PlainMamba

Improving Non-Hierarchical Mamba in Visual Recognition

Chenhongyi Yang*[1], Zehui Chen*[2], Miguel Espinosa*[1,4], Linus Ericsson[1], Zhenyu Wang[3], Jiaming Liu[3], Elliot J. Crowley[1]

[1]University of Edinburgh, [2]University of Science and Technology of China, [3]Peking University, [4]SENSE CDT

BMVC 2024

github.com/ChenhongyiYang/PlainMamba

*We present **PlainMamba**: a simple non-hierarchical state space model (SSM) designed for **general** visual recognition.*

Why

- Transformers are expensive: quadratic cost of attention.
- State-Space Models (SSM) provide long context lengths with linear complexity to input sequence length.
  - Recently, Mamba architecture was able to scale to sizes and performances of modern transformer-based LLMs
- There is now a need for adapting Mamba into the visual domain

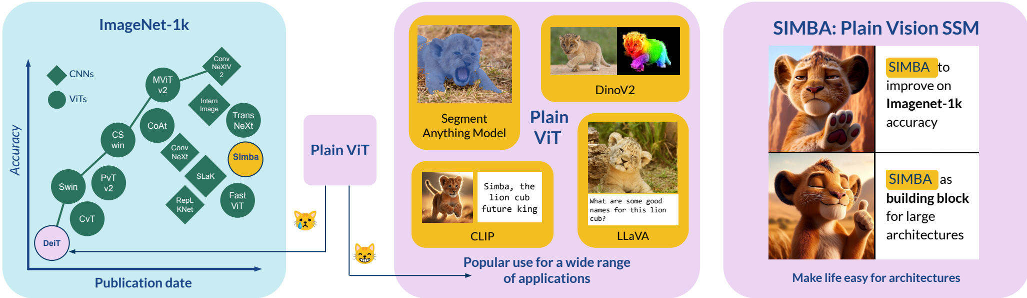

Figure 1: Although hierarchical encoders demonstrate superior accuracy, the **plain non-hierarchical models have had more widespread use because of their simple structure.** We investigate the potential of the plain Mamba model in visual recognition.

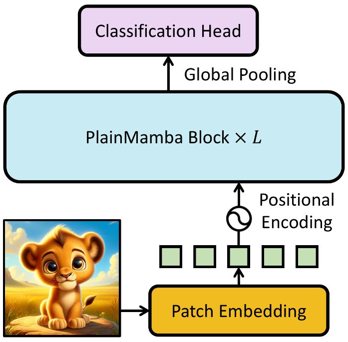

Figure 2a. Architecture of PlainMamba

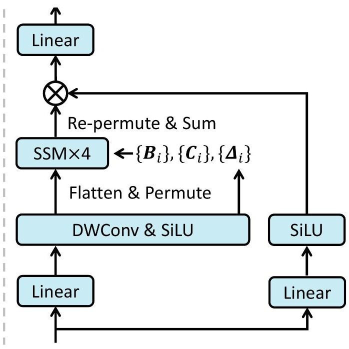

Figure 2b. PlainMamba Block Architecture

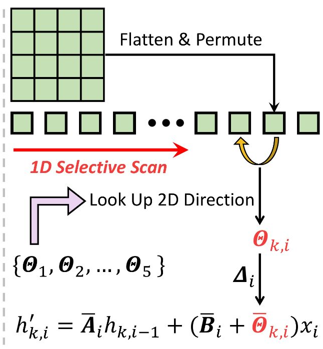

Figure 2c. Direction-Aware Updating

Method

Architecture

- PlainMamba has a **non-hierarchical** design:
  - A convolution tokenizer
  - A stack ***L* identical PlainMamba blocks**
  - A task-specific head for downstream tasks
- The model maintains:
  - **constant width** throughout the layers
  - **constant feature resolution**

  making it **easy to reuse** and **easy to scale**.

