Flickering Adversarial Attacks against Video Recognition Networks

Abstract

- present a manipulation scheme for fooling video classifiers
- introduce a flickering temporal perturbation
- unnoticeable by human observers
- implementable in the real world

Background

- Deep Neural Networks (DNNs) have shown phenomenal performance in a wide range of tasks, such as image classification
- Despite their success, DNNs have been found vulnerable to adversarial attacks
- Many works have shown that a small (sometimes imperceptible) perturbation added to an image, can make a given DNNs prediction false
- In video action recognition networks temporal information is of the essence in categorizing actions, in addition to per-frame image classification

FAA

- Aim:attack the video action recognition task by applying a uniform RGB perturbation to each frame, thus constructing a temporal adversarial pattern
- Threat model :I3D
- DataSet:the Kinetics-400 Human Action Video Dataset

Video action recognition function

For a given X, through the identification function $F_\theta(X) = y$, the probability distribution y on the output data domain.The model F implicitly depends on some parameters θ that are fixed during the attack.The classifier assigns the label

$$A_\theta(X) = \arg\max_i y_i$$

$$input : X = [x_1, x_2, ..., x_T] \in R^{T*H*W*C}$$

$$T - frames$$

$$H - rows$$

$$W - columns$$

$$C - ColorChannel$$

Methodology

Make a perturbation δ,

$$\delta = [\delta_1, \delta_2, ..., \delta_T] \in R^{T*H*W*C}$$

$$\hat{X} = X + \delta (\hat{x}_i = x_i + \delta_i)$$

$$s.t. A_\theta(X) \neq A_\theta(\hat{X})$$

and X and $\hat{X}$ look similar,
in particular xi and $\hat{x}_i$ look similar

Objective function

- The parameter λ weights the relative importance of being adversarial and the regularization terms
- The set of functions Dj(·) controls the regularization terms that allows us to achieve better imperceptibility for the human observer.
- The parameter β weights the relative importance of each regularization term
- m > 0 is the desired margin of the original class probability below the adversarial class probability

$$\arg\min_{\delta} \lambda \sum_{j} \beta_j D_j(\delta) + \frac{1}{N} \sum_{n=1}^{N} \ell(F_\theta(X_n + \delta), t_n)$$

$$s.t. \hat{x}_i \in [V_{\min}, V_{\max}]^{H*W*C}$$

$$\ell(y,t) = \max(0, \min(\frac{1}{m} \ell_m(y,t)^2, l_m(y,t))$$

$$\ell_m(y,t) = y_t - \max_{i \neq t}(y_i) + m$$

Regularization terms

THICKNESS REGULARIZATION

This loss term forces the adversarial perturbation to be as small as possible in gray-level over the three color channels (per-frame), having no temporal constraint and can be related to the ″ thickness″ of the adversarial pattern

$$D_1(\delta) = \frac{1}{3T} \| \delta \|_2^2$$

ROUGHNESS REGULARIZATION

We introduce temporal loss functions which incorporate two different terms

$$D_2(\delta) = \frac{1}{3T} \left\| \frac{\partial \delta}{\partial t} \right\|_2^2 + \frac{1}{3T} \left\| \frac{\partial^2 \delta}{\partial t^2} \right\|_2^2$$

Implementation Details

- Discard all misclassified videos
- Roll operator

Roll(X, τ) produce the time shifted tensor, whose elements are τ-cyclic shifted along the first axis (time)

$$Roll(X,\tau) = [x_{(\tau \bmod T)+1}, x_{(1+\tau \bmod T)+1}, x_{(T-1+\tau \bmod T)+1}]$$

$$\frac{\partial X}{\partial t} = Roll(X,1) - Roll(X,0)$$

$$\frac{\partial^2 X}{\partial t} = Roll(X,-1) - 2Roll(X,0) + Roll(X,1)$$

Metrics

- Fooling ratio
- Mean Absolute Perturbation per-pixel

$$thickness_{gl}(\delta) = \frac{1}{3T}\|\delta\|_1^2$$

- Mean Absolute Temporal-diff Perturbation perpixel

$$roughness_{gl}(\delta) = \frac{1}{3T}\left\|\frac{\partial\delta}{\partial t}\right\|_1^2$$

Single Video Attack

- we have selected a random sub-sample from the validation section of the kinetics dataset and solve Objective function for each one of the selected sample (N = 1)

- the results of a single-video attack using β1= β2= 0.5, reaching 100% fooling ratio with low roughness and thickness values.

