Exploring: Masked Autoencoders Are Scalable Vision Learners

Alejandro García
19990722-5552
algc@kth.se

Altan Senel
19990526-8398
altans@kth.se

Christos Frantzolas
19980124-4618
frant@kth.se

Abstract

This paper explores the possibility of reproducing the results of the Masked Autoencoders paper [1] using a Tiny ImageNet dataset. We focus on reproducing the results of the ablation study regarding the masking ratio, decoder depth and width and data augmentation. We also add minor ablation studies on the regularization techniques used during the finetuning. We find that, although some modifications to the original training recipes are necessary, the pretrained weights obtained from the MAE method produce better top-1 accuracy when fine-tuning (72.71%) than starting from a random initial state (62.19%). However, we obtain worse results than state-of-the-art ViT methods on this dataset. We conclude by presenting useful techniques for ViT training in low data regimes, and by demonstrating the benefits of using a novel distillation method during MAE pretraining proposed in [2], which resulted in a top-1 accuracy of 76.23%. The code of the methods and experiments can be found in `https://github.com/altansnl/exploring-mae-vision-learners`.

1 Introduction

We can see that the number of published papers related to Machine Learning scales exponentially [3]. This fact makes the task of reproducibility highly valuable in order to assess whether the results proposed by the paper come from the proposed method or if there are some additional factors.

In this project, we tried to reproduce the results presented in [1]. However, due to our time and resource constraints, we used a Tiny ImageNet [4], so we will test whether some of the results mentioned in the original paper are reproducible in a low data regime. In addition, we will try to reproduce the results of the ablation study regarding the masking ratio, decoder depth and width, and data augmentation. Furthermore, we add some minor ablation studies on the regularization techniques used during the fine-tuning.

As a result, we have seen that even in a decreased resolution subsampled version of the original dataset, we had to modify some parts of their proposed training recipes to produce valuable results. Nevertheless, after the proper hyperparameter tuning, we see that using the pre-trained weights obtained from the MAE method produces better top-1 ACC while fine-tuning (72.71%) than starting from a random initial state (62.19%). However, we observe that we obtain worst results than state-of-the-art (92.0% [5]) ViT methods on this dataset. Hence we state some helpful techniques in ViT training in low data regimes. We reproduce a novel distillation method during the MAE pretraining proposed in [2], obtaining better top-1 ACC results (76.23%).

2 Related Work

The original paper introducing Transformers for computer vision tasks [6] adapted the Transformer architecture preferred for Natural Language Processing tasks. In order to exploit the attention mechanism employed by transformer models, the authors first divide the input images into rectangular patches of a smaller size, flatten them and utilize them as "image tokens". Each patch is then embedded using a projection layer. Even though transformer models are not explicitly structured to handle image data (such as CNNs, which exploit several image characteristics, such as shared weights, shift invariance, and local structure of image features), the self-attention mechanism allowed for this family of models to dominate vision task benchmarks in the last two years (2020-2022). Nevertheless, because of their structural agnosticism towards image data and their large number of parameters, training transformer models for vision tasks requires employing a variety of regularization techniques (data augmentations, weight decay) and pre-training regimes to achieve adequate performance.

Masked language modeling (MLM) and autoregressive methods such as BERT and GPT are highly successful for pre-training in Natural Language Processing (NLP). These techniques involve withholding a segment of the input sequence and then training the model to predict the missing content accurately. Such methods have been demonstrated to scale to tremendous sizes, and a vast amount of evidence suggests that the pre-trained representations they generate can generalize to many different downstream tasks. The MAE method [1] attempts to adapt this technique from the language to the image domain.

Knowledge Distillation (KD) is a widely used technique of model compression, which allows the model to obtain the robust capabilities of a large model while still keeping a fast speed of inference like that of a small model. The original KD method [7] works by reducing the KL divergence between the soft logits of the teacher and student models. Since then, various more sophisticated KD approaches have emerged, divided into two main categories: logit distillation and intermediate representation distillation.

