Multimodal Attack Detection for Action Recognition Models

Furkan Mumcu, Yasin Yilmaz
University of South Florida
4202 E Fowler Ave, Tampa, FL 33620
{furkan, yasiny}@usf.edu

Abstract

Adversarial machine learning attacks on video action recognition models is a growing research area and many effective attacks were introduced in recent years. These attacks show that action recognition models can be breached in many ways. Hence using these models in practice raises significant security concerns. However, there are very few works which focus on defending against or detecting attacks. In this work, we propose a novel universal detection method which is compatible with any action recognition model. In our extensive experiments, we show that our method consistently detects various attacks against different target models with high true positive rates while satisfying very low false positive rates. Tested against four state-of-the-art attacks targeting four action recognition models, the proposed detector achieves an average AUC of 0.911 over 16 test cases while the best performance achieved by the existing detectors is 0.645 average AUC. This 41.2% improvement is enabled by the robustness of the proposed detector to varying attack methods and target models. The lowest AUC achieved by our detector across the 16 test cases is 0.837 while the competing detector's performance drops as low as 0.211. We also show that the proposed detector is robust to varying attack strengths. In addition, we analyze our method's real-time performance with different hardware setups to demonstrate its potential as a practical defense mechanism.

1. Introduction

Adversarial machine learning attacks have been a very popular research topic since the introduction of the Fast Gradient Sign Method (FGSM) [7]. While the initial research on these attacks focused on image classification models, many recent works showed that adversarial attacks are effective against video understanding models including action recognition [14, 27, 32], object detection [11, 13, 37], and anomaly detection [18]. Different architectural types such as convolutional neural networks (CNNs) and vision transformers (ViTs) are shown to be vulnerable against adversarial attacks [17, 28].

Increasing numbers of successful adversarial attacks in a broad range of applications against various architectures raise real-world security concerns. However, compared to the attacks, defense or detection mechanisms against adversarial attacks have not been studied widely. Although there are several defense methods proposed for image recognition models, there is very limited work on defending against attacks targeting video recognition models. In this paper, we propose a *Vision-Language Attack Detection (VLAD)* mechanism, which is the first vision-language based detection method for action recognition models.

Adversarial attacks are achieved by creating input data with perturbations which are not obvious to human eye, but result in errors for the machine learning algorithms. For images, this perturbation can be applied to the whole input [7, 16] or a specific part of the input in the form of pixels or patches [22]. For videos, in addition to the spatial domain, attackers also have the opportunity to make perturbations in the temporal domain, which makes defending against video attacks a more challenging task. Compared to the image models, there are much less defense methods for action recognition models. Advit [30] is an adversarial frame detection method which proposes using optical flow to generate pseudo frames and then evaluating the classifier output consistency between the original input frames and the pseudo frames. Shuffle [9] tries to defend against adversarial attacks by shuffling the input frames randomly before feeding into the action recognition model. In our extensive experiments we show that neither of these two methods perform well under many attack scenarios.

Attacks against videos can be achieved in various ways due to the temporal nature of videos. In addition to possible perturbations in one frame (e.g., pixel, patch, all frame), one might combine these methods in many ways along the temporal dimension. Therefore, we believe detecting adversarial attacks against videos can be accomplished by observing the inputs and the model outputs with a separate-modality subsystem.

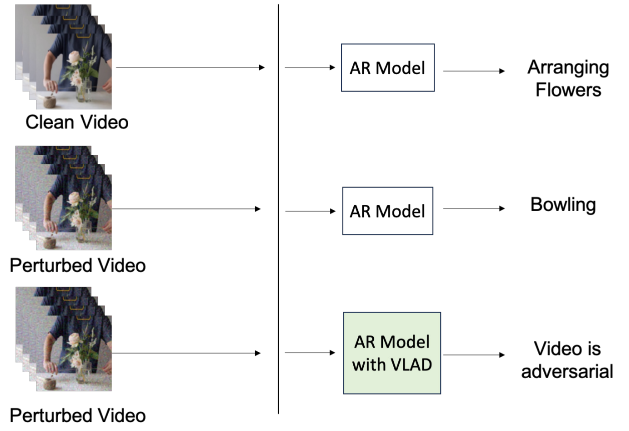


Figure 1. The proposed vision-language attack detection (VLAD) method cooperates with the action recognition (AR) model to detect adversarial videos.

In our approach, we use a vision-language model (VLM) as an observing subsystem. VLMs are gaining popularity with the introduction of CLIP [24] and are being used for many other applications such as video action recognition [19, 29], object detection [36] and video anomaly detection [35]. VLM's multimodal processing of visual and language features provide additional information space for detecting attacks based on only visual features. Leveraging the capability of VLMs to make connections between texts and images, our method utilizes the context of video by computing the similarity scores between the video frames and the action class labels. Then, it decides if the video is adversarial or clean based on the consistency/inconsistency between the predictions by the action recognition model and the similarity scores obtained by VLM.

