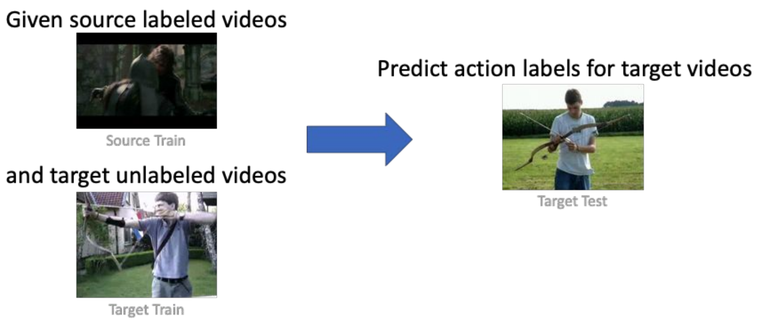
Adversarial Cross-Domain Action Recognition with Co-Attention

Boxiao Pan*, Zhangjie Cao*, Ehsan Adeli, Juan Carlos Niebles

Motivation and Idea

Motivation:

1. Transfer learning for video action recognition;
2. Temporal misalignment problem leads to suboptimal performance.



1. Transfer learning setup

2. Temporal misalignment problem

Idea:

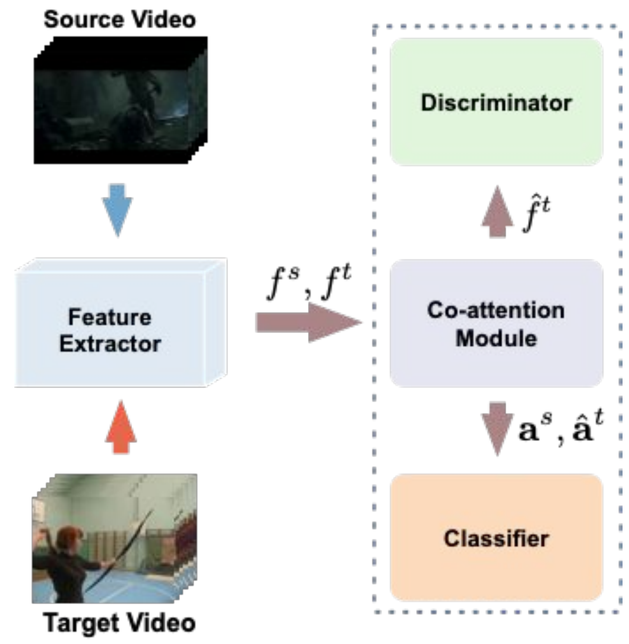
Find the semantically similar video segments **across domain** and align them (**co-attention**).

Model

Pipeline:

1. Feature extraction with standard backbone models (e.g., C3D / TSN);
2. Co-attention module aligns the features temporally, and generates co-attention weighting scores;
3. Discriminator achieves distribution matching with aligned features;
4. A classifier is simultaneously trained to predict the action class.

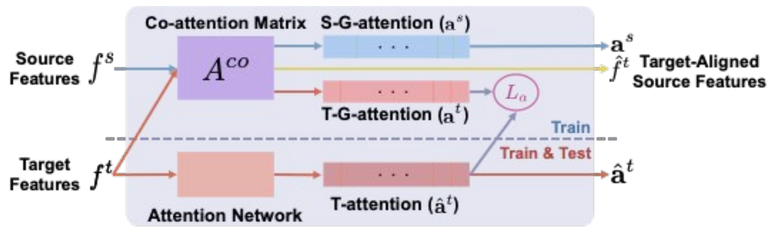
Model



Model overview

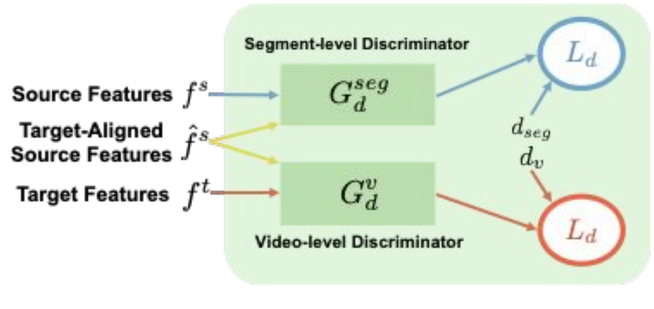
Co-attention Module:

1. Co-attention matrix: assigns high scores to segments that are both **important** action indicators and **common** across domains;
2. Co-attention vectors: row / column summation of co-attention matrix, which are used during classification;
3. Attention network: predicts target co-attention vector at test time (no access to source videos during then).



Discriminator:

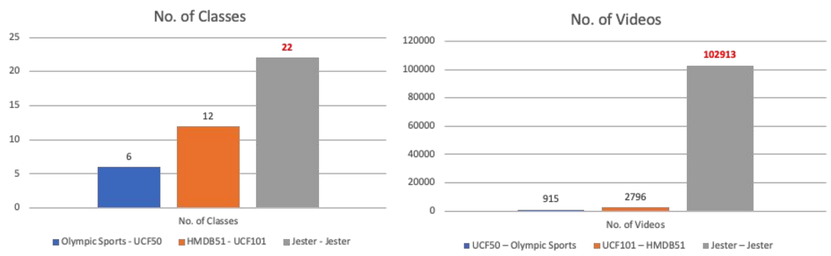
1. Video-level discriminator: traditional distribution matching;
2. Segment-level discriminator: ensures that the aligned features follow the source feature distribution.



Experiments: Quantitative

Datasets:

1. Olympic Sports - UCF50;
2. HMDB51 - UCF101;
3. Jester - Jester.



	Domain Shift
Olympic Sports – UCF50	Appearance + Dynamics
HMDB51 – UCF101	Appearance + Dynamics
Jester – Jester	Appearance + Dynamics + **Action**

Observations:

1. TCoN outperforms all previous methods on all datasets, with the largest margin on Jester;
2. Optical flow model consistently outperforms RGB model.

Method	(HMDB51 → UCF101)$_1$			UCF50 → Olympic.Sports			Olympic.Sports → UCF50			Jester (S) → Jester (T)		
	RGB	Flow	R + F	RGB	Flow	R + F	RGB	Flow	R + F	RGB	Flow	R + F
TSN (Wang et al. 2016)	82.10	76.86	83.11	80.00	81.82	81.75	76.67	73.34	74.47	51.70	49.89	50.56
CMFGLR (Tang et al. 2016)	85.14	78.45	84.85	81.06	79.64	80.23	77.43	77.05	78.89	52.52	54.34	53.36
DAAA (Jamal et al. 2018)	88.36	89.93	91.31	88.37	88.16	89.01	86.25	87.00	87.93	56.45	55.92	57.63
CDAN (Long et al. 2018)	90.09	90.96	91.86	90.65	90.46	91.77	90.08	90.13	90.57	58.33	55.09	59.30
TCoN (ours)	**93.01**	**96.07**	**96.78**	**93.91**	**95.46**	**95.77**	**91.65**	**93.77**	**94.12**	**61.78**	**71.11**	**72.24**

Comparison with prior works

Method	Jester (S) → Jester (T)		
	RGB	Flow	R + F
TCoN - SAdNet	61.23	68.23	71.13
TCoN - TAdNet	58.76	64.56	65.48
TCoN - CoAttn	57.25	56.93	57.95
TCoN - Attn	59.03	62.74	63.13
TCoN	**61.78**	**71.11**	**72.24**

Ablation study

Experiments: Qualitative

Co-attention Matrix Visualization:

We visualize the co-attention matrix for a pair of source-target videos. Only the segments that are both important and common across domains will be paid high attention to.