NEURAL INFORMATION PROCESSING SYSTEMS

VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

Zhan Tong[1,2] Yibing Song[2] Jue Wang[2] Limin Wang[1,3]

[1] State Key Laboratory for Novel Software Technology, Nanjing University
[2] Tencent AI Lab [3] Shanghai AI Lab

南京大学 NANJING UNIVERSITY

MCG MULTIMEDIA COMPUTING GROUP 媒体计算研究组

Tencent AI Lab 腾讯人工智能实验室

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

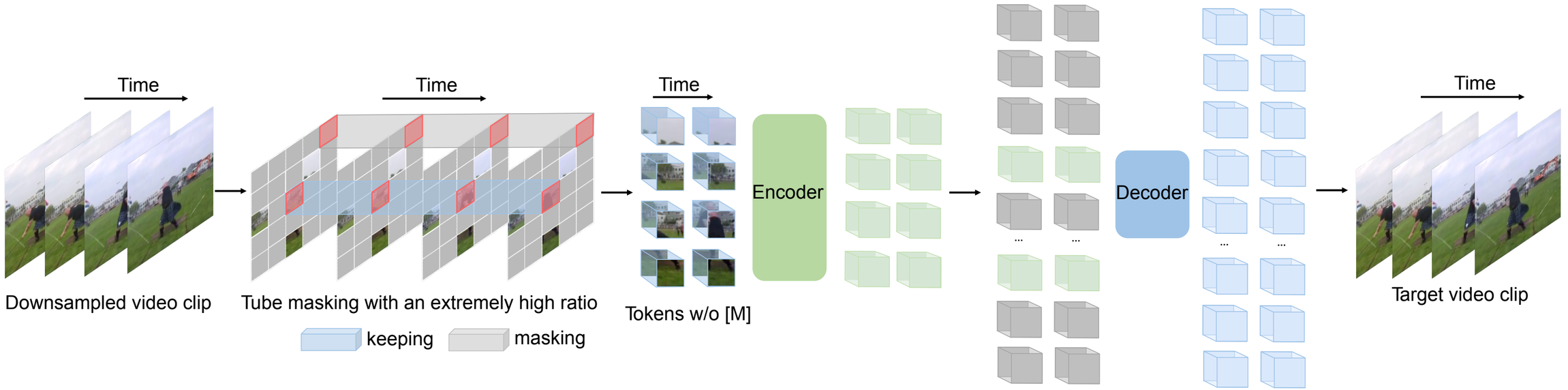
Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

How to efficiently train a vanilla ViT on the video dataset itself without any pre-trained model or extra data?

VideoMAE

- Our VideoMAE attempts to solve it in two aspects
  - Self-supervised pre-training with masked autoencoder
  - A new masking strategy: tube masking with an extremely high ratio



Overall VideoMAE

NEURAL INFORMATION PROCESSING SYSTEMS

⇥ and eventually, VideoMAE is

- a **simple**, **data-efficient** method for **self-supervised video pre-training**
  - with **high** performance and **no** extra data **required**

