Masked Autoencoders Are Scalable Vision Learners

(CVPR 2022)

Paper Authors:

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick

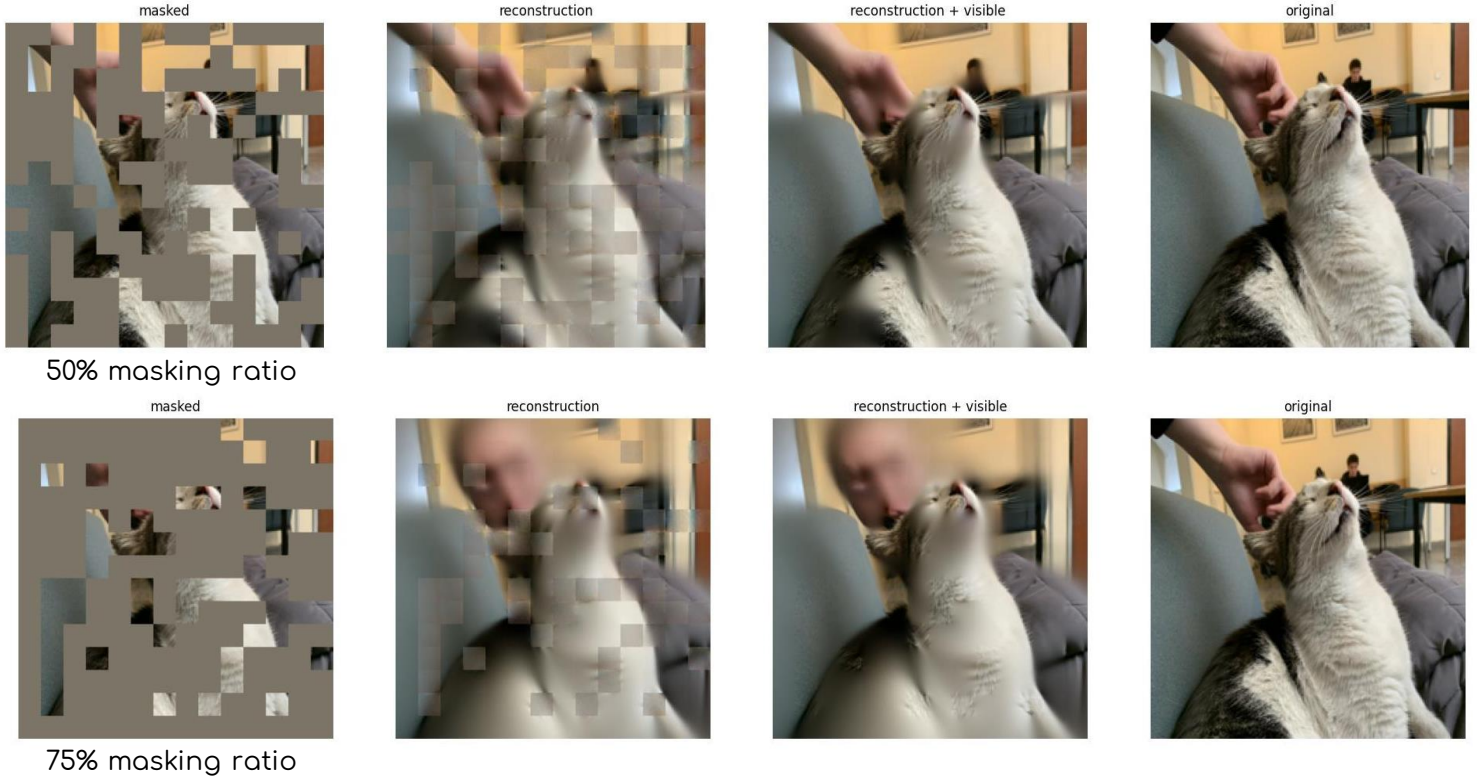
Presentation author:

Umut Özyurt

MIDDLE EAST TECHNICAL UNIVERSITY

This presentation was prepared for the CENG502 - Advanced Deep Learning course for the Spring 2024 term.
It is part of the coursework assigned by our lecturer, Prof. Sinan Kalkan.

Reconstruction example of METU CENG's cat, "Java"



Reconstruction example of METU CENG's cat, "Java"

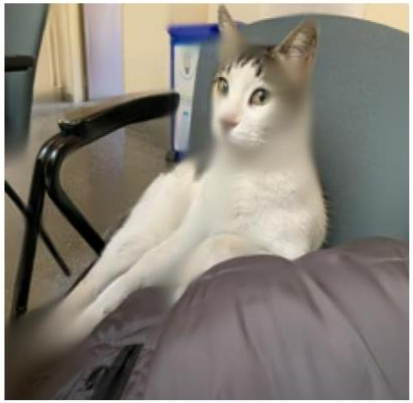
50% masking ratio

Reconstruction example of METU CENG's cat, "Java"

masked

reconstruction

reconstruction + visible

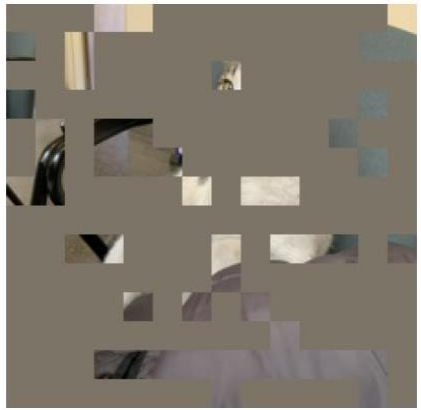


original

50% masking ratio

Reconstruction example of METU CENG's cat, "Java"

masked



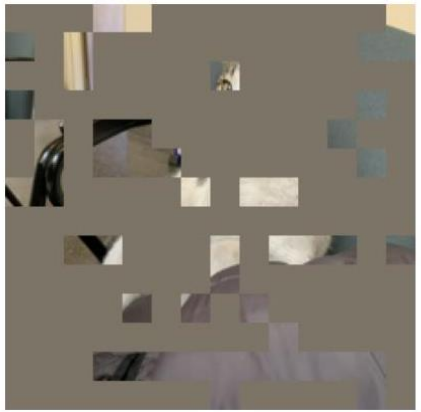
reconstruction

reconstruction + visible

75% masking ratio

Reconstruction example of METU CENG's cat, "Java"

masked



reconstruction

reconstruction + visible

original

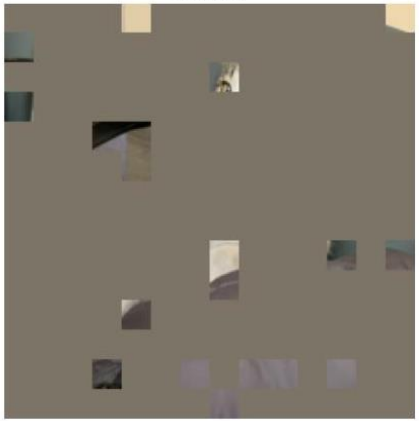
75% masking ratio

Reconstruction example of METU CENG's cat, "Java"

