End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Presented by Bang, Lorry, Luchao, Xinyu

UNC | COLLEGE OF ARTS AND SCIENCES
Computer Science

Battle

Paper #1 (ours):

1. works with **uncurated** dataset
   a. HW100 with **136M** clips
2. new loss function to handle misalignment for **in-the-wild** dataset
3. end-to-end training from scratch **without** pretraining

Paper #2:

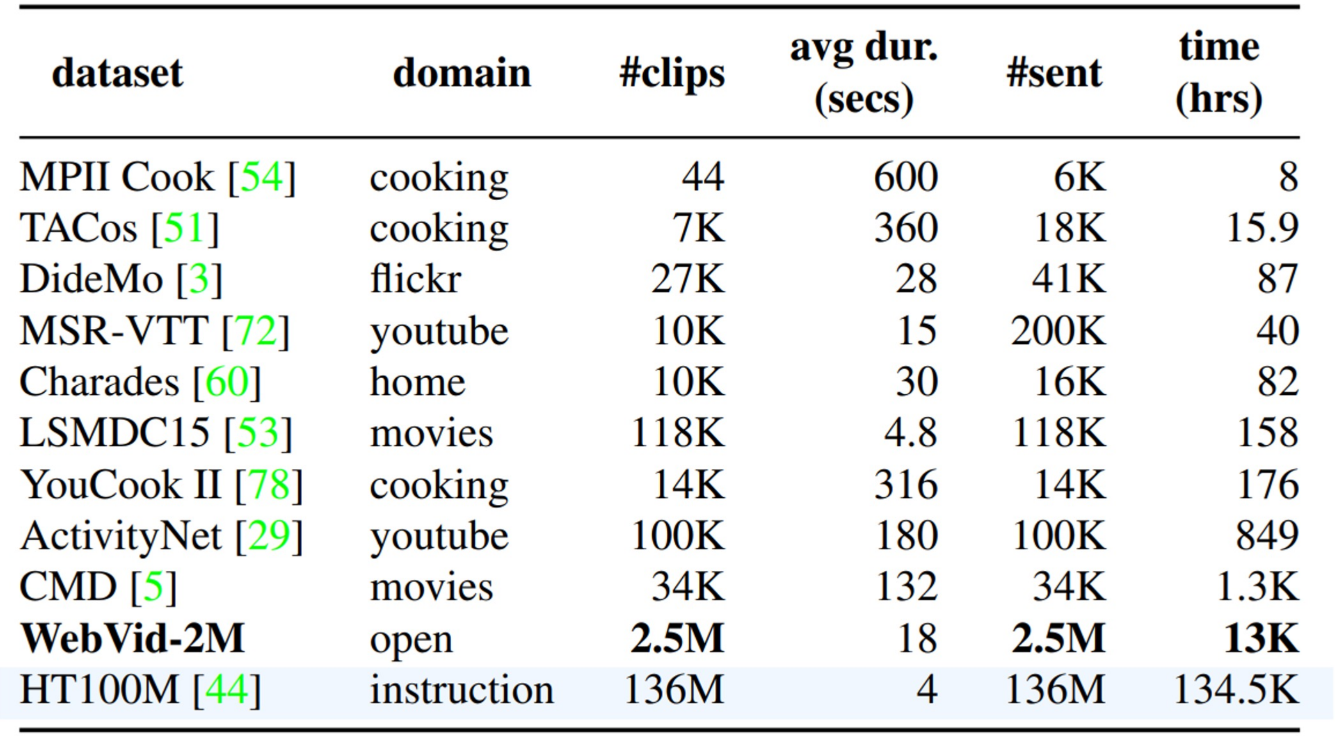
1. works with new dataset
   a. WebVid2M with **2.5M** clips
2. requires alignment between manually generated captions and visual content
3. **pretrained** on ImageNet-21k using ViT

Battle

Similarities:

1. Same corresponding author Andrew Zisserman
    a. paper #1 2020 and followup work paper #2 2021
2. Similar statistics
    a. paper #1 has ~600 citations and ~200 github stars
    b. paper #2 has ~400 citations and ~300 github stars

