VideoMamba: State Space Model for Efficient Video Understanding

Appendix

https://github.com/OpenGVLab/VideoMamba

A More Results

In Table I, we present additional results on the Kinetics-400 dataset. These results clearly demonstrate that our SSM-based model outperforms all previous attention-based methods. We observe consistent performance improvements with increasing resolution and frame count.

B More Implementation Details

B.1 Training Details

We sparsely sample frames from the raw videos as in TSN [82] for all the datasets. Table II details the masked pretraining hyperparameters. For the unmasked multi-modality pretraining, we load the pretrained model and train it for an additional epoch with a learning rate of 8e-5. Moreover, Tables III, IV, V, and VI show the training details for the different datasets used for fine-tuning.

B.2 Dataset Descriptions

We show the statistics of multi-modality datasets in Table VII, and single-modality datasets in Table VIII.

C Discussions

C.1 Technical Contributions

To the best of our knowledge, VideoMamba is the first purely SSM-based model for video understanding. Similar to the success of TimeSformer, the first purely attention-based model, we aim to offer a simple yet powerful Mamba architecture that paves the way for future advancements in long-video understanding. Applying Mamba to video understanding presents unique challenges, which we have thoroughly investigated, including various spatiotemporal scanning methods, masking strategies, and text integration. For example, while tube masking

Table I: More results on scene-related Kinetics-400. "*iso.*" means isotropic architecture without downsampling layers.

Arch.	Model	*iso.*	Extra Data	Input Size	#Param (M)	FLOPs (G)	K400 Top-1	K400 Top-5
Supervised: *Those models with extra data are under supervised training.*								
Trans.	STAM [64]	✓	IN-21K	64×224^2	121	1040×1×1	79.2	-
	TimeSformer-L [4]	✓	IN-21K	96×224^2	121	2380×3×1	80.7	94.7
	ViViT-L [2]	✓	IN-21K	16×224^2	311	3992×3×4	81.3	94.7
	Mformer-HR [60]	✓	IN-21K	16×336^2	311	959×3×10	81.1	95.2
SSM	VideoMamba-Ti	✓	IN-1K	8×224^2	7	9×3×4	76.9	92.9
	VideoMamba-Ti	✓	IN-1K	16×224^2	7	17×3×4	78.1	93.5
	VideoMamba-Ti	✓	IN-1K	32×224^2	7	34×3×4	78.8	93.9
	VideoMamba-Ti	✓	IN-1K	64×224^2	7	69×3×4	79.6	94.2
	VideoMamba-Ti	✓	IN-1K	64×384^2	7	202×3×4	80.3	94.8
	VideoMamba-S	✓	IN-1K	8×224^2	26	34×3×4	79.3	94.2
	VideoMamba-S	✓	IN-1K	16×224^2	26	68×3×4	80.8	94.8
	VideoMamba-S	✓	IN-1K	32×224^2	26	135×3×4	81.5	95.2
	VideoMamba-S	✓	IN-1K	64×224^2	26	271×3×4	81.8	95.3
	VideoMamba-S	✓	IN-1K	64×384^2	26	395×3×4	82.7	95.6
	VideoMamba-M	✓	IN-1K	8×224^2	74	101×3×4	80.6	94.6
	VideoMamba-M	✓	IN-1K	16×224^2	74	202×3×4	81.9	95.4
	VideoMamba-M	✓	IN-1K	32×224^2	74	403×3×4	82.4	95.7
	VideoMamba-M	✓	IN-1K	64×224^2	74	806×3×4	82.8	96.0
	VideoMamba-M	✓	IN-1K	64×384^2	74	2368×3×4	**83.3**	**96.1**
Self-supervised: *For UMT, the CLIP-400M is used in pretrained teacher.*								
Trans.	ST-MAE-B_{1600e} [19]	✓		16×224^2	87	180×3×7	81.3	94.9
	VideoMAE-S_{2400e} [75]	✓		16×224^2	22	57×3×5	79.0	93.8
	VideoMAE-B_{1600e} [75]	✓		16×224^2	87	180×3×5	81.5	95.1
	UMT-B_{800e} [43]	✓	CLIP-400M	8×224^2	87	180×3×5	**85.7**	**97.0**
SSM	VideoMamba-M_{800e}	✓	CLIP-400M	8×224^2	74	101×3×4	82.0	95.4
	VideoMamba-M_{800e}	✓	CLIP-400M	16×224^2	74	202×3×4	83.4	95.9
	VideoMamba-M_{800e}	✓	CLIP-400M	32×224^2	74	403×3×4	83.9	96.2
	VideoMamba-M_{800e}	✓	CLIP-400M	64×224^2	74	806×3×4	84.3	96.6
	VideoMamba-M_{800e}	✓	CLIP-400M	64×384^2	74	2368×3×4	85.0	96.9