Our contributions can be summarized as follows:

- We propose a universal attack detection method which can easily work with any action recognition model.
- To the best of our knowledge, this is the first method that leverages a vision-language model for context awareness against adversarial machine learning attacks.
- We benchmark our detection with extensive experiments and analyze its effectiveness compared to existing defense methods. Experimental results show that our method exhibits a 41.2% performance improvement relative to the state-of-the-art detector and robustness to varying attack methods and strengths.
- The real-time operation of the proposed method is demonstrated with several GPUs.

2. Related Work

Adversarial machine learning attacks have been investigated for several years since the introduction of FGSM [7]. The first adversarial attacks were introduced for images under the black-box and white-box settings. In the white-box setting (e.g., FGSM [7], CW [1], PGD [16]), an attacker has full access to the model including its architecture and parameters, while in the black-box setting (e.g., ZOO [2], NES [10], SPSA [26]), the attacker has access only to the output labels or probabilities.

After the increasing popularity of adversarial attacks on images, attacks against action recognition models have also been investigated broadly in recent years. Attacks like SparseAttack [32], Geo-Trap [14] or HeuristicAttack [27] are query-based black-box attacks against video action recognition models. Although it is possible to adapt white-box image attacks to action recognition models, there are also several adversarial attacks designed for action recognition models. One Frame Attack (OFA) [8] tries to attack models by perturbing one frame of the input. Flickering Attack (Flick) [23] tries to attack by changing the RGB stream of the input videos. In our experiments we test our defense method against white-box attacks since they provide more flexibility and effectiveness for adversaries, which make them more challenging to detect.

However, defense against adversarial attacks is not sufficiently investigated despite the increasing numbers and variety of attacks. Several defense methods have been developed for image models. [33] aims to improve classification performance by adding an adversarial attack module and a data augmentation module to the model. [34] proposes a defense method where they analyze the common information between clean and perturbed data. [20] tries to remove perturbations with the help of adaptive compression and reconstruction. [31] implements random resizing to inputs to achieve robustness. [3] uses compression of JPEG images to avoid any possible perturbations. Some adversarial benchmark datasets, such as [21], were proposed to evaluate the robustness against adversarial attacks.

There is not much research on defense against adversarial video attacks. Advit [30] is the first defense method introduced for videos. It generates pseudo frames using optical flow and evaluates the consistency between the outputs for original inputs and pseudo frames to detect attacks. Shuffle [9] is a recent defense method which aims to increase the robustness of action recognition models to attacks by randomly shuffling the input frames. They show that the predictions for clean videos are not significantly affected by shuffling while the adversarial effects of attacks are mitigated. Although Shuffle [9] is not an attack detection method, we show that it can be used for attack detection and compare our method with it.

Vision language models are gaining popularity in recent years with the introduction of CLIP [24]. Many video understanding tasks benefit from VLMs including object detection [15, 36], video action recognition [19, 29] and video anomaly detection [35]. In our work, we present the first usage of VLMs for adversarial video detection.

Algorithm 1 Vision-Language Attack Detection (VLAD)

```
Input: input video $X$, word vector $w$ containing class labels, action recognition model $g$, vision language model $vlm$, threshold $h$.
Output: detection result $d$.
 1: action recognition model receives the input video, returns the classification probabilities $p_a \leftarrow g(X)$
 2: $vlm$ calculates the similarity scores between the input video frames and word vector, $\{s_i\} \leftarrow vlm(X, w)$
 3: take the average of similarity scores over frames
    $s \leftarrow \frac{1}{N}\sum_{i=1}^{N} s_i$
 4: apply softmax to similarity scores to get the context probabilities $p_c \leftarrow \text{softmax}(s)$
 5: Calculate detection score
    $\alpha \leftarrow \frac{1}{2}[D_{KL}(p_a||p_c) + D_{KL}(p_c||p_a)]$
 6: if $\alpha > h$ then
 7:     $d \leftarrow$ adversarial
 8: else
 9:     $d \leftarrow$ not adversarial
10: end if
11: return $d$
```

3. Method

3.1. Threat model

An action recognition (AR) model $g(\cdot)$ takes a video $X \in \mathbb{R}^{N\times H\times W\times C}$ as an input, which is a sequence of N frames, each of which consists of $H \times W$ pixels and C channels. Denoting the true label of the input with y, the true classification by the action recognition model is given by $g(X) = y$. However, an adversary can attack the system by generating an adversarial version X^{adv} of the input, which might be classified as $g(X^{adv}) = y'$, where $y \neq y'$. We consider the most challenging case for the detector, the white-box attack scenario, in which the attacker has full access to the AR model. The detector has also access to the predicted class probability vector $p_a = [p_1, p_2, p_3, ..., p_M]$, where M is the number of class labels.

3.2. Proposed Method: VLAD

There are many ways to generate successful video perturbation from changing only one frame to perturbing all frames. Due to the vastness of attack space and the typical obliviousness of the action recognition models to the attack strategy, a defense mechanism should not use any bias regarding the attacks. Therefore, we propose a universal detection mechanism which does not rely on any assumption about the attack method or action recognition model, and hence can work with any model to detect a broad range of adversarial attacks (Fig. 1).