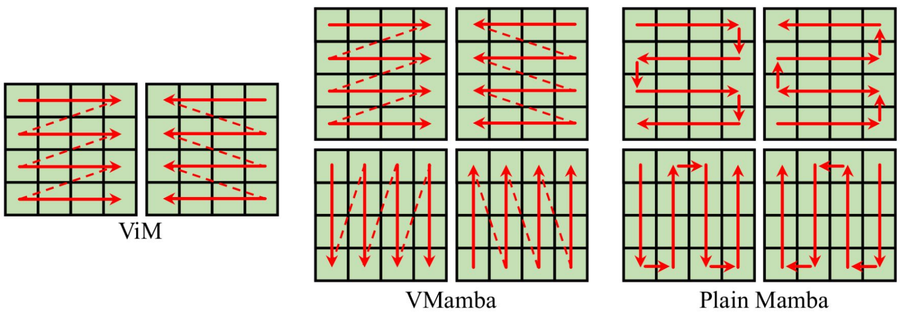

Figure 3. Comparison between our **Continuous 2D Scanning** and the selective scan orders in ViM and VMamba

PlainMamba Block

- We introduce the **PlainMamba** block:
  - **Continuous 2D Scanning process**
    - to improve spatial continuity by ensuring adjacency of tokens
  - **Direction-Aware Updating**
    - to discern spatial relations of tokens by encoding directional information
  - Removal of special tokens (CLS)

Results

- We evaluate PlainMamba on downstream tasks:classification, semantic segmentation, object detection
- We provide three PlainMamba model variants for different parameter sizes: 7M, 25M, 50M.
- We observe efficiency gains (FLOPs, Mb) for high-resolution inputs compared to vision transformers.

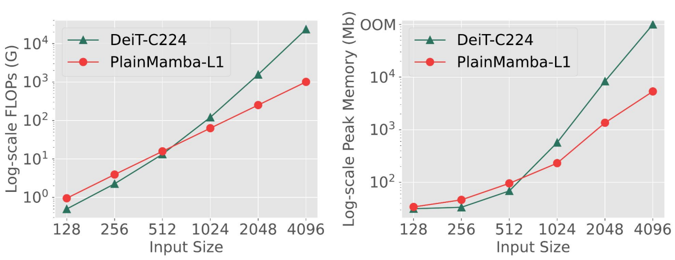

Figure 4. **Efficiency comparison** between PlainMamba and DeiT

Backbone	Hierarchical	Params	FLOPs	mIoU
ViM-T [110]	✗	13M	-	41.0
ViM-S [110]	✗	46M	-	44.9
LocalVim-T [45]	✗	36M	181G	43.4
LocalVim-S [45]	✗	58M	297G	46.4
VMamba-T [59]	✓	55M	964G	47.3
VMamba-S [59]	✓	76M	1081G	49.5
PlainMamba-L1	✗	35M	174G	44.1
PlainMamba-L2	✗	55M	285G	46.8
PlainMamba-L3	✗	81M	419G	49.1

Figure 5. Comparison between PlainMamba and SSMs on **ADE20K Semantic Segmentation**

Model	Hierarchical	Params	FLOPs	Top-1
Transformer				
DeiT-Tiny [82]	✗	5M	1.3G	72.2
DeiT-Small [82]	✗	22M	4.6G	79.9
DeiT-Base [82]	✗	86M	16.8G	81.8
Swin-Tiny [60]	✓	29M	4.5G	81.3
Swin-Small [60]	✓	50M	8.7G	83.0
PVT-Tiny [88]	✓	13M	2G	75.1
PVT-Small [88]	✓	25M	3.8G	79.8
PVT-Medium [88]	✓	44M	6.7G	81.2
Focal-Tiny [103]	✓	29M	4.9G	82.2
Focal-Small [103]	✓	51M	9.1G	83.5
State Space Modeling				
ViM-T [110]	✗	7M	-	76.1
ViM-S [110]	✗	26M	-	80.5
LocalViM-T [45]	✗	8M	1.5G	76.2
LocalViM-S [45]	✗	28M	4.8G	81.2
Mamba-ND-T [49]	✗	24M	-	79.2
Mamba-ND-S [49]	✗	63M	-	79.4
S4ND-ViT-B [64]	✗	89M	-	80.4
S4ND-ConvNeXt-T [64]	✓	30M	-	82.2
VMamba-T [59]	✓	22M	5.6G	*82.2
VMamba-S [59]	✓	44M	11.2G	*83.5
PlainMamba-L1	✗	7M	3.0G	77.9
PlainMamba-L2	✗	25M	8.1G	81.6
PlainMamba-L3	✗	50M	14.4G	82.3

Figure 5. Comparison between PlainMamba and SSMs on **ImageNet-1K** (*denotes best epoch result)

Centre for Satellite Data in Environmental Science

Acknowledgements

Funding for this research is provided in part by a studentship from the School of Engineering at the University of Edinburgh, the SENSE - Centre for Satellite Data in Environmental Science CDT, and an EPSRC New Investigator Award (EP/X020703/1).