3 Methods

In the paper referenced by our project [1], Masked Autoencoders (MAE) are an autoencoding method for reconstructing the original signal from its fragmentary observation. It follows the same pattern as other autoencoders, having an encoder that maps the observed input into a latent vector and a decoder that translates the latent dimension back into the original signal. In addition, the input to the encoder network is masked, meaning that a high proportion of the input tokens are not given to the encoder. However, unlike traditional autoencoders, this approach uses an asymmetric design where the encoder only works on the partial, seen signal (without any mask symbols) and a lightweight decoder that reconstructs the entire signal from the latent representation and mask symbols. Reconstructing a heavily masked input image is used for self-supervised pre-training.

The authors of the original study pre-trained ViT models (used as the encoder network of the MAE architecture) using the ImageNet-1k dataset [8]. The reconstruction target was the Mean Square Error for each pixel between the masked patches of the input image and the decoder output. By setting the ratio of masked to unmasked patches (masking ratio) at a high value (the optimal was found to be 75%), the researchers ensured that the latent representation mapping learned by the encoder is complex enough to capture meaningful features of the images. The pre-trained models were fine-tuned on ImageNet-1k for image classification. The authors conducted several experiments varying the masking ratio, decoder architecture (width and depth), mask token, reconstruction target, mask sampling strategy, data augmentation, and training schedule.

Considering the heavy computational requirements for pre-training the models, we adjusted the experiments and architecture to our given resources. According to the authors of the MAE paper, the wall-clock time of pre-training the MAE (ViT-L on ImageNet-1k for 800 epochs) was 15.4 hours on 128 TPU-v3 cores with Tensorflow. Since this was unattainable for us, we made the following adjustments:

- We only conducted experiments with the ViT-B model with 86M parameters (compared to ViT-L at 307M and ViT-H at 632M).
- We used a smaller dataset (Tiny ImageNet), with 13 times fewer images of low resolution. Causing the following changes in the training:
  - We reduced the patch size to 8 to have a similar proportion of image information per patch in these lower-sized images.

  - We use a batch size of 128 instead of the accumulated 1024 batch size as used in the original paper. This is due to the reduced data set and our computational limitations.
  - We changed the pre-training epoch count to 500, with 50 warm-up epochs. Similarly, in fine-tuning, we train for 300 epochs and with ten warm-up epochs.
- We only focused on a handful of experiments, detailed below.

We used a standard pre-training setup as our reference point (a masking ratio of 75%, with a decoder of 8 layers with a width of 512 channels). The experiments we conducted (varying from the standard regime) were the following:

- Varying the masking ratio (using 50% and 90% instead of 75%).
- Varying the decoder number of layers (using one instead of 8 layers).
- Varying the decoder width (using 128 channels instead of 512).
- Using the 3-Augment data augmentation [9] in the pretraining.
- Fine-tuning on Tiny Imagenet with two different regularization regimes (called small and full augmentation).
- Training a ViT-B model on Tiny ImageNet without any pre-training.
- Combining knowledge distillation with the MAE method, including fine-tuning the resulting model. Note that this is beyond the scope of the original paper and was acquired from [2].

The 3-Augment data augmentation combines greyscale, polarization, and gaussian blur. It has been presented in [9] (an updated paper on DEiT) as a better way to enforce ViT models to focus on shapes rather than colors.

Lastly, in the study [2], they employ KD to improve the self-supervised pre-training method in [1], a computationally efficient knowledge distillation framework called Distilled MAE (DMAE) is employed. This works by minimizing the distance between the intermediate encoder feature maps of a teacher model and the student model - as an additional regularization loss added to the pixel reconstruction loss.

However, to use the pre-trained weights of the ViT-L model from the MAE original implementation as done in [2], we had to add image transformations before feeding the batch to the teacher model. Namely, we resize the image from 64*64 to 128*128 and add padding to get a 224*224 image size. This way, we can have the same number of patches inside the image in our 8-pixel patch-sized Vit-B and 16-pixel patch-sized Vit-L. Furthermore, after the patch embedding layer in the teacher model, we remove the tokens corresponding to padding, obtaining the one-to-one token correspondence between teacher and student needed for computing the distillation loss.