An overview of the proposed detection method is depicted in Fig. 2. In our detection mechanism, in parallel with the action recognition model, we apply a VLM to obtain context probabilities. We feed two inputs to the VLM to get the context similarity scores for each frame. Taking the input video frames $f = [f_1, f_2, \ldots, f_N]$ and the class labels $w = [w_1, w_2, \ldots, w_M]$ VLM outputs the similarity score matrix $S \in \mathbb{R}^{N\times M}$ for each input video X.

The similarity scores $s_i = [s_{i1}, s_{i2}, \ldots, s_{iM}]$ for each frame i are averaged to obtain the similarity vector $s \in \mathbb{R}^M$ of the video:

$$s = \frac{1}{N}\sum_{i=1}^{N} s_i. \tag{1}$$

Next, the softmax function is applied to the video similarity score to obtain the context probabilities:

$$p_c = \text{softmax}(s) = [p_{c1}, p_{c2}, \ldots, p_{cM}]. \tag{2}$$

A final detection score α is calculated by getting the average of forward and reverse KL divergences between p_a and p_c:

$$\alpha = \frac{1}{2}[D_{KL}(p_a||p_c) + D_{KL}(p_c||p_a)]. \tag{3}$$

The detection score α is expected to be low for clean inputs and high for adversarial inputs. A threshold h is decided by calculating the detection scores for a set of clean videos, $\beta = [\alpha_1, \alpha_2, \ldots, \alpha_K]$, where K is the number of clean videos and the scores in β are sorted in ascending order. Threshold h is selected as the θth percentile of the clean training scores:

$$h = \beta[\lfloor K\theta/100 \rfloor], \tag{4}$$

where $\lfloor\cdot\rfloor$ denotes the floor operator, and $\beta[i]$ denotes the ith element of β.

After obtaining the detection score for an input, decision d is made as follows:

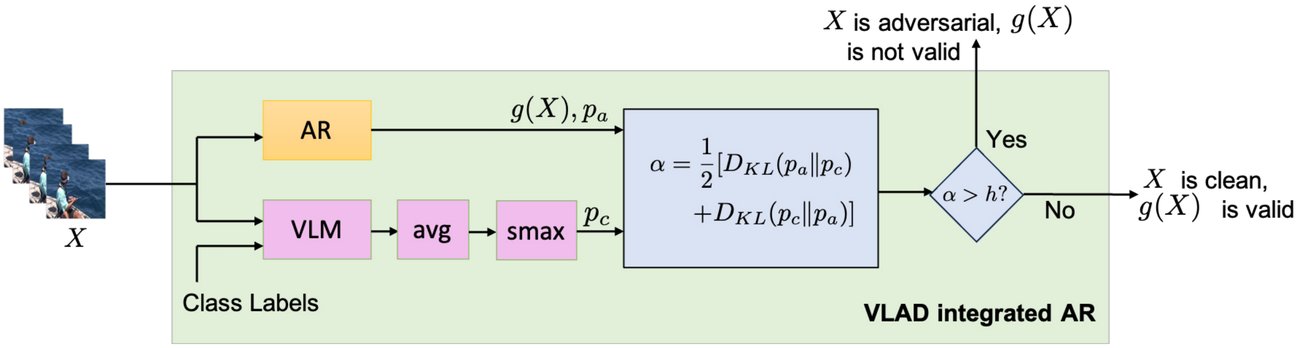
$$d = \begin{cases} X \text{ is adversarial} & \text{if } \alpha > h \\ X \text{ is not adversarial} & \text{if } \alpha \leq h, \end{cases} \tag{5}$$

where α is computed as in Eq. (3).

3.3. Applicability and Implementation

Our proposed method is highly compatible with the existing action recognition models since it does not make any architectural changes to the existing models and it does not require any neural network training. VLAD only needs the classification probabilities from the action recognition model. Algorithm 1 describes the overall workflow of VLAD. [1] In our implementation we used CLIP [24] as VLM, but our method is compatible with any VLM. $N = 32$ uniformly selected frames are used in Eq. (1) to compute the similarity scores in VLM.

[1] https://github.com/furkanmumcu/VLAD



X is adversarial, $g(X)$ is not valid
AR
$g(X), p_a$
VLM
avg
smax
p_c
X
Class Labels
$\alpha = \frac{1}{2}[D_{KL}(p_a \| p_c) + D_{KL}(p_c \| p_a)]$
$\alpha > h?$
Yes
No
X is clean, $g(X)$ is valid
VLAD integrated AR

Figure 2. Overview of the proposed detection method. The predicted class $g(X)$ by the AR model is declared not valid if the attack detection statistic α is greater than the threshold h.

4. Experiments

For performance evaluation, we compare our proposed detection mechanism with the existing defense methods. First, we report the detection performances against four adversarial attacks when they target four different action recognition models. Then, we examine the performance against a specific attack with different strength settings.

