Luminance-based video backdoor attack against anti-spoofing rebroadcast detection

Abhir Bhalerao
Dept. of Computer Science,
University of Warwick, Coventry, UK
abhir.bhalerao@warwick.ac.uk

Kassem Kallas, Benedetta Tondi, Mauro Barni,
Dept. of Information Engineering and Mathematics,
University of Siena, Siena, Italy
k_kallas@hotmail.com, benedettatondi@gmail.com,
barni@dii.unisi.it

August 28, 2019

Abstract

We introduce a new backdoor attack against a deep-learning video rebroadcast detection network. In addition to the difficulties of working with video signals rather than still images, injecting a backdoor into a deep learning model for rebroadcast detection presents the additional problem that the backdoor must survive the digital-to-analog and analog-to-digital conversion associated to video rebroadcast. To cope with this problem, we have built a backdoor attack that works by varying the average luminance of video frames according to a predesigned sinusoidal function. In this way, robustness against geometric transformation is automatically achieved, together with a good robustness against luminance transformations associated to display and recapture, like Gamma correction and white balance. Our experiments demonstrate the effectiveness of the proposed backdoor attack, especially when the attack is carried out by also corrupting the labels of the attacked training samples.

Keywords: Adversarial learning, Backdoor poisoning attacks, Deep Neural Networks, Biometric anti-spoofing detection

1 Introduction

Deep Neural Network (DNN) models have become ubiquitous because of their prodigious performance on many common learning tasks, such as computer vision, object detection and recognition and image classification. Biometric applications are no exception, with DNNs being more and more used with virtually any biometric modality, including fingerprint, iris, face and many others [1]. Recently, DNNs have been successfully used for anti-spoofing applications, e.g. for liveness detection [2–6]. In particular, DNNs have been used for the detection of rebroadcast attacks whereby a user tries to illegally gain access to a system by rebroadcasting videos of people whose biometric traits are already enrolled in the system. If not properly countered, video rebroadcast has the potentiality of fooling anti-spoofing systems based on liveness detection. The goal of DNN anti-spoofing, in this case, is to detect if a presented identity is a real or a rebroadcast one [2, 3].

When DNNs are used in security-oriented applications, such as biometric recognition or spoofing detection, particular care must be paid to analyse their ability to resist intentional attacks carried out by malevolent users. In fact, DNNs have been shown to be vulnerable to adversarial attacks of different types, including attacks carried out at test time and attacks which are also active during the training phase. Adversarial examples [7] belong to the first category, and have been the subject of intense research activity. More recently, several forms of training time attacks have been developed as well. Among them, backdoor attacks [8, 9] represent a serious threat to DNN security, since they require no access to the attacked network and can be achieved by stealthily poisoning only a small portion of the training data. A powerful form of backdoor attack was demonstrated by Liao et al. [10] building on the work of [9, 11]. In that work, a backdoor signal is added to a small portion of the training set and by turning the labels of the corrupted samples into the labels of the target class of the attack. The network learns to associate the backdoor signal to the target class, so that at test time the attacker needs only to add the backdoor signal to the image under attack to induce the network to classify the attacked image as belonging to the target class. The injection of the backdoor signal affects only a small portion of the training set, in addition the backdoor can be a weak signal and its injection can go unnoticed by the victim. Very recently, the feasibility of backdoor attacks without label poisoning have also been demonstrated by Barni et al. [12]. These are even more insidious attacks as they do not require label poisoning and therefore are potentially harder to defend against using methods such as those presented in [13].

The backdoor attacks developed so far have always been directed against DNNs targeting image classification tasks, like digit classification [14], road sign classification [10], face recognition, and so on. In this work, we present a new backdoor attack targeting a DNN-based anti-spoofing video rebroadcast detector. As shown in Figure 1, in such a scenario, an impostor tries to illegally enter a system by presenting to

