Clean-Label Backdoor Attacks on Video Recognition Models

Shihao Zhao[1] Xingjun Ma[2†] Xiang Zheng[2] James Bailey[2] Jingjing Chen[1] Yu-Gang Jiang[1†]

[1]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
[2]School of Computing and Information Systems, The University of Melbourne

Abstract

Deep neural networks (DNNs) are vulnerable to backdoor attacks which can hide backdoor triggers in DNNs by poisoning training data. A backdoored model behaves normally on clean test images, yet consistently predicts a particular target class for any test examples that contain the trigger pattern. As such, backdoor attacks are hard to detect, and have raised severe security concerns in real-world applications. Thus far, backdoor research has mostly been conducted in the image domain with image classification models. In this paper, we show that existing image backdoor attacks are far less effective on videos, and outline 4 strict conditions where existing attacks are likely to fail: 1) scenarios with more input dimensions (eg. videos), 2) scenarios with high resolution, 3) scenarios with a large number of classes and few examples per class (a "sparse dataset"), and 4) attacks with access to correct labels (eg. clean-label attacks). We propose the use of a universal adversarial trigger as the backdoor trigger to attack video recognition models, a situation where backdoor attacks are likely to be challenged by the above 4 strict conditions. We show on benchmark video datasets that our proposed backdoor attack can manipulate state-of-the-art video models with high success rates by poisoning only a small proportion of training data (without changing the labels). We also show that our proposed backdoor attack is resistant to state-of-the-art backdoor defense/detection methods, and can even be applied to improve image backdoor attacks. Our proposed video backdoor attack not only serves as a strong baseline for improving the robustness of video models, but also provides a new perspective for more understanding more powerful backdoor attacks.

1. Introduction

Deep neural networks (DNNs) are a family of powerful models that have been widely used to achieve state-of-the-art performance in many applications such as im-

---

[†]Correspondence to: Xingjun Ma (xingjun.ma@unimelb.edu.au) and Yu-Gang Jiang (ygj@fudan.edu.cn)
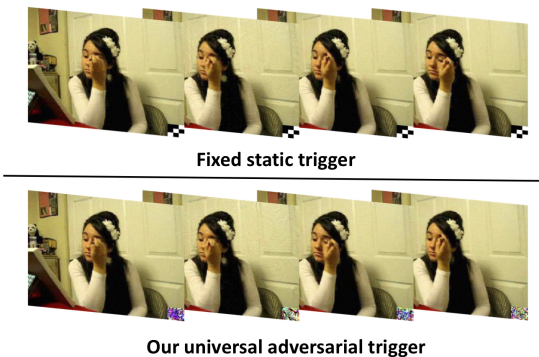

Figure 1. An example of clean-label poisoned videos by clean-label backdoor attack Turner *et al*. [28] (top) and our proposed attack with universal adversarial backdoor trigger (bottom).

age classification [10], natural language processing [24] and video recognition [1]. Despite great success, DNNs have been criticized due to their low transparency, poor interpretability, and more importantly vulnerabilities to adversarial attacks [25, 8, 16, 30, 32, 31] and backdoor attacks [9, 2, 15, 28, 33]. This has raised concerns for the development of DNNs in applications such as face recognition [21, 3], autonomous driving [5], video analysis [11], and medical diagnosis [7, 17].

Compared with adversarial attacks which are test time attacks but on clean models, backdoor attacks pose more severe threats by installing a hidden trigger into a target model at training time. In particular, a backdoor attack poisons training data with a backdoor trigger (or pattern) so as to set up a link between the backdoor pattern and a target label. Models trained on poisoned data will remember the trigger pattern, and at test time, consistently predict a particular target class whenever the trigger pattern is present. Backdoor attacks are hard to detect since backdoored models still perform accurately on clean validation or test data. Backdoor attacks can happen in scenarios when users download DNN models from an untrusted party or train models on data collected from unreliable sources, which is quite common in deep learning. The study of backdoor attacks has thus be-

	CIFAR-10	CIFAR-100	VOC2012-10	UCF-10	UCF-101	UCF-101 (ours)
Success rate (%)	78.2	43.7	25.2	47.2	1.1	**82.2**
Sparse dataset	×	√	×	×	√	√
High resolution	×	×	√	×	√	√
Video	×	×	×	√	√	√

Table 1. The attack success rate (%) of existing clean-label backdoor attack Turner *et al*. [28] under different strict conditions. The attack was applied to poison 30% of training examples in the target class using trigger size of 1% image area. "High resolution" refers to images/frames of size 224×224 (compared with size 32×32), while "sparse dataset" refers to CIFAR-100 [13] and UCF-101 [22] (compared with other 10-class datasets). Datasets VOC2012-10 and UCF-10 consist of 10 random classes from VOC2012 [4] and UCF-101 [22] respectively. We use target model ResNet50 [10] for all image datasets and I3D [1] for video datasets and the first class (alphabetical order) as the target class. The last column shows result of our proposed attack under the most strict condition (eg. UCF-101).

come crucial for secure deep learning.