is effective in VideoMAE, it performs poorly with VideoMamba. Instead, our row masking and attention masking strategies prove more suitable.

Our extensive experiments highlight VideoMamba's four core strengths. Compared to TimeSformer, VideoMamba shows significant improvements, achieving **+2.6%** and **+5.9%** on K400 and SthSthV2, respectively, with similar computational requirements. To foster further research, we have made all the code and models publicly available.

C.2 Potential Overfitting Issues

As a global operator, SSM converges faster than convolution, requiring a higher DropPath rate than attention (*e.g.*, 0.5 for VideoMamba-M and 0.1 for ViT-B). Additionally, Mamba architectures tend to be deeper, such as VideoMamba-M with 32 layers compared to ViT-B's 12 layers, posing optimization challenges for larger models. However, our self-distillation technique enables effective training of larger models. VideoMamba-B (∼100M parameters) and VideoMamba-L

Table II: Masked pre-training settings. I-image, **V**-video, **T**-text.

config	Single-Modality		Multi-Modality
	SthSthV2	**K400**	**5M & 17M & 25M**
optimizer	*AdamW*		*AdamW*
optimizer momentum	β_1, β_2=0.9, 0.95		β_1, β_2=0.9, 0.999
weight decay	0.05		0.05
learning rate schedule	*cosine decay*		*cosine decay*
learning rate	1.2e-3		4e-4
minimal learning rate	1e-5		4e-6
batch size	2048		2048 **I**, 2048 **V**
warmup epochs	40		1
total epochs	800		10
mask ratio	80%		50% **I**, 80% **V**, 50% **T**
input frame	8		8
drop path	0.4		0.25
flip augmentation	*no*	*yes*	*yes*
augmentation	*MultiScaleCrop*[0.66, 0.75, 0.875, 1]		*MultiScaleCrop*[0.5, 1]

Table III: Training settings for ImageNet-1K.

config	224×224	448×448	512×512
optimizer	*AdamW*		
optimizer momentum	β_1, β_2=0.9, 0.999		
weight decay	0.1(Ti), 0.05(S,M)	1e-8	1e-8
learning rate schedule	*cosine decay*		
base learning rate	5e-4	5e-6	5e-6
minimal learning rate	1e-5	5e-6	5e-6
base batch size	512		
repeated augmentation	*no*(Ti), *yes*(S,M)		
warmup epochs	5(Ti,S), 30(M)	5	2
total epochs	300	30	10
drop path	0(Ti), 0.15(S), 0.5(M)		
label smoothing	0.1		
cutmix	1.0		
augmentation	*RandAug*(7, 0.25)(Ti), *RandAug*(9, 0.5)(S,M)		

(∼300M parameters) achieve superior performance (**83.0%** and **83.9%**) compared to VideoMamba-S (82.8%), demonstrating its scalability.

Interestingly, training VideoMamba from scratch with masked modeling shows normal convergence, as shown in Table 3 and 4. We hypothesize that masking increases task difficulty, which helps mitigate potential overfitting issues.

Table IV: Training settings for Kinetics-400. "†" means masked pretraining.

config	224×224	384×384
optimizer	*AdamW*	
optimizer momentum	β_1, β_2=0.9, 0.999	
weight decay	0.1(Ti), 0.05(S,M,M†)	1e-8
learning rate schedule	*cosine decay*	
base learning rate	4e-4(Ti,S), 2e-4(M), 1e-4(M†)	5e-6
minimal learning rate	1e-6	
base batch size	256	
repeated augmentation	2	
warmup epochs	5	2
total epochs	70(Ti), 50(S,M), 45(M†)	10
drop path	0.1(Ti), 0.35(S), 0.8(M), 0.4(M†)	
layer-wise lr decay	0.75(S,M,M†), 0.8(M†)	
flip augmentation	*yes*	
label smoothing	0.1	
cutmix	1.0	
augmentation	*RandAug*(7, 0.25)(Ti), *RandAug*(9, 0.5)(S,M,M†)	