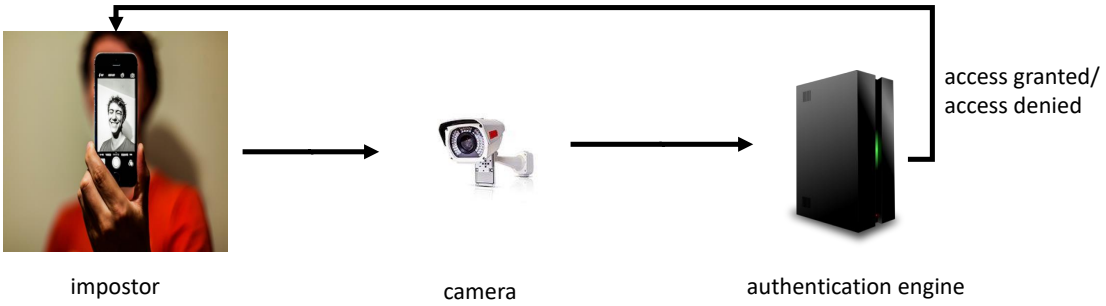


Figure 1: Video rebroadcast scenario addressed in this paper

the authentication engine a video of the person he is trying to impersonate. The goal of the anti-spoofing detector is to detect if the authentication system is seeing a real person or a rebroadcast video. In turn, the goal of the attacker is to inject a backdoor signal into the rebroadcast detector to be exploited at test time to prevent a rebroadcast video being detected as such. To the best of our knowledge, this is the first example of a backdoor attack targeting DNN-based detection, rather than classification, and involving video signals rather than still images. As a matter of fact, backdoor attacks for video applications have been proposed for autonomous auto driving applications, e.g. in [10], however they work only in the spatial domain without taking into account the temporal dimension of the video. On the contrary, the backdoor signal presented in this paper is a temporal video sequence, exploiting the temporal correlation of videos. To the best of our knowledge, this is also the first time that a backdoor attack is used to attack an anti-spoofing system[1].

Creating a backdoor attack working in the scenario described above presents some additional challenges with respect to the attacks considered so far. First of all, the backdoor signal must include a temporal dimension, since it is arguable that any DNN-based video rebroadcast detector will strongly rely on the temporal characteristics of the input signal. More importantly, at test time the backdoor signal must survive a number of transformations linked to the rebroadcast operation itself. These transformations include geometric transformations, motion of the rebroadcasting device, impact of ambient light, brightness changes caused by the rebroadcasting and acquisition devices etc. The solution we propose to cope with the above additional difficulties consists of injecting the backdoor signal by imperceptibly modifying the average luminance of the rebroadcast video following a slowly varying sinusoidal wave. In such a way, the backdoor signal is intrinsically immune to geometric transformations, moreover, previous works in digital watermarking have shown that slowly varying changes of average frame luminance can survive the heavy distortions intro-

[1]We believe that our attack can be easily extended to target more general person identification systems [17, 18]

duced by digital-to-analog and analog-to-digital transformations associated to video recapture [15, 16]. As we will show, corrupting a portion of the training set with our luminance-based backdoor signal effectively induces the network to associate the presence of the backdoor signal to the desired video class (in our case a pristine non-rebroadcast video), hence allowing the attacker to exploit the backdoor to evade the anti-spoofing control at test time. Noticeably, the attack is carried out without making any assumptions about the DNN architecture used for rebroadcast detection.

Throughout the paper, we will focus mainly on the case of backdoor injection with label poisoning (as in [8–10, 14, 19]), since in this case the attack requires that only a small percentage of the training data is corrupted, nevertheless, we will also report some results regarding the more challenging scenario of backdoor injection without label poisoning [12, 20].

2 Attack Formulation'

Let $f(X)$ be the decision function learned by a convolutional neural network, supervised by a training set of data-label pairs $\{X_i, l_i\}$. The learned discrimination function is optimized by stochastic gradient descent to 'to minimize the average loss $\sum_{X_i \in D_n} \mathcal{L}(f(X_i), l_i)$, where $\mathcal{L}$ is the cross entropy loss over the training set D_n consisting of n training videos with the corresponding labels. Furthermore, an unseen test data set, T_m of m samples, is available.

2.1 Backdoor attacks with Label Poisoning

In the following we use t to indicate the label of the target class of the attack (in our case pristine videos) and $\bar{t}$ for the label of the complementary class (in our case rebroadcast videos). A backdoor is added by using a function, $B(X_i, \Delta)$, to modify a proportion α of data samples $X_i \in D_n$ with labels $l_i = \bar{t}$. The function is applied with strength Δ and the corresponding labels of the poisoned data are modified to the targeted class: $l_i \rightarrow t$. Then the poisoned model, $f(X)$, is trained. The model can be attacked at test time, by poisoning some or all of the test set T_m, by introducing a backdoor with the same function $B(X_i, \Delta_T)$, where X_i now belongs to T_m and the strength Δ_T may be greater than that used during training, i.e. $\Delta_T \geq \Delta$. We anticipate that using a backdoor signal with larger strength during testing will improve the effectiveness of the attack, without affecting the stealthiness of the attack at training time. The backdoor would be injected into samples, X_i in T_m for which l_i is not equal to the target class, i.e. $\bar{t}$; the goal of course being to force X_i to be misclassified as belonging to the target class, t. (Left-hand side, Figure 2)

