Actor-agnostic Multi-label Action Recognition with Multi-modal Query

Anindya Mondal*[1,2,3] Sauradip Nag*[1,2,5] Joaquin M Prada[1,4] Xiatian Zhu[1,2,3] Anjan Dutta*[1,2,3,4]

[1]University of Surrey [2]CVSSP [3]Surrey Institute for People-Centred AI [4]School of Veterinary Medicine

[5]iFlyTek-Surrey Joint Research Center on AI

*Authors have equal contributions

ICCV23 PARIS

Paper Code

TL;DR

- **Task:** Proposing a generalized model for multi-label action recognition in humans and animals.
- **Action:** Formulating a novel Multi-modal Semantic Query Network (MSQNet) model in a transformer-based object detection framework (e.g., DETR) for action recognition. It leverages visual and textual cues to eliminate the actor-specific information required in prior art.
- **Results:** MSQNet outperforms the prior art on five single and multi-label datasets by up to **50%**.

Motivation

- **Different actors** (*e.g.* humans and animals) required **separate action recognition models**.
- Those models had **customized designs** to accommodate **specific pose information** of the actors.
- Incorporating **multiple actors** in a **single model** is **challenging** since **animals** exhibit **different shapes, sizes, and appearances**.
- Previous models primarily focused on **single-label action classification**, whereas **multiple actions** usually occur in a **single video** in the **real world**.

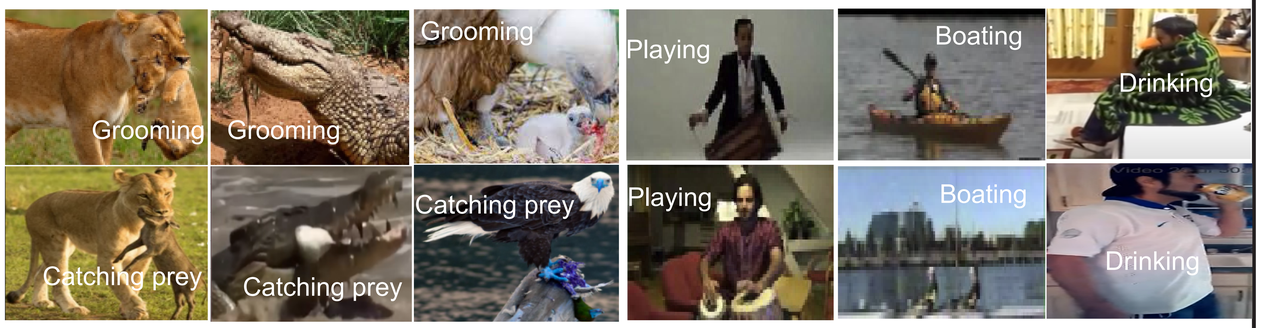
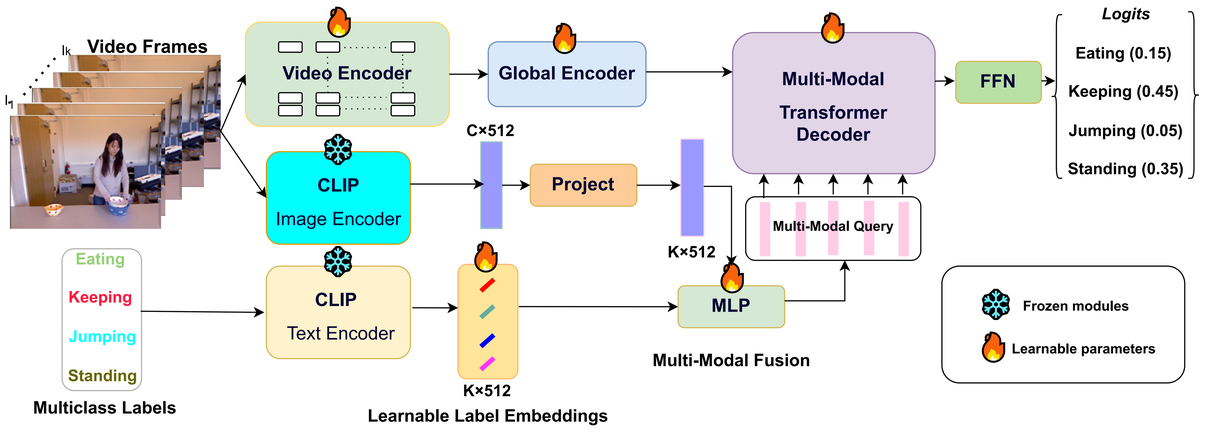