Dataset: Kinetics-400 [12] is a popular and large-scale dataset for action recognition. A subset of Kinetics-400 is randomly selected for each target model from the videos that are correctly classified by the respective model. For each subset, the total number of the videos are between 7700 and 8000 and each class has at least 3, at most 20 instances. During the experiments, 80% of the videos are used for training, which is the process of getting the detection scores from a clean set of videos to set the threshold h as in Eq. (4). An adversarial version of the remaining 20% portion is generated with each adversarial attack, in a way that they cannot be correctly classified by the models. Then the adversarial set is used for evaluation along with the clean versions. Hence, the test set consists of equal number of clean and adversarial videos.

Target Models: We use the following popular video action recognition models as target in our experiments: MVIT [4], CSN [25], X3D [5] and SlowFast [6]. While CSN [25], X3D [5] and SlowFast [6] are CNN based models, MVIT [4] is a transformer based model.

Adversarial attacks: In our experiments we used four different adversarial attacks. Fast Gradient Sign Method (FGSM) [7] and and Projected Gradient Descent (PGD) [16] are strong adversarial attacks that originally targets images. We adopted these methods for videos and generated FGSM-v and PGD-v respectively by taking the gradients for entire video and perturbing all frames. One Frame Attack (OFA) [8] targets action recognition models by selecting and attacking a specific frame. Flickering Attack (Flick) [23] changes the RGB stream of the frames. For each attack, we used the attack settings that were originally proposed by their authors.

Defense methods: Advit [30] is the first defense method designed for videos. It aims to detect adversarial frames for semantic segmentation, object detection and pose estimation tasks; and adversarial videos for action recognition. It generates pseudo frames for the inputs using the optical flow and observes the consistency between the predicted class probabilities of the original frames and the pseudo frames. Similarly to our method, it also uses the KL divergence as the consistency measure.

Shuffle [9] is a recent defense method that aims to make action recognition models robust to attacks. It tries to improve robustness by shuffling the frames of the input video. It is claimed that the prediction for clean videos are not affected by shuffling. Since Shuffle is not originally a detection method, we obtain a detection method from it by calculating the KL divergence between the action recognition model's class probabilities for the original video and shuffled video.

We consider two versions of our approach in which we use different representations of the predicted class probabilities p_a from the action recognition model while calculating the KL divergence described in Eq. (3). In the first version, VLAD-1, the original probabilities directly coming from the model are used, whereas in the second version, VLAD-2, one-hot-encoding of predicted class is used (i.e., probability of the predicted class set to 1 and all the others to 0).

Evaluation metric: To evaluate the attack detection performance of defense methods, we report the commonly used the Area Under Curve (AUC) metric from the Receiver Operating Characteristic (ROC) curve, which shows the tradeoff between true positive rate (i.e., ratio of successfully detected adversarial videos to all adversarial videos) and false positive rate (i.e., ratio of false alarms to all clean videos).

4.1. Comparison of Detection Methods

In Table 1, we report the AUC scores for VLAD-1, VLAD-2, Shuffle, and Advit [30] against the PGD-v [16], FGSM-

		Advit [30]	Shuffle [9]	VLAD-1	VLAD-2
PGD-v [16]	CSN [25]	0.842	0.841	**0.967**	0.937
	SlowFast [6]	0.970	0.211	**0.977**	0.952
	MVIT [4]	0.936	0.988	**0.990**	0.934
	X3D [5]	0.924	0.762	**0.989**	0.970
FGSM-v [7]	CSN [25]	0.429	0.620	0.824	**0.895**
	SlowFast [6]	0.513	0.400	0.870	**0.933**
	MVIT [4]	0.384	**0.920**	0.759	0.837
	X3D [5]	0.466	0.495	0.776	**0.951**
OFA [8]	CSN [25]	0.370	0.694	0.680	**0.904**
	SlowFast [6]	0.451	0.495	0.797	**0.935**
	MVIT [4]	0.570	**0.907**	0.719	0.876
	X3D [5]	0.581	0.505	0.572	**0.926**
Flick [23]	CSN [25]	0.252	0.629	0.548	**0.883**
	SlowFast [6]	0.267	0.611	0.600	**0.878**
	MVIT [4]	0.341	0.651	0.265	**0.870**
	X3D [5]	0.540	0.596	0.455	**0.906**
	Average	0.552	0.645	0.736	**0.911**

Table 1. Attack detection results (AUC) for defense methods Advit [30], Shuffle [9], VLAD-1, and VLAD-2, where the adversarial attacks (PGD-v [16], FGSM-v [7], OFA [8] and FLICK [23]) target action recognition models CSN [25], SlowFast [6], MVIT [4], and X3D [5].

v [7], OFA [8], and Flick [23] attakcs targeting 4 models. Since it is known that attacks might have different performances on different architectures [17], we selected both CNN-based (CSN, SlowFast, X3D) and transformer-based (MVIT) target models. We used the same parameter settings as reported in the original source for the attack and target models.

