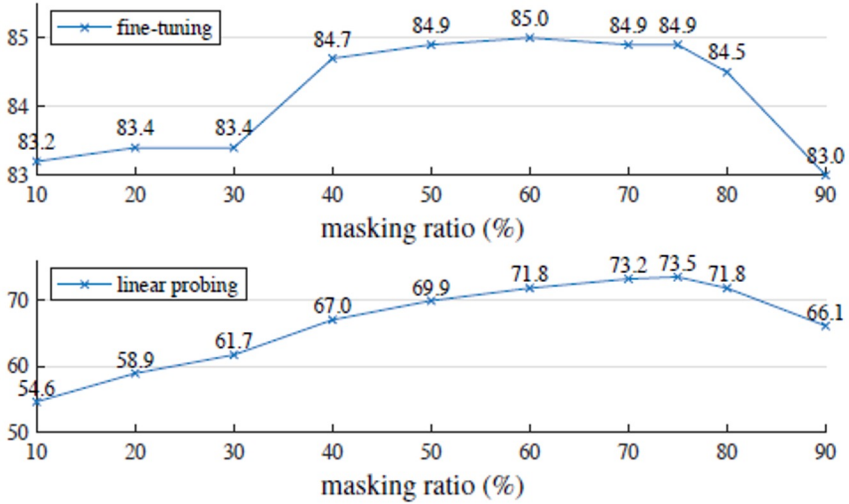


Figure 5. **Masking ratio**. A high masking ratio (75%) works well for both fine-tuning (top) and linear probing (bottom). The y-axes are ImageNet-1K validation accuracy (%) in all plots in this paper.

ImageNet-1K Results

Decoder Design

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.

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).

Single-block decoder perform well with fine-tuning
Decoder: 8 blocks, 512-d width, only 9% FLOPs per token vs Encoder: 24 blocks, 1024-d

ImageNet-1K Results

Mask token

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

(c) **Mask token**. An encoder without mask tokens is more accurate and faster (Table 2).

An important design of our MAE is to skip the mask token [M] in the encoder and apply it later in the lightweight decoder.

ImageNet-1K Results

Reconstruction target

case	ft	lin
pixel (w/o norm)	84.9	73.5
pixel (w/ norm)	**85.4**	**73.9**
PCA	84.6	72.3
dVAE token	85.3	71.6

(d) **Reconstruction target**. Pixels as reconstruction targets are effective.

ImageNet-1K Results

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.

MAE works well using cropping-only augmentation, either fixed-size or random-size (both having random horizontal flipping), behaves decently even if using no data augmentation (only center-crop, no flipping).

Experiment - ImageNet-1K

Mask sampling strategy

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.

- Block-wise: harder task, degrade at 75%,blurrier reconstruction
- Grid-wise: easier task,lower quality representation, sharper reconstruction
- Random: Higher mask ratio, faster and accurate

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.

ImageNet-1K Results

Training schedule

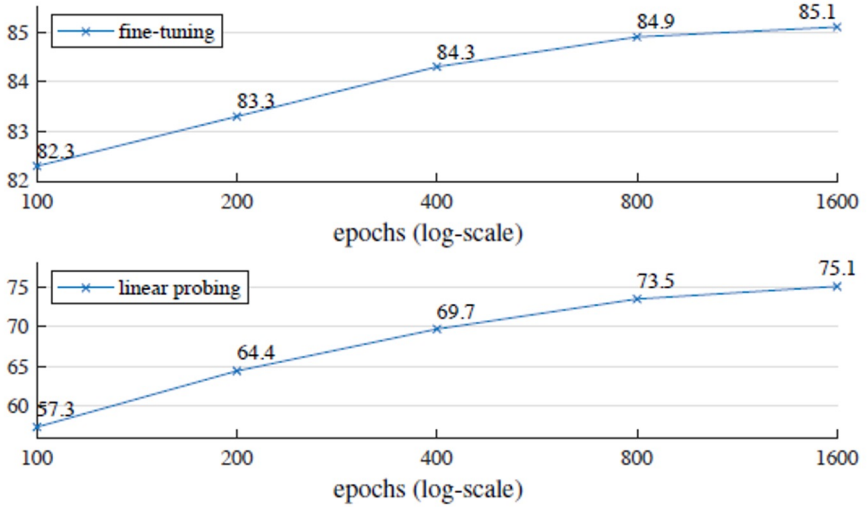


Figure 7. **Training schedules**. A longer training schedule gives a noticeable improvement. Here each point is a full training schedule. The model is ViT-L with the default setting in Table 1.