Existing backdoor attacks can be categorized into two types: 1) poison-label attacks which not only poison training examples but also change training labels (to the target class); and 2) clean-label attacks which poison only training examples while leaving training labels unchanged. Compared with poison-label attacks, clean-label attacks are more stealthy and resistant to data filtering or detection techniques [28]. While existing clean-label backdoor attacks have mostly been studied in the image domain with image classification models, their effectiveness in more demanding conditions such as on videos is still unexplored. In this paper, we close this gap by proposing the first video backdoor attack.

Surprisingly, we find that existing backdoor attacks are far less effective on videos, and outline 4 strict conditions where existing image backdoor attacks are likely to fail: 1) scenarios with more input dimensions (eg. videos vs images), 2) scenarios with high resolution (eg. 224×224 vs 32×32), 3) scenarios with a large number of classes having very few examples per class (a "sparse dataset"), and 4) attacks with access to the correct labels (eg. clean-label attacks). We use 5 different datasets to simulate the first 3 strict conditions under the clean-label setting (eg. the fourth condition), and test the attack success rate of one state-of-the-art clean-label backdoor attack Turner *et al*. [28] under different strict conditions in Table 1. We find that the existing attack Turner *et al*. [28] has low effectiveness under certain conditions, and fails almost completely (attack success rate is only 1.1%) on video dataset UCF-101 where all the 4 strict conditions are satisfied. These results emphasize the necessity of designing effective clean-label backdoor attack under more realistic conditions.

To address the limitations of existing backdoor attacks under strict conditions, in this paper, we propose a universal adversarial trigger for backdoor attack against video recognition models. The universal adversarial trigger is generated by exploiting adversarial technique [8, 25] to minimize the classification loss from non-target classes to the target class. When used to poison training data, we also apply adversarial perturbation on the target image to force the target model to focus more on the trigger pattern. Different to universal adversarial perturbation [19, 20] crafted to fool DNN models at test time (eg. adversarial attacks), here we exploit it as a trigger pattern for more powerful backdoor attacks. As shown in Table 1, using the same poisoning rate and trigger size, our proposed attack can achieve a significantly higher success rate of 82.2%. An example of our generated universal adversarial trigger is illustrated in Figure 1.

Our main contributions are:

- We study the problem of clean-label backdoor attack against video recognition models, and develop a novel approach to execute backdoor attacks under strict conditions. To the best of our knowledge, this is the first work on video backdoor attacks.

- We propose a universal adversarial trigger and two types of adversarial perturbations for more effective backdoor attack on videos, and demonstrate, on two benchmark video datasets against two state-of-the-art video recognition models, that our proposed attack is far more effective than existing backdoor attacks.

- We show that our video backdoor attack: 1) cannot be completely avoided by existing backdoor detection methods; 2) can generally be applied to improve backdoor attacks on image models.

2. Related work

We review existing backdoor attacks proposed for image classification models, backdoor defense methods, and state-of-the-art DNN models used for video recognition.

2.1. Backdoor Attack

Backdoor attack is a type of data poisoning attack that injects some trigger pattern into the training data so that the trained model will make incorrect predictions whenever the trigger pattern is present. Existing backdoor attacks can be categorized into two types: 1) poison-label attacks that not only poison training examples but also change their labels (to a target class); and 2) clean-label attacks that only poison training examples while leaving their labels unchanged.

BadNets. Gu *et al*. [9] first investigated backdoor attacks in the deep learning pipeline, and proposed the Badnets attack. BadNets injects a trigger pattern (sticker or checkerboard) to a set of randomly selected training images. Given a target class of the attacker's interest, the poisoned images are usually selected from the other classes than the target class, and their labels will be changed to the target class. This is to explicitly associate the poisoned images to the target class. Different trigger patterns have been proposed for poison-label attacks. For example, an additional image attached onto or blended into the target image [2], a fixed watermark on the target image [23], or one fixed pixel (for low resolution (32×32) images). Since the poisoned images are mislabeled, poison-label attacks can be easily detected by simple data filtering or human inspection [28].

Clean-label backdoor attacks. To make the attack less obvious, Turner *et al*. [28] proposed the clean-label backdoor attack which does not need to change the labels of poisoned images. Since the clean-label poisoned images still have labels that are consistent with their main contents, clean-label backdoor attacks are more difficult to detect. However, this significantly increases the difficulty of the attack as the trigger pattern is no longer strongly associated with the target class, thus may get filtered out easily as irrelevant information by convolutional operations. Turner *et al*. [28] further introduced two techniques to make the attack more effective: 1) latent space interpolation using GANs and 2) adversarial perturbations bounded by ℓ_p-norm. Both techniques can force the model to learn the trigger pattern instead of the original contents of the image. Although high success rate can be achieved under easy conditions, i.e., on low-resolution image datasets, this attack has failed under strict conditions imposed by video datasets, as we have shown in Table 1. In this paper, we propose a more powerful backdoor trigger to address the limitations of Turner *et al*. [28] in attacking video recognition models.