90% masking ratio

Reconstruction example of METU CENG's cat, "Java"

masked



reconstruction

reconstruction + visible

original

90% masking ratio

Overview of the presentation

- ➔ Introduction
- ➔ Related work
- ➔ Approach and method
- ➔ Pre-training experiments
- ➔ Transfer Learning Experiments
- ➔ Discussion and Conclusion
- ➔ Q & A

1-
Introduction

Dataset needs in Computer Vision

For a vision model to perform well, datasets should:

- ➔ Contain a massive amount of samples
- ➔ Be labeled
- ➔ Despite gigantic datasets, pre-training methods do not scale efficiently.

Dataset needs in Computer Vision

For a vision model to perform well, datasets should:

- Contain feasibly large amount of samples
- Be labeled or unlabeled (pre-training)
- With datasets like imagenet-1k (*), ~~does not~~ pre-training methods scale efficiently.

*ImageNet: A Large-Scale Hierarchical Image Database: https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf

Vision vs NLP

- We can adopt pre-training techniques from NLP (like BERT*), which scales well.
- Caution! Information density is different!
- Architecture differences.
- Decoder goal differences.

* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805

Vision vs NLP

- Caution! Information density is different!

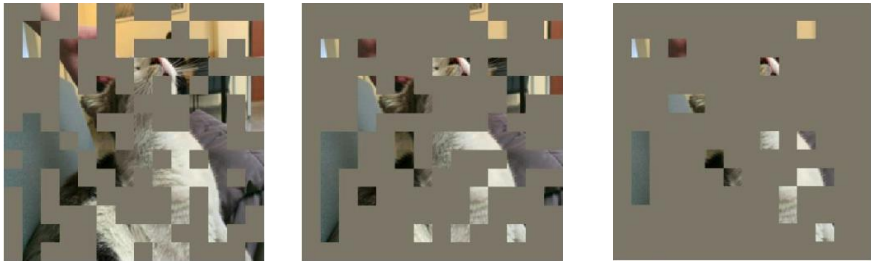


Figure 1. Masked images of METU CENG's cat "Java" at masking ratios of 50%, 75%, and 90% (from left to right).

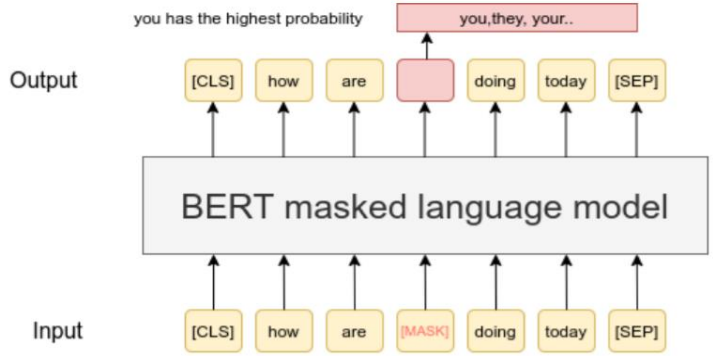


Figure 2. Masked input sentence for an NLP pre-training task.

* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: https://arxiv.org/pdf/1810.04805

Figure 2: https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

Vision vs NLP

- The nature of inputs are different in terms of information density.
- In NLP, signals (words) are highly semantic and rich in information.

Figure 1. Masked images of METU CENG's cat "Java" at masking ratios of 50%, 75%, and 90% (from left to right).

Imagine masking high ratios in NLP

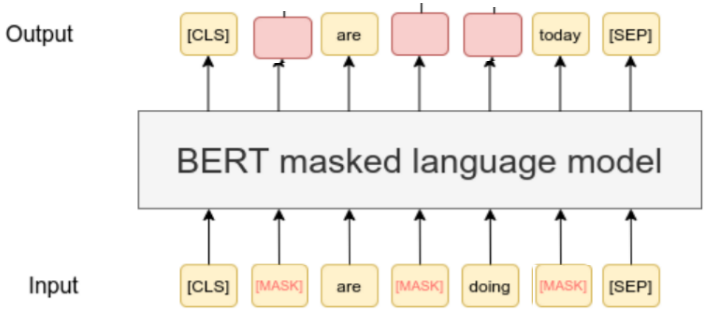


Figure 2. Masked input sentence for an NLP pre-training task. (60% masked) *(edited)*

* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: https://arxiv.org/pdf/1810.04805

Figure 2 (edited): https://www.sbert.net/examples/unsupervised_learning/MLM/README.html

The paper's main goal

- Design a method to learn visual representations.
- Utilize unlabeled datasets for this process.
- Ensure scalable performance.
- Ultimately, provide a backbone for other tasks through transfer learning.

2– Related Work

Masked Language Modeling

- Masking some of the input, learning to predict the masked content.
- In NLP tasks, masked language modeling works well on pre-training.
- They provide high scalability**

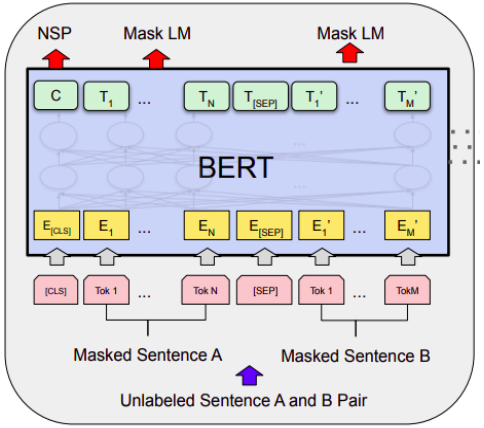


Figure 3. BERT* pre-training

* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/pdf/1810.04805

** Language Models are Few-Shot Learners: https://arxiv.org/pdf/2005.14165

(Denoising) Autoencoders

- Autoencoders learn representations by mapping input to a latent space, and attempting to reconstruct it using this latent representation.