Table V: Training settings for SthSthV2. "†" means masked pretraining.

config	224×224	288×288
optimizer	*AdamW*	
optimizer momentum	β_1, β_2=0.9, 0.999	
weight decay	0.1(Ti), 0.05(S,M,M†)	1e-8
learning rate schedule	*cosine decay*	
base learning rate	4e-4(Ti,S,M) 1e-4(M†)	5e-6
minimal learning rate	1e-6	
base batch size	256	
repeated augmentation	2	
warmup epochs	5	2
total epochs	35(Ti), 30(S,M,M†)	10
drop path	0.1(Ti), 0.35(S), 0.8(M), 0.4(M†)	
layer-wise lr decay	0.75(S,M,M†), 0.8(M†)	
flip augmentation	*no*	
label smoothing	0.1	
cutmix	1.0	
augmentation	*RandAug*(7, 0.25)(Ti), *RandAug*(9, 0.5)(S,M,M†)	

Table VI: Training settings for Breaskfast, COIN and LVU. "†" means masked pretraining. We directly sample the frames from the raw video sparsely.

config	BreakFast & LVU	COIN
optimizer	*AdamW*	
optimizer momentum	β_1, β_2=0.9, 0.999	
weight decay	0.1(Ti), 0.05(S,M,M†)	
learning rate schedule	*cosine decay*	
base learning rate	2e-4	
minimal learning rate	1e-6	
base batch size	256	
repeated augmentation	2	
warmup epochs	5	
total epochs	70(Ti), 50(S,M), 45(M†)	40(Ti), 35(S), 30(M,M†)
drop path	0.1(Ti), 0.35(S), 0.8(M), 0.4(M†)	
layer-wise lr decay	0.75(S,M,M†), 0.8(M†)	
flip augmentation	*yes*	
label smoothing	0.1	
cutmix	1.0	
augmentation	*RandAug*(7, 0.25)(Ti), *RandAug*(9, 0.5)(S,M,M†)	

Table VII: Statistics of multi-modality datasets.

Dataset	#image/video	#text	Type
COCO	113K	567K	image
Visual Genome	100K	768K	image
SBU Captions	860K	860K	image
CC3M	2.88M	2.88M	image
CC12M	11.00M	11.00M	image
WebVid-2M	2.49M	2.49M	video
WebVid-10M	10.73M	10.73M	video
5M corpus = CC3M+WebVid-2M	5.37M	5.37M	video+image
17M corpus = 5M+COCO+VG+SBU+CC12M	17.44M	18.57M	video+image
25M corpus = 17M+WebVid-10M−WebVid-2M	25.68M	26.81M	video+image

Table VIII: Statistics of single-modality datasets.

Dataset	#video			#text			Avg Video
	Train	Val	Test	Train	Val	Test	Length (s)
Image Classification							
ImageNet-1K	1,281,167	50,000	100,000	-	-	-	-
Short-term Action Recognition							
Kinetics-400	240,436	19,787	-	-	-	-	10
Something-Something V2	168,913	24,777	-	-	-	-	4
Long-term Action Recognition							
Breakfast	1,577	-	410	-	-	-	137
COIN	9,026	-	2,796	-	-	-	142
LVU	7,619	1,666	1,551	-	-	-	134
Relation	138	49	41	-	-	-	127
Speak	871	196	188	-	-	-	133
Scene	514	107	81	-	-	-	132
Director	680	163	107	-	-	-	137
Genre	2807	569	584	-	-	-	130
Writer	748	174	168	-	-	-	142
Year	725	163	141	-	-	-	133
Like	658	142	139	-	-	-	159
View	478	103	102	-	-	-	112
Video-Text Retrieval							
MSRVTT	7,010	-	1,000	140,200	-	1,000	15
DiDeMo	8,496	1,094	1,036	8,496	1,094	1,036	29
ActivityNet	10,009	4,917	-	10,009	4,917	-	180
LSMDC	101,055	-	1,000	101,055	-	1,000	5
MSVD	1,200	100	670	1,200	100	670	15