Poster SESSION: 2.2, POSTER ID: 174b

Hierarchical Self-Supervised Representation Learning for Movie Understanding

AWS AI Labs

Fanyi Xiao, Kaustav Kundu, Joseph Tighe, Davide Modolo




Movie understanding

- Movies are:
  - a lot more complex than short YouTube videos (e.g., Kinetics)
  - a lot fewer (MovieClips: 3k *vs* Kinetics 650k)
  - classic video action models are not enough (e.g., SlowFast, I3D, etc.)
  - require reasoning at many levels

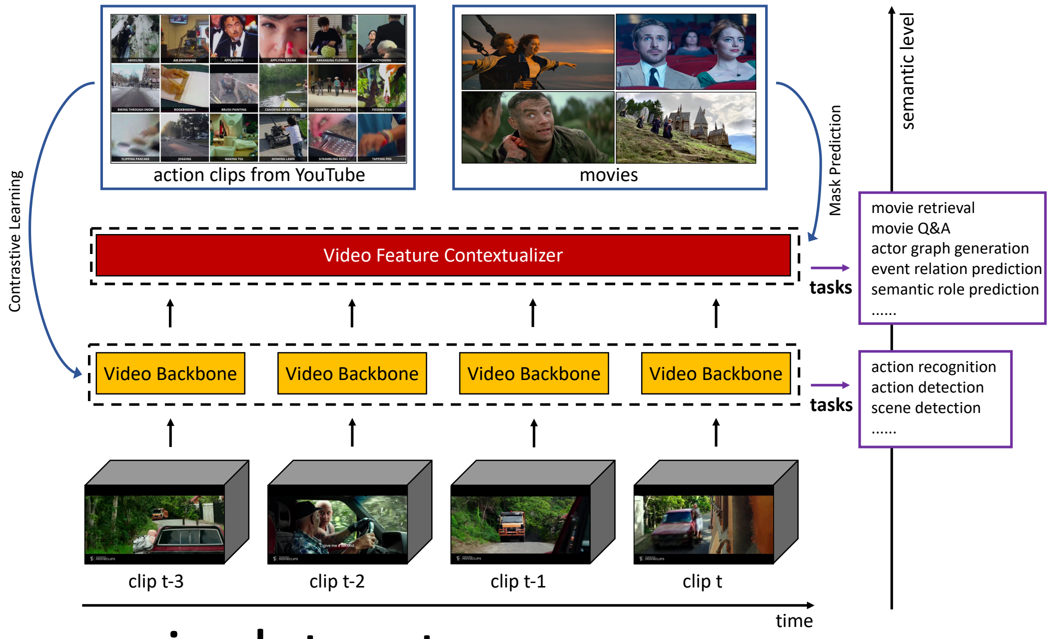
[From MovieNet dataset]

From simple low-level actions → *Hugging*

To high-level semantic narratives → *The actors are sad because the boat is sinking and they don't know if they will survive*

Recent advances on movie understanding

- VidSitu's hierarchical movie model [1]
  - Low-level video backbone encoder
  - Higher-level transformer contextualizer
  - Fully-supervised learning



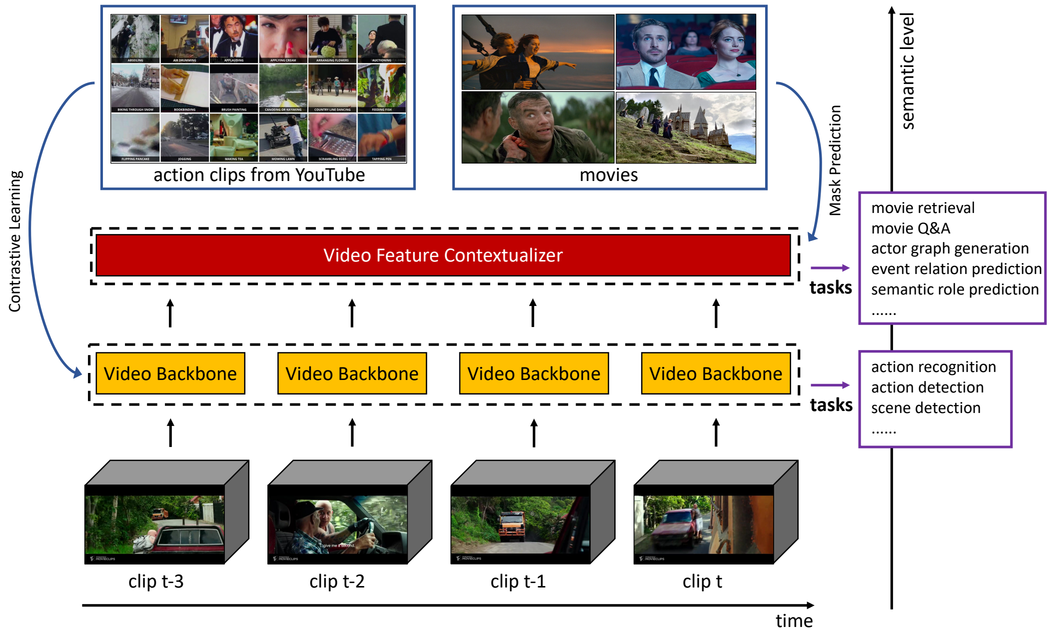
- Challenges:
  - Extremely difficult to annotate large-scale movie datasets