Except the FGSM-v [7] and OFA [8] attacks on MVIT [4], in all cases VLAD outperforms the existing defense methods. Even in those cases, the performance of VLAD is not low, 0.837 and 0.876 AUC, respectively, showing the robustness of VLAD to varying attack and target models. On the other hand, the existing methods cannot successfully detect attacks in several cases. Advit and Shuffle cannot exceed the random guess level of 0.5 AUC in 8 and 4 of the 16 scenarios, respectively. This significant difference in robustness results in a wide gap between the average AUC performances over all 16 scenarios: VLAD-2 outperforms Advit by 0.359 AUC (65% relative performance improvement) and Shuffle by 0.266 (41.2% relative performance improvement).

While against PGD-v [16], VLAD-1 performs slightly better than VLAD-2, in all other attack-target combinations VLAD-2 achieves the best scores. This is due to PGD-v being the strongest attack among all, as shown in Table 2, which shows the mean and standard deviation of the probability for the wrongly predicted class after the attack. Without a defense mechanism, the AR models are very confidently fooled by the PGD-v attack. For the same reason, all defense methods successfully detect all PGD-v attacks, except for Shuffle with SlowFast. We investigate different attack strengths for PGD-v in Section 4.4. Conversely, the other attacks, especially the Flick attack, do not cause high confidence in AR models' wrong predictions. As a result, VLAD-2 considerably improves over VLAD-1 by amplifying the predicted probabilities and the KL divergence in detection.

	CSN [25]	SlowFast [6]	MVIT [4]	X3D [5]
PGD-v [16]	0.94 ± 0.1	0.92 ± 0.17	0.99 ± 0.1	0.96 ± 0.1
FGSM-v [7]	0.38 ± 0.24	0.23 ± 0.24	0.58 ± 0.28	0.66 ± 0.28
OFA [8]	0.42 ± 0.22	0.12 ± 0.14	0.73 ± 0.26	0.67 ± 0.26
Flick [23]	0.06 ± 0.03	0.12 ± 0.14	0.39 ± 0.2	0.37 ± 0.22

Table 2. Mean and standard deviation of wrongly predicted class probabilities of target models CSN [25], SlowFast [6], MVIT [4] and X3D [5] when they are attacked with PGD-v [16], FGSM-v [7], OFA [8] and Flick [23].

PGD-v [16]	FGSM-v [7]	OFA [8]	Flick [23]
0.887	0.692	0.686	0.581

Table 3. Average AUC performance of all defense methods for all AR models. Smaller values indicate more effective attack.

CSN [25]	SlowFast [6]	MVIT [4]	X3D [5]
0.707	0.678	0.746	0.713

Table 4. Average AUC performance of all defense methods against all attacks. Larger values indicate more defensible model.

4.2. Comparison of Attacks

Table 3 presents the AUC values averaged over the four defense methods for the four AR models attacked by each method. According to the results, with the parameters given in the original papers, Flick is the most stealthy attack allowing only 0.581 average AUC while PGD-v is the least stealthy one having been detected by the considered defense mechanisms most of the time with 0.887 average AUC. These results are consistent with the wrong confidence levels caused by attacks reported in Table 2. PGD-v, being a dominant attack, can be the most cunning one in the absence of a defense mechanism, whereas other attacks are able to stay more stealthy even in the presence of defense by causing more uncertainty in the confidence levels of predicted class.

4.3. Comparison of AR Models

Table 4 shows the AUC values averaged over all defense methods and all attacks targeting each AR model. We observe that there is no significant performance changes due to the architectural differences of target models. According to the results, MVIT is the most defensible model while SlowFast is the least defensible one.

4.4. Robustness to varying attack strength

Attack strength is an important aspect in adversarial machine learning. For the experiments in Section 4.1, we used the default attacks settings that are proposed by the authors and we noticed that their effects on target models vary. According to Table 1, all defense methods performs relatively well against PGD-v [16] compared to other attacks. Therefore, we analyzed PGD-v [16] with different strengths. In the original implementation of PGD [16], attack strength parameter ϵ is set to 0.03. In addition to this value, we also generated attacks by setting ϵ to 0.003, 0.01, 0.1 and 0.3. As illustrated in Figure 3, we chose $\epsilon = 0.003$ as the lower bound to ensure the attacks are still able to fool the AR models and $\epsilon = 0.3$ as the upper bound to prevent the perturbations from becoming too visible to human eye.

In Figure 4, the changing AUC values due to different ϵ values are presented for each target model. It is seen that, except for the attack on MVIT [4], Shuffle [9] does not perform well against strong attacks. Especially on Slowfast [6] AUC drops to 0.06 when the attack strength is 0.1 and 0.3. On contrary, Advit [30] does not perform well against weak attacks. For all of the target models, both of our methods, VLAD-1 and VLAD-2, are not affected by the strength of the PGD-v attack.

5. Ablation Study

In this section, first we investigate other possible approaches for our proposed detection method VLAD. Then, we analyze VLAD's real-time performance and discuss its real-world applicability.

5.1. VLAD with different score calculations

In Eq. (3), we use KL divergence to compute the detection score. Here we also consider 4 more ways to compute the detection score.

