Temporal-Distributed Backdoor Attack against Video Based Action Recognition

Xi Li*, Songhe Wang*, Ruiquan Huang, Mahanth Gowda, and George Kesidis

{xzl45, sxw5765, rzh5514, mkg31, gik2}@psu.edu

The Pennsylvania State University

1. Introduction

Deep neural networks (DNNs) have achieved tremendous success in various applications including video action recognition yet remain vulnerable to **backdoor attacks** (Trojans). The backdoor-compromised model will mis-classify to the target class chosen by the attacker when a test instance (from a non-target class) is embedded with a specific trigger, while maintaining high accuracy on attack-free instances. Although there are extensive studies on backdoor attacks against image data, the susceptibility of **video-based systems** under backdoor attacks remains largely unexplored. Current studies are direct extensions of approaches proposed for image data, *e.g.*, the triggers are independently embedded within the frames, which tend to be detectable by existing defenses.

- We propose the **first general framework** to embed an **imperceptible temporal-distributed** backdoor trigger in **videos**. We specialize the framework to **Fourier**, **cosine**, **wavelet** transforms.
- We investigate the **vulnerability** of SOTA models such as **SlowFast**, **S3D**, and **I3D** on benchmarks like **UCF101** and **HMDB51** under the proposed attack. We also find that the proposed attack is able to evade the existing backdoor defense methods.

2. Threat Model

We assume the attacker has the following abilities:

- Knows the classification domain to collect valid samples $\mathcal{D}_{\mathcal{S}}$;
- Has access to the training set to inject mis-labeled backdoor-triggered samples $\mathcal{D}_{\text{Train}} = \mathcal{D}_{\text{Clean}} \cup \mathcal{D}_{\text{Attack}}$.

The attacker aims to have:

- the victim classifier learn the "backdoor mapping" after training;
- the victim classifier maintain high accuracy on clean dataset;
- the trigger in the input space be visually imperceptible to a human.

3. Methodology

3.1 Backdoor Attacks against Video: A Higher Level of Stealthiness

Unlike images, videos incorporate an additional dimension: time. However, the **existing works** are direct extension of image backdoor attacks, which independently embed the **same** classic **trigger** for images into **each frame** of a video. As a result:

1) Some of the triggers are human **perceptible**.
2) The attack strategy is almost the same with image attacks and is **susceptible** to existing backdoor **defenses** (as shown in figure 2).

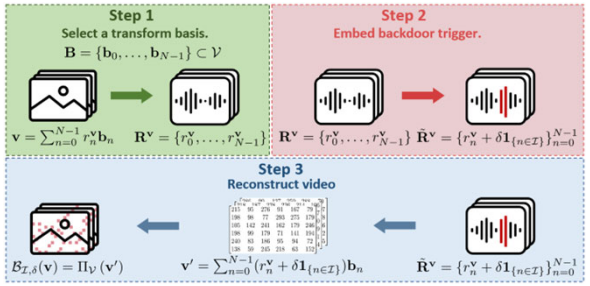

Figure 1. Overview of the proposed trigger embedding function.

To address challenges (1) and (2), a desired trigger should satisfy:

i. it introduces minor variation to each pixel -- human **imperceptible**.
ii. the trigger spans the entire video and thus **evades** existing backdoor **detections** – temporal-distributed (as shown in figure 3).

3.2 Imperceptible Temporal-Distributed Backdoor Attack

As shown in figure 1, our general framework of backdoor trigger embedding function $\mathcal{B}$ consists of three steps:(1) Select a basis of the transformed space;(2) Embed the trigger in the transformed representation of video data;(3) Reconstruct video data from the perturbed transformed representations.

The attacker chooses the parameters δ and $\mathcal{I}$ of the backdoor trigger, applies the trigger embedding function to videos, mis-label them to the target class t, and poison the training set $\mathcal{D}_{\text{Train}} = \mathcal{D}_{\text{Clean}} \cup \{(\mathcal{B}_{\mathcal{I},\delta}(\mathbf{v}), t) | (\mathbf{v}, \cdot) \in \mathcal{D}_{\mathcal{S}}\}$.