- Denoising autoencoders (DAE) (*) are autoencoders that corrupt the input and learn to reconstruct it.

- Masking some of the input can be regarded as a form of corruption. Hence, MAE is a type of DAE, albeit with differences

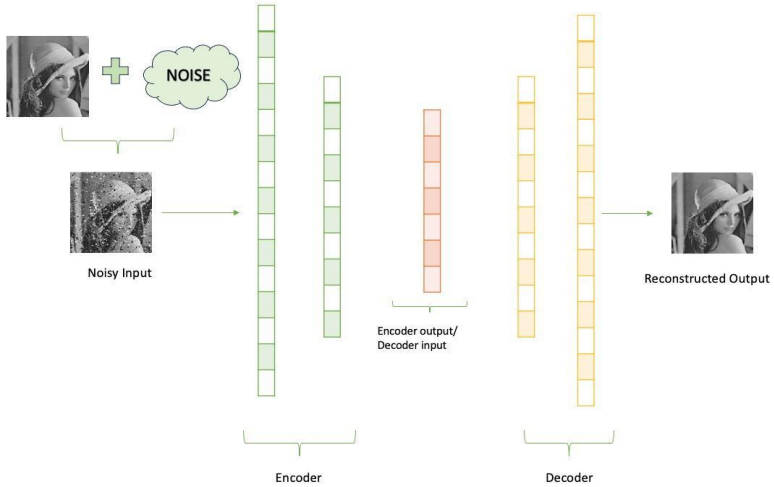


Figure 4. DAE example

Figure 4: https://www.geeksforgeeks.org/denoising-autoencoders-in-machine-learning

*Extracting and composing robust features with denoising autoencoders: https://www.cs.toronto.edu/~larocheh/publications/icml-2008-denoising-autoencoders.pdf

Masked image encoding

- Pixel masking is used as a noise to learn better representations in DAEs.(*)

- The original ViT paper (**) uses masked patch prediction as a pre-training, method, replacing approximately 45% of patches with mask embeddings or any random patch embedding.

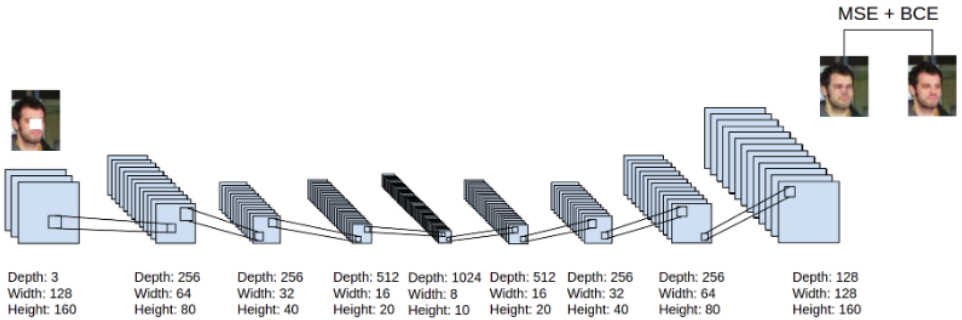


Figure 5. example of a DAE using pixel masking as a noise

*: Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion: https://www.jmlr.org/papers/volume11/vincent10a/vincent10a.pdf

**: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale: https://arxiv.org/pdf/2010.11929

METU

Figure 5: https://faculty.cc.gatech.edu/~hays/7476/projects/Avery_Wenchen/

Self-supervised learning

- Self-supervised learning aims to learn representations without using labels.

- For pre-training, various self-supervised techniques have been used, such as contrastive learning (MoCo v3) (*), which compares two crops.

*: An Empirical Study of Training Self-Supervised Vision Transformers: https://arxiv.org/pdf/2104.02057