We can measure the success of the poisoning attack, hereafter referred to as the

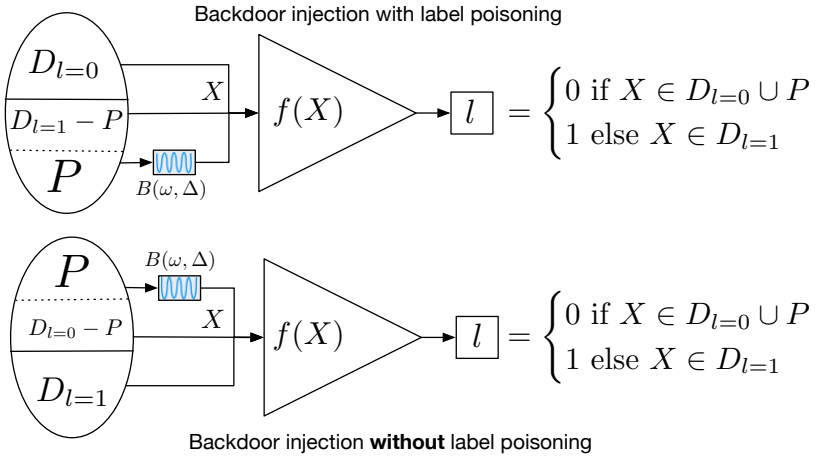


Figure 2: Diagram of two modes of backdoor injection: with and without label poisoning. When backdoors are injected with label poisoning into a proportion α of the data $D_{l=1}$(top), the target label is changed from class $l = 1$ to $l = 1$. Poisoned data is denoted as the set of samples P. When injected without label poisoning (bottom), the target class data is attacked but the label is not altered. In both cases, at test time, if successful, the backdoor injection into data of the spoof class $D_{l=1}$ should produce a prediction of the target label. Injection parameters are the backdoor signal frequency (ω), its strength (Δ, Δ_T), and the proportion of poisoned data, α.

attack success rate (ASR), as:

$$\text{ASR} = \frac{\sum_{X_j \in P}(f(X_j) \equiv t)}{N_P} \tag{1}$$

where P is the set of poisoned test samples and N_P is the number of samples in P.

2.2 Backdoor attacks without Label Poisoning

In this case, the backdoor function, $B(X_i, \Delta)$ is used on $X_i \in D_n$, s.t. $l_i = t$, for a proportion of α samples of the targeted class, t. So for example, with a two class model with labels $[0, 1]$, should we want to attack label $\bar{t} = 1$ (to force test samples to be misclassified as label 0), we must inject a backdoor into training samples of class $t = 0$.

At testing time, to attack the classifier, the backdoor signal is applied to a proportion α_T of test samples, $x_i \in T$, targeting images of class, $\bar{t}$, with amplitude Δ_T. A successful attack would turn the label x_i to be equal to the target class t. (Bottom part of Figure 2)

2.3 Proposed Video Backdoor Attack Signal

The goal of our attack is to attain high attack success rates with a stealthy (imperceptible) backdoor signal which can be easily applied to video sequences. The attack must also not greatly impact the classification of the targeted class (real videos) and yet be effective at reclassifying rebroadcast video sequences as real. We excluded the idea of introducing a spatial pattern, which might be easily detected over multiple video frames, and might also be affected by geometric transformation during video rebroadcast. Our attack method also assumes no knowledge of the attacked network, other than the anti-spoofing detection system uses video sequences.

Motivated by the need for the backdoor to work effectively in video sequences and for it to be relatively imperceptible, we designed a backdoor that introduces temporal changes to a video sample. In particular, the attack is designed to be intrinsically robust to geometric transformations. To do so, we were inspired by a similar approach used successfully in video watermarking (e.g. [15]) whereby the watermark was embedded into the video by modulating the mean video illumination, at some frequency related to the video frame-rate.

Our luminance-based backdoor function is defined as follows. Let $X = \{x_j\}$, $j = 1, N$ be a set of N consecutive video frames. The mean intensity of the frame x_j is changed by applying the same sinusoidal intensity change to all the pixels in the frame according to the following expression:

$$B(x_j, \Delta; \omega) = (1 - \Delta)x_j + \Delta \sin(2\pi\omega j/\text{FPS})x_j \tag{2}$$

where FPS is the frame rate (frame per second) of the video, Δ the amplitude of the backdoor signal, ω the temporal frequency (in Hz) of the sinusoidal backdoor signal. All pixel values in of a frame, x_i, are first weighted down by the factor $1 - \Delta$ and then the sinusoidal value of amplitude Δ is added. The resulting video sequences is a modified video sequence with a mean intensity variation in the range $[1 - 2\Delta, 1]$ as illustrated in Figure 3.

