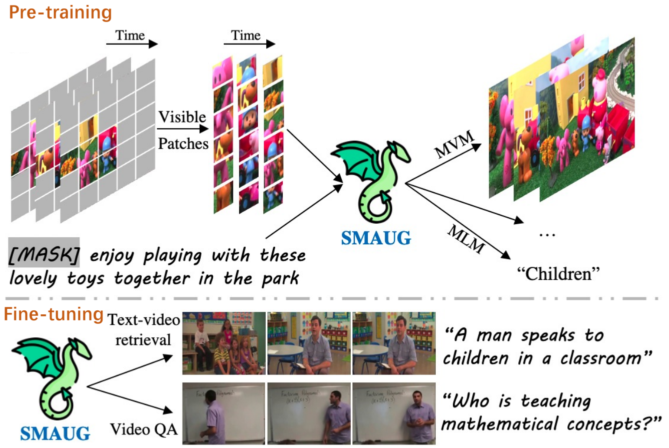
SMAUG: Sparse Masked Autoencoder for Efficient Video-Language Pre-training

Yuanze Lin[1], Chen Wei[2], Huiyu Wang[2], Alan Yuille[2], Cihang Xie[3]

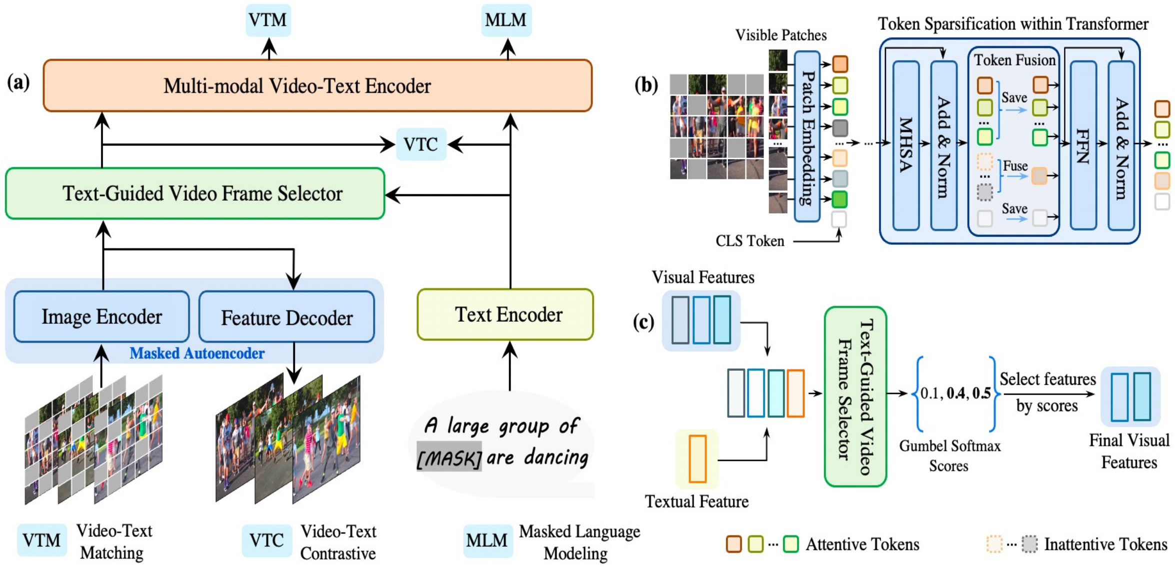
[1]University of Washington, [2]Johns Hopkins University , [3]UC Santa Cruz

Motivation

- Previous methods require **heavy computation** for pre-training.
- Can we **guarantee performance** while **significantly reducing costs**?
- **Masked Autoencoders (MAE)** can offer a decent solution.


Pre-training
Time
Time
Visible Patches
SMAUG
MVM
MLM
...
[MASK] enjoy playing with these lovely toys together in the park
“Children”
Fine-tuning
Text-video retrieval
SMAUG
Video QA
“A man speaks to children in a classroom”
“Who is teaching mathematical concepts?”

Our proposed SMAUG



(a) SMAUG method: adopt MAE to extract features and reconstruct original pixels.

(b) Token sparsification: reduce spatial redundancies for visible patches.

(c) Frame selection: take visual and textual features as inputs and outputs the selected frames by the scores.