4 Data

The dataset we used was Tiny Imagenet [4]. The dataset contains 100000 training images of size 64*64 pixels belonging to 200 classes. It also includes 50 validation and 50 test images for each class. The benchmark for Tiny Imagenet classification using a ViT-B model used in this project is given by a DeiT-B/16 (Data-Efficient Image Transformers) [10] (87.29% [11]). This method proposes a way of incorporating distillation on ViTs while doing supervised training by adding a "distillation_token" and using some specific setups that help training with small models or with low data regimes.

However, the current state-of-the-art benchmark for the dataset is [5], with a reported top-1 validation accuracy of 92%. The researchers trained an ensemble of 5 DeiT-B/16-D models trained using a novel method called "Overfitting with Conditional Diffusion models" (OCD for short).

5 Experiments and findings

As previously mentioned, we will try different fine-tuning setups. We can see that in the original paper, the authors use several regularization techniques in order to improve training stability and performance. However, we will focus on the effect of Mixup, CutMix, and LabelSmoothing, which combines images of the dataset and its corresponding labels to obtain a more generalizable network. Hence the finetuning without those regularization techniques will be called "small augmentation set-up", and the one with all the augmentations mentioned in the paper will be called "full augmentation".

Our main reproducibility experiment is the ablation study of the decoder structure. We can see in Table 1 that we will focus on testing the decoder depth and width in the "standard" set-up as well as in the "extremely reduced" set-ups. All the reported results are the finetuning top-1 ACC. Furthermore, in order to have our own local benchmark, we finetuned a vanilla ViT-B/16 with random initialization and obtained 57.24% ACC with the small augmentation set-up and 62.19% ACC with the full augmentation set-up.

blocks	small aug	full aug
1	67.56	**72.71**
8	66.76	72.58

(a) Decoder depth

dim	small aug	full aug
128	68.01	72.47
512	66.76	**72.58**

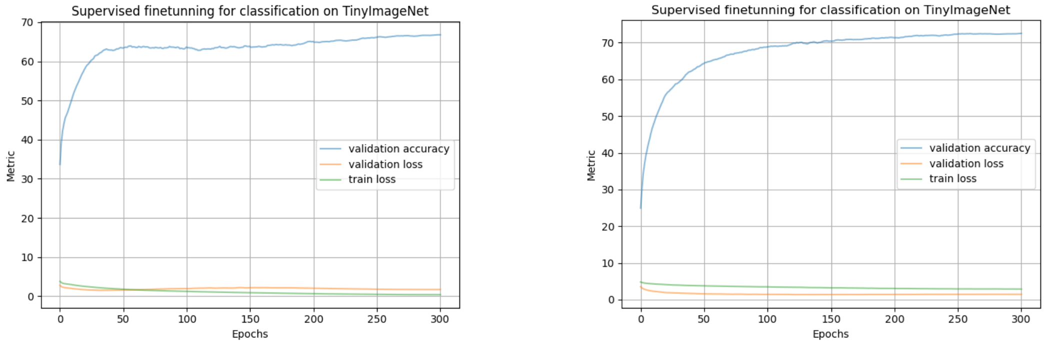
(b) Decoder width

Table 1: Decoder structure ablation study

Starting with the decoder depth, we can see in Table 1a that we obtain similar ACC when using one and eight blocks in the decoder, as in the original paper. However, we can observe that we obtain slightly better ACC with one block when using the proper complete augmentation setup. We believe that the reason could be the simplification of Imagenet to a lower-resolution data set (i.e. Tiny Imagenet). In addition, the original paper mentioned that having a higher number of decoder blocks will force the last decoder layers to learn the reconstruction task, leaving the end of the encoder with a more abstract representation. However, due to our simplified dataset, we believe that the encoder can reach a relevant abstract representation of the image objects with fewer layers (due to smoother edges and textures caused by downscaling). Therefore, when we use one decoder block, we can prevent the network from learning abstract representations by forcing the last encoder layers to help with the reconstruction since the abstract features will be in the middle layers of the encoder.

Due to our time limitations, we could not test this hypothesis. However, we think it could be done similarly to [12] where by taking gradient steps to maximize feature activations starting from random noise, they produce an analysis of the learned content in several ViT models at each layer (see Figure 4).

