Hierarchical Self-supervised Representation Learning for Movie Understanding

Fanyi Xiao, Kaustav Kundu, Joseph Tighe, Davide Modolo
AWS AI Labs

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Motivation

Learning good representations for movies is challenging!

1) Annotations are scarce compared to typical action datasets:
3K movies in MovieClips *VS* 650k YouTube clips in Kinetics

2) Movies are very complex!
- Classic video action models are not enough (SlowFast, I3D, …)
- Requires reasoning at many levels:

Simple low-level actions:	Hugging
High-level complex semantic narratives:	*The actors are afraid because the ship is sinking*

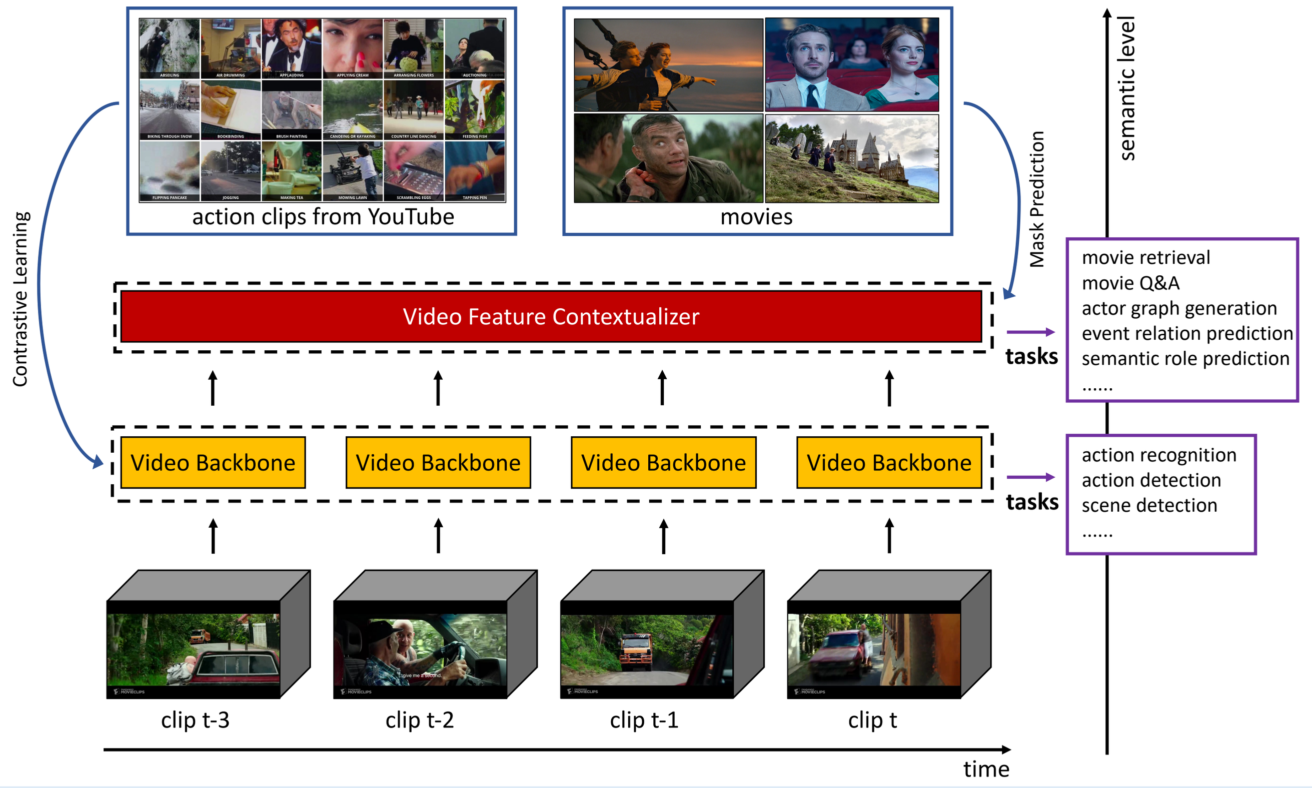
Previous Work

\+ VidSitu [1] proposed a hierarchical movie model (CVPR 21)
\+ Important progress in movie understanding
\- Trained in fully-supervised manner
\- Limited scalability due to the need for expensive annotations