3 Experimental Setup

The proposed video backdoor attack is used to attack a model for anti-spoofing detection. The attacked model, $f(X)$ is a convolutional neural network trained to minimize average loss, $\sum_{i \in V} \mathcal{L}(X_i, l_i)$, where $\mathcal{L}$ is the categorical cross entropy loss function, over the validation data $V \subset D_n$, where the label can be $l = [0, 1]$, for real (0) and spoof videos (1). Each X is a sequence of 12 frames taken at steps of 2 frames, from videos sampled at 24 frames per seconds consisting of cropped faces, resized to64×64 pixels with 3 channels (RGB) per frame. The model input size is $12 \times 64 \times 64 \times 3$ and the model output size is 2.

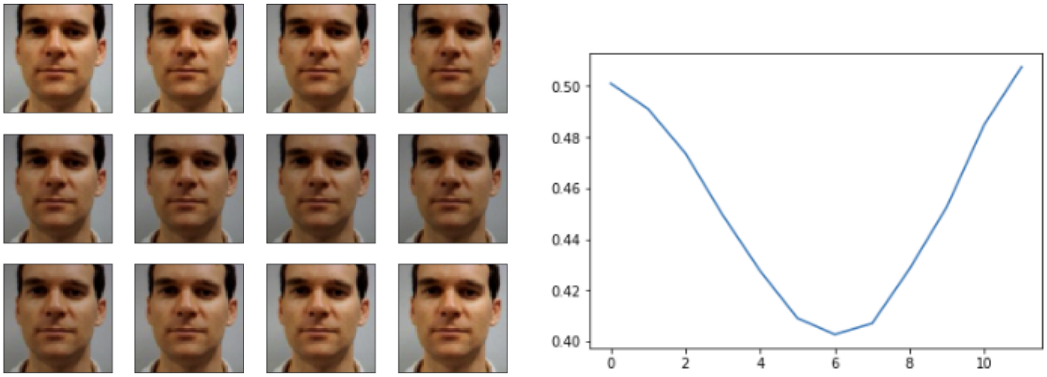


Figure 3: Example of mean values plot of a sequences and frame block (12 frames) for $\Delta = 0.1$ (corresponding to an intermediate strength among the values considered in this paper)'.

3.1 Model Architecture

The input sequence is split in groups of 3 overlapping frames, and each group of 3 frames is fed into a pair of 3D convolutional layers with 8 and 16, $3 \times 3 \times 3$ kernels each (Figure 4). Each layer is followed by batch normalisation, a $1 \times 2 \times 2$ max-pooling. The activation for the convolutions is a ReLU function. The frame-grouped outputs are flattened and reshaped into a time series of 3D features. These are then fed into an LSTM layer (with 6 units)[2]. The convolutional-LSTM feature extraction is followed by a 16 wide, sigmoid activation, dense layer and a final single neuron output with sigmoid activation.

To train the model, we used a SGD optimiser with learning rate (LR) of 0.01 and decay of 1e-6, with Nesterov momentum of 0.9. A binary cross-entropy loss is minimised and we use an accuracy metric to judge its performance. We use a 0.35 dropout rate on the 3D-CNN layers followed by batch normalisation and recurrent drop-out at rate 0.35 in the LSTM layer. We train with a batch size of 128 and pick the model with the smallest validation loss over a maximum of 25 epochs. The training data consists of 14385, 12 frame sequences and 14272 validation sequences. For testing, we used 4644 sequences.

3.2 Dataset

We used the IDIAP REPLAYATTACK anti-spoof video dataset [21] which consists of 1,300 video clips of attack attempts on 50 different identities. The size of all videos is 320×240 with a duration of about 9 seconds at 25 frames per second. Various types of re-broadcast attacks (using iPhone and iPad) and print attacks are included in the dataset.