3- Approach and Method

Overall Architecture

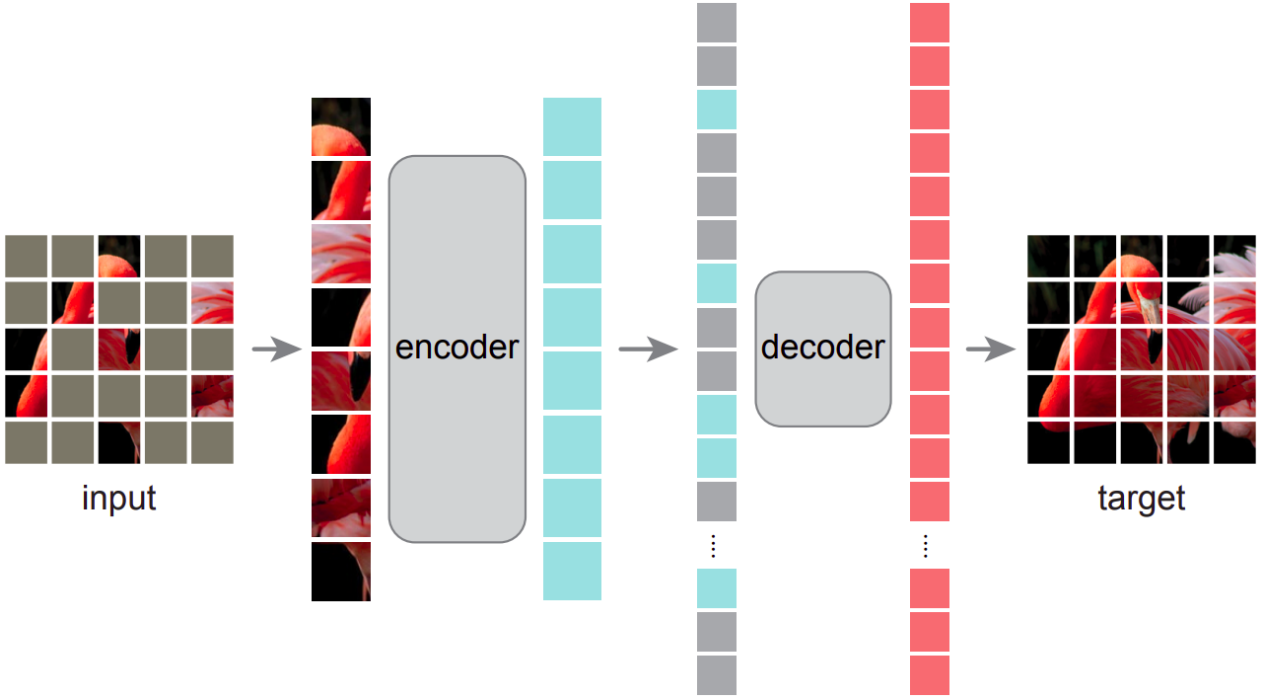


Figure 6. MAE architecture from the paper

Figure 6 (from the presented paper):
https://arxiv.org/pdf/2111.06377

Masking Method

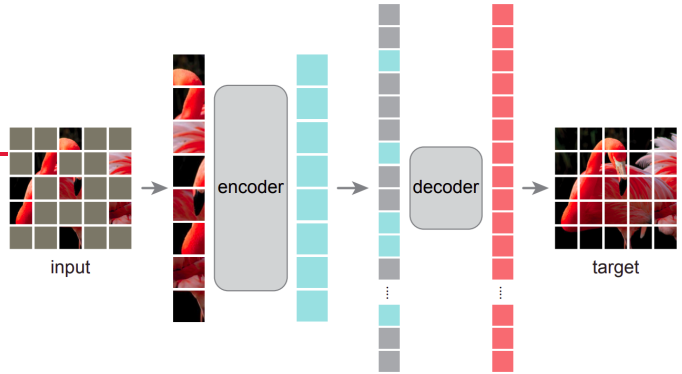


Figure 6. MAE architecture from the paper

- Similar to the original ViT (*), they divide the input to non-overlapping patches.
- They uniformly sample a very large proportion (75%) of these patches and mask them, which is a departure from previous methods.
- They do not use mask tokens in masking operation, differing from the previous work.
- Masking a high ratio makes predicting task challenging and meaningful.
- Also, not using mask tokens and picking a high sampling ratio enables the encoder to be very large.

Figure 6 (from the presented paper): https://arxiv.org/pdf/2111.06377

* *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale:* https://arxiv.org/pdf/2010.11929

Mae Encoder

- As stated, the encoder only operates on unmasked patches (25%).
- It embeds the patches by linear projection and add positional embeddings.
- It uses transformer blocks.
- Since the input patches is a small part of the input, the encoder is chosen to be very large. (encoder has over 9 times computations per token vs decoder)

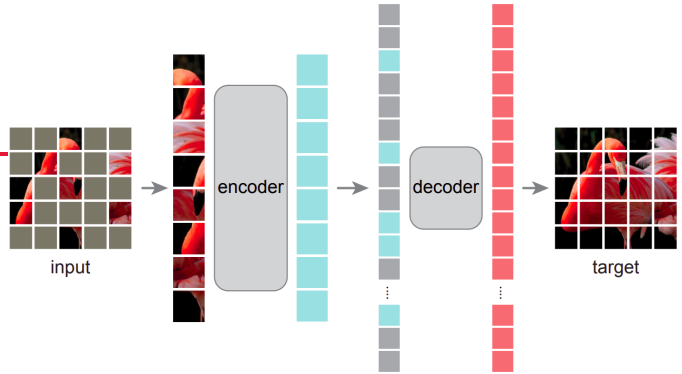


Figure 6. MAE architecture from the paper

Figure 6 (from the presented paper): https://arxiv.org/pdf/2111.06377

Mae Decoder

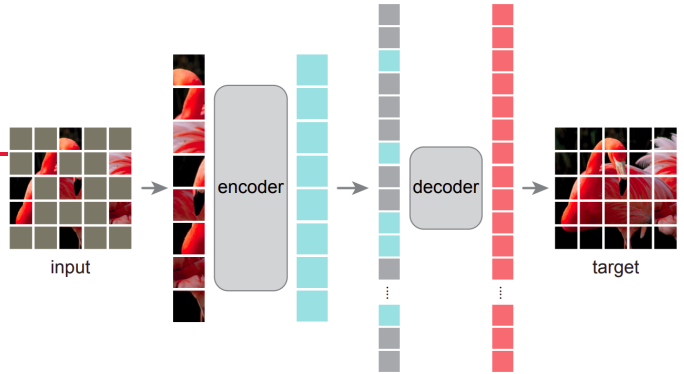


Figure 6. MAE architecture from the paper

- Merges encoder outputs with the shared mask tokens in previously masked places, adding positional encodings to them.
- It also uses transformer blocks, followed by a linear projection for finalizing pixel reconstruction.
- It is solely responsible for reconstruction, meaning it is not used post-training. Hence, it is independent of the encoder design, making it flexible.
- Experiments made with light-weight decoders, significantly reducing the training time. Also, this choice implies an asymmetrical autoencoder design.
(This is different than the previous work)

Figure 6 (from the presented paper): https://arxiv.org/pdf/2111.06377

Reconstruction Target

- Basically, the MSE (mean squared error loss) in the pixel space between input image and reconstructed image is adopted.

- Loss is only computed for masked patches, following the BERT (*)

- They also tried to normalize all patches with their mean and std. Using normalized pixels as the reconstruction target enhanced the representation quality.

* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding: https://arxiv.org/pdf/1810.04805

4-
Pre-training Experiments

Baseline Model and Dataset

- ViT (*) Large model is used as a baseline. (Very big and normally prone to overfitting.)
- Pre-training is made with ImageNet-1K (IN1K) (**) dataset.
- Baseline MAE outperforms the trained from scratch ViT (*) model, even with a good training technique.

scratch, original	scratch, our impl.	baseline MAE
76.5	82.5	84.9

Figure 7. top-1 validation accuracy on IN1K

* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale: https://arxiv.org/pdf/2010.11929

** ImageNet: A Large-Scale Hierarchical Image Database: https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf

Figure 7 (from the presented paper): https://arxiv.org/pdf/2111.06377

Linear Probing vs Full Fine-tuning

- In linear probing, the pre-trained model is fixed, and only one linear layer is added at the end, to predict the labels (or produce the output). This method is used to assess the quality of representations from a pre-trained feature extraction model. (*)
- In fine-tuning, pre-trained model is further trained (not fixed), and one or more layers, possibly with non-linearities are added.
- Hence, linear probing provides a measure of representation quality of a pre-training in restricted conditions, while fine-tuning exploits models near-true potential to adopt for new tasks.
- In linear probing, since only 1 linear layer is trained, is fast to train and less computationally expensive.

* Self-Supervised Learning of Pretext-Invariant Representations: https://arxiv.org/pdf/1912.01991

Ablation Experiments

8 topics are considered in ablation studies:

- ➔Masking ratio
- ➔Decoder depth
- ➔Decoder width
- ➔Mask token (used or not in encoder)
- ➔Reconstruction target
- ➔Data augmentation
- ➔Mask sampling method
- ➔Training schedule

Masking Ratio

- High masking ratios (≈75%) works well for both cases, considering the information density difference with NLP.
- Choosing high masking ratios increases the speed in both training and inference time since encoders computations

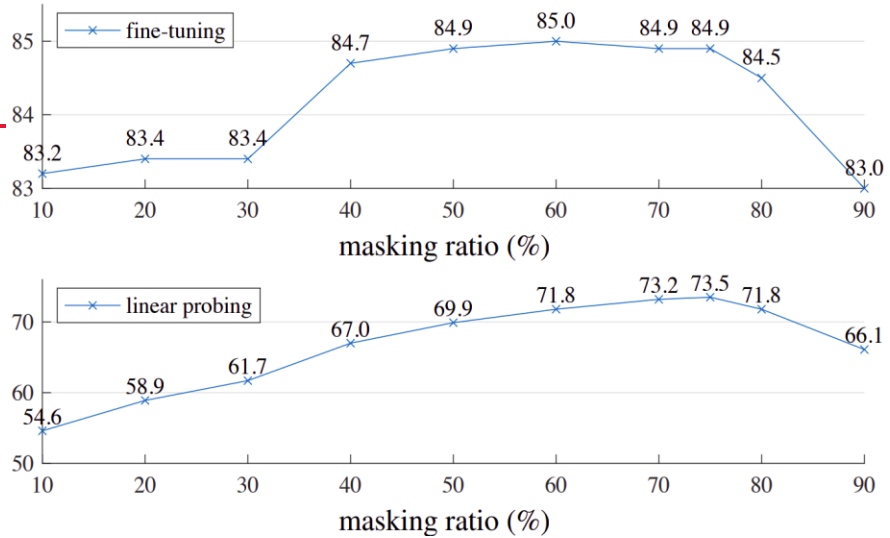


Figure 8. Effect of masking ratio on fine-tuning and linear probing top-1 accuracies on IN1K

- From 10% to 90%, all masking ratios produce better results when compared to trained from scratch (82.5% accuracy)

Figure 8 (from the presented paper): https://arxiv.org/pdf/2111.06377

Decoder Depth

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

Figure 9. Effect of decoder depth on fine-tuning and linear probing top-1 accuracies on IN1K

- Using a decoder with an 8-block depth is effective for both fine-tuning and linear probing.
- Decoder depth is not critical for fine-tuning.
- However, decoder performance deteriorates with shallow decoders. If the decoder is not deep enough, then the encoders output will possibly be less abstract and more focused on reconstruction task.
- Only 1 decoder block is enough for fine-tuning task. This can speed up the training even more if needed!

Figure 9 (from the presented paper): https://arxiv.org/pdf/2111.06377

Decoder Width

- Choosing a decoder width of 512 is a viable option for both fine-tuning and linear probing.
- Narrower decoders also yield strong performance in fine-tuning tasks.
- These depth and width choices makes the decoder lightweight, which has 9% FLOPs per token vs ViT-L (24 blocks, 1024-d)
- From now on, 8-blocks 512-width decoder is used if not specified.

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

Figure 10. Effect of decoder width on fine-tuning and linear probing top-1 accuracies on IN1K

Figure 10 from the presented paper):
https://arxiv.org/pdf/2111.06377

Mask Token (for encoder)

- Not using mask tokens on encoder is one of the key-features of the paper.

case	ft	lin	FLOPs
encoder w/ [M]	84.2	59.6	3.3×
encoder w/o [M]	**84.9**	**73.5**	**1**×

Figure 11. Effect of using mask token on encoder on fine-tuning and linear probing top-1 accuracies on IN1K

- It reduces overall FLOPs by 3.3x which significantly reduces the training time. (using 1-block decoder further boosts the speed!)
- Moreover, accuracy considerably drops, especially in linear probing when mask tokens are used.
- Also, memory usage is reduced which makes training larger models possible or increase batch-size for even faster trainings.

encoder	dec. depth	ft acc	hours	speedup
ViT-L, w/ [M]	8	84.2	42.4	-
ViT-L	8	84.9	15.4	2.8×
ViT-L	1	84.8	11.6	**3.7**×
ViT-H, w/ [M]	8	-	119.6†	-
ViT-H	8	85.8	34.5	3.5×
ViT-H	1	85.9	29.3	**4.1**×

Figure 12. Effects of not using mask tokens and decoder depth as training times and fine-tuning top-1 accuracies on IN1K

Figure 11 and 12 (from the presented paper): https://arxiv.org/pdf/2111.06377

Reconstruction Target

case	ft	lin
pixel (w/o norm)	84.9	73.5
pixel (w/ norm)	**85.4**	**73.9**
PCA	84.6	72.3
dVAE token	85.3	71.6

Figure 13. Effect of reconstruction target on fine-tuning and linear probing top-1 accuracies on IN1K

- Using pixel MSE loss with normalization improves accuracy.
- In PCA, they perform PCA and use largest 96 coefficients as the target.
- They also compared the target in BEiT (*), using dVAE (**) as the tokenizer, where the MAE decoder is predicts token indices using cross-entropy loss.
- The BEiT(*) target has no advantages to normalized pixel MSE loss. Moreover, it is more complex and slower since dVAE (**) tokenizer is large (40% of ViT-L).