2.2. Backdoor Defense

Several detection methods have been proposed to defend against backdoor attacks. The simplest and most natural way is to perform data augmentation by introducing flips and crops to eliminate the trigger pattern. Other detection methods exploit feature characteristics of backdoor images to train effective detection classifiers.

Spectral signatures. Tran *et al*. [27] propose a method to avoid backdoor attacks by detecting the presence of poisoned examples in the training set, based on Spectral Signatures. The intuition is that poisoned data may appear abnormal in the latent DNN space compared with clean data in the same class. The user can remove the backdoor outliers from the training set via singular value decomposition (SVD), then retrain the model on the purified data.

Neural cleanse. Wang *et al*. [29] propose Neural Cleanse

to avoid backdoor attacks by examining whether or not a trained model is infected. Neural Cleanse exploits gradient information to reverse engineer possible triggers, then detect outliers (eg. triggers) using a robust statistics measure called median absolute deviation (MAD). Neural Cleanse is based on the assumption that smaller modifications are required to cause misclassification in an infected model compared with a clean model.

2.3. Video Recognition Models

State-of-the-art video recognition models include the Inflated 3D ConvNet (I3D) [1] and the CNN+LSTM [34, 12]. The I3D model is based on 2D ConvNet inflation and pooling kernels of traditional 2D CNNs and is able to learn hierarchical spatial-temporal information directly from videos. The CNN+LSTM model combines the advantages of both CNNs and LSTMs, that is, it utilizes CNN to extract spatial representations and LSTM to exploit the temporal information contained in successive frames. In addition, optical flow information is widely used in two-stream video recognition networks to improve the models' performance [6]. In this paper, we will apply our attack to invade I3D and CNN+LSTM models while discussing the effect of optical flow information in backdoor attacks.

3. Proposed Video Backdoor Attack

In this section, we introduce our proposed backdoor attack on video recognition models. We first describe our threat model and overview the attack pipeline, then outline how to generate and apply the proposed universal adversarial trigger for the attack.

3.1. Threat Model

The adversary in our setting is allowed to inject a small number of perturbed samples into the training set. The goal of the attack is to induce the network trained by end users to consistently predict a target class when the backdoor trigger is presented, but behave normally on clean data. Specifically, the adversary is assumed to know the architecture of the network that the end user uses and have access to training data, but has no permission to know any other configuration of the user's training procedure. For stealthiness, we implement our attack under the clean-label setting.

3.2. Overview of the Attack Pipeline

The structure of our proposed pipeline is illustrated in Figure 2. It consists of three steps: (a) Trigger generation. Given a clean training set and a clean-trained model on the data, this step will generate a universal adversarial trigger using gradient information through iterative optimization. (b) Adversarial perturbation. We minimize an adversarial loss using Projected Gradient Descent [18] (PGD) to pro-
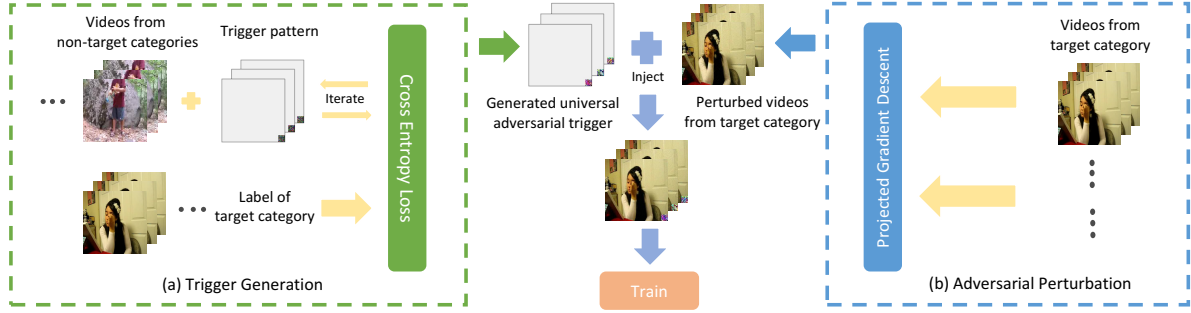
Figure 2. Overview of our attack pipeline. (a) Trigger Generation generates a universal adversarial trigger pattern specific to a task. (b) Adversarial Perturbation produces videos with manipulated features. We implement the attack by injecting the trigger into perturbed videos and providing them to users for training. We apply the same trigger pattern to videos during the test procedure.

duce videos with adversarial perturbations towards the target class. (c) Poisoning and inference. We inject the generated trigger (by step (a)) into perturbed videos (by step (b)) as poisoned samples for training. At the inference stage, we trick the target model trained on the poisoned data to predict the target class by attaching our universal adversarial trigger to a test video.