We can see in Table 1b that in the case of the full augmentation setup, we can use the lower dimensional tokens in the decoder without a significant performance drop. Furthermore, we observed that we could speed up the training considerably with the lower dimensional embeddings, so it could be highly beneficial to use it in computationally constrained setups.



(a) Using the small augmentation set-up

(b) Using the full augmentation set-up

Figure 1: Finetuning with the standard network configuration

In general, we observe that the use of Mixup, CutMix, and LabelSmoothing helps the network by preventing overfitting. As shown in Figure 1 (similar behavior in all small augmentation experiments), we start overfitting after 50 epochs, so the obtained ablation results show which setups help the network memorize slightly better the training dataset. However, we observed that using extra augmentation in the pretrain, such as 3-Augment, is not beneficial (72.51%), as stated in the original paper.

We can see in Figure 2 that in the case of MAE with the full regularization set-up we obtain the same trend as the original paper. We observe that we obtain higher ACC when we have a masking ratio of 75% than when we have a smaller masking ratio of 50%. Therefore, confirming that the assumptions made in the original paper regarding the difference between pretraining a language model with a masking procedure (e.g., BERT) and a ViT still hold in lower-resolution images.

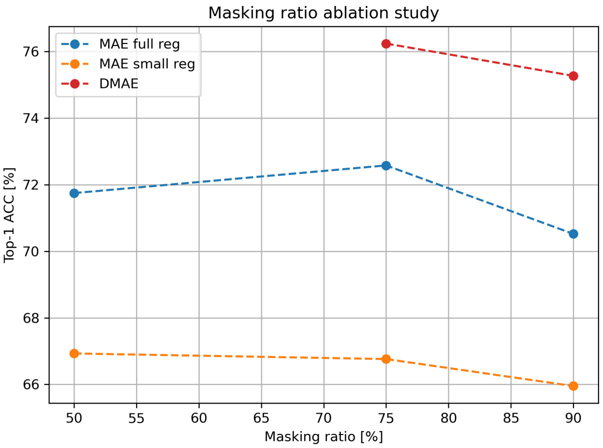
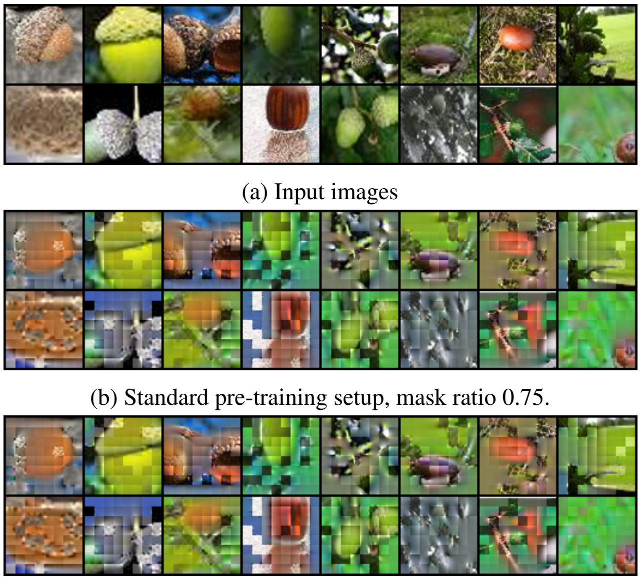


Figure 2: Masking ablation study of the top 1 ACC when fine-tuning after pretraining

Lastly, we observe that pretraining the ViT with the Distilled MAE (DMAE) method and finetuning with the complete augmentation set-up produces a significant increase in top-1 ACC. Furthermore, as mentioned in [2], when using the DMAE method, we can use a higher masking ratio (e.g., 90%) without as much performance drop as the standard MAE method. Surprisingly, we can see in Figure 3 that the reconstruction when using DMAE is almost identical to using the standard MAE (additional reconstruction experiments can be found in the Annex). We believe this is because, in DMAE, we align the features from the 3/4 depth of both the student and the teacher models, which, as we previously mentioned, might be mainly involved in obtaining an abstract representation of images. Therefore intuitively, the DMAE network, in our case, should have learned an adaptation of the MAE ViT-L abstract features towards these new low-resolution images, which are more useful for downstream tasks but not that much for reconstruction.