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

⇥ Leading performance on Something-Something V2

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
TEINet$_{En}$ [39]	ResNet50$_{\times 2}$	ImageNet-1K	✓	8+16	99×10×3	50	66.5	N/A
TANet$_{En}$ [40]	ResNet50$_{\times 2}$		✓	8+16	99×2×3	51	66.0	90.1
TDN$_{En}$ [74]	ResNet101$_{\times 2}$		✓	8+16	198×1×3	88	69.6	92.2
SlowFast [22]	ResNet101	Kinetics-400	✓	8+32	106×1×3	53	63.1	87.6
MViTv1 [21]	MViTv1-B		✓	64	455×1×3	37	67.7	90.9
TimeSformer [6]	ViT-B	ImageNet-21K	✓	8	196×1×3	121	59.5	N/A
TimeSformer [6]	ViT-L		✓	64	5549×1×3	430	62.4	N/A
ViViT FE [3]	ViT-L	IN-21K+K400	✓	32	995×4×3	N/A	65.9	89.9
Motionformer [50]	ViT-B		✓	16	370×1×3	109	66.5	90.1
Motionformer [50]	ViT-L		✓	32	1185×1×3	382	68.1	91.2
Video Swin [38]	Swin-B		✓	32	321×1×3	88	69.6	92.7
VIMPAC [64]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	68.1	N/A
BEVT [76]	Swin-B	IN-1K+K400+DALLE	✗	32	321×1×3	88	70.6	N/A
MaskFeat↑$_{312}$ [79]	MViT-L	Kinetics-600	✓	40	2828×1×3	218	75.0	95.0
VideoMAE	ViT-B	Kinetics-400	✗	16	180×2×3	87	69.7	92.3
VideoMAE	ViT-L	Kinetics-400	✗	16	597×2×3	305	74.0	94.6
VideoMAE	ViT-S	*no external data*	✗	16	57×2×3	22	66.8	90.3
VideoMAE	ViT-B		✗	16	180×2×3	87	70.8	92.4
VideoMAE	ViT-L		✗	16	597×2×3	305	74.3	94.6
VideoMAE	ViT-L		✗	32	1436×1×3	305	**75.4**	**95.2**

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

⇥ Leading performance on Kinetics-400

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
NL I3D [77]	ResNet101	ImageNet-1K	✓	128	359×10×3	62	77.3	93.3
TANet [40]	ResNet152		✓	16	242×4×3	59	79.3	94.1
TDN$_{En}$ [74]	ResNet101		✓	8+16	198×10×3	88	79.4	94.4
TimeSformer [6]	ViT-L	ImageNet-21K	✓	96	8353×1×3	430	80.7	94.7
ViViT FE [3]	ViT-L		✓	128	3980×1×3	N/A	81.7	93.8
Motionformer [50]	ViT-L		✓	32	1185×10×3	382	80.2	94.8
Video Swin [38]	Swin-L		✓	32	604×4×3	197	83.1	95.9
ViViT FE [3]	ViT-L	JFT-300M	✓	128	3980×1×3	N/A	83.5	94.3
ViViT [3]	ViT-H	JFT-300M	✓	32	3981×4×3	N/A	84.9	95.8
VIMPAC [64]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	77.4	N/A
BEVT [76]	Swin-B	IN-1K+DALLE	✗	32	282×4×3	88	80.6	N/A
MaskFeat↑352 [79]	MViT-L	Kinetics-600	✗	40	3790×4×3	218	87.0	97.4
ip-CSN [68]	ResNet152	*no external data*	✗	32	109×10×3	33	77.8	92.8
SlowFast [22]	R101+NL		✗	16+64	234×10×3	60	79.8	93.9
MViTv1 [21]	MViTv1-B		✗	32	170×5×1	37	80.2	94.4
MaskFeat [79]	MViT-L		✗	16	377×10×1	218	84.3	96.3
VideoMAE	ViT-S	*no external data*	✗	16	57×5×3	22	79.0	93.8
VideoMAE	ViT-B		✗	16	180×5×3	87	81.5	95.1
VideoMAE	ViT-L		✗	16	597×5×3	305	85.2	96.8
VideoMAE	ViT-H		✗	16	1192×5×3	633	**86.6**	**97.1**
VideoMAE↑320	ViT-L	*no external data*	✗	32	3958×4×3	305	86.1	97.3
VideoMAE↑320	ViT-H		✗	32	7397×4×3	633	**87.4**	**97.6**