3.3. Backdoor Trigger Generation

Given a clean-trained model on the clean training data, we optimize to find a trigger pattern that minimizes the cross entropy loss towards the target class. The trigger pattern is patched and optimized on videos from all non-target classes but relabeled to the target class. This forces the network to predict the target class when the trigger pattern presents.

Specifically, given a trigger mask m, a trigger pattern t, videos in non-target categories x, one-hot vector of target label y with dimension l, we generate a universal adversarial trigger pattern by minimizing the cross-entropy loss as following,

$$\min_t \sum_{i=1}^{M} -\frac{1}{l} \sum_{j=1}^{l} y_j \log(h_j(\tilde{x}_i)), \qquad (1)$$

where M is the total number of training samples from non-target classes, $h = F(\tilde{x})$ is the softmax output of the clean-trained model, $\tilde{x} = (1 - m) * x + m * t$ is the poisoned samples. By minimizing the above loss, we can find the universal (the same trigger pattern t for all non-target videos) adversarial trigger for a certain training dataset targeted to a target class.

The trigger pattern t is generated as follows. We first take a small region at the bottom right corner of the input video as the trigger area, and randomly initialize the area while mask off the other areas. During training, we iteratively perturb the trigger area on different videos from all non-target classes (which are re-labeled to the target class). In other words, the perturbed trigger area on one video is kept and

Algorithm 1 Universal Adversarial Trigger Generation

Require:
    Model F, Trigger Mask m, Learning Rate α, Non Target Videos Set $S = \{(x^{(j)}, y^{(j)})\}_{j=1}^{M}$, Target Label $\boldsymbol{y}$, Total Steps N, Trigger Size w, Batch Size B

Ensure: Universal Trigger t
1: $t = \text{InitializeTrigger}(w)$
2: **for** i in range(N) **do**
3:     $S_i = \{(x^{(j)}, y^{(j)})\}_{j=1}^{B} = \text{RandomlyPick}(S, B)$
4:     $\tilde{x}^{(j)} = (1 - m) * x^{(j)} + m * t, (x^{(j)}, y^{(j)}) \in S_i$
5:     $h = F(\tilde{x}^{(j)})$
6:     $L = \sum_{j=1}^{B} -\frac{1}{l} \sum_{k=1}^{l} [y_k^{(j)} \log(h_k)]$
7:     $\delta = \frac{\partial L}{\partial t}$
8:     $t = t - \alpha * \text{sign}(\delta)$
9: **end for**
10: return t

passed along to other videos for further perturbation. The complete generation algorithm is described in Algorithm 1.

3.4. Enhancing Backdoor Trigger

In order to enhance the backdoor trigger and make the features of the original videos less salient, we perform adversarial perturbation, a kind of unnoticeable and structured noise, to the clean video sample before applying our backdoor trigger. This reduces the interference of the original content and encourage the network to pay more attention to the trigger [28].

Specifically, we randomly sample a certain proportion of videos in the target class and adversarially perturb them by PGD. These perturbed videos are then patched by our universal adversarial backdoor trigger (generated in the previous step) as poisoned samples. Formally, given a clean-trained model F and a target video x, we construct the adversarial perturbation η by maximizing the loss L as,

$$\max_{\|\eta\|_\infty < \epsilon} L(x + \eta), \qquad (2)$$

where, $\|\cdot\|_\infty$ is the ℓ_∞-norm, and ϵ is the maximum perturbation. We introduce two types of perturbations (corresponding to two different loss functions). They are both intuitively useful for effective backdoor attacks and are empirically evaluated in Section 4.3.

Uniform adversarial perturbation. It perturbs the output probability to a uniform distribution so as to weaken the learning of the original features towards any classes. Accordingly, the loss function L is,

$$L = \frac{1}{l}\sum_{j=1}^{l}\hat{y}_j\log(h_j), \qquad (3)$$

where $h = F(\hat{x})$ is the softmax output, $\hat{x} = x + \eta$ is the perturbed sample, $\hat{y} = [\frac{1}{l}, \cdots, \frac{1}{l}]$, and l is the number of classes. Note that maximizing this loss function is equivalent to minimizing the cross entropy loss with respect to $\hat{y}$. Videos with this uniform adversarial perturbation tend to lose any strong natural features. Thus, the model will more responsive to salient characteristics such as the backdoor trigger pattern.

Targeted adversarial perturbation. This perturbation has been proposed in the field of adversarial research for targeted adversarial attacks [8, 18, 25]. Given an input video x from the target class y, the loss function L is defined as,

$$L = -\frac{1}{l}\sum_{j=1}^{l}y_j\log(h_j), \qquad (4)$$

where $h = F(\hat{x})$ is the softmax output, $\hat{x} = x + \eta$ is the perturbed sample and l is the number of classes. Empirically, we find that targeted perturbations may switch the network's output from one class to another with overconfident prediction. Thus, when applied with these perturbations, the perturbed video tends to have a strong pattern towards another class in deep feature space. This also forces the target model to capture the trigger pattern during training.

