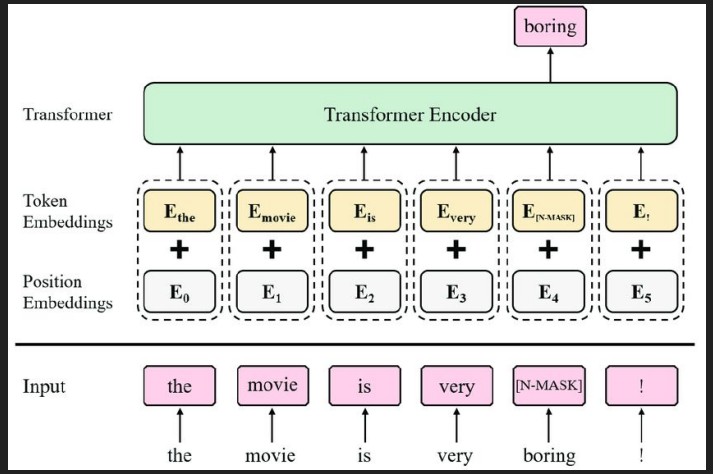
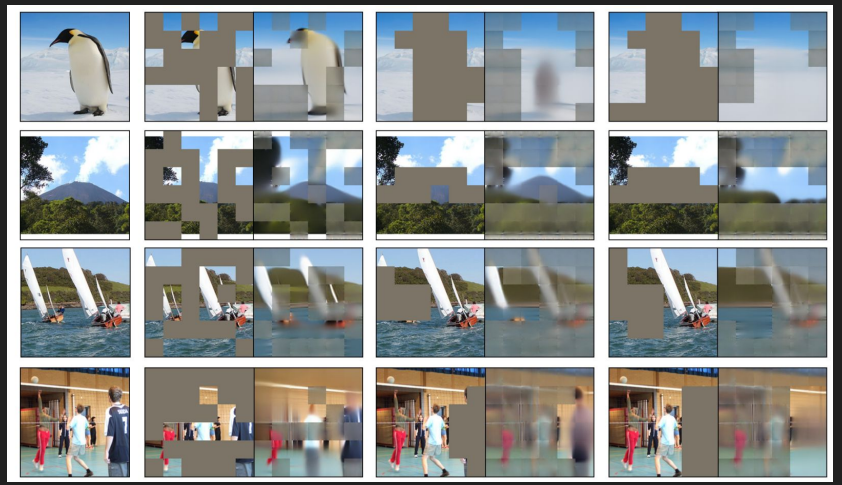
SimMIM: a Simple Framework for Masked Image Modeling

- Finding what is important with masked image modeling
- Similar to BERT style pretraining but for images
  - Fill in the blank

Masked language modeling



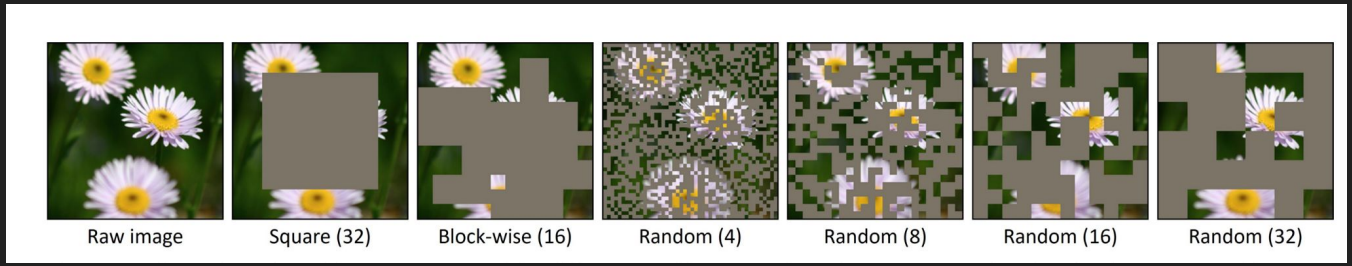
Masked image modeling



Primary Findings

- Directly learning to regress all masked pixels can very powerful
- Not necessary to do more complicated setups with patch classification
- Not necessary to use a heavy decoder, single linear layer is fine
- Need a much higher masking rate than in language
    - in language models 15% is usually a good baseline, in vision models they found it useful to go all the way up to 80%
- Quality of prediction isnt really that important
    - Can train a larger decoder head at a higher resolution that yields better inpainting but doesnt transfer to new tasks any better

Results



Mask Type	Masked patch size	Mask ratio	Top-1 acc (%)
square	32	0.11 (2×2)	82.6
	32	0.25 (3×3)	82.5
	32	0.44 (4×4)	82.5
block-wise	16/32	0.4	82.7/82.7
	16/32	0.6	82.6/82.6
	16/32	0.8	82.4/82.5
random	4/8/16/32	0.4	81.9/82.0/82.4/82.9
	4/8/16/32	0.6	82.0/82.1/82.7/82.8
	4/8/16/32	0.8	82.1/82.4/82.8/82.4
	64	0.1	82.6
	64	0.2	82.6
random	32	0.1	82.7
	32	0.2	82.8
	32	0.3	82.8
	32	0.4	82.9
	32	0.5	**83.0**
	32	0.6	82.8
	32	0.7	82.7
	32	0.8	82.4
	32	0.9	82.4

Head	#params	Training costs	Top-1 acc (%)
Linear	89.9M	1×	82.8
2-layer MLP	90.9M	1.2×	82.8
inverse Swin-T	115.2M	1.7×	82.4
inverse Swin-B	174.8M	2.3×	82.5

Table 2. Ablation on different prediction heads. A simple linear layer performs the best with lower training costs.

Results

Prediction method doesn't really matter and you can predict tiny images and get similar results

Image size (ratio of inputs)	6^2 (1/32)	12^2 (1/16)	24^2 (1/8)	48^2 (1/4)	96^2 (1/2)	192^2 (1/1)
Top-1 acc (%)	82.3	82.7	82.8	82.7	82.8	82.8

Table 3. Ablation on different prediction resolutions. A moderately large resolution (no less than $1/16$ all perform well.

Scope to predict	Top-1 acc (%)
masked area	82.8
full image	81.7

Loss	Pred. Resolution	Top-1 acc (%)
Classification		
8-bin	192^2	82.7
8-bin	48^2	82.7
256-bin	192^2	N/A
256-bin	48^2	82.3
iGPT cluster	192^2	N/A
iGPT cluster	48^2	82.4
BEiT	-	82.7
Regression		
ℓ_2	192^2	82.7
smooth-ℓ_1	192^2	82.7
ℓ_1	192^2	82.8
ℓ_1	48^2	82.7
ℓ_1	6^2	82.3