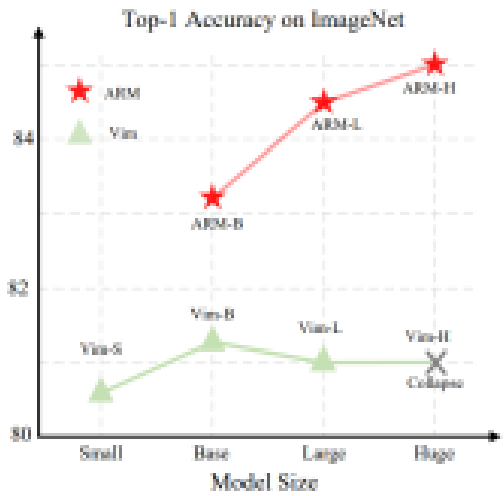
Autoregressive Pretraining with Mamba in Vision

Sucheng Ren[1] Xianhang Li[2] Haoqin Tu[2] Feng Wang[1] Fangxun Shu[3] Lei Zhang[3]
Jieru Mei[1] Linjie Yang[4] Peng Wang[4] Heng Wang[4] Alan Yuille[1] Cihang Xie[1]
[1] Johns Hopkins University [2]UC Santa Cruz [3]Alibaba Group [4]ByteDance

Scaling Mamba in Vision



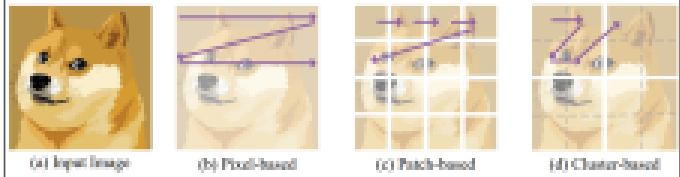
(1) attempts to scale the Vision Mamba (Vim) under supervised conditions often lead to either performance plateauing or even training collapse when pushed to very large sizes

(2) we primarily focus on the autoregressive pretraining paradigm for self-supervised visual representation learning

Autoregressive Pretraining with Mamba in Vision

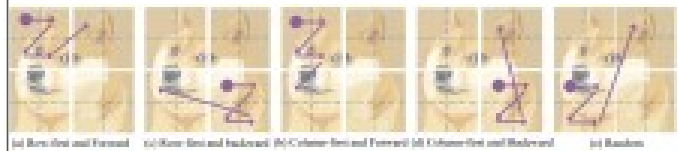
We propose ARM and study various designs in autoregressive modeling:

Prediction Unit



we propose grouping spatially adjacent patches into larger clusters to serve as the prediction unit

Order



Unlike the 1D sentences in NLP, which inherently have a clear sequence order. We explore four primary prediction orders:
1) Row-first and forward orders the clusters row by row processing from the first to the last cluster within each row sequentially
2) Row-first and backward orders the clusters row by row but inverts the processing direction.
3) Column-first and forward organizes the clusters column by column, processing sequentially within each column from top to bottom
4) Column-first and backward similarly sequences the clusters column by column but inverts the processing direction

Experiments

We conduct experiments on ImageNet-1K classification.

Model	Token Mixer	Image Size	Params. (M)	Throughput (image/s)	Top-1 (%)
Base-size models					
RegNetY-16G	3D Conv.	224^2	84	670	82.9
DeiT-B	Attention	224^2	21	1073	81.2
Vim-B	Mamba	224^2	98	958	81.2
MambaMLP-B	Mamba	224^2	85	1381	83.2
ViTamin-B	Mamba+3D Conv.	224^2	101	535	82.9
ARM-B	Mamba	224^2	85	1381	83.2
ARM-B	Mamba	384^2	85	449	84.2
ARM-B ‡	Mamba	448^2	85	96	84.9
Large-size models					
Vim-M	Mamba	224^2	340	345	81.9
MambaMLP	Mamba	224^2	207	445	81.4
ARM-L	Mamba	224^2	207	445	84.5
ARM-L	Mamba	384^2	207	154	85.1
Huge-size models					
Vim-H	Mamba	224^2	755	216	collapsed
ARM-H	Mamba	224^2	662	275	85.0
ARM-H	Mamba	384^2	662	94	85.5

(1) We are the first scaling Mamba to huge size.

Ablation Study

Num of Prediction unit	Cluster size	Top-1 (%)
0 (Supervised)	N/A	81.2
144 (GPT)	1×1 (Pixel)	79.8
4	96×96	82.0
9	64×64	82.5
16	48×48	82.2
36	32×32	81.9
144	16×16	81.7

(1) Number of prediction units: the number of the prediction units set to 9.

Order	Direction	Top-1 (%)
Raw-first	Forward	82.5
Raw-first	Backward	82.3
Column-first	Forward	82.5
Column-first	Backward	82.4
Random	Random	81.5

(2) Prediction order: no significant difference with predefined order.