MASKED AUTOENCODERS ARE SCALABLE VISION LEARNERS

Nam Trinh Tuan

University of Information Technology, Ho Chi Minh City, Vietnam

What ?

We introduce a scalable self-supervised framework for computer vision (MAE), in which we:

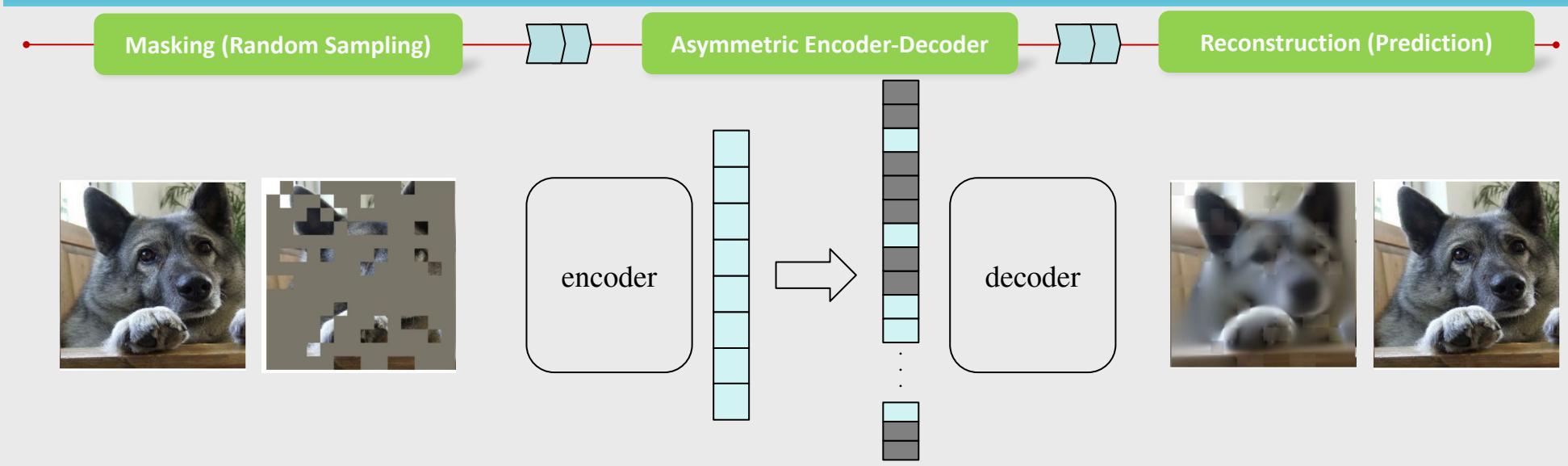
- Propose an asymmetric architecture to mask 75% of the image and reconstruct missing pixels.
- Accelerate training by 3x or more since the encoder operates only on visible patches.
- Achieve state-of-the-art accuracy (87.8%) and scalability on large models (e.g., ViT-Huge).

Why ?

We developed this framework to:

- **Overcome** the data-hungry nature of deep learning models without relying on massive labeled datasets.
- **Address** the heavy spatial redundancy in images that makes standard masking tasks too trivial.
- **Enable** efficient training of high-capacity models that generalize well to downstream tasks.

Overview



Description

1. Masking & Encoding Strategy

- **Patching**: Divide the image into regular non-overlapping patches.
- **75% Masking Ratio**: Randomly mask a high proportion (75%) of the input image.
- **Goal**: This eliminates spatial redundancy, forcing the model to learn holistic, semantic representations rather than just extrapolating from neighbors.
- **Efficient Encoding:** The ViT encoder operates *only* on the visible subset (~25%) of patches, utilizing no mask tokens.

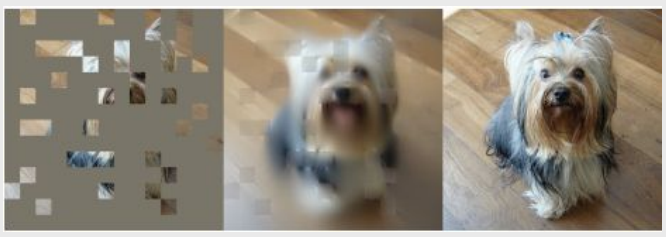


Figure 1. Random masking strategy. With a 75% masking ratio, the majority of visual information is removed, leaving only a small subset of visible patches for the Encoder.

2. High-Quality Reconstruction

- **Decoder Input**: The decoder receives the full set of tokens: latent features from the encoder plus learnable mask tokens.
- **Pixel-wise Prediction**: The model reconstructs the missing pixel values for each masked patch.
- **Semantic Understanding**: Results show the model can infer complex shapes and objects (gestalt) even when 75% of the visual information is missing.

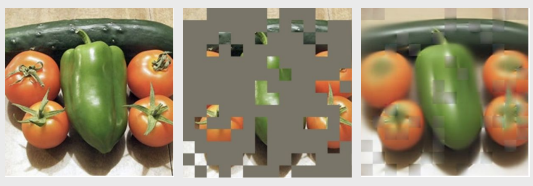


Figure 2. Reconstruction results on ImageNet validation images. Despite missing 75% of the input, MAE infers plausible holistic structures. Left: Masked input. Middle: MAE reconstruction. Right: Ground truth.

3. Performance & Scalability

- **3x Speedup:** Training is accelerated by 3x or more because the encoder processes only 25% of the patches.
- **State-of-the-Art Accuracy:** A vanilla ViT-Huge model achieves 87.8% accuracy on ImageNet-1K, outperforming previous methods using only ImageNet-1K data.
- **Scalability:** Performance consistently improves with larger models (ViT-Large/Huge) and longer training schedules without saturation, surpassing supervised pre-training.

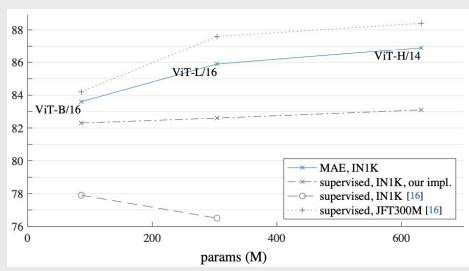


Figure 3. Scalability comparison. MAE pre-training (blue line) consistently outperforms supervised pre-training (gray lines) as the model size increases, achieving 87.8% accuracy with ViT-Huge.

NII

Nam Trinh Tuan – University of Information Technology, Ho Chi Minh City, Vietnam
Email : namtrt.20@grad.uit.edu.vn