This paper

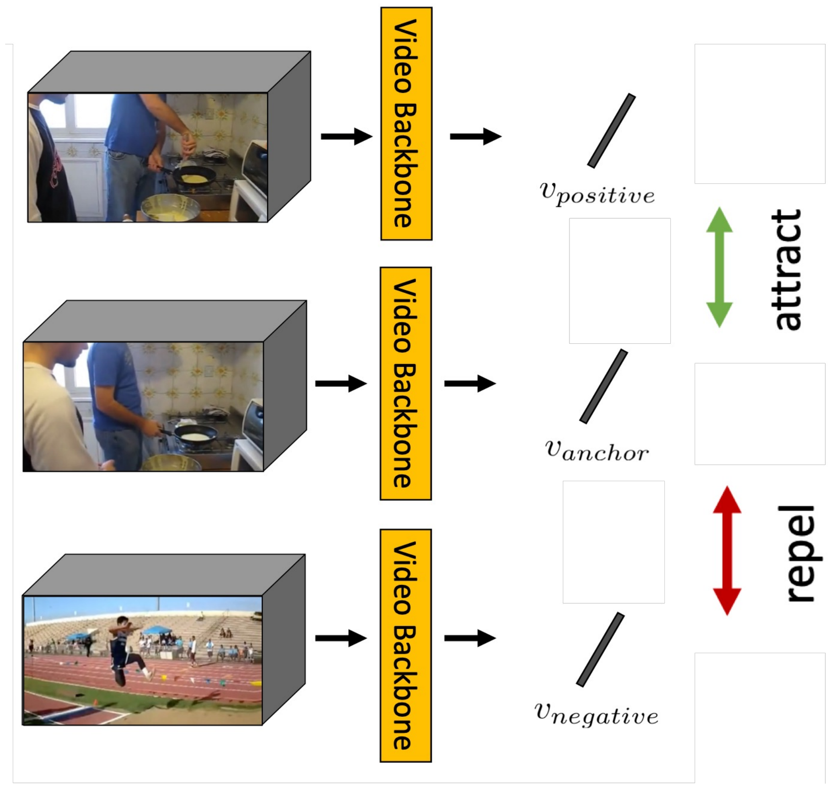
<u>Self-supervised pre-training</u> of a hierarchical movie model:
1. *a low-level backbone*
2. *a high level contextualizer*

Benefits:
- More specialized representations for each level
- Reduced data needs → Each level is trained on its own dataset
- SOTA performance

