Masked Autoencoders Are Scalable Vision Learners

Introduction:

The authors of this paper propose a scalable and efficient Masked Autoencoder (MAE) approach for computer vision. Noting that such autoencoders have been lagging in vision tasks compared to Natural Language Processing (NLP), they present an asymmetric design, where the encoder operates only on visible patches and a lightweight decoder reconstructs the full image from the latent representation and mask tokens. They employ a high masking ratio, which challenges the model and reduces redundancy. The authors approach significantly decreases pre-training time and memory consumption and improves performance in comparison to supervised pre-training methods. This MAE model, which can be efficiently implemented without specialized sparse operations, is proposed as a powerful and scalable tool for visual representation learning.

Approach:

These are the details of the approach taken to develop the proposed Masked Autoencoder (MAE) and how it operates.

Masking: The first step in the process involves dividing the image into non-overlapping patches. These patches are then randomly sampled, and the unsampled patches are “masked” or removed. The selection process is randomized to avoid a center bias and to reduce redundancy. This approach of using high masking ratios creates a challenging task for the model as it cannot easily extrapolate information from visible neighboring patches.

MAE Encoder: The encoder is a Vision Transformer (ViT) that only operates on visible, unmasked patches. Each patch is embedded via a linear projection with added positional embeddings, and then processed through a series of Transformer blocks. Since the encoder only operates on a fraction of the full set (25% for example), it can be trained efficiently with reduced computational and memory requirements.

MAE Decoder: The MAE decoder takes as input the full set of tokens, including both the encoded visible patches and the mask tokens (which indicate the presence of a missing patch to be predicted). The decoder also has a series of Transformer blocks, but these are designed to be lighter and less resource-intensive than the encoder’s. The decoder is used only during the pre-training phase to perform the image reconstruction task.

Reconstruction Target: The MAE reconstructs the input image by predicting the pixel values for each masked patch. The decoder’s output is reshaped to form a reconstructed image, and the loss function computes the Mean Squared Error (MSE) between the reconstructed and original images. In the paper, the authors also suggest a variant where the target of reconstruction is the normalized pixel values of each masked patch, which showed improved representation quality in their experiments.

Implementation: The authors point out that the implementation of MAE pre-training is straightforward and does not require specialized sparse operations. It involves generating a token for every input patch, randomly shuffling these tokens, and removing a portion based on the masking ratio. This is followed by

appending mask tokens to the list of encoded patches, unshuffling the full list to align all tokens with their targets, and applying the decoder to this full list.

They highlight that their approach doesn’t require any specialized sparse operations and their MAE architecture that achieves better performance than supervised pre-training, is efficient and scalable, and is straightforward to implement.

ImageNet Experiments:

This is a detailed summary of an experiment conducted on self-supervised pre-training of a model called "Masked Autoencoder" (MAE) using the ImageNet-1K (IN1K) training set. They use Vision Transformer Large (ViT-Large) as a baseline in the experiment. After pre-training, the model is evaluated using supervised training techniques. The main highlights of the experiment are:

The use of a masking ratio: The authors found that a surprisingly high masking ratio (75%) works best for both linear probing and fine-tuning. This ratio contrasts with the typical masking ratio (15%) used in BERT models. They also found that masking ratios for computer vision tasks are generally between 20% and 50%. The accuracy and efficiency of fine-tuning depend heavily on the pre-training phase.

Decoder design: The researchers examined the effects of decoder depth and width. A deep decoder is crucial for linear probing, but its depth is less important if fine-tuning is used. A small decoder can still perform well and speed up training.

Mask tokens: The use of mask tokens in the encoder leads to a decrease in accuracy. By removing the mask token from the encoder, the training computation is significantly reduced, leading to a considerable speedup in the training process.

Reconstruction target: They experimented with different reconstruction targets and found that using pixels without normalization works best. A pixel-based MAE is simpler than tokenization and doesn't require an additional pre-training stage.

Data augmentation: They found that their MAE works well with cropping-only augmentation. Interestingly, it even performs reasonably well without any data augmentation, which is significantly different from contrastive learning methods that heavily rely on data augmentation.

Mask sampling strategy: Simple random sampling worked best for their MAE. It allowed for a higher masking ratio, which provided both speedup and better accuracy.

Training schedule: The accuracy improved with longer training schedules. They observed that the model hadn't reached saturation even after 1600 epochs of training.

The experiment concluded that the MAE model can be scaled up easily with steady improvements from larger models. The MAE was more accurate, simpler, and faster than other self-supervised models. The researchers also found that partial fine-tuning could significantly improve the accuracy of the model.

Transfer Learning Experiments:

The authors in this text are evaluating their model's performance in several different domains through transfer learning.

"Object detection and segmentation": The authors fine-tuned their model (MAE or Masked Autoencoder) on the COCO dataset using Mask R-CNN for object detection and segmentation. This resulted in their MAE model performing better than models pre-trained using supervised learning methods.

"Semantic segmentation": They experimented on the ADE20K dataset using UperNet for semantic segmentation. Similar to the results in object detection and segmentation, the MAE pre-training outperformed supervised pre-training.

"Classification tasks": They transferred their pre-trained model to the iNaturalist and Places tasks. Their method performed better as the model size increased, outperforming the previous best results by large margins.

"Pixels vs. tokens": This refers to the choice of using individual pixels or groups of pixels (tokens) as the smallest unit of information in the model. They found that while using dVAE tokens (discrete VAE tokens, a certain method of tokenizing images) is better than using unnormalized pixels, it's statistically similar to using normalized pixels. This suggests that tokenization (at least using dVAE) is not necessary for their MAE.

Discussion & Broader Impacts:

Despite images not having a natural semantic decomposition like language, their Masked Autoencoder (MAE) manages to generate complex reconstructions, indicating an understanding of numerous visual concepts.

The authors highlight that their model, like any AI model, could reflect and perpetuate biases in the training data, leading to potential negative societal impacts. They caution that the model's ability to create non-existent content could have ethical implications and urge further research in these areas.

Summary:

The research paper revolves around the implementation of a simple self-supervised method called the Masked Autoencoder (MAE), which is inspired by successful self-supervised techniques in Natural Language Processing (NLP). The authors apply this method to computer vision, specifically to the ImageNet dataset, and explore its benefits for transfer learning.

Their MAE works by masking random patches in an image and then reconstructing those patches. The primary goal is to understand the image on a pixel level rather than attempting to recognize semantic entities.

The paper shows that this method allows for excellent scalability and demonstrates strong performance across various tasks. Despite the MAE working on a non-semantic level, it is capable of inferring complex, holistic reconstructions. This suggests that the MAE has learned numerous visual concepts or semantics, potentially due to a rich hidden representation inside the model.