(a) Input images

(b) Standard pre-training setup, mask ratio 0.75.

(c) Distilled MAE, masking ratio 0.75.

Figure 3: Reconstructions after 500 pretrain epochs

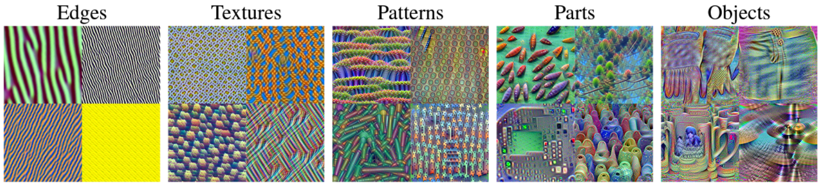


Figure 4: Progression for visualized features of ViT B-32. Source: [12]

6 Challenges

Some difficulties were encountered while reimplementing the original paper. Most of our issues came from the authors relying on the work done in [10, 13] without clearly stating which parts they used from them. For example, while implementing the MAE model, we observed that the original authors used some specific weight initialization not mentioned in the paper. Namely, they use a truncated normal distribution to initialize the cls_token as previously done in [10] to stabilize the training. Furthermore, they use Xavier initialization for all Linear layers instead of He initialization as done in their JAX implementation. We believe this could be due to using the GELU activation function in the blocks. Also, they initialize the Convolution weights of Timm's Patch Embedding layer with Xavier uniform instead of the standard He initialization. However, we were unable to track this last change from previous work, which hinders its reproducibility.

Moreover, to have a feasible training time, we had to use the same "implementation tricks" as the original authors. For example, we used the Pytorch Automatic Mixed Precision Package, which optimizes the floating point precision used in each model component. Furthermore, in the original author's implementation, they mention that using the cls_token for fine-tuning produces some numerical instabilities in Pytorch. At the same time, the paper states that global pooling and cls_token are equally preferable. Hence, we opted to use global pooling to avoid such issues.

Lastly, as mentioned before, due to our time and computational resources, we had to use ViT-B on Tiny Imagenet. Therefore, we had to adapt some hyperparameters to fit our computation capabilities and the properties of this new data set. However, we could not perform proper hyper-parameter tuning other than a few educated trials. The main reason is that grid search and similar search strategies for hyper-parameter tuning require many training runs. Furthermore, since we evaluated our pre-training performance by fine-tuning for a downstream task, we have different parameters associated with them. Therefore, we can only see the effect of changing the hyper-parameters of pre-training with fine-tuning. Nevertheless, if proper hyperparameter tuning on the DMAE method is done, we could possibly surpass the 80% top-1 ACC.

7 Ethical consideration

An ethical consideration that we came across is whether doing these kinds of reproducibility experiments that require high computational costs is worth the carbon footprint that it produces.

8 Conclusion & Self Assessment

In this project, we have seen that even with a smaller dataset and a smaller architecture, we have confirmed that masked auto-encoders are efficient self-supervised learners and a self-supervised pre-trained vision transformer does considerably better than a randomly initialized one.

Similar to the original paper, we have observed that heavy regularization is essential to avoid overfitting. In our experiments with and without MixUp, CutMix, and LabelSmoothing, we have seen that these data augmentation techniques helped the downstream task gain top-1 accuracy of around 5%. Additionally, we had the best results in the case of a 75% mask ratio, similar to the original paper, confirming it is kind of a "sweet spot" in terms of the information masked away.

Since our image resolution is smaller than the original paper, we used a smaller patch size. We showcased that scaling the patch size with respect to the image dimensionality is viable. Unlike the paper, a decoder of depth 1 had the best results for our experiments. On top of the paper, we have also implemented knowledge distillation and confirmed that it is a viable strategy for improving the self-supervised pretraining performance. Distillation also made the usage of high mask ratios feasible.

In terms of education, we have learned how to use transformers for computer vision tasks. In addition, we gained knowledge and intuition about state-of-the-art self-supervised learning methods. We also gained experience performing deep learning reproducibility experiments with limited resources.