[2]Implimented using Keras `keras.layers.LSTM`

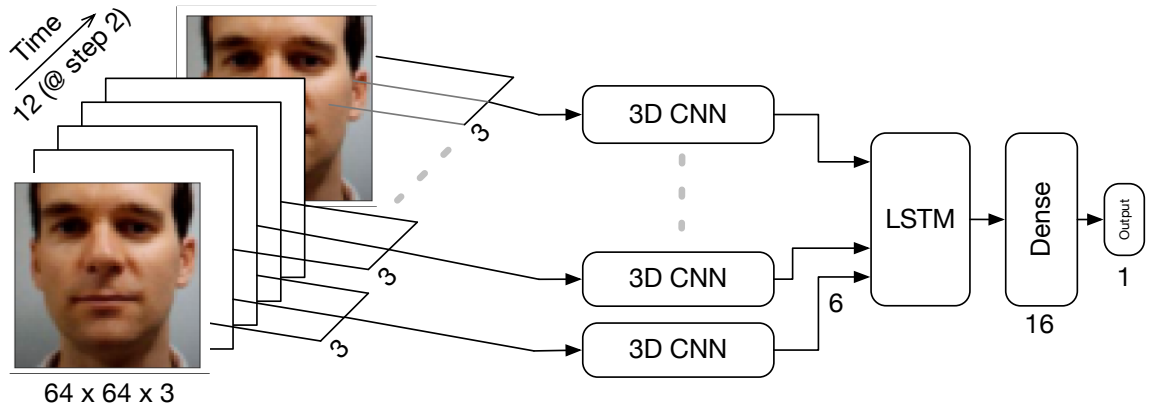


Figure 4: Diagram of architecture. Groups of 3 frames are input into the 3D-CNN feature detectors which have two convolutional layers each consisting of 8 and 16, $3 \times 3 \times 3$ filters, and $1 \times 2 \times 2$ max-pooling. The LSTM layer has 6 units (hidden layers).

Before feeding the sequence into our model, we crop faces across the frames, taking the bounding region across 24 frames with a border of 16 pixels, and then resizing the cropped faces to a size of 64×64 pixels. This strategy ensures that any head or replay device motion is captured in the frame sequence.

The trained model on pristine data, without introduction of backdoors, has a validation accuracy of 97.5%. On test data, its precision was 99.6% with a recall of 96.5% producing an error of 1.2% on real sequences and 3.5% on spoofed ones.

4 Experimental Evaluation of Video Backdoors

Backdoors are injected across the entire video sequence for each identity, prior to decimation into frame sub-sequence blocks. This is because the backdoor signal is temporally related to the frame-rate of the captured videos and similarly during attack, it must be added prior to any rebroadcast. The sub-sequence decimation is a function of the data preprocessing related to the model architecture.

During training, an α proportion of the data is poisoned: for backdoors with label poisoning, spoofed sequences and their labels are poisoned; for backdoors without label poisoning, original live sequences are poisoned. The attack signal amplitude is Δ during training and Δ_T during testing. The attack signal frequency was set to ω.

We performed a number of experiments with variations of α, ω, Δ, Δ_T to determine what was the relationship between the proportion of training data with backdoors (in both scenarios) and the attack success rate. We also tested the impact of the signal amplitude, Δ, during training; and how much greater Δ_T had to be to make the attack effective (especially in the no-label poisoning case). For testing, we

used $\alpha_T = 50\%$ so we could assess any side effect on non-attacked data.

Finally, we ran some experiments to investigate the immunity of the backdoor injection attack to geometric and contrast modification transformations, since these are the transformations typically introduced during a rebroadcast attack.

4.1 Backdoor attacks with label poisoning

1. We first evaluated the impact that the attack proportion α has on ASR. Figure 5(a) shows the ASR versus the test-time backdoor amplitude Δ_T for various values of α, fixing the training backdoor strength $\Delta = 0.05$. The ASR increases rapidly with the strength of the backdoor at run-time and with the proportion of the training data poisoned. The performance can be compared with the results obtained by adding the backdoor on a pristine model trained on uncorrupted samples (blue curve). The success of the attack on the pristine model for the larger values of Δ_T may be explained by the distortion of the test data caused by such high backdoor amplitudes, which ultimately induce classification errors even in the absence of training corruption.

2. We repeated the experiments with significantly lower attack proportions, namely $\alpha = \{1, 3, 5\}\%$, with varying attack strength $\Delta_T = [0.05, 0.15]$ and $\omega = 1$. Figure 5(b) shows that a relatively small proportions of the training data is enough for a successful attack, especially when the strength of the backdoor signal is increased at test time.

3. Next, we varied the (training time) backdoor strength over a range of values in the interval [0.025, 0.20], fixing $\omega = 1$ and $\alpha = 10\%$. Figure 5(c) shows that relatively stealthy attacks of $\Delta = 0.025$ can be effective. Increasing the backdoor poison amplitude achieves a commensurate increase in ASR. Increasing Δ above about 0.10 has diminishing returns at test time.

4. Eventually, we considered the effect of varying the frequency of the backdoor signal, with $\omega \in [0.25, 6]$, $\alpha = 20\%$, $\Delta = 0.2$. Figure 5(d) shows how frequencies of 4Hz or greater (up to Nyquist rate of 6 Hz) achieve 100% attack success rates on test data. From these results, we can conclude that higher frequencies appear to have more attack potency although the effect is not strictly linear with frequency.