(d) Pre-training objectives: $\mathcal{L} = \mathcal{L}_{vtm} + \mathcal{L}_{mlm} + \mathcal{L}_{vtc} + \mathcal{L}_{mvm}.$

Experiments

Results on text-to-video retrieval task

Method	PT Datasets	#Frame	MSRVTT R@1	MSRVTT R@5	MSRVTT R@10	DiDeMo R@1	DiDeMo R@5	DiDeMo R@10	ActivityNet Cap R@1	ActivityNet Cap R@5	ActivityNet Cap R@10
Pre-trained with >100M video-text pairs											
HT100M [39]	HT100M	16	14.9	40.2	52.8	-	-	-	-	-	-
HERO [29]	HT100M	310	20.5	47.6	60.9	-	-	-	-	-	-
MMT [14]	HT100M	1K/-/3K	26.6	57.1	69.6	-	-	-	28.7	61.4	94.5
AVLNet [44]	HT100M	-	27.1	55.6	66.6	-	-	-	-	-	-
SupportSet [41]	HT100M	-	30.1	58.5	69.3	-	-	-	-	-	-
VideoCLIP [60]	HT100M	960	30.9	55.4	66.8	-	-	-	-	-	-
VIOLET [12]	YT180M+5M	4	34.5	63.0	73.4	32.6	62.8	74.7	-	-	-
All-in-one [53]	HT100M+WebVid	9	34.4	65.4	75.8	32.7	61.4	73.5	22.4	53.7	67.7
Pre-trained with <100M video-text pairs											
ClipBERT [27]	COCO + VG	16/16/8	22.0	46.8	59.9	20.4	48.0	60.8	21.3	49.0	63.5
Frozen [2]	5M	4	31.0	59.5	70.5	31.0	59.8	72.4	-	-	-
ALPRO [28]	5M	8	33.9	60.7	73.2	35.9	67.5	78.8	-	-	-
Singularity [26]	5M	1	36.8	65.9	75.5	47.4	75.2	84.0	43.0	70.6	81.3
Singularity [26]	17M	1	41.5	68.7	77.0	53.9	79.4	86.9	47.1	75.5	85.5
Ours	**5M**	**1**	**40.6**	**67.6**	**77.5**	**49.2**	**76.7**	**85.6**	**44.8**	**72.2**	**82.7**
Ours	**17M**	**1**	**44.0**	**70.4**	**78.8**	**55.6**	**80.8**	**88.4**	**49.2**	**76.9**	**86.8**

Influences of different components

Masking Ratio	PT Time	MSRVTT R@1	MSRVTT R@5	MSRVTT R@10
0%	74.9 hours	**40.7**	**66.6**	**76.7**
10%	70.8 hours	40.5	66.3	76.5
25%	64.6 hours	40.1	65.7	76.0
50%	50.4 hours	39.3	65.4	75.6
65%	**44.3 hours**	38.1	64.2	75.0

Keeping Rate	PT Time	MSRVTT R@1	MSRVTT R@5	MSRVTT R@10
0.6	**44.8 hours**	37.5	64.2	74.6
0.7	47.6 hours	38.2	64.7	75.3
0.8	50.4 hours	39.3	65.4	75.6
0.9	53.8 hours	**39.8**	**65.9**	**76.0**

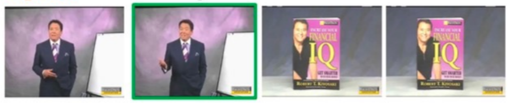
Frame Selection	PT Time	MSRVTT R@1	MSRVTT R@5	MSRVTT R@10
Single-frame selection				
1	**50.4 hours**	39.3	65.4	75.6
4→1	75.3 hours	**40.6**	**67.6**	**77.5**
Multiple-frame selection				
4→2	**77.8 hours**	41.2	67.8	77.9
4→3	80.2 hours	41.5	68.0	78.0
4	82.5 hours	41.7	68.3	78.3
8→4	100.3 hours	**42.8**	**69.2**	**79.5**

MAE	VTS	FS	PT Time	MSRVTT R@1	MSRVTT R@5	MSRVTT R@10
✗	✗	✗	144.5 hours	42.6	68.8	79.2
✓	✗	✗	93.5 hours	42.0	68.4	78.7
✓	✓	✗	82.5 hours	41.6	67.9	78.3
✓	✓	✓	**77.8 hours**	41.2	67.8	77.9

Analysis & Visualization

Method	PT Time	MSRVTT R@1	MSRVTT R@5	MSRVTT R@10
Singularity* [26]	83.4 hours	36.8	65.9	75.5
Ours*	**75.3 hours**	**40.6**	**67.6**	**77.5**
Singularity+ [26]	285.3 hours	41.5	68.7	77.0
Ours+	**198.2 hours**	**44.0**	**70.4**	**78.8**

Caption: A man is talking



Green box: the selected frame

Caption: Aerial shot of tractor raking grass for combine to silage

Prediction for masked patches