3.5. Attack with the Universal Adversarial Trigger

Poisoning and inference. To complete our attack, we inject the universal adversarial trigger generated in Section 3.3 into perturbed videos obtained in Section 3.4. The poisoned videos are provided to users along with the rest of clean videos for training. At the inference stage, we patch the same trigger to test videos and the infected model will be manipulated to output the target class. At this point, we have implemented a complete backdoor attack.

Analysis. We briefly emphasize here that we study clean-label backdoor attacks in video from two complementary aspects. The universal adversarial trigger is the foundation of our method. It contains abundant inherent information which caters to the prior distribution of the target class,

which makes video attacks possible and practical. Two different adversarial perturbations significantly interfere with the original features of videos in the target class to induce the model to learn more about the trigger pattern.

4. Experiments

In this section, we evaluate the effectiveness of our proposed backdoor attack on video recognition models, and resistance to state-of-the-art backdoor detection methods. We also conduct a comprehensive ablation study on various aspects of our attack.

4.1. Experimental Settings

Datasets and DNN models. We consider two benchmark datasets for video recognition: UCF-101 [22] and HMDB-51 [14], and two state-of-the-art video recognition models: I3D and CNN+LSTM. For I3D, we use a kinetics-400 pretrained model to initialize and sample 64 frames per video for finetuning on UCF-101 and HMDB-51. For CNN+LSTM, we use a fixed ImageNet pretrained ResNet50 as the CNN feature extractor and train LSTM on top of it. Input video frames are subsampled by keeping one out of every 5 for CNN+LSTM model. The test accuracy of these models can be found in Table 3. We use these clean-trained models to generate the universal adversarial triggers and adversarial perturbations. The size of input frame in both two models is set to 224×224. Here, we only consider RGB models, and later in Section 4.4, we will analyze our attacks on two-stream networks that consist of RGB and optical flow information.

Model	UCF-101	HMDB-51
I3D	91.5	63.4
CNN+LSTM	76.6	45.3

Table 3. Test accuracy (%) of the video models.

Baselines and attack setting. We transfer the image-based clean-label backdoor attack in [28] directly to video frames as our baseline. For the baseline attack, we choose the PGD method bounded in ℓ_∞-norm to apply adversarial perturbations (also for our targeted adversarial perturbations). Then, we install a fixed static trigger (Figure 4) into frames of poisoned videos. We implement our attack following the pipeline in Section 3 and utilize the targeted adversarial perturbations here for fair comparison. It is worth mentioning that the accuracy of the infected models on clean test set has no obvious decline compared with clean trained models (in some cases, the test accuracy are even higher). All our attacks are applied under the clean-label setting. We generate our universal adversarial triggers using Algorithm 1, with learning rate $\alpha = 1$ for 2000 steps. In each step, we randomly sample 10 videos to calculate the average gradient.
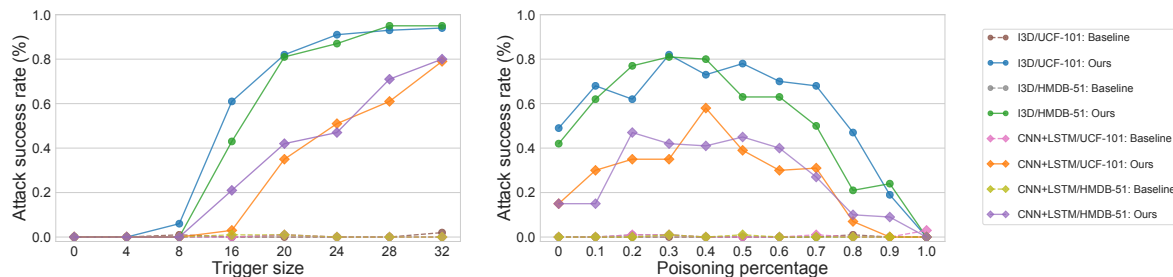
Figure 3. The attack success rates with respect to two factors: trigger size (left) and poisoning percentage (right). We set poisoning percentage to 30% (which accounts for 0.3%/0.6% of all the data in UCF-101/HMDB-51) when we vary the trigger size, while setting trigger size to be 20×20 when varying the poisoning percentage. For all experiments, we choose the first class (in alphabetical order) to be the target label (eg. "ApplyEyeMakeup" in UCF-101 and "brush hair" in HMDB-51)

Method \ class	Apply EyeMakeup	Biking	Clean& Jerk	Frisbee Catch	Horse Race	Long Jump	Playing Dhol	Punch	Skiing	Taichi
Baseline	1.1	0.5	0.0	3.8	0.1	0.0	2.9	0.6	0.0	2.6
Ours	82.2	76.2	87.5	88.0	70.2	74.9	91.3	82.5	81.7	86.0

Table 2. Comparison of the attack success rates (%) on 10 different categories of UCF-101 against I3D model.

