[26th European Conference on Artificial Intelligence (ECAI2023)]

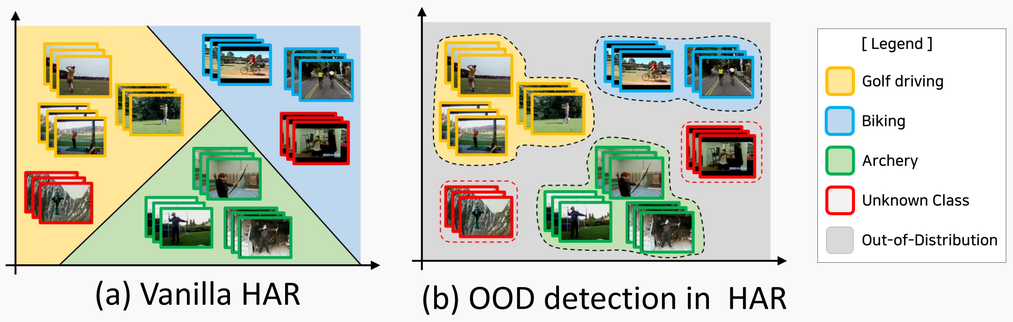
A Simple Debiasing Framework for Out-of-Distribution Detection in Human Action Recognition

Minho Sim, Young-Jun Lee, Dongkun Lee, Jongwhoa Lee, and Ho-Jin Choi
{smh3946, yj2961, hagg30, jongwhoa.lee, hojinc} @ kaist.ac.kr
Knowledge Engineering & Artificial Intelligence (KEAI) Lab
School of Computing, KAIST

I. Introduction

- **Human Action Recognition (HAR)**
  - HAR aims to recognize actions of an individual or a group of people
  - Earlier studies focused on:
    - Modalities: RGB frames, Optical flows, Human skeletons, etc.
    - Architectures: CNN+LSTM, 3D-CNN, GCN, Vision transformer, etc.

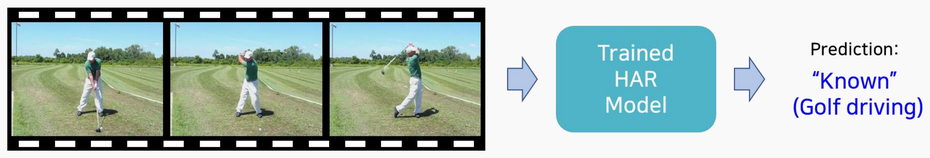
- **Out-of-Distribution (OOD) Detection in HAR**
  - Typical HAR model is trained based on closed-world assumption
    - The model can only make predictions with known labels
  - In a real-world scenario, HAR is essentially an open set problem [Scheirer`2012]
    - OOD detection in HAR aims to detect actions from "unknown classes".



(a) Vanilla HAR (b) OOD detection in HAR

II. Background & Research objective

- **Static Bias Problem** [Choi`2019]
  - Actions occur in specific scene contexts (e.g., playing golf on a grass field)
  - HAR model is easily biased towards static cues in the video clip
    - Cannot focus on the temporal dynamics of human actions
    - Can produce incorrect prediction for unknown actions with similar background



(a) Correct detection for known action "Golf driving"

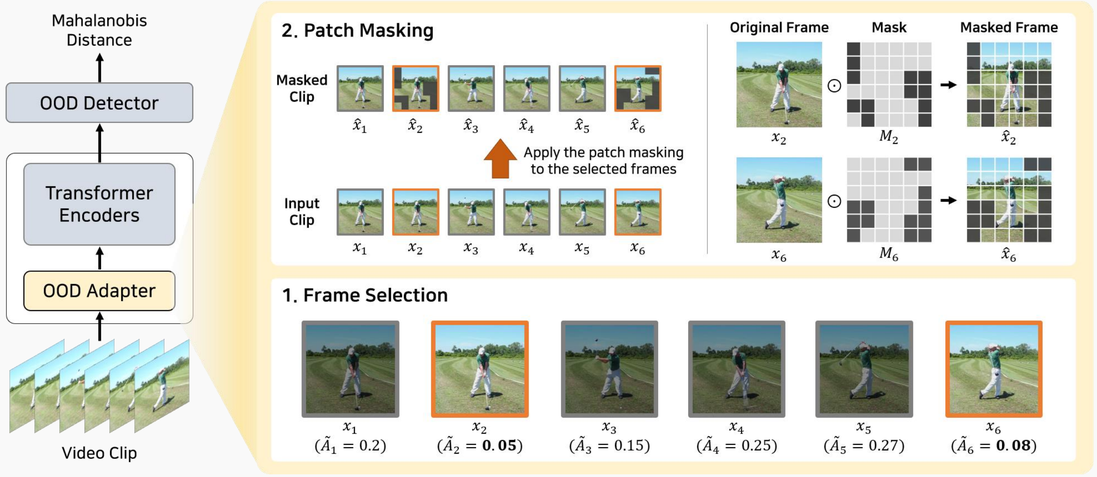
(b) Incorrect detection for unknown action "Kick ball"