Figure 13 (from the presented paper): https://arxiv.org/pdf/2111.06377

* BEIT: BERT Pre-Training of Image Transformers: https://arxiv.org/pdf/2106.08254

** Zero-Shot Text-to-Image Generation: https://arxiv.org/pdf/2102.12092

Data Augmentation

- The method is robust to not using augmentations, which is significantly differs from the contrastive learning that heavily relies on augmentation.

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

Figure 14. Effect of data augmentation on fine-tuning and linear probing top-1 accuracies on IN1K

- Using random size (or even fix size) cropping increases the accuracies.

- Adding color jittering drops the accuracy.

Figure 14 (from the presented paper): https://arxiv.org/pdf/2111.06377

Mask Sampling Method

- Works best with the random sampling.
- Block sampling is better as 50% ratio, but still, it is worse.
- Grid sampling makes the task easier with a lower training loss and better reconstruction. However, the representation quality is worse.

case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

Figure 15. Effect of decoder depth on fine-tuning and linear probing top-1 accuracies on IN1K

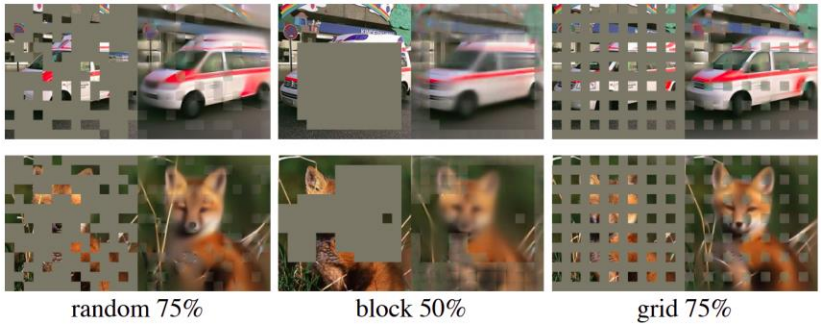


Figure 16. Sampling method visualizations

Figure 15 and 16 (from the presented paper): https://arxiv.org/pdf/2111.06377

Training Schedule

- No saturation is observed even after 1600 epochs.
- This trend demonstrates that the method is highly scalable for extended training periods.

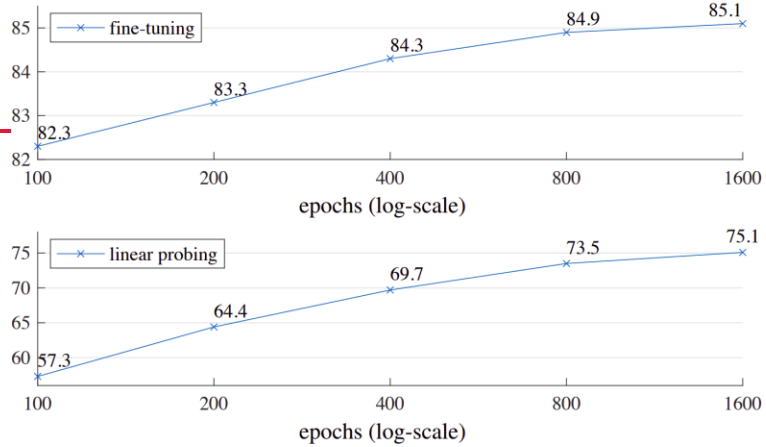


Figure 17. Effect of training schedule on fine-tuning and linear probing top-1 accuracies on IN1K (each point is a full training)

- To compare, MoCo v3 (*) saturates around 300 epochs.
- MAE only sees 25% of the patches, where contrastive learning methods see 200% (two-crop) or even more.

Comparison with Previous Results

- Even when trained with robust techniques, ViT-L exhibits worse performance and not lacks the scalability in a supervised setting.
- MAE can easily scale up.
- SOTA (87.8%) on methods which only uses IN1K dataset. Previous best (*) is 82.1%.

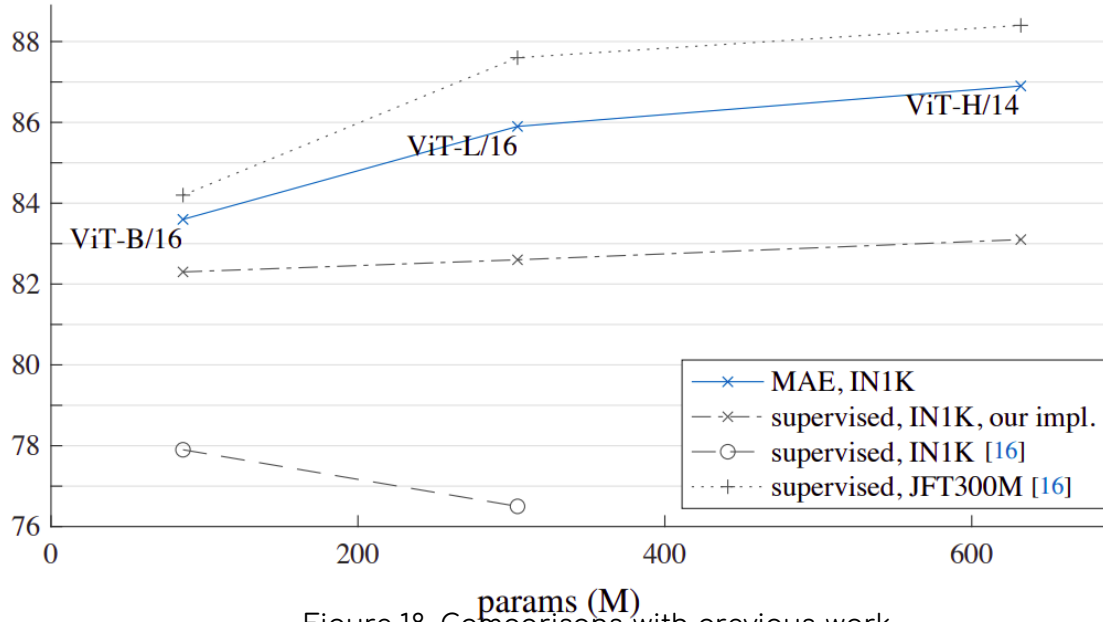


Figure 18. Comparisons with previous work on accuracy and scalability

- Supervised pre-training with JFT300M dataset is more accurate but regarding the pre-training dataset size, results are comparable. Also, trends are similar.