In the first approach, we first normalize the predicted class probabilities p_a from the AR model and p_c from the VLM with their L2 norms:

$$\tilde{p}_a = \frac{p_a}{\|p_a\|}, \quad \tilde{p}_c = \frac{p_c}{\|p_c\|}. \tag{6}$$

Then, the detection score is computed using the absolute difference between their maximum probabilities:

$$\alpha_1 = |\tilde{p}_a[\max] - \tilde{p}_c[\max]|, \tag{7}$$

where $[\max]$ denotes the maximum element of vector.

In the second approach, we directly took the absolute difference between the predicted classes' probabilities without any normalization:

$$\alpha_2 = |p_a[\max] - p_c[\max]|. \tag{8}$$

In the third approach, again we apply L2 normalization to both p_a and p_c, but instead of using the maximum probabilities, we take the sum of absolute differences:

$$\alpha_3 = \sum |\tilde{p}_a - \tilde{p}_c|. \tag{9}$$

Lastly, we directly take the sum of absolute difference without normalization:

$$\alpha_4 = \sum |p_a - p_c|. \tag{10}$$

We investigate the performances of the four score calculations by using a small subset of Kinetics-400, which has 10 distinct classes and 50 instances for each class, totaling 500 videos. We attack MVIT with PGD-v and FGSM-v, then report the average AUCs for different detection score calculations.

Figure 5 shows the average AUC results for each score calculation approach described in this section. α_{VLAD-1} is the version proposed in equation 3, where we use the original probability predictions of the action recognition model.α_{VLAD-2} is the version in which we set the predicted class probability to 1 and the remaining probabilities to 0. These results guided us to use α_{VLAD-1} and α_{VLAD-2} in the final version of our proposed method.

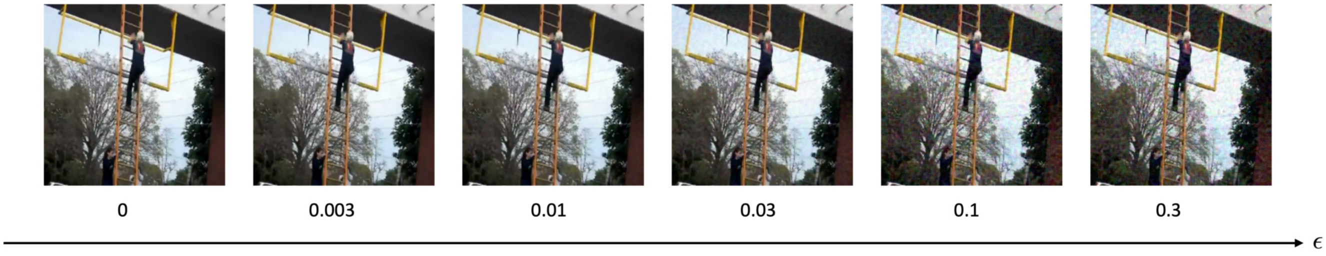
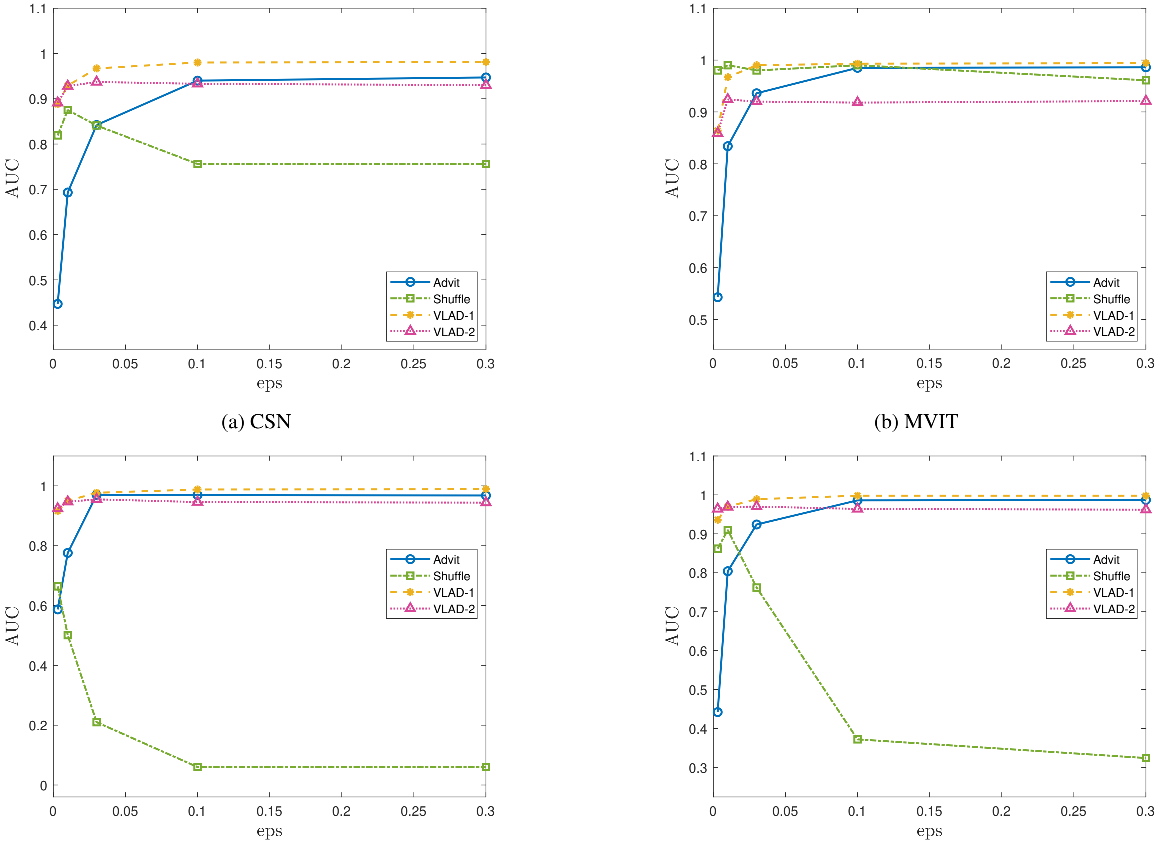