- **Contributions**
  - Propose a simple debasing framework for OOD detection in HAR which can alleviate the static bias problem
  - With attention-based video masking, our framework consistently boosts the performance of various OOD detection methods while achieving SOTA results on challenging benchmarks
  - Extensive experiments and analyses demonstrate the validity of our framework and the effect of static bias on OOD detection in HAR

III. Method

- **A Simple Debiasing Framework for OOD detection in HAR**
  - Attention map can be used as a reasonable guide to determine whether the patch contains action-related objects
  - Introduce "Adapter-based Video Masking" which masks less-attended patches to mitigate the static bias problem

x_i the i-th frame in the given video clip — $\hat{x}_i$ a masked frame for the i-th input frame
$\tilde{A}_i$ an attention map of the i-th frame — M_i a mask of the i-th frame



- **Extracting Attention Maps**
  - For each frame and patch, we compute how much model attends to each frame and patch using Attention Rollout [Abnar`2020]
    - Spatial attention map:

$$A(s_i) = Q(s_i) \cdot K(s_i)^T, i \in \{1, 2, \cdots, n_s\}$$
$$\tilde{A}_s = A(s_1)A(s_2)\cdots A(s_{n_s}), \tilde{A}_s \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}}$$

    - Temporal attention map:

$$A(t_i) = Q(t_i) \cdot K(t_i)^T, i \in \{1, 2, \cdots, n_t\}$$
$$\tilde{A}_t = A(t_1)A(t_2)\cdots A(t_{n_t}), \tilde{A}_t \in \mathbb{R}^T$$

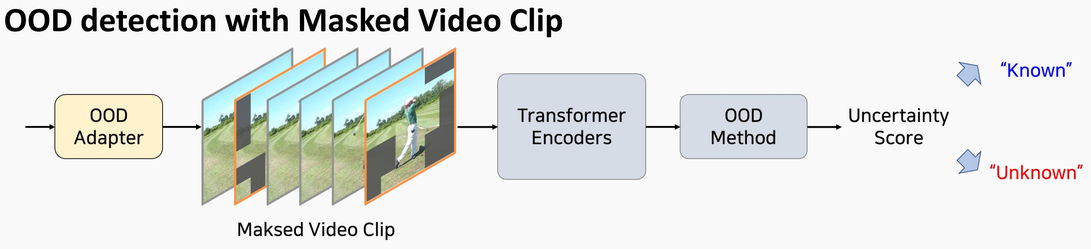
- **Adapter-based Video Masking**
  - Adopt a two-stage adapter consisting of frame selection and patch masking
  - The adapter works in a coarse-to-fine manner, first selecting frames and then performing fine-grained patch masking on each selected frame.
    - Frame selection:

$$F = \text{lt–threshold}(\tilde{A}_t, \gamma_t)$$

    - Patch masking:

$$J_t = \text{lt–threshold}(\tilde{A}_s, \gamma_s), \qquad M\left[\left\lfloor\frac{J_{t_i}}{p}\right\rfloor, mod(J_{t_i}, p)\right] = 1$$

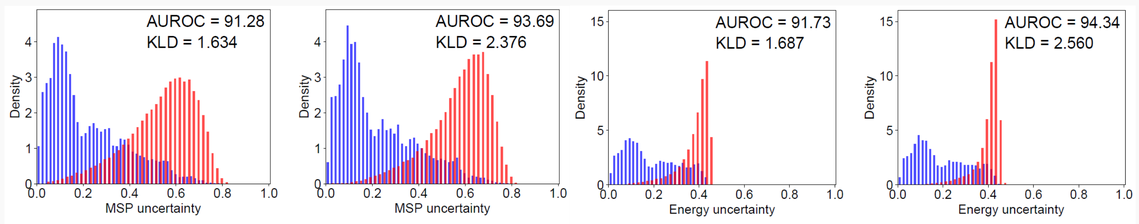
- **OOD detection with Masked Video Clip**



IV. Experiments & Anaylsis

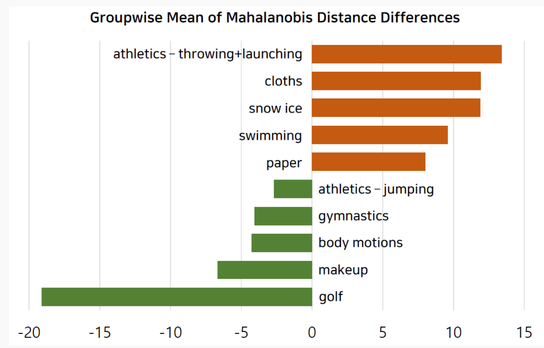
- **OOD Detection in HAR Results**
  - Test the proposed on two challenging benchmarks: UCF-101 (in) vs. HMDB-51 (out) and UCF-101 (in) vs. MiT-v2 (out)
  - Consistent increase in KL-divergence shows that our method effectively enhances the ID/OOD separability