Experiments

→ **Leading performance on AVA v2.2**

Method	Backbone	Pre-train Dataset	Extra Labels	$T \times \tau$	GFLOPs	Param	mAP
supervised [22]	SlowFast-R101	Kinetics-400	✓	8×8	138	53	23.8
CVRL [53]	SlowOnly-R50	Kinetics-400	✗	32×2	42	32	16.3
ρBYOL$_{\rho=3}$ [23]	SlowOnly-R50	Kinetics-400	✗	8×8	42	32	23.4
ρMoCo$_{\rho=3}$ [23]	SlowOnly-R50	Kinetics-400	✗	8×8	42	32	20.3
MaskFeat$\uparrow$312 [79]	MViT-L	Kinetics-400	✓	40×3	2828	218	37.5
MaskFeat$\uparrow$312 [79]	MViT-L	Kinetics-600	✓	40×3	2828	218	38.8
VideoMAE	ViT-S	Kinetics-400	✗	16×4	57	22	22.5
VideoMAE	ViT-S	Kinetics-400	✓	16×4	57	22	28.4
VideoMAE	ViT-B	Kinetics-400	✗	16×4	180	87	26.7
VideoMAE	ViT-B	Kinetics-400	✓	16×4	180	87	31.8
VideoMAE	ViT-L	Kinetics-400	✗	16×4	597	305	34.3
VideoMAE	ViT-L	Kinetics-400	✓	16×4	597	305	37.0
VideoMAE	ViT-H	Kinetics-400	✗	16×4	1192	633	**36.5**
VideoMAE	ViT-H	Kinetics-400	✓	16×4	1192	633	**39.5**
VideoMAE	ViT-L	Kinetics-700	✗	16×4	597	305	**36.1**
VideoMAE	ViT-L	Kinetics-700	✓	16×4	597	305	**39.3**

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

⇥ Leading performance on UCF101 and HMDB51

Method	Backbone	Extra data	Frames	Param	Modality	UCF101	HMDB51
OPN [35]	VGG	UCF101	N/A	N/A	V	59.6	23.8
VCOP [82]	R(2+1)D	UCF101	N/A	N/A	V	72.4	30.9
CoCLR [29]	S3D-G	UCF101	32	9M	V	81.4	52.1
Vi^{2}CLR [18]	S3D	UCF101	32	9M	V	82.8	52.9
VideoMAE	ViT-B	*no external data*	16	87M	V	**91.3**	**62.6**
SpeedNet [5]	S3D-G	Kinetics-400	64	9M	V	81.1	48.8
VTHCL [84]	SlowOnly-R50	Kinetics-400	8	32M	V	82.1	49.2
Pace [73]	R(2+1)D	Kinetics-400	16	15M	V	77.1	36.6
MemDPC [28]	R-2D3D	Kinetics-400	40	32M	V	86.1	54.5
CoCLR [29]	S3D-G	Kinetics-400	32	9M	V	87.9	54.6
RSPNet [12]	S3D-G	Kinetics-400	64	9M	V	93.7	64.7
VideoMoCo [45]	R(2+1)D	Kinetics-400	16	15M	V	78.7	49.2
Vi^{2}CLR [18]	S3D	Kinetics-400	32	9M	V	89.1	55.7
CVRL [53]	SlowOnly-R50	Kinetics-400	32	32M	V	92.9	67.9
CVRL [53]	SlowOnly-R50	Kinetics-600	32	32M	V	93.6	69.4
CVRL [53]	Slow-R152 (2×)	Kinetics-600	32	328M	V	94.4	70.6
CORP$_f$ [32]	SlowOnly-R50	Kinetics-400	32	32M	V	93.5	68.0
ρSimCLR$_{\rho=2}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	88.9	N/A
ρSwAV$_{\rho=2}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	87.3	N/A
ρMoCo$_{\rho=2}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	91.0	N/A
ρBYOL$_{\rho=2}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	92.7	N/A
ρBYOL$_{\rho=4}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	94.2	72.1
VideoMAE(Ours)	ViT-B	Kinetics-400	16	87M	V	**96.1**	**73.3**

VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

Code is available at
https://github.com/MCG-NJU/VideoMAE

NEURAL INFORMATION PROCESSING SYSTEMS