MAE

Masked Autoencoders Are Scalable Vision Learners (2022) 1k

We mask random patches of the input image and reconstruct the missing pixels.

- First, we develop an asymmetric encoder-decoder architecture, with an encoder that operates only on the visible subset of patches (without mask tokens), along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens.
- Second, we find that masking a high proportion of the input image, e.g., 75%, yields a nontrivial and meaningful self-supervisory task.

Accelerate training (by 3× or more) and improve accuracy.

Self-supervised pre-training in NLP: based on autoregressive language modeling in GPT and masked autoencoding in BERT

Problem: a missing patch can be recovered from neighboring patches with little high-level understanding of parts, objects, and scene; Solution: mask a very high portion of random patches

MAE encoder. A ViT but applied only on visible, unmasked patches.

MAE decoder. Input is the full set of tokens consisting of (i) encoded visible patches, and (ii) mask tokens (shared, learned vector that indicates the presence of a missing patch to be predicted). Add positional embeddings to all tokens. Use Transformer blocks.

MAE decoder is only used during pre-training to perform the image reconstruction task (only the encoder is used to produce image representations for recognition).

Reconstruction target. Our loss function computes the mean squared error (MSE) between the reconstructed and original images in the pixel space. We compute the loss only on masked patches, similar to BERT.

Experiment:

(i) end-to-end fine-tuning

(ii) linear probing