ImageNet-1K Results

Comparisons with self-supervised methods

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.

Comparisons with supervised pre-training

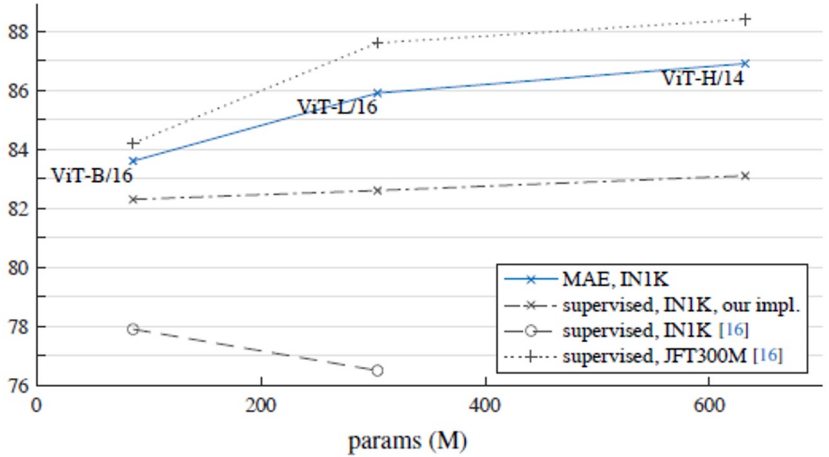


Figure 8. **MAE pre-training *vs.* supervised pre-training**, evaluated by fine-tuning in ImageNet-1K (224 size). We compare with the original ViT results [16] trained in IN1K or JFT300M.

ImageNet-1K Results

Partial fine-tuning

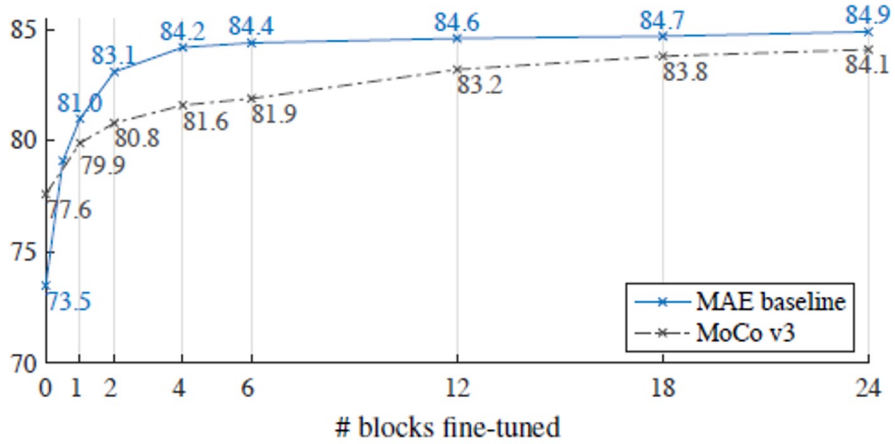


Figure 9. **Partial fine-tuning** results of ViT-L w.r.t. the number of fine-tuned Transformer blocks under the default settings from Table 1. Tuning 0 blocks is linear probing; 24 is full fine-tuning. Our MAE representations are less linearly separable, but are consistently better than MoCo v3 if one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

Experiment - Transfer Learning

Object detection and segmentation

		AP^{box}		AP^{mask}	
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 implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

Experiment - Transfer Learning

Semantic segmentation

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

Table 5. **ADE20K semantic segmentation** (mIoU) using UperNet. BEiT results are reproduced using the official code. Other entries are based on our implementation. Self-supervised entries use IN1K data *without* labels.