Illustration of large action variation across different actors (e.g., animals and humans). Such differences often motivate the development of actor-specific action recognition models, such as using actor-specific pose estimation.

Multi-modal Semantic Query Network



Our proposed Multi-modal Semantic Query Network (MSQNet) for multi-modal multi-label action recognition has three main components:

- a **spatio-temporal video encoder**, that extracts the spatio-temporal features from an input video,
- a **vision-language query encoder**, that merges the visual and textual information,
- a **multi-modal decoder**, transforms the video encoding to make multi-label classification with a feed-forward network (FFN).

We evaluate MSQNet on **three multi-label (Charades, Hockey, and Animal Kingdom)** and **two single-label (Thumos 14 and HMDB51)** action recognition datasets.

Quantitative Performance

Charades					Thumos 14				
Method	Backbone	Pretrain	MMQ	*mAP*	Method	Backbone	Pretrain	MMQ	*Accuracy*
Zhang et al. CVPR '21	Nonlocal-101	-	No	44.20	Zhao et al. ICCV '17	C3D	-	No	45.42
Fan et al. ICCV '21	SlowFast	K600	No	43.90	Xu et al. TPAMI '19	C3D	-	No	57.19
Wang et al. ArXiv '21	ViT-B	-	No	44.30	Lin et al. ICCV '19	C3D	S1M	No	62.12
MSQNet	ViT-B	K400	No	43.99	MSQNet	C3D	K400	No	67.47
MSQNet	TS	K400	No	44.11	**MSQNet**	TS	K400	Yes	**79.71**
MSQNet	TS	K400	Yes	**47.57**	**MSQNet**	TS	K400	Yes	**83.16**

Animal Kingdom					Hockey				
Method	Backbone	Pretrain	MMQ	*mAP*	Method	Backbone	Pretrain	MMQ	*Multilabel Accuracy*
Ng et al. CVPR '22	X3D	-	No	25.25	Carbonneau. ETS'17	-	-	No	90.00
Ng et al. CVPR '22	I3D	-	No	16.48	Zhang et al. CVPR '21	CSN-152	-	No	96.30
MSQNet	I3D	-	No	**55.59**	MSQNet	I3D	-	No	93.29
MSQNet	TS	K400	No	**71.63**	**MSQNet**	TS	K400	No	93.45
MSQNet	TS	K400	Yes	**73.10**	**MSQNet**	TS	K400	Yes	**96.95**

HMDB51

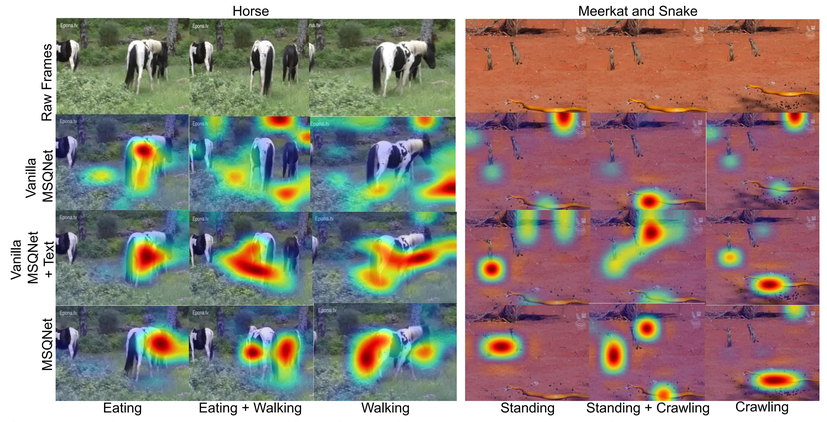
Method	Backbone	Pretrain	MM	Accuracy
Wu et al. CVPR '23	ViT	WIT-400M	Yes	84.31
Kalfaoglu et al. CVPR '23	R(2+1)D	IG65M	No	85.10
Wang et al. CVPR '23	ViT	K400/K600	No	88.10
MSQNet	TS	K400	Yes	**93.25**