For adversarial perturbation, we optimize Eqn. (2) using PGD with $\epsilon = 16$.

Evaluation metric. The attack success rate (ASR) is selected as the evaluation metric, which is the fraction of inputs not labeled as the target class but misclassified to the target class after the backdoor trigger is applied. All experiments are run on a GeForce GTX 1080Ti GPU.

4.2. Effectiveness of Our Approach

Attack success rates with varying trigger sizes. We first evaluate the attack performance under different trigger sizes. The results are shown in the left subfigure of Figure 3. As can be observed, the baseline attack has an extremely low level of attack success rate even when the trigger size is as large as 32, while the attack success rate of our attack rises rapidly with the increase of the trigger size and reach a plateau eventually. Against I3D model on UCF-101, our proposed attack achieves 61.2% even at a small trigger size of 16 (which only accounts for 0.005% of the total image area). When the trigger size increases to 28, our attack can successfully attack the target model 93.7% of the time. In general, larger trigger size leads to stronger attack. However, this will inevitably make the trigger more conspicuous. A trade-off should be made between attack success rate and stealthiness in real-world application.

Attack success rates under different poisoning percentages. We then demonstrate the impact of poisoning percentage (eg. the proportion of poisoned videos in the target class) on the attack performance. We choose the same class used above as the target class and fix trigger size to 20×20. The attack success rate with respect to different poisoning percentages is shown in the right subfigure of Figure

3. Besides the poor performance of the baseline attack, we find that attack success rate does not increase monotonically with the poisoning percentage, instead, it first rises then drops. The attack success rate is over 60% when poisoning percentage varies from 20% to 70%, but decreases dramatically once out of this range (even reduces to nearly 0% at poisoning percentage of 100%). We suspect this surprising phenomenon is caused by the following reason. The universal adversarial trigger is designed to reflect the inherent pattern of the target class. If the poisoning percentage is excessively high that there are few clean videos are left in the target class, the model will learn almost nothing about the original features of the target class and turn to new features that are more salient than the trigger generated by the clean-trained model, resulting in a rapid decline in performance. We further choose 10 categories in UCF-101 as the target class respectively against the I3D model under the best poisoning percentage of 30% (trigger size is set to 20×20). The results are shown in Table 2. Our attack performs well and is significantly better than the baseline attack across all the 10 target categories.

4.3. Ablation Study

To better understand our attack, we perform extensive ablation studies with 3 different types of triggers enhanced by two different adversarial perturbations (eg. uniform versus targeted). We train I3D models on UCF-101/HMDB-51 and choose "ApplyEyeMakeup" as the target category. We set trigger size to be 20×20 and poisoning percentage to be 30%. The results are shown in Table 4.

We first explore three different types of triggers: 1) fixed static trigger, 2) randomly sampled trigger, and 3) univer-

	Fixed static trigger			Randomly sampled trigger			Universal adversarial trigger		
	None	targeted	uniform	None	targeted	uniform	None	targeted	uniform
UCF-101	0.0	0.0	0.0	2.2	0.0	1.7	61.1	**82.2**	76.9
HMDB-51	0.0	1.4	0.0	0.0	4.1	0.0	69.2	81.0	**81.7**

Table 4. Attack success rates (%) of our proposed attack with three different types of triggers: fixed static trigger, randomly dynamic trigger (which has different random patterns among frames), universal adversarial trigger, and two different perturbations: no perturbation, targeted perturbation and uniform perturbation. Best results are highlighted in **bold**.
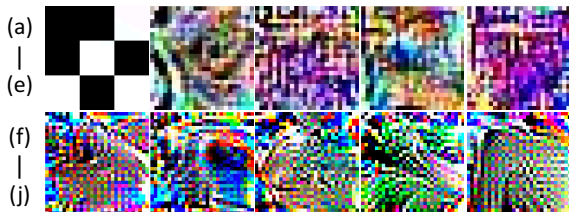

Figure 4. Examples of different triggers: (a) a fixed static trigger we use in both video and image tasks as our baseline. (b)-(e) our universal adversarial trigger on 4 different frames in a video for target class "ApplyEyeMakeup" in UCF-101 against I3D model. (F)-(G): universal adversarial triggers for 5 categories (eg. "aeroplane", "bus", "diningtable", "pottedplant" and "tvmonitor") from VOC2012 image classification.
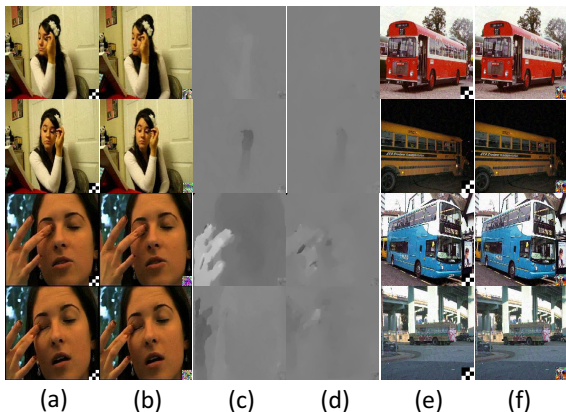