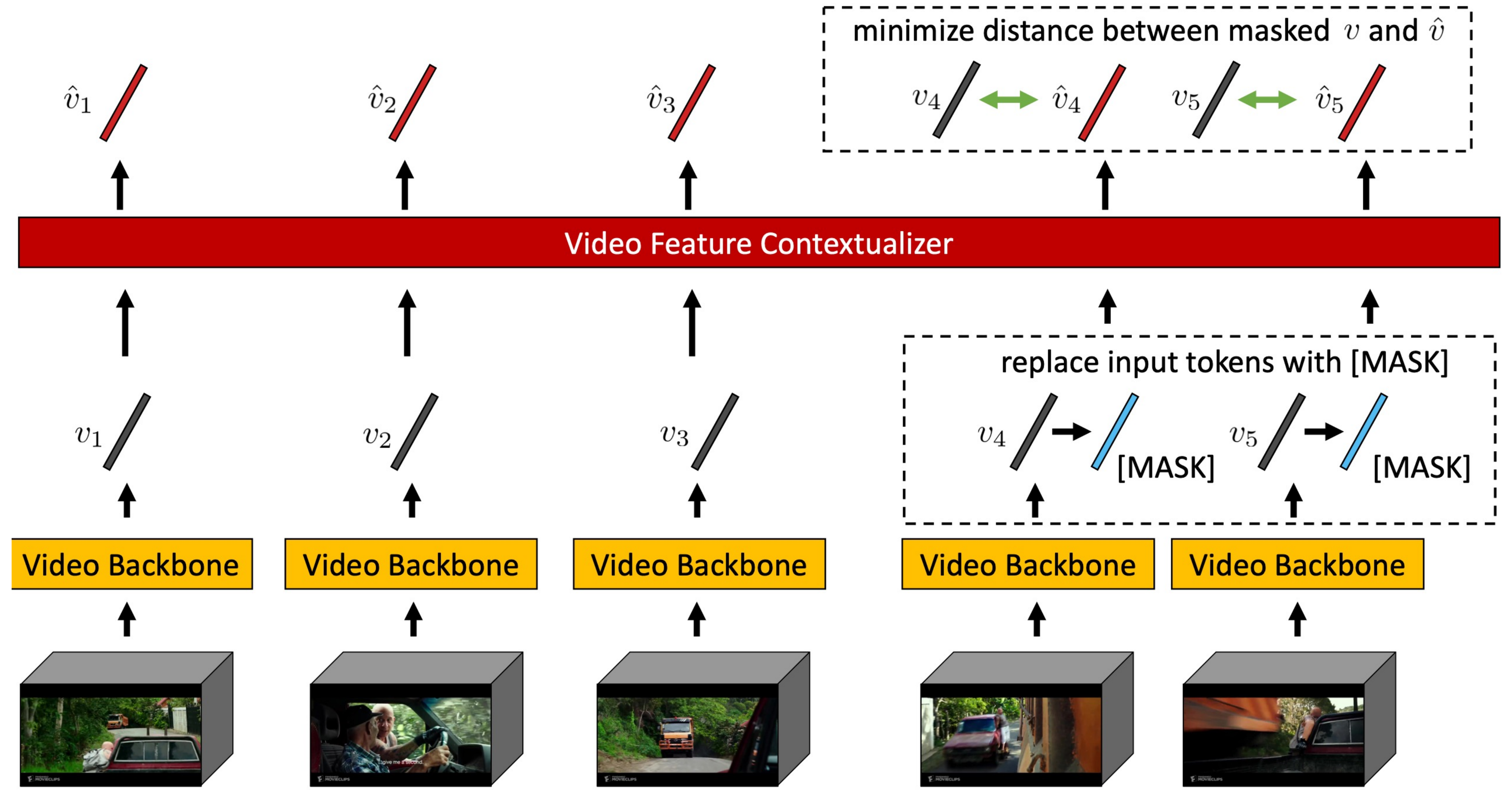


Our self-supervised pre-training for hierarchical movie models

1. **Low-level backbone**
- Extracts *low-level appearance* and *motion cues* for people, objects and scenes from raw pixels
- High capacity models
- Trained on large amount of YouTube videos (e.g., Kinetics)
- Trained using <u>contrasting learning objective</u>



2. **High-Level contextualizer**
- Learn to contextualize the neighboring low-level visual tokens
- Low capacity
- Trained on small amount of movie data with stronger semantic and temporal structures (e.g., VidSitu [1], LVU [4])
- Trained using <u>mask prediction objective</u>



References

[1] "Visual Semantic Role Labeling for Video Understanding", CVPR 2021.
[2] "Spatiotemporal contrastive video representation learning", CVPR 2021.
[3] "Modist: Motion distillation for self-supervised video representation learning", arXiv:2106:09703.
[4] "Towards long-form video understanding", CVPR '21.

Results on VidSitu [1]

Semantic Role Prediction

- *Goal*: predict semantic role labels for each verb. For example, for a verb, "throw", the person is the *agent*, "ball" is the patient and "gym" is the scene.

Model	Pre-training		CIDEr	CIDEr-verb	CIDEr-arg	ROUGE-L	LEA
	Low-level backbone (K400)	Contextualizer					
VidSitu [1]	Supervised Training	None	51.4	59.7	47.3	41.7	46.0
Ours	Contrastive Learning [2]	None	54.4	63.2	47.6	41.8	46.3
Ours	Contrastive Learning [2]	Mask Prediction (MovieClips)	**61.2**	**69.2**	**55.0**	**43.4**	**47.8**

Event Relation Prediction

- *Goal*: 4-way classification problem between four relation types: "A is enabled by B", "A is a reaction to B", "A causes B", and "A is unrelated to B".

Model	Pre-training		Mean Accuracy	Top1 Accuracy
	Low-level backbone (K400)	Contextualizer		
VidSitu [1]	Supervised Training	None	34.0	40.7
Ours	Contrastive Learning [3]	None	34.7	**41.8**
Ours	Contrastive Learning [3]	Mask Prediction (MovieClips)	**35.3**	**41.6**

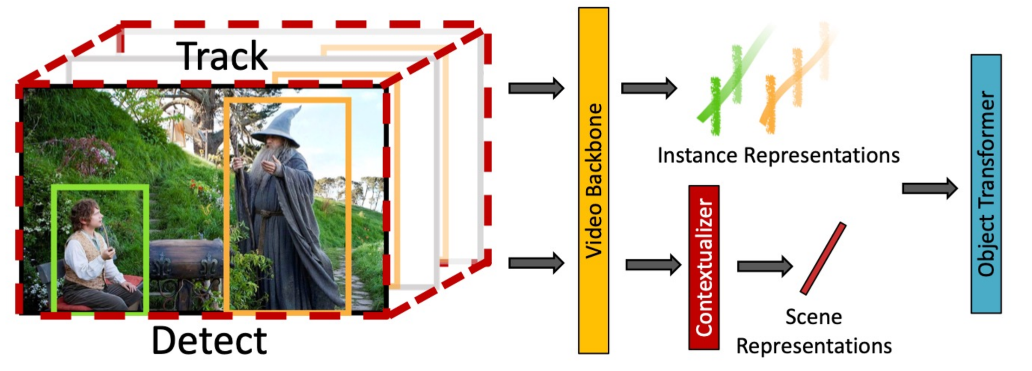
Verb Prediction

- *Goal*: predict action classes (among 1560 classes) for short video segments. Example classes: *look*, *talk*, *walk*, etc.

Model	Pre-training		Acc@1	Acc@5	Recall@5
	Low-level backbone (K400)	Contextualizer			
VidSitu [1]	Supervised Training	None	39.3	69.3	**18.7**
Ours	Contrastive Learning [3]	None	43.0	73.2	17.5
Ours	Contrastive Learning [3]	Mask Prediction (MovieClips)	**44.7**	**74.4**	**18.4**

Results on LVU [4]

Goal: Predict 9 diverse tasks for user engagement, movie meta data classification and content understanding



Our <u>Object transformer ++</u> design

Instance model	Scene model	Top1 Rank	Mean rank
Object Transformer [4]	None	1/9	3.2
Contrastive Learning [2]	None	0/9	3.9
None	Supervised	2/9	4.4
None	Ours	3/9	2.9
Object transformer ++	Ours	**4/9**	**2.3**