Masked Autoencoders Are Scalable Vision Learners
Enhanced Representation Learning through High-Ratio Image Masking and Efficient Architecture

Rohit Suhas Madke
Jio Institute
Mumbai

Rutika Bankar
Jio Institute
Mumbai

Guna Shekhar
Jio Institute
Mumbai

Uziar Kazi
Jio Institute
Mumbai

Abstract

This research explores Masked Autoencoders (MAE) as a scalable, self-supervised learning method for computer vision tasks. MAE leverages a straightforward yet robust process: it randomly masks portions of input images and tasks the model with reconstructing these missing parts, fostering deeper visual representation learning. This approach utilizes an asymmetric encoder-decoder structure—where the encoder focuses on the unmasked patches and the decoder reconstructs the image using both the latent features from the encoder and the masked tokens. With this design, high masking rates (up to 75%) significantly improve computational efficiency, reducing training time by over three times while enhancing accuracy. The model demonstrates strong generalization on ImageNet-1K benchmarks and performs better than supervised models on downstream tasks, validating the effectiveness of high-capacity models trained on masked autoencoding.

1. Introduction

The expansion of deep learning has given rise to increasingly complex models, which, while powerful, often require vast quantities of labeled data. This dependency poses a challenge in computer vision, as annotating large datasets can be costly and time-consuming. In the natural language processing (NLP) field, self-supervised learning has revolutionized this dependency. For instance, models like BERT utilize masked language modeling to learn from unlabeled text, where certain words are masked, and the model learns to predict them based on surrounding context. The success of these masked autoencoding methods in NLP inspired interest in applying similar concepts to vision.

However, adapting masked autoencoding directly to visual data has unique challenges. First, image data differs structurally from language; it is typically more spatially redundant, meaning that some missing image patches can be inferred from nearby regions without understanding high-level concepts. Additionally, vision models historically relied on convolutional neural networks (CNNs), which are not as naturally suited to masked token-based learning as transformers are. The introduction of Vision Transformers (ViTs) closed some of these gaps, enabling architectures that better handle masked patches.the model with reconstructing these areas. This approach prompts the model to grasp the context and structure within images, reducing redundancy and enhancing

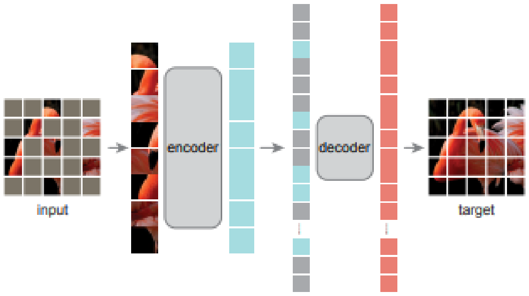

Fig no 1
Autoencoders Img

Our work addresses these challenges by proposing a Masked Autoencoder (MAE) approach. MAE leverages high masking rates (around 75%) to increase task difficulty and reduce redundancy. This forces the model to learn a more meaningful understanding of the content in the unmasked portions of images. The MAE approach reduces computational complexity by processing only a fraction of the input data, enabling it to scale to larger model sizes. Our method achieves state-of-the-art accuracy on ImageNet-1K and surpasses traditional supervised pre-training on various downstream tasks, showing promise for self-supervised learning in vision at scale.

2. Related Work

1. Masked Language Models

Self-supervised language models like BERT and GPT pioneered the concept of predicting masked content within input data, allowing models to generalize across tasks with minimal labeled data. Their success has fueled interest in adapting masked prediction to other domains.

2. Self-Supervised Learning in NLP

Self-supervised methods like BERT and GPT in NLP model the prediction of masked words to capture nuanced language understanding without labeled data. These methods, based on transformers, have shown impressive scalability and generalization, setting a foundation for vision adaptations.

3. Denoising Autoencoders in Vision

Denoising autoencoders (DAEs) have laid the groundwork for masked models by training networks to reconstruct corrupted inputs. Early vision methods, such as [6] and [46], used masking as a form of corruption, focusing on pixel prediction and inpainting large regions of missing content. Our approach builds on these methods with a novel encoder-decoder design and high masking rates for effective scalability.

4. Transformer-Based Image Masking

Following the success of transformers in NLP, visual transformers (ViTs) have been adapted for image tasks. Works such as BEiT and iGPT explored masked image encoding with transformers, predicting discrete tokens or pixels. However, the potential for high-capacity representation with efficient masking ratios remains largely unexplored.

5. Contrastive Learning in Computer Vision

Contrastive learning approaches, such as SimCLR and MoCo, use data augmentation and pairwise comparisons to model image similarity and dissimilarity, yielding impressive performance in self-supervised learning. Despite their success, contrastive methods depend heavily on complex augmentations and high computational costs, contrasting with our MAE approach that uses masking as the primary augmentation.