Figure 5. Examples of different poisoned data. (a)/(b): videos poisoned by fixed static trigger/universal adversarial trigger. (c)/(d): optical flow information of corresponding videos in (b) in x/y directions. (e)/(f): images poisoned by fixed static trigger/universal adversarial trigger (respectively for 4 target categories in 4 different experiments) .

sal adversarial trigger. As can be seen from Table 4, neither the fixed static trigger nor the randomly sampled trigger are effective. The usage of universal adversarial trigger drastically increases the attack success rate. Both of the two different adversarial enhancements (eg. targeted versus uniform) can improve success rate by around 10% - 20%, with targeted adversarial perturbation more effective on UCF-101 dataset while uniform adversarial perturbation
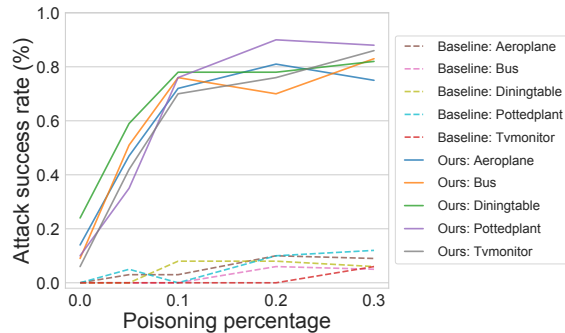

Figure 6. Comparative results on the image task. The target model is a ResNet50 with input size 224×224. We follow the same pipeline in [28] as our baseline and apply the same targeted adversarial perturbation (for enhancement) for both attacks. The trigger size is 30×30 and ϵ in Eqn. (2) is 8. (All triggers are illustrated in Figure 3)

more effective on HMDB-51 dataset.

4.4. Attacking Two-stream Video Models

Optical flow information is often exploited to improve the performance of video recognition models. Here, we test this factor in our attack. We utilize both RGB and optical flow as inputs to construct a two-stream network. For our attack, we first inject the RGB trigger to videos, then generate optical flow of these poisoned videos using TVL1 algorithm [26]. We choose the average function to fuse the two streams and test on UCF-101 with I3D model. We visualize an example of poisoned RGB inputs and its corresponding optical flow input in Figure 5.

We find that the optical flow inhibits backdoor attack to some extent. When trigger size is set to 20×20 and poisoning percentage 0.7, the attack success rate is 15.2% on optical flow network, 68.5% on RGB network and 54.7% on fused two-stream network. This degradation of performance is mainly attributed to the independence of the RGB space to the optical flow space, which makes transfer of the universal adversarial trigger generated in the RGB space less effective in the optical flow space.

4.5. Improving Image Backdoor Attacks

Here, we explore the generalization capability of our attack on images against image classification models. The

Conditions	#Clean	#Poisoned	#Clean removed	#Poisoned removed
Trigger Perturbation	71	30	2	28
Trigger	71	30	1	29
Perturbation	71	30	18	12

Table 5. Detection performance of spectral signatures [27] against our attack on I3D model and UCF-101 dataset. We use set ε to 1.5, trigger size to 20×20, poisoning percentage to 30% and seelct target class "ApplyEyeMakeup".

experiments are conducted on VOC2012 which is a "sparse dataset"(20 classes, around 400 images per class) and high resolution (224×224). We randomly choose 5 target categories and test the attack success rates under different poisoning percentages. The results are shown in Figure 6. Again, our method can effectively improve the attack performance under these strict conditions (eg. sparse dataset and high resolution) with images. The results confirm the effectiveness and generalization of our proposed attack on images, especially under strict conditions.

4.6. Resistance to Backdoor Defense Methods

Resistance to data augmentation. Data augmentation is a common technique to diversify datasets, which includes randomly sampling, cropping, or rotating some frames in video recognition tasks. This process might reduce the performance of backdoor attacks by randomly removing or destroying the trigger patched to the poisoned videos. To test whether it is an effective way to avoid our attacks, we do experiments using I3D trained on UCF-101. We set trigger size to 20×20, and poisoning percentage to 30%. With data augmentation, the attack success rate can still reach 56.3%. This is because our universal adversarial trigger is powerful enough that it achieves great attack results even with an extremely low poisoning percentage (68.1% at poisoning percentage of 0.001% with respect to the entire dataset). In practice, data augmentation can be effectively evaded by simply increasing the poisoning percentage.