4.2 Effect of Geometric and Contrast Transformations

As we said in the introduction, one of the main challenges associated to a backdoor attack against a rebroadcast detection anti-spoofing system, is that the backdoor signal should survive the digital-to-analog and analog-to-digital transformations implied by rebroadcast and recapture. In this section, then, we show the results that we have

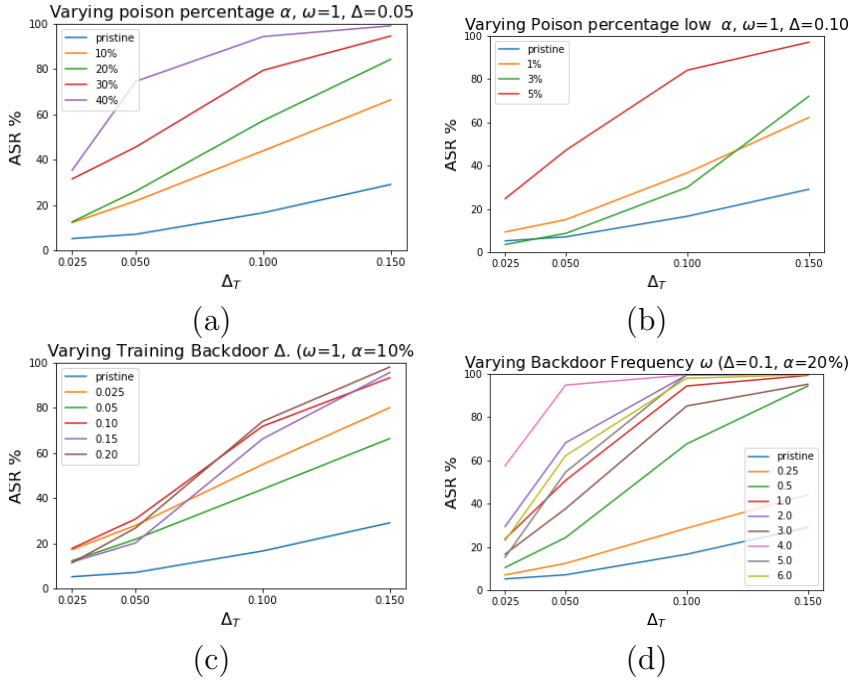
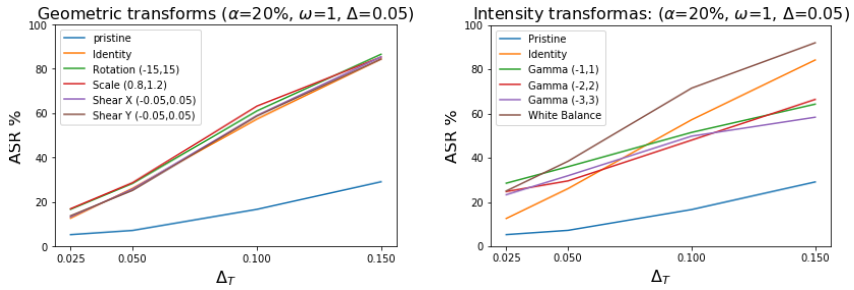


Figure 5: Backdoor attacks with label poisoning: (a) Effect of varying attack proportion α; (b) Effect of varying attack proportion for low α; (c) Effect of varying attack backdoor strength Δ of poisoned data; (d) Effect of varying attack backdoor frequency.



(a) (b)

Figure 6: (a) Effect of geometric transformations on backdoor signal. (b) Effect of contrast transformations on backdoor.

obtained by simulating the most common transformations associated to rebroadcast and recapture, namely geometric and contrast transformations.

The transformations were applied to the signal after the introduction of the backdoor but before the data was cropped. This simulates the situation of a rebroadcast attack where the display device (e.g. a mobile phone) is hand-held and the resulting image may be rotated, scaled (zoomed near and far from the authentication camera), or sheared (not held in a plane strictly parallel to the imaging plane). To simulate varying exposure and camera capture characteristics, we induced contrast changes to the videos: gamma correction and white-balance correction (applied framewise). The results of our simulations are discussed below.

1. Figure 6(a) shows the effect of geometric transformations (rotation, scaling, shear). As expected, the ASR is unaffected over a range of random transformations selected from uniform parameter distributions. Specifically, we used rotations in the range $[-15, 15]$ degrees; image zoom/scaling in the range $[0.8, 1.2]$ and X and Y shears in the range $[-0.05, 0.05]$.