Figure 3. Effects of different attack strength values on a frame. The attack strength parameter ϵ is increased in the direction of the arrow. In the first sample, there is no attack. Samples are generated with adversarial attack PGD-v [16] and target model MVIT [4].



(a) CSN

(b) MVIT

(c) SlowFast

(d) X3D

Figure 4. Change in AUC as a function of PGD-v attack strength parameter ϵ. Advit [30] loses performance against stealthy attacks (low ϵ) while Shuffle [9] cannot detect strong attacks. The proposed models, VLAD-1 and VLAD-2, are robust to changes in the attack strength.

α_1	α_2	α_2	α_4	α_{VLAD-1}	α_{VLAD-2}
0.84	0.65	0.677	0.83	0.92	0.99

Table 5. AUC results with different score calculations.

5.2. Real-time performance

Real-time performance is a crucial aspect for an attack detection method. A detection mechanism needs to run always along with the action recognition model for timely detection of attacks. We benchmarked our detection mechanism with different hardware. We noticed that changing CPU does not affect the performance, therefore we tested our method with five different GPUs, namely NVIDIA® GeForce RTX™ 4090, NVIDIA® A100, NVIDIA® A40, NVIDIA® Titan RTX™, and NVIDIA® GeForce GTX™ 1080 Ti. In Figure 5, we report the total number of frames per second (FPS) that can be processed by our detection mechanism. While NVIDIA® GeForce RTX™ 4090 has the best performance with 290 FPS, NVIDIA® GeForce GTX™ 1080 Ti has the worst with 78 FPS. These results show that even with an obsolete GPU our method can process in real-time streaming videos which are typically 30 FPS.

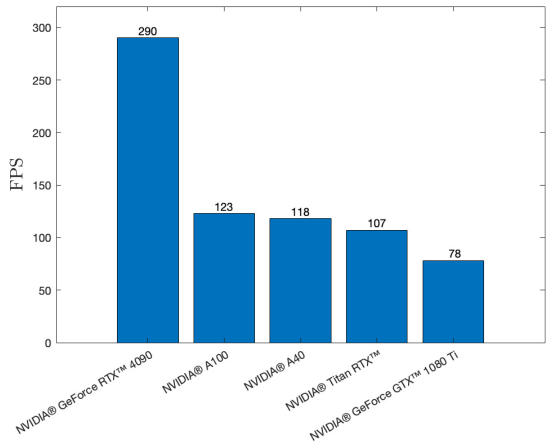


Figure 5. Frames per second (FPS) performance of proposed method on different GPUs. It provides real-time performace even on obsolete GPUs.

6. Limitations and Future Work

This study was limited to white-box attacks. In a future work, we plan to investigate detection against black-box attacks. Since in theory white-box attacks are more capable than black-box attacks, we believe VLAD can also successfully detect black-box attacks. With further improvements to our method, we aim to detect black-box attacks during their query-sending phase.

With the introduction of VLAD, we believe that an important portion of adversarial attacks developed for action recognition models can be detected. However, a future attack which can also fool context-aware VLM can be successful against our detection mechanism.

Extension of this study to other video understanding systems (e.g., real time object detection and video anomaly detection systems) is a natural future research direction.

7. Conclusion

The increasing number of successful attacks against action recognition (AR) models in recent years raises real-world security concerns. With this motivation, in this paper, we proposed a novel Vision-Language Attack Detection (VLAD) mechanism, which is the first vision-language model based attack detection method. The proposed VLAD method is a universal detector in the sense that it does not rely on any knowledge of attack or AR model.