Experiment - Transfer Learning

Classification tasks

dataset	ViT-B	ViT-L	ViT-H	ViT-H_{448}	prev best
iNat 2017	70.5	75.7	79.3	**83.4**	75.4 [55]
iNat 2018	75.4	80.1	83.0	**86.8**	81.2 [54]
iNat 2019	80.5	83.4	85.7	**88.3**	84.1 [54]
Places205	63.9	65.8	65.9	**66.8**	66.0 [19]†
Places365	57.9	59.4	59.8	**60.3**	58.0 [40]‡

Table 6. **Transfer learning accuracy on classification datasets**, using MAE pre-trained on IN1K and then fine-tuned. We provide system-level comparisons with the previous best results.

†: pre-trained on 1 billion images. ‡: pre-trained on 3.5 billion images.

Experiment - Transfer Learning

Pixels vs. tokens.

	IN1K			COCO		ADE20K	
	ViT-B	ViT-L	ViT-H	ViT-B	ViT-L	ViT-B	ViT-L
pixel (w/o norm)	83.3	85.1	86.2	49.5	52.8	48.0	51.8
pixel (w/ norm)	83.6	85.9	86.9	50.3	53.3	48.1	53.6
dVAE token	83.6	85.7	86.9	50.3	53.2	48.1	53.4
△	0.0	-0.2	0.0	0.0	-0.1	0.0	-0.2

Table 7. **Pixels *vs.* tokens** as the MAE reconstruction target. △ is the difference between using dVAE tokens and using normalized pixels. The difference is statistically insignificant.

Experiment - Some Examples

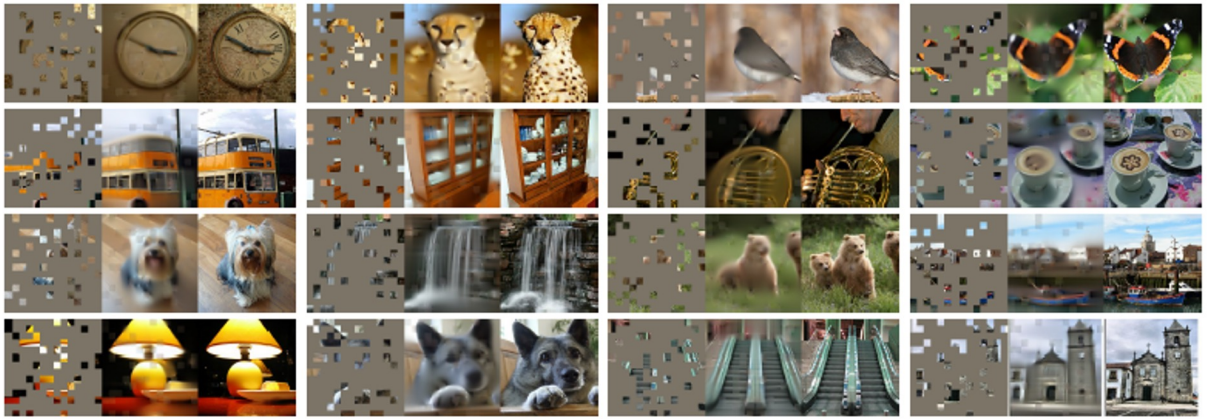


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.
†*As no loss is computed on visible 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.*

Figure 3. Example results on COCO validation images, using an MAE trained on ImageNet (the same model weights as in Figure 2). Observe the reconstructions on the two right-most examples, which, although different from the ground truth, are semantically plausible.

Experiment - Key findings

- MAE pre-trained on ImageNet1K outperform supervised counterpart.
- High masking ratio is crucial to the results, it also significant decrease the computation cost during training.
- MAE is more data efficient, as shown in Table 6, MAE pre-trained on ImageNet1K outperform the previous best results which were pre-trained on billions of images.
- MAE works very well on transfer tasks.
- Compared to contrast learning, MAE requires very little data argumentation.
- MAE perform worse on linear probe compared to contrast learning counterpart. It implies that the features learned from MAE is less linearly separable.

Summary

- Milestone work of applying masked autoencoding on vision, equivalent BERT for NLP.
- Simple architecture, high training efficiency, superior results, easy to scale.
- Thorough evaluations on ImageNet ablation experiments.
- Why MAE works so well is still not fully understood, especially theory analysis is lacking.