Fully supervised setting

Split	Method	Thumos 14	Charades	HMDB51
Reported	Wang et al. ArXiv '21	-	21.1	58.70
	Bain et al. ArXiv '22	-	25.8	-
	Wu et al. CVPR '23	-	-	61.40
50% Seen	Vanilla MSQNet	49.37	15.87	45.66
	Vanilla MSQNet + Text Init	53.76	18.30	51.22
	MSQNet	**63.98**	**30.91**	**59.24**
75% Seen	Vanilla MSQNet	52.02	17.43	48.37
	Vanilla MSQNet + Text Init	60.28	18.62	59.58
	MSQNet	**75.33**	**35.59**	**69.43**

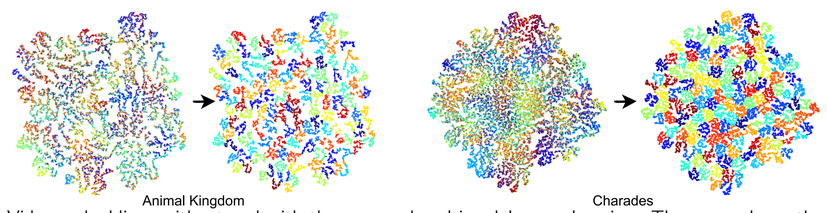
Zero-shot setting

Qualitative Results



Attention rollout on sample videos from Animal Kingdom [1] showing raw frames, heatmap with only bare backbone, with uni-modal prompt, and MSQNet.

Feature Visualization



Video embeddings without and with the proposed multimodal query learning. The arrow shows the transition.

Ablation Studies

Text Encoder	AK (mAP)	Charades (mAP)
BERT	69.81	43.57
CLIP	**71.63**	**44.11**

Effect of text encoding

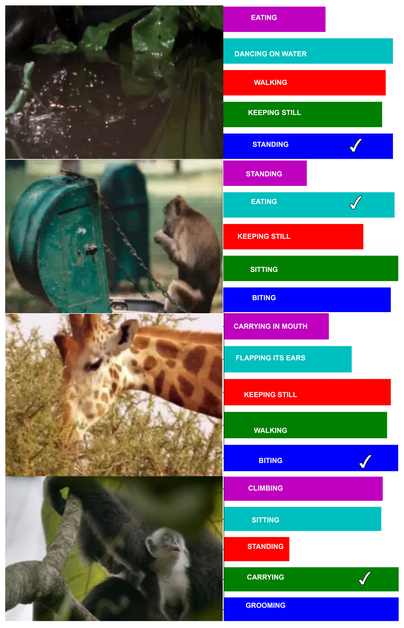
Backbone	Animal Kingdom (mAP)	Charades (mAP)
VideoMAE	71.19	41.69
TimeSformer	**73.10**	**47.57**

Performance of MSQNet using different backbones for the video encoder

# Frames	AK (mAP)	Charades (mAP)
8	67.74	43.06
10	**69.73**	**45.33**
16	**73.10**	**47.57**

Performance of MSQNet with a different number of frames

Top-5 predictions



References

1. Ng *et al.* Animal kingdom: A large and diverse dataset for animal behavior understanding. CVPR, 2022.
2. Wang *et al.* Actionclip: A new paradigm for video action recognition. ArXiv, 2021.
3. Zhang *et al.* Multi-label activity recognition using activity-specific features and activity correlations. CVPR, 2021.
4. Bertasius *et al.* Is space time attention all you need for video understanding? ICML, 2021.
5. Bangalath *et al.* Bridging the gap between object and image-level representations for open vocabulary detection. NeurIPS, 2022