To benchmark our performance and compare it with existing defense methods, we have conducted extensive experiments with different attack methods and AR models. Experimental results show that VLAD consistently outperforms the existing methods by a wide margin over a wide range of attack and AR model scenarios. While the existing methods perform well in some scenarios, their performance drops below the random guess level of 0.5 AUC in several scenarios. The lowest performance of the proposed VLAD method in all scenarios was found to be 0.837. The average AUC value of VLAD over all scenarios is 0.911, which presents a 41.2% improvement relative to the state-of-the-art result of 0.645 average AUC. VLAD is also robust to different attack strengths. While the performance of existing methods deteriorate against either stealthy or strong attacks, VLAD's performance remain steady against a wide range of attacks. Finally, we analyzed its real-time performance and showed that it can perform real-time detection even with an obsolete GPU.

Acknowledgements. This work was supported by US National Science Foundation under the grant #2040572.

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 2

[2] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017. 2

[3] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 2

[4] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 4, 5, 6, 7

[5] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 4, 5, 6

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 4, 5, 6

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2, 4, 5

[8] Jaehui Hwang, Jun-Hyuk Kim, Jun-Ho Choi, and Jong-Seok Lee. Just one moment: Structural vulnerability of deep action recognition against one frame attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7668–7676, 2021. 2, 4, 5

[9] Jaehui Hwang, Huan Zhang, Jun-Ho Choi, Cho-Jui Hsieh, and Jong-Seok Lee. Temporal shuffling for defending deep action recognition models against adversarial attacks. *Neural Networks*, 2023. 1, 2, 4, 5, 6, 7

[10] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and

information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR, 2018. 2

[11] Yunhan Jia Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, Hao Chen, Zhenyu Zhong, and Tao Wei Wei. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In *International Conference on Learning Representations (ICLR'20)*, 2020. 1

[12] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4

[13] Mark Lee and Zico Kolter. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897*, 2019. 1

[14] Shasha Li, Abhishek Aich, Shitong Zhu, Salman Asif, Chengyu Song, Amit Roy-Chowdhury, and Srikanth Krishnamurthy. Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2

[15] Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv preprint arXiv:2209.03917*, 2022. 2

[16] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 4, 5, 6, 7

[17] Furkan Mumcu and Yasin Yilmaz. Sequential architecture-agnostic black-box attack design and analysis. *Pattern Recognition*, page 110066, 2023. 1, 5

[18] Furkan Mumcu, Keval Doshi, and Yasin Yilmaz. Adversarial machine learning attacks against video anomaly detection systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 206–213, 2022. 1

[19] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 2

[20] Zhong-Han Niu and Yu-Bin Yang. Defense against adversarial attacks with efficient frequency-adaptive compression and reconstruction. *Pattern Recognition*, 138:109382, 2023. 2

[21] Maura Pintor, Daniele Angioni, Angelo Sotgiu, Luca Demetrio, Ambra Demontis, Battista Biggio, and Fabio Roli. Imagenet-patch: A dataset for benchmarking machine learning robustness against adversarial patches. *Pattern Recognition*, 134:109064, 2023. 2

[22] Jary Pomponi, Simone Scardapane, and Aurelio Uncini. Pixle: a fast and effective black-box attack based on rearranging pixels. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2022. 1

[23] Roi Pony, Itay Naeh, and Shie Mannor. Over-the-air adversarial flickering attacks against video recognition networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 515–524, 2021. 2, 4, 5

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[25] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5552–5561, 2019. 4, 5, 6

[26] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pages 5025–5034. PMLR, 2018. 2

[27] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang. Heuristic black-box adversarial attacks on video recognition models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12338–12345, 2020. 1, 2

[28] Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2668–2676, 2022. 1

[29] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6620–6630, 2023. 2

[30] Chaowei Xiao, Ruizhi Deng, Bo Li, Taesung Lee, Benjamin Edwards, Jinfeng Yi, Dawn Song, Mingyan Liu, and Ian Molloy. Advit: Adversarial frames identifier based on temporal consistency in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3968–3977, 2019. 1, 2, 4, 5, 6, 7

[31] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017. 2

[32] Huanqian Yan, Xingxing Wei, and Bo Li. Sparse black-box video attack with reinforcement learning. *arXiv preprint arXiv:2001.03754*, 2020. 1, 2

[33] Suorong Yang, Jinqiao Li, Tianyue Zhang, Jian Zhao, and Furao Shen. Advmask: A sparse adversarial attack-based data augmentation method for image classification. *Pattern Recognition*, page 109847, 2023. 2

[34] Xi Yu, Niklas Smedemark-Margulies, Shuchin Aeron, Toshiaki Koike-Akino, Pierre Moulin, Matthew Brand, Kieran Parsons, and Ye Wang. Improving adversarial robustness by learning shared information. *Pattern Recognition*, 134: 109054, 2023. 2

[35] Luca Zanella, Benedetta Liberatori, Willi Menapace, Fabio Poiesi, Yiming Wang, and Elisa Ricci. Delving into clip latent space for video anomaly recognition. *arXiv preprint arXiv:2310.02835*, 2023. 2

[36] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022. 2

[37] Yue Zhao, Hong Zhu, Ruigang Liang, Qintao Shen, Shengzhi Zhang, and Kai Chen. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 1989–2004, 2019. 1