In our project proposal, we argued that reproducing the results of the ablation study of the original paper, considering the resource limitations, would get us in the range of C-D. Furthermore, on top of the original paper, we have implemented knowledge distillation, showcasing significant performance improvements. Unlike the paper, in our ablation studies, we experimented with not using MixUp, CutMix and LabelSmoothing; and observed their effect on the downstream performance and overfitting. Additionally, we have implemented a different augmentation schema from a recent paper [9]. On top of that, making the MAE learn sufficiently with smaller-resolution images required additional complexities not described in the original paper. Based on these arguments, this reproducibility project would get a grade in the range B-A.

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[2] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers, 2022.

[3] Philippe Fournier-Viger. Too many machine learning papers?

[4] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.

[5] Shahar Shlomo Lutati and Lior Wolf. Ocd: Learning to overfit with conditional diffusion models. *arXiv preprint arXiv:2210.00471*, 2022.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022.

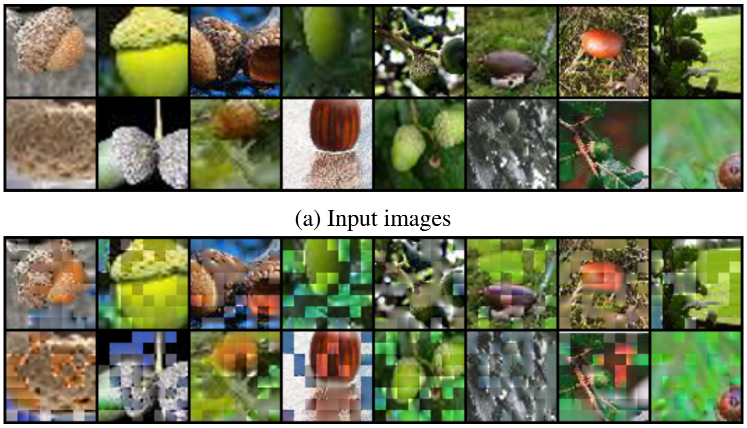
[10] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers amp; distillation through attention, 2020.

[11] Ching-Hsun Tseng, Hsueh-Cheng Liu, Shin-Jye Lee, and Xiaojun Zeng. Perturbed gradients updating within unit space for deep learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2022.

[12] Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. What do vision transformers learn? a visual exploration, 2022.

[13] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2021.

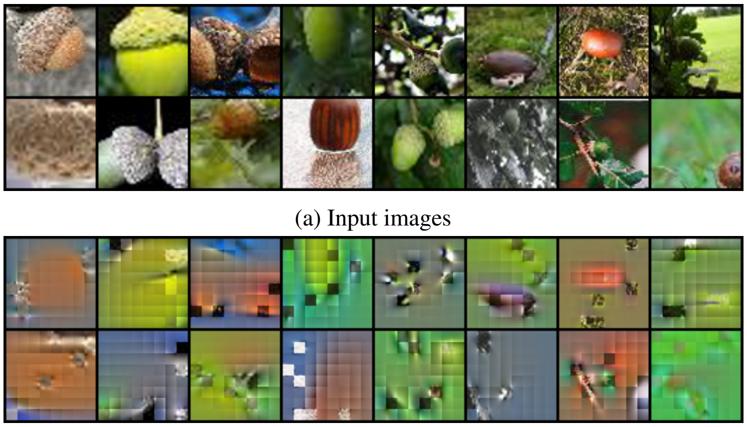
A Annex



(a) Input images

(b) Reconstruction, epoch 500

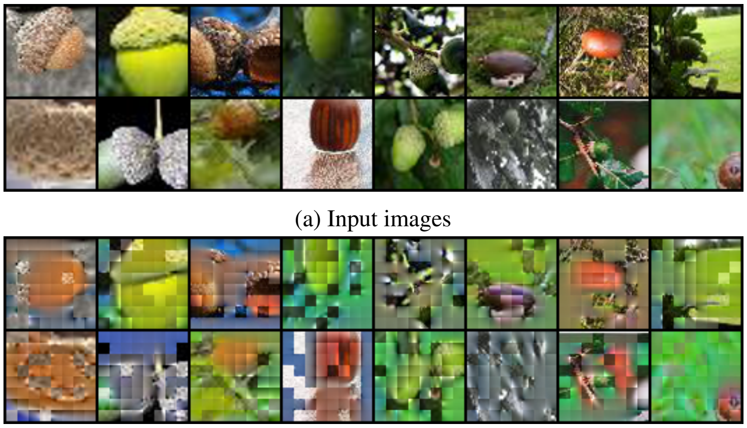
Figure 5: Masking ratio 0.5.