OOD Deteciton Methods	Metrics	UCF-101 (in) + HMDB-51 (out)			UCF-101 (in) + MiT-v2 (out)		
		Original	Ours	Diff.	Original	Ours	Diff.
MSP [14]	AUROC	83.625 ± 0.0005	**85.807** ± 0.0034	2.181	91.280 ± 0.0011	**93.695** ± 0.0008	2.415
	AUPR	81.561 ± 0.0005	**84.188** ± 0.0044	2.626	90.125 ± 0.0010	**92.660** ± 0.0008	2.535
	FPR95	58.187 ± 0.0214	**47.354** ± 0.0100	10.833	34.843 ± 0.0064	**26.182** ± 0.0069	8.660
Energy [25]	AUROC	83.937 ± 0.0052	**86.013** ± 0.0037	2.075	91.731 ± 0.0011	**94.345** ± 0.0008	2.614
	AUPR	82.609 ± 0.0070	**85.197** ± 0.0050	2.588	91.355 ± 0.0010	**93.950** ± 0.0008	2.595
	FPR95	59.049 ± 0.0126	**47.536** ± 0.0108	11.512	34.501 ± 0.0076	**25.194** ± 0.0064	9.306
Mahalanobis [22]	AUROC	80.884 ± 0.0058	**85.319** ± 0.0032	4.435	90.560 ± 0.0011	**93.539** ± 0.0009	2.979
	AUPR	78.845 ± 0.0083	**85.582** ± 0.0051	6.736	89.866 ± 0.0012	**92.630** ± 0.0009	2.764
	FPR95	77.490 ± 0.0109	**68.073** ± 0.0088	9.416	37.497 ± 0.0040	**26.965** ± 0.0051	10.532

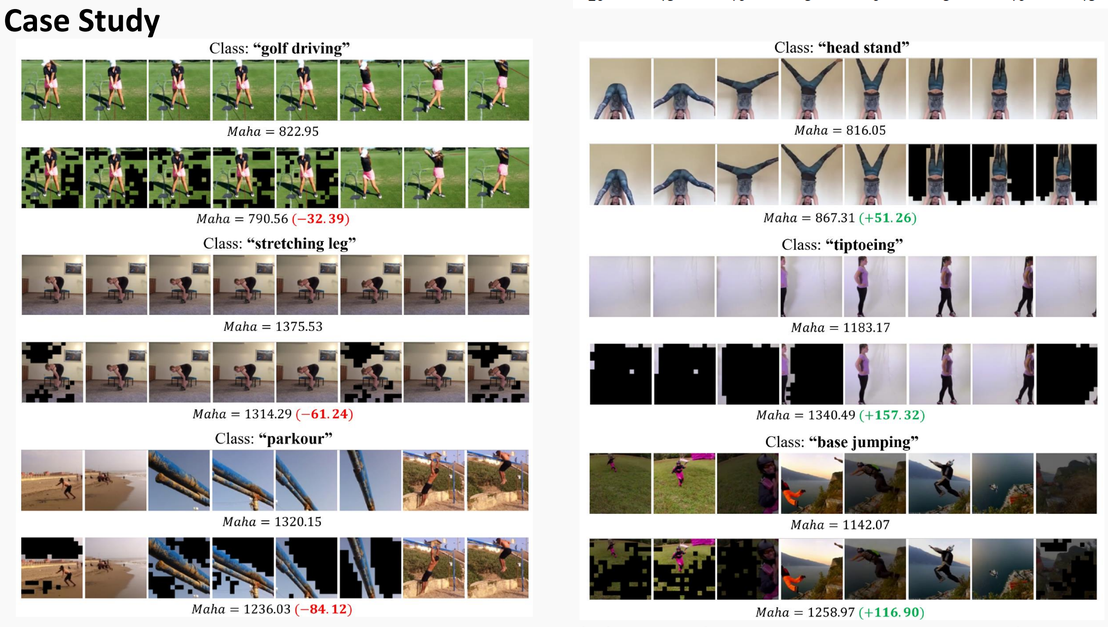


(a) Original (w/ MSP) (b) Ours (w/ MSP) (c) Original (w/ Energy) (d) Ours (w/ Energy)

- **Groupwise Analysis**
  - Adopt parent-child groupings of the Kinetics-400 dataset
  - Compute classwise median of the difference in Mahalanobis distance after applying video masking
  - Our method is effective in the groups where temporal dynamic is essential



- **Case Study**



(a) ID (Kinetics-400) cases (b) OOD (Kinetics-600 exclusive) cases