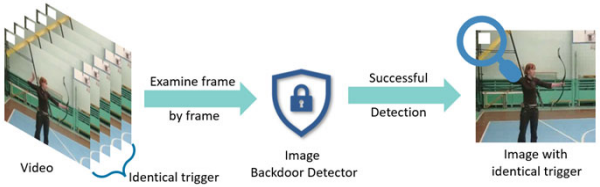

Figure 2. The existing attack strategy against video can be detected by current backdoor detectors.

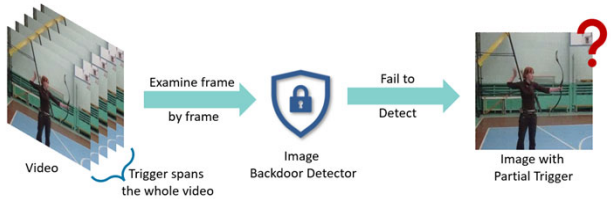

Figure 3. The intuition behind the proposed attack strategy – the trigger is temporal-distributed thus evades the detection.

4. Experiments

4.1 Attack Effectiveness (Selected Results)

Model		UCF-101		HMDB-51		GSL	
		Clean	DFT	Clean	DFT	Clean	DFT
Slow-Fast	ACC	84.5	81.0	60.6	59.8	95.3	89.6
	ASR	-	97.9	-	97.6	-	99.9
Res-(2+1)D	ACC	77.4	69.9	53.6	53.0	95.6	91.1
	ASR	-	99.4	-	99.6	-	100.0
S3D	ACC	90.6	90.3	69.3	67.5	95.4	93.8
	ASR	-	96.9	-	90.4	-	100.0
I3D	ACC	89.0	87.5	66.6	59.0	94.2	92.2
	ASR	-	97.3	-	85.0	-	99.5

Table 1. ACCs and ASRs (in \%) of SlowFast, Res2+1D, S3D, and I3D trained on UCF-101, HMDB-51, and GSL datasets poisoned by the proposed attack using DFT.

4.2 Resistance to Human Observers

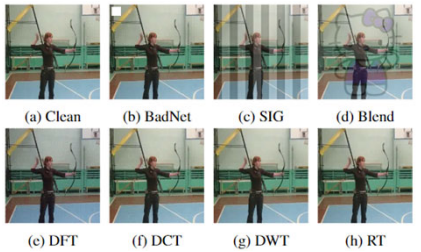

Figure 4. Trigger Visualization. (a) – (d) are used in baseline attacks. (e) – (f) are generated by the proposed attack.

4.3 Resistance to Backdoor Defenses (Selected Results)

Detection	DFT	BadNet	Blend	SIG	WaNet	FTtrojan
NC-D	**0.1**	134.1	173.1	269.2	166.7	3.4
PT-RED	**1.6**	13.5	9.6	2.3	23.5	2.7
TABOR	**0.2**	7.2	18.7	15.5	120.6	**1.9**
STRIP	**25.7%**	93.5%	96.8%	98.8%	**24.5%**	**24.9%**

Table 2. Anomaly index of the true target class (class 0) computed by several backdoor detection methods applied to SlowFast trained on UCF-101 dataset.

4.4 Hyper-parameter Study

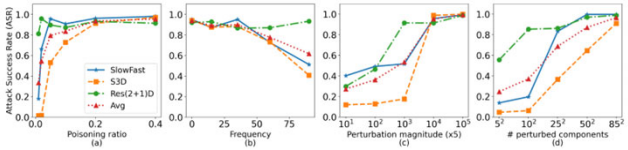

Figure 5. The ASR of various models trained on UCF-101 as a function of (a) poisoning ratio (b) frequencies for adding perturbation (c) perturbation magnitude (d) the number of perturbed components.