3. Methodology

The MAE framework is a simple yet powerful self-supervised approach to visual learning, leveraging an asymmetric encoder-decoder structure to process and reconstruct masked image patches efficiently.

3.1. Image Masking

The input image is split into non-overlapping patches, a standard procedure in Vision Transformers. Random masking removes a high proportion of these patches (typically 75%), leaving only a small, sparse set of visible patches for the encoder. This high masking ratio forces the model to reconstruct significant missing portions, requiring it to learn a holistic understanding of the content. The masking strategy is random, ensuring that no specific region of the image is consistently visible or masked, which encourages the model to generalize across varied masking patterns.

3.3. Asymmetric Encoder-Decoder Design

The encoder is a Vision Transformer (ViT) adapted to process only visible, unmasked patches, drastically reducing computational load. After encoding, mask tokens are appended to the encoded patches before passing through the lightweight decoder, which reconstructs the full image from these inputs. This separation between encoder and decoder minimizes the computational demands on the encoder, allowing it to operate on sparse inputs while delegating the reconstruction task to the decoder**.**

Reconstruction Target and Loss Function

The MAE reconstructs the input image by predicting pixel values for each masked patch. Unlike some approaches that predict discrete tokens, pixel-based reconstruction ensures that the model learns fine-grained visual details. The model calculates the mean squared error (MSE) loss between the reconstructed and original pixel values, focusing on masked patches. This targeted loss calculation is more efficient and promotes learning rich, high-level features.

Efficient Implementation

The model's implementation does not require specialized sparse operations. Masked patches are removed at the token generation stage, and visible patches are shuffled to prevent biases. The encoder operates only on the visible tokens, and the decoder processes the unshuffled list, making this method both computationally and memory efficient, ideal for scaling large models.

4. Experiments

4.1. Datasets

The experiments conducted validate the scalability and effectiveness of the MAE model, demonstrating its performance in both pre-training and downstream tasks.

Pre-training on ImageNet-1K

Self-supervised pre-training was conducted on ImageNet-1K, followed by two evaluation methods: fine-tuning (where all weights are updated) and linear probing (where only a linear layer is trained on frozen features). Surprisingly, the model performed best with a high masking ratio of 75%, unlike language models like BERT, which typically use a 15% masking ratio. This demonstrates the need for higher task difficulty in vision to capture rich representations.

Ablation Studies

Comprehensive ablation experiments were conducted to determine the best model configuration. These studies revealed that an eight-layer decoder with 512-dimensional width was optimal, providing high accuracy with reduced training time. Different reconstruction targets, such as token-based and pixel-based methods, were compared, with normalized pixel values showing the best accuracy for visual tasks. Additionally, excluding mask tokens from the encoder improved accuracy and training speed, emphasizing the importance of asymmetry in encoder-decoder roles.

Transfer Learning

The MAE model was tested on various downstream tasks, including object detection and segmentation using the COCO dataset and semantic segmentation on ADE20K. In each case, MAE pre-training outperformed supervised models, showing a 2-4 point improvement in accuracy. This result highlights the generalization capability of MAE pre-training and its effectiveness for diverse vision tasks.

Comparative Analysis with Supervised and Contrastive Models

Comparing MAE with supervised and contrastive models (e.g., MoCo and BEiT) highlighted its efficiency and accuracy gains, particularly with larger models. Our experiments showed that MAE training with high masking is simpler and faster than token-based methods and achieves strong accuracy, particularly for high-capacity models.

5. Conclusion

The MAE approach exemplifies how simple, scalable methods can reshape the landscape of computer vision. Drawing inspiration from self-supervised learning in NLP, MAE effectively addresses the unique challenges in visual data with its innovative asymmetric architecture and high masking ratios. By training on masked autoencoding, our model generalizes well across diverse tasks and datasets, achieving a trajectory similar to NLP's advancement with BERT and GPT.

The success of MAE underscores the potential for scalable, self-supervised learning methods to surpass supervised pre-training in vision applications. Our findings open avenues for future research in masked autoencoding, emphasizing that even simple, pixel-based reconstruction can yield high-performance representations when approached with an efficient encoder-decoder design. This approach could lead to greater advancements in self-supervised learning for large-scale visual models, promoting further exploration of vision transformers and masked modeling at even larger scales..

5.1. Paper ID

Make sure that the Paper ID from the submission system is visible in the version submitted for review (replacing the “****” you see in this document).

References

[1] **[**Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In Advances in Neural Information Processing Systems, pages 153–160,2007.

[2] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. Generative pretraining from pixels. *ICML*, 2020.

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[4] Bao, H., Dong, L., & Wei, F. BEiT: BERT pre-training of image transformers. *arXiv:2106.08254*, 2021

[5] Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. *ICML*, 2008..