2. Figure 6(b) presents the results we obtained by applying non-linear contrast changes: gamma correction over a range of random uniform values, and white-balancing. Gamma correction was tested over a range $[-3, 3]$ using:

$$G(x;\gamma) = \begin{cases} x^{-\frac{1}{\gamma}}, & \text{if } \gamma < 0 \\ x, & \text{if } \gamma = 0 \\ x^{\gamma} & \text{else.} \end{cases} \tag{3}$$

Gamma correction marginally worsens the ASR but conversely, white-balancing operation marginally improves the ASR as it has the effect of stretching the backdoor signal magnitude with the overall image contrast.

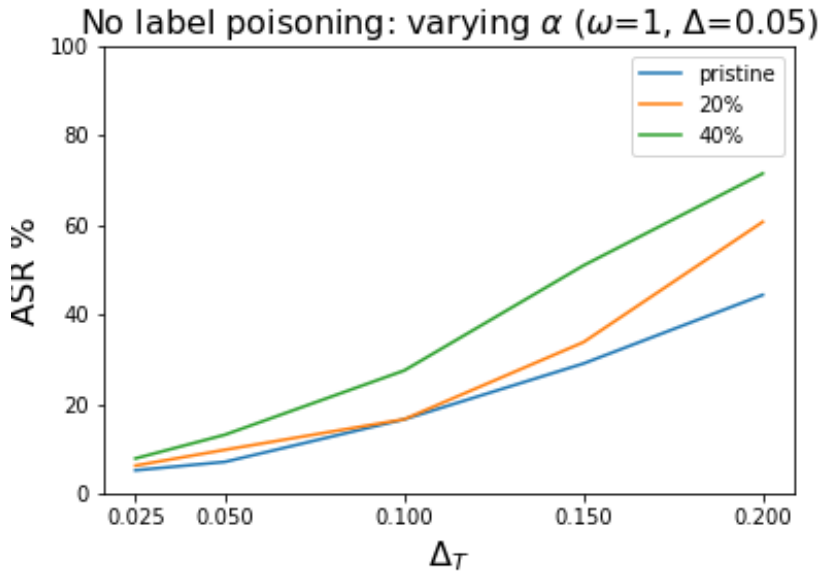


Figure 7: ASR with no label poisoning with two different poison percentages.

4.3 Backdoor attacks without label poisoning

We experimented with various combinations of attack proportion, frequency and backdoor amplitude to maximise the attack success rates in the case of attacking the network without label poisoning. As observed in [12], the attack proportion has to be considerably greater when the labels are not poisoned, which is exactly what we confirmed with our video backdoor attack. We had to increase the attack proportions to beyond $\alpha = 20\%$ to see a significant effect on the ASR.

Additionally, we did not find that the frequency and amplitude parameters which worked best in the label poisoning cases applied as well to the non-label poisoning scenario. So for example, $\omega > 1$ did not see a corresponding increase in ASR, and similarly neither did $\Delta > 0.1$. The only stable trend we discovered was an increase in ASR with attack proportion, α.

Figure 7 shows the best results we obtained by attacking the model without label poisoning. We found that to achieve ASRs above the baseline, we had to use a high attack proportion, $\alpha = 20\%$ or more but with a low signal amplitude, $\Delta = 0.05$. An ASR of over 50% can be achieved with $\alpha = 50\%$ and $\Delta_T = 0.15$.

5 Discussion and Concluding Remarks

We have introduced a novel illumination-based video backdoor attack against deep anti-spoofing rebroadcast detection systems. The attack has a number of interesting properties, including imperceptibility and robustness against geometric transformations, and to a good extent, intensity transformations. These robustness charac-

teristics make the attack suitable for use in rebroadcast attacks when display device motion and image contrast may not be controllable. We have demonstrated that when the video backdoor is embedded into the training data, it can be used to change classifier decisions with minimal of data poisoning, when the corresponding labels are also poisoned, and with some success when they are not (without label poisoning).

When labels are poisoned, the experiments demonstrate that increasing backdoor frequency and amplitude make the attacks more powerful, and in all cases, increasing the backdoor amplitude at test time increases the attack success rate.

In our experiments, we discovered that very low attack proportions (as low as 3%) is sufficient to attack the model when labels are poisoned (much larger attack proportions are necessary when labels are not poisoned, in line with the findings in [12]). This could be a function of the type of backdoor signal we have employed and as yet we do not have a method to generate the optimal signal which satisfies all the requirements of low poison percentage, stealth, geometric invariance and results in the highest ASRs. Furthermore, thus far, we have only tested the method on a 2-class model and we have found it fairly easy to attack the spoof-class, as we see some attack success without data poisoning (on the pristine model), when at test time the backdoor amplitude is increased. Only by testing a similar attack method on a multi-class video classification problem will we see whether the method generalises to other deep video classification models.