Dataset



dataset	domain	#clips	avg dur. (secs)	#sent	time (hrs)
MPII Cook [54]	cooking	44	600	6K	8
TACos [51]	cooking	7K	360	18K	15.9
DideMo [3]	flickr	27K	28	41K	87
MSR-VTT [72]	youtube	10K	15	200K	40
Charades [60]	home	10K	30	16K	82
LSMDC15 [53]	movies	118K	4.8	118K	158
YouCook II [78]	cooking	14K	316	14K	176
ActivityNet [29]	youtube	100K	180	100K	849
CMD [5]	movies	34K	132	34K	1.3K
WebVid-2M	open	**2.5M**	18	**2.5M**	**13K**
HT100M [44]	instruction	136M	4	136M	134.5K

Frozen in Time:
A Joint Video and Image Encoder for End-to-End Retrieval

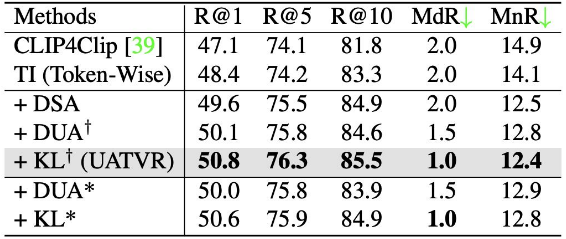
ICCV 2021

Authors: Max Bain, Arsha Nagrani, Gul Varol, Andrew Zisserman

Presenters: Ziyang Wang, Han Wang, Han Lin

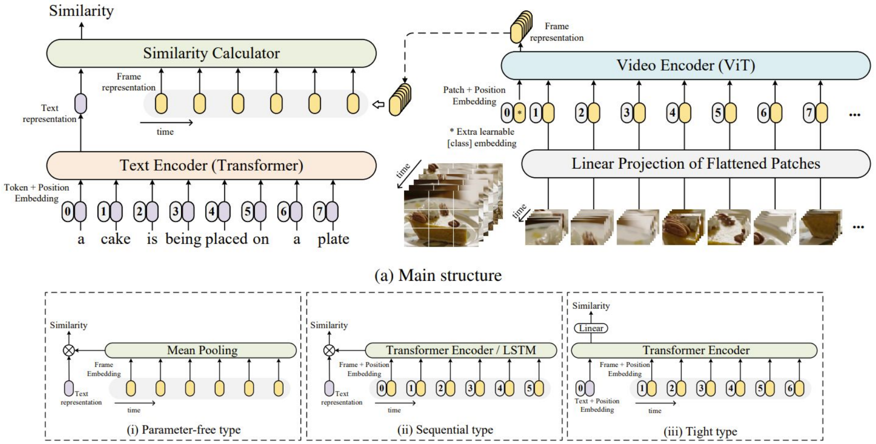
Advantages of Frozen in Time

1) Unified framework on visual information **VS** video-only
2) Collect WebVid2M dataset with clean caption **VS** train from noisy data
3) Inspiration for future works (case study of ICCV23 video-text retrieval papers)
    a) MIL-NCE (1/7)
    b) Image-text learning to video-text learning (7/7)



Methods	R@1	R@5	R@10	MdR↓	MnR↓
CLIP4Clip [39]	47.1	74.1	81.8	2.0	14.9
TI (Token-Wise)	48.4	74.2	83.3	2.0	14.1
+ DSA	49.6	75.5	84.9	2.0	12.5
+ DUA†	50.1	75.8	84.6	1.5	12.8
+ KL† (UATVR)	**50.8**	**76.3**	**85.5**	**1.0**	**12.4**
+ DUA*	50.0	75.8	83.9	1.5	12.9
+ KL*	50.6	75.9	84.9	**1.0**	12.8

Table 1. Ablation study of different components. † denotes the implementation with MIL-NCE contrast and * is implemented with soft contrastive loss via Monte-Carlo estimation [45].



(a) Main structure

(b) Similarity calculator

[1] UATVR: Uncertainty-Adaptive Text-Video Retrieval, Fang et al. ICCV23
[2] CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval, Luo et al.