Figure 18 (from the presented paper): https://arxiv.org/pdf/2111.06377

METU

*VOLO: Vision Outlooker for Visual Recognition: https://arxiv.org/pdf/2106.13112

Comparison with BEIT (*) and MoCo v3 (**)

- MAE is more accurate, faster and simpler.
- MAE reconstructs pixels, whereas BEIT (*) predict tokens.
- MAE is 3.5 times faster per epoch than BEIT.

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

Figure 19. Comparisons of MAE with BEIT and MoCo v3

- MAE result in the figure is for 1600 epoch pre-training. Still, pre-training time is less than other methods. MAE and MoCo v3 (**) has the training durations of 31 hours and 36 hours, respectively.

Figure 19 (from the presented paper): https://arxiv.org/pdf/2111.06377

METU

** BEIT: BERT Pre-Training of Image Transformers: https://arxiv.org/pdf/2106.08254*

*** An Empirical Study of Training Self-Supervised Vision Transformers: https://arxiv.org/pdf/2104.02057*

Partial fine-tuning

- It is clear that linear probing and fine-tuning results are uncorrelated.
- Linear probing cannot measure how well the representations can be used with non-linear combinations, which is important. The linear separability of the features is not the only thing for measuring the representation quality.
- Also, linear probing is not highly correlated with transfer learning (*), which is the ultimate goal of pre-training.
- So, authors also used partial fine-tuning, where some of the last layers are trained while others stay fixed.

* *Exploring Simple Siamese Representation Learning:* https://arxiv.org/pdf/2011.10566

MoCo v3 (*) vs MAE with Partial Fine-Tuning

- As it can be seen, MoCo v3 (*) performs better on linear probing. However, when allowing the models some last layers to relearn, there is a significant accuracy improvement in the favor of MAE.
- This means MAE features are less linearly separable than MoCo v3 (*), but they are still stronger representations when a non-linear head will be tuned.

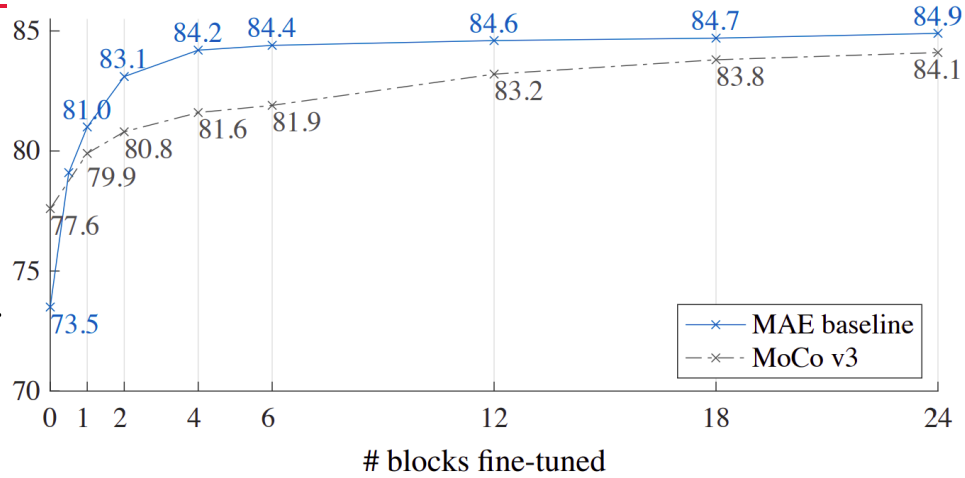


Figure 20. Comparisons of MAE with MoCo v3 on partial fine-tuning

Figure 20 (from the presented paper): https://arxiv.org/pdf/2111.06377

* *An Empirical Study of Training Self-Supervised Vision Transformers:* https://arxiv.org/pdf/2104.02057

5-
Transfer Learning Experiments

Object Detection and Segmentation

- The authors fine-tune the Mask R-CNN (*) on the COCO (**) dataset for object detection and segmentation tasks.

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Figure 21. Comparisons of MAE with BEIT and MoCo v3 on COCO detection and segmentation.

- When using only IN1K dataset for pre-training, MAE surpasses all other supervised and self-supervised pre-training techniques.

Figure 21 (from the presented paper): https://arxiv.org/pdf/2111.06377

** Mask R-CNN: https://arxiv.org/pdf/1405.0312*

*** Microsoft COCO: Common Objects in Context: https://arxiv.org/pdf/1405.0312*

Semantic Segmentation

- The authors used UperNet (*) for the semantic segmentation on the ADE20K (**) dataset.

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

Figure 22. Comparisons of MAE with BEIT and MoCo v3 on iNat and Places semantic segmentation task.

- Results demonstrate that MAE outperforms all supervised and self-supervised pre-training techniques once again. These results are consistent with the former object detection and segmentation task.

Figure 22 (from the presented paper): https://arxiv.org/pdf/2111.06377

METU

** Unified Perceptual Parsing for Scene Understanding: https://arxiv.org/pdf/1807.10221*

*** Semantic Understanding of Scenes through the ADE20K Dataset: https://arxiv.org/pdf/1608.05442*

Classification

- The authors used the iNaturalist (*) and Places (**) tasks for comparison.

dataset	ViT-B	ViT-L	ViT-H	ViT-H_{448}	prev best
iNat 2017	70.5	75.7	79.3	**83.4**	75.4 [55]
iNat 2018	75.4	80.1	83.0	**86.8**	81.2 [54]
iNat 2019	80.5	83.4	85.7	**88.3**	84.1 [54]
Places205	63.9	65.8	65.9	**66.8**	66.0 [19]†
Places365	57.9	59.4	59.8	**60.3**	58.0 [40]‡

Figure 23. Comparisons of MAE with BEIT and MoCo v3 on iNat and Places classification task.