- Our solution:
  - Self-Supervised pre-training

[1] "Visual Semantic Role Labeling for Video Understanding", CVPR 2021

This paper

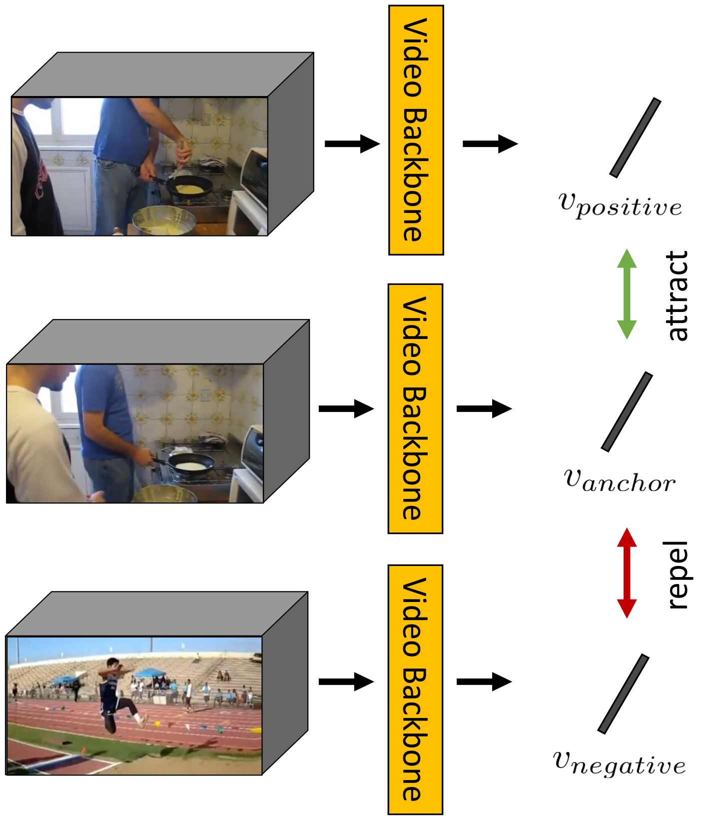
- Sequentially pretrain
  a low-level backbone and
  a high-level contextualizer
  in a self-supervised manner



- Benefits:
  - Each level can specialize better
  - Reduces data needs → Each level can be trained on its own dataset
  - SOTA performance

This paper: Low-level backbone

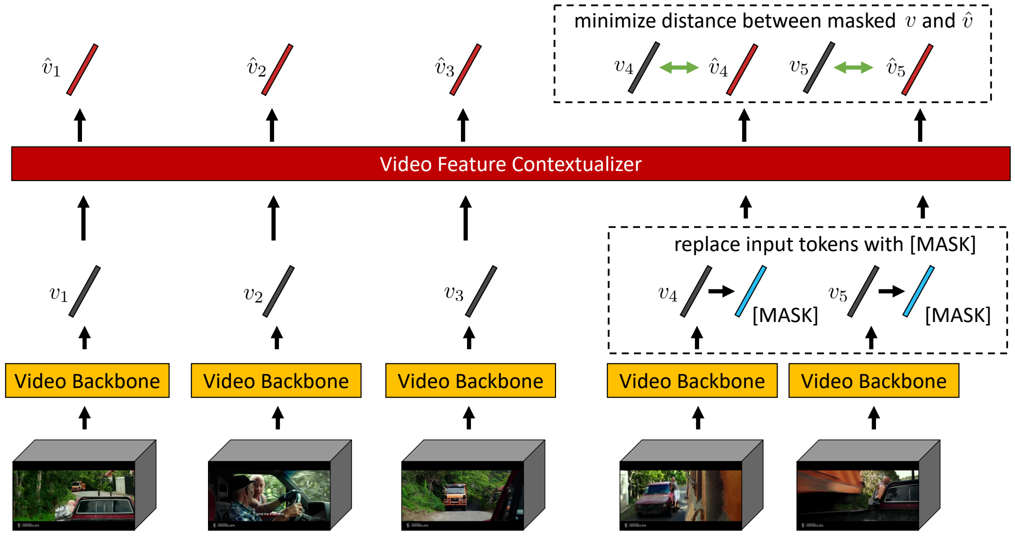
- Extracts low-level appearance and motion cues for people, objects and scenes from raw pixels
- High capacity
- Trained on a large amount of YouTube videos (e.g., Kinetics)
- Trained using contrastive learning objective



contrastive learning for video feature backbone

This paper: High-Level Contextualizer

- Learn to contextualize the neighboring low-level visual tokens
- Low capacity
- Trained on a small amount of movie data with stronger semantic and temporal structures (e.g., VidSitu, LVU)
- Trained using mask prediction objective



mask prediction for video feature contextualizer

Results on VidSitu: Semantic Role Prediction

- Goal: predict various semantic role labels for each verb. For example, the agent (“person”) and the patient (”ball”) of the verb (“throw”), as well as other attributes like the scene where the verb is happening (“a gym”)

Model	Pre-training		CIDEr	
	Backbone	**Contextualizer**		
VidSitu paper [1]	Fully-Supervised (K400)	None	47.06	
This paper	Self-Supervised (K400)	Self-Supervised (VS)	60.34	*(+13.28)*
This paper	Self-Supervised (K400)	Self-Supervised (LVU)	**61.18**	***(+14.12)***

[1] “Visual Semantic Role Labeling for Video Understanding”, CVPR 2021

Results on VidSitu: 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		Top1-Acc
	Backbone	**Contextualizer**	
VidSitu paper [1]	Fully-Supervised (K400)	None	39.91
This paper	Self-Supervised (K400)	Self-Supervised (VS)	**41.62** ***(+1.71)***

[1] “Visual Semantic Role Labeling for Video Understanding”, CVPR 2021

Poster SESSION: 2.2, POSTER ID: 174b

Thank you!

CVPR JUNE 19-24 2022

NEW ORLEANS • LOUISIANA

aws