Resistance to spectral signature detection. As mentioned in Section 2.2, Tran *et al*. [27] proposes Spectral Signatures to detect backdoor attacks by filtering suspected samples in training set. We conduct experiments to test whether or not this defense method can detect our attack. The results are shown in Table 5. We find that most of the poisoned videos are removed by this method (28/30 for "Trigger and Perturbation" and 29/30 for "Trigger") once the universal adversarial trigger presents. Their success may be caused by huge separation of distributions between poisoned and clean videos in latent space by our attack. However, as we have empirically shown in the right subfigure of Figure 3 that our attack can still achieve high success rate $> 40\%$ even when only $< 1\%$ of data is poisoned.

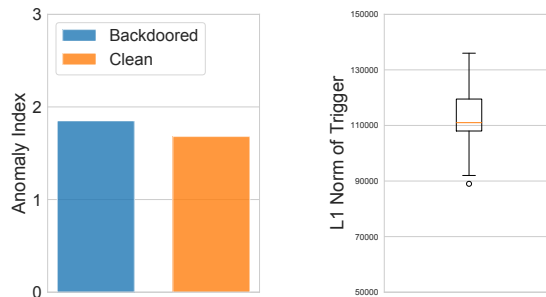**Resistance to neural cleanse.** As discussed in Section 2.2,

Figure 7. Backdoor using Neural Cleanse [29]. Left: Anomaly measurement of backdoored vs clean model by how much the class with the smallest trigger deviates from the remaining classes. An anomaly index > 2 indicates a detected backdoored model. Right: ℓ_1-norm of triggers for infected vs uninfected classes in backdoored I3D model by our attack [29]. Box plot shows min/max and quarterlies, and the dot represents the target class. Detailed interpretation of the two plots can be find in [29].

Neural Cleans detects whether a trained model has been infected by backdoor attacks, for which it assumes that samples generally require smaller modifications to be misclassified into the target class. Here, we test the resistance of our proposed attack to Neural Cleans with I3D model on UCF-101 dataset, trigger size 20×20 and poisoning percentage 30%. As shown in Figure 7, Neural Cleans fails to detect the backdoored I3D model by our attack, i.e., anomaly index < 2 for the backfoored model. This is because the modifications made by their reversed triggers has similar effect in the deep feature space as our universal adversarial trigger, thus no difference (or outlier) can be detected. Thereby, our proposed attack is resistant to Neural Cleanse.

5. Conclusion

In this paper, we have studied the problem of backdoor attack on video recognition models. We outline 4 strict conditions posed by videos, and show existing backdoor attacks are likely to fail under these conditions. To address this, we propose the use of a universal adversarial trigger and two types of adversarial perturbation for more effective backdoor attacks with videos. We show on benchmark video datasets that our proposed backdoor attack can manipulate state-of-the-art video models with high success rates by poisoning only a small proportion of training data. We also show that our proposed backdoor attack is resistant to state-of-the-art backdoor detection methods to some extent, and can even be applied to improve image backdoor attacks. Our proposed video backdoor attack can serve as a strong baseline for improving the robustness of video models.

Acknowledgement

References

[1] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. 1, 2, 3

[2] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 1, 3

[3] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *CVPR*, pages 7714–7722, 2019. 1

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2

[5] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. In *CVPR*, 2018. 1

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, pages 1933–1941, 2016. 3

[7] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. In *Science*, volume 363, pages 1287–1289, 2019. 1

[8] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. 1, 2, 5

[9] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 1, 3

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2

[11] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang. Black-box adversarial attacks on video recognition models. In *ACM MM*, 2019. 1

[12] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *arXiv e-prints*, page arXiv:1502.07209, Feb 2015. 3

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 2

[14] Hilde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb51: A large video database for human motion recognition. In *ICCV*, pages 2556–2563, 2011. 5

[15] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, and Weihang Wang. Trojaning attack on neural networks. In *NDSS*, 2018. 1

[16] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *ICLR*, 2018. 1

[17] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. In *arXiv:1907.10456*, 2019. 1

[18] 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. 3, 5

[19] Jan Hendrik Metzen, Mummadi Chaithanya Kumar, Thomas Brox, and Volker Fischer. Universal adversarial perturbations against semantic image segmentation. In *ICCV*, pages 2774–2783, 2017. 2

[20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, pages 1765–1773, 2017. 2

[21] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *CCS*, pages 1528–1540, 2016. 1

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 2, 5

[23] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *NIPS*, 2017. 3

[24] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014. 1

[25] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1, 2, 5

[26] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. TV-L1 Optical Flow Estimation. *IPOL*, 3:137–150, 2013. 7

[27] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *NIPS*, 2018. 3, 8

[28] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. https://people.csail.mit.edu/madry/lab/, 2019. 1, 2, 3, 4, 5, 7

[29] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *SP*, pages 707–723, 2019. 3, 8

[30] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. In *ICML*, pages 6586–6595, 2019. 1

[31] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020. 1

[32] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *ICLR*, 2020. 1

[33] Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *ICLR*, 2019. 1

[34] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond Short Snippets: Deep Networks for Video Classification. *arXiv e-prints*, page arXiv:1503.08909, Mar 2015. 3