- Especially on iNat (*), MAE has a very large gap with the previous SOTA results.
- On Places (**) task, previous SOTA were obtained with pre-training on billions of images.

- Also, MAE shows significant scaling behavior.

Figure 23 (from the presented paper): https://arxiv.org/pdf/2111.06377

METU

** The iNaturalist Species Classification and Detection Dataset: https://arxiv.org/pdf/1707.06642*

*** Learning Deep Features for Scene Recognition using Places Database: https://papers.nips.cc/paper_files/paper/2014/file/3fe94a002317b5f9259f82690aeea4cd-Paper.pdf*

Pixels vs Tokens

- In transfer learning, using dVAE(*) tokens as the target is yields better results then the unnormalized pixel target.

	IN1K			COCO		ADE20K	
	ViT-B	ViT-L	ViT-H	ViT-B	ViT-L	ViT-B	ViT-L
pixel (w/o norm)	83.3	85.1	86.2	49.5	52.8	48.0	51.8
pixel (w/ norm)	83.6	85.9	86.9	50.3	53.3	48.1	53.6
dVAE token	83.6	85.7	86.9	50.3	53.2	48.1	53.4
△	0.0	-0.2	0.0	0.0	-0.1	0.0	-0.2

Figure 24. Effect of using pixels vs tokens as target

- However, per-patch normalization with a pixel target can be applied, rendering the use of dVAE (*) tokens redundant.

Figure 24 (from the presented paper): https://arxiv.org/pdf/2111.06377

** Zero-Shot Text-to-Image Generation:* https://arxiv.org/pdf/2102.12092

6-
Discussion
And
Conclusion

Summary

- LLM pre-training methods are promising for vision domain and they can be adopted regarding the information difference.
- Masking a high proportion in vision works well.
- Encoder not uses mask tokens. Thus, it can be very large.
- Decoder can be lightweight, further speeding up the training.
- Linear probing alone is not enough to measure feature quality.

* *Exploring Simple Siamese Representation Learning:* https://arxiv.org/pdf/2011.10566

References (1 of 4)

Bao, H., Dong, L., & Wei, F. (2021). BEiT: BERT Pre-Training of Image Transformers. *ArXiv:2106.08254 [Cs]*. https://arxiv.org/abs/2106.08254

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Hesse, C. (2020). Language Models are Few-Shot Learners. *Arxiv.org*. https://arxiv.org/abs/2005.14165

Chen, X., & He, K. (2020). Exploring Simple Siamese Representation Learning. *ArXiv:2011.10566 [Cs]*. https://arxiv.org/abs/2011.10566

Chen, X., Xie, S., & He, K. (2021, August 16). *An Empirical Study of Training Self-Supervised Vision Transformers*. ArXiv.org. https://doi.org/10.48550/arXiv.2104.02057

Chen, X., Xie, S., & He, K. (2021, August 16). *An Empirical Study of Training Self-Supervised Vision Transformers*. ArXiv.org. https://doi.org/10.48550/arXiv.2104.02057

Devlin, J., Chang, M.-W., Lee, K., Google, K., & Language, A. (n.d.). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://arxiv.org/pdf/1810.04805

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv:2010.11929 [Cs]*.

References (2 of 4)

Fei-Fei, L., Deng, J., & Li, K. (2010). ImageNet: Constructing a large-scale image database. *Journal of Vision*, *9*(8), 1037–1037. https://doi.org/10.1167/9.8.1037

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). Masked Autoencoders Are Scalable Vision Learners. *ArXiv:2111.06377 [Cs]*. https://arxiv.org/abs/2111.06377

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Lawrence, Z. C., & Dollár, P. (2014). *Microsoft COCO: Common Objects in Context*. ArXiv.org. https://arxiv.org/abs/1405.0312

Misra, I., & van der Maaten, L. (2019). Self-Supervised Learning of Pretext-Invariant Representations. *ArXiv:1912.01991 [Cs]*. https://arxiv.org/abs/1912.01991

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (n.d.). *Zero-Shot Text-to-Image Generation*. https://arxiv.org/pdf/2102.12092

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The iNaturalist Species Classification and Detection Dataset. *ArXiv:1707.06642 [Cs]*. https://arxiv.org/abs/1707.06642

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). *Extracting and Composing Robust Features with Denoising Autoencoders*. https://www.cs.toronto.edu/~larocheh/publications/icml-2008-denoising-autoencoders.pdf

References (3 of 4)

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., & Sun, J. (2018, July 26). *Unified Perceptual Parsing for Scene Understanding*. ArXiv.org. https://doi.org/10.48550/arXiv.1807.10221

Yuan, L., Hou, Q., Jiang, Z., Feng, J., & Yan, S. (2022). *VOLO: Vision Outlooker for Visual Recognition*. 1–13. https://doi.org/10.1109/tpami.2022.3206108

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (n.d.). *Learning Deep Features for Scene Recognition using Places Database*. https://papers.nips.cc/paper_files/paper/2014/file/3fe94a002317b5f9259f82690aeea4cd-Paper.pdf

References for figures (4 of 4)

- Figure 1:
  From local experiments with the model from: https://github.com/facebookresearch/mae
- Figure 2:
  https://www.sbert.net/examples/unsupervised_learning/MLM/README.html
- Figure 3:
  https://arxiv.org/pdf/1810.04805
- Figure 4:
  https://www.geeksforgeeks.org/denoising-autoencoders-in-machine-learning/
- Figure 5:
  https://faculty.cc.gatech.edu/~hays/7476/projects/Avery_Wenchen/
- Figure 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24:
  From the presented paper, Masked Autoencoders Are Scalable Vision Learners: https://arxiv.org/abs/2111.06377

Attribution

Keep this page and the following notice AS IS.

This presentation has been prepared with METU Presentation Template by Devrim Çavuşoğlu licensed under the CC BY-SA 4.0. Also read, the full LICENSE content.

Refer to github.com/devrimcavusoglu/metu-presentation-template

Thank you for listening
Q & A