(a) Input images

(b) Reconstruction, epoch 500

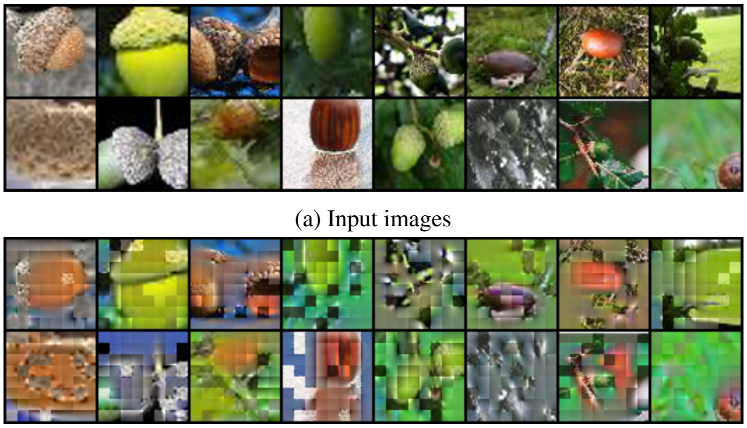
Figure 6: Masking ratio 0.9.



(a) Input images

(b) Reconstruction, epoch 500

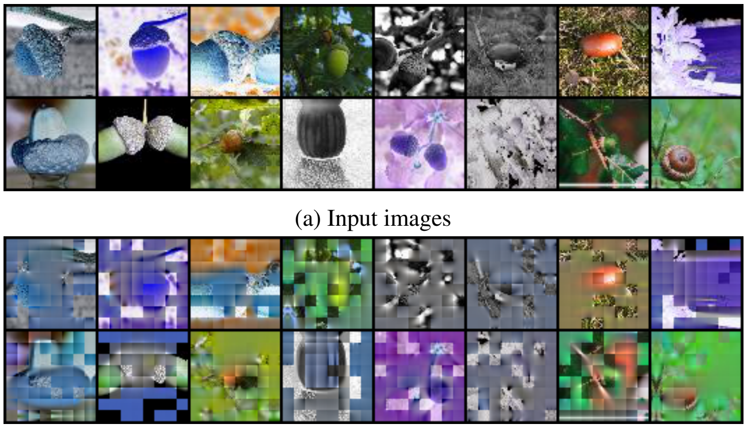
Figure 7: Decoder embedding width 128.



(a) Input images

(b) Reconstruction, epoch 500

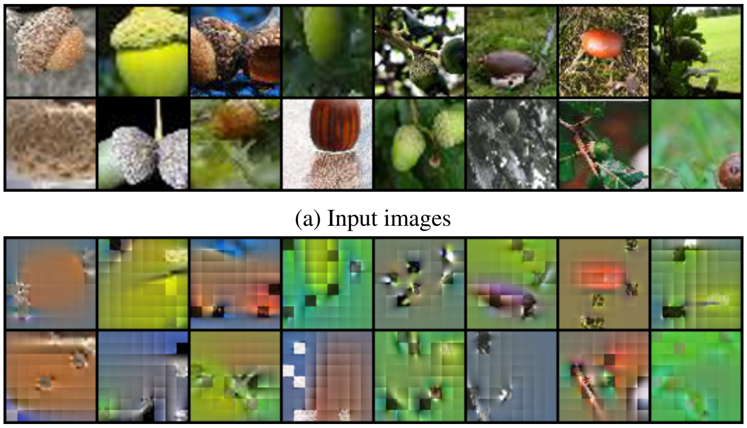
Figure 8: Decoder depth 1.



(a) Input images

(b) Reconstruction, epoch 500

Figure 9: Using 3-Augment



(a) Input images

(b) Reconstruction, epoch 500

Figure 10: Distilled MAE, masking ratio 0.9.