We believe that this type backdoor signal which modulates the video illumination in time, might also be turned into a physical attack, perhaps introduced stealthily during biometric identity enrolment and/or identity verification by a physical alteration of the environmental lighting conditions.

References

[1] K. Sundararajan and D. L. Woodward. "Deep Learning for Biometrics: A Survey", ACM Computing Surveys, Vol. 51, No. 3, 2018.

[2] J. Yang, Z. Lei, and S. Z. Li. "Learn convolutional neural network for face anti-spoofing". arXiv:1408.5601, 2014.

[3] Z. Xu, S. Li and W. Deng. "Learning Temporal Features Using LSTM-CNN Architecture for Face Anti-spoofing", in 3rd IAPR Asian Conference on Pattern Recognition, 2015.

[4] N. Lakshminarayana, N. Narayan, N. Napp, S. Setlur, and V. Govindaraju, "A discriminative spatio-temporal mapping of face for liveness detection", in Proc. IEEE Int. Conf. Identity, Secur. Behavior Anal. (ISBA), Feb. 2017, pp. 1-7.

[5] J. Gan, S. Li, Y. Zhai, and C. Liu, "3D convolutional neural network based on face anti-spoofing", in Proc. 2nd Int. Conf. Multimedia Image Process. (ICMIP), Mar. 2017, pp. 1-5.

[6] H. Li, P. He, S. Wang , et al. "Learning generalized deep feature representation for face anti-spoofing". IEEE Transactions on Information Forensics and Security, 2018, 13(10): 2639-2652.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus. "Intriguing properties of neural networks". arXiv preprint arXiv:1312.6199, 2013.

[8] Y. Ji, X. Zhang, and T. Wang, "Backdoor attacks against learning systems", in 2017 IEEE Conference on Communications and Network Security (CNS). IEEE, 2017, pp. 1-9.

[9] Y. Liu, S. Ma, Y. Aafer, W-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks", in Proc. 25th Annual Network and Distributed System Security Symposium, NDSS 2018, 2018.

[10] C. Liao, H. Zhong and A. C. Squicciarini and S. Zhu and D. J. Miller, "Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation", CoRR, arXiv, 2018.

[11] X Chen, C Liu, B Li, K Lu and D Song. "Targeted backdoor attacks on deep learning systems using data poisoning." arXiv preprint arXiv:1712.05526 (2017).

[12] M. Barni, K. Kallas and B. Tondi. "A New Backdoor Attack in CNNs by Training Set Corruption Without Label Poisoning", Proceedings of 2019 IEEE Int. Conf. on Image Processing, ICIP 2019, arXiv:1902.11237.

[13] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Malloy and B. Srivastava. "Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering", In AAAI Workshop on Artificial Intelligence and Safety, CEUR Workshop Proceedings, 2019.

[14] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks". IEEE Access, 7, 2019, 47230-47244.

[15] J. Haitsma and T. Kalker, "A watermarking scheme for digital cinema", Proceedings of 2001 International Conference on Image Processing, Thessaloniki, Greece, 2001, pp. 487-489 vol.2.

[16] Y. Zhao and R. L. Lagendijk, "Video watermarking scheme resistant to geometric attacks", Proceedings of 2002 International Conference on Image Processing, Rochester, NY, USA, 2002.

[17] D. S. Trigueros, L. Meng, and M. Hartnett. “Face Recognition: From Traditional to Deep Learning Methods”. 2018. arXiv preprint arXiv:1811.00116.

[18] N. McLaughlin, J. Martinez del Rincon and P. Miller. “Recurrent convolutional network for video-based person re-identification”. In Proc. of the IEEE Conf. on computer vision and pattern recognition 2016, CVPR 2016.

[19] H. Li, S. Wang, and A. C. Kot, “Image recapture detection with convolutional and recurrent neural networks”, Electron. Imag., vol. 2017, no. 7, pp. 87-91, 2017.

[20] M. Alberti, V. Pondenkandath, M. Wursch, M. Bouillon, M. Seuret, R. Ingold, and Marcus Liwicki, “Are you tampering with my data?”, in Computer Vision, ECCV 2018 Workshops, 2019.

[21] I. Chingovska, A. Anjos and S. Marcel. “On the Effectiveness of Local Binary Patterns in Face Antispoofing”, Proceedings of IEEE BIOSIG, 2012.