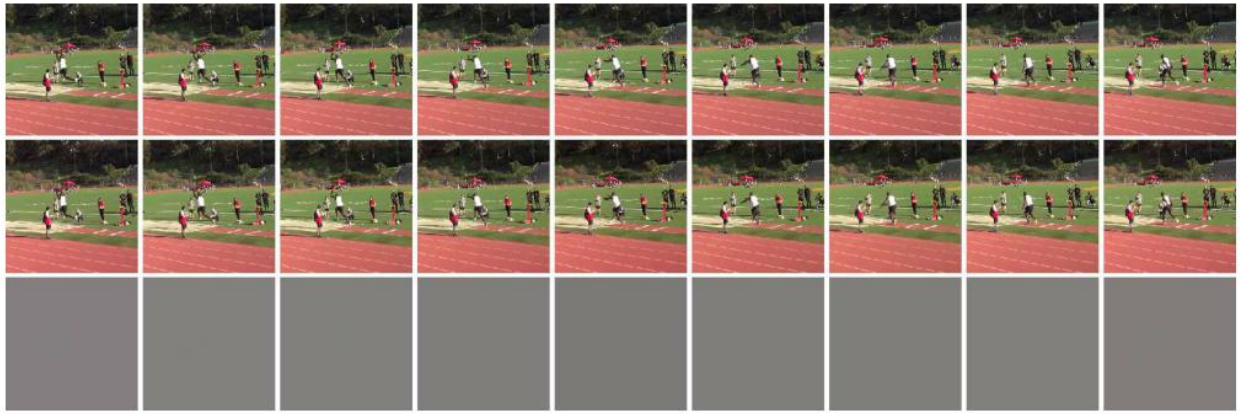


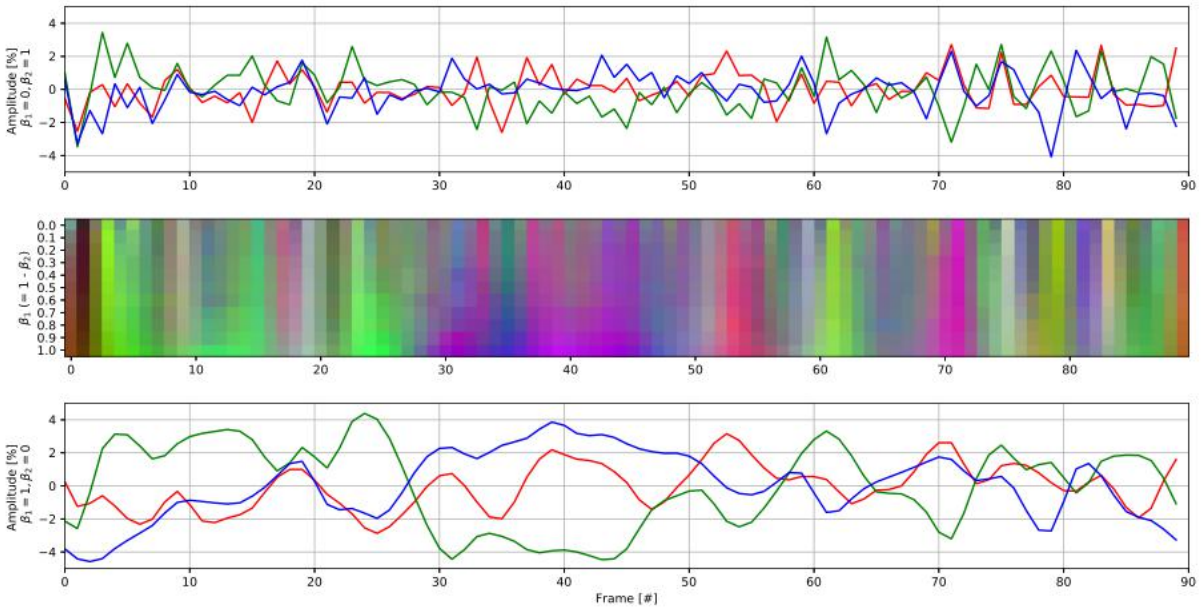
Figure 1. Top: Consecutive frames of the original video from the "Triple jump" category. Middle: Consecutive frames of the misclassified adversarial video. The adversarial perturbation is practically unnoticed by the human observer. Bottom: The flickering adversarial perturbation of each frame. The perturbation is a constant offset applied to the entire frame. Due to the fact that the perturbation can be negative, it is displayed over gray background. The gentle hue changes displayed at each frame are the adversarial pattern.

THICKNESS and ROUGHNESS

β1= 1, β2= 0:
The signal of the RGB channels fluctuates strongly with a thickness value of 0.87% and a roughness of 1.24%

β1= 0,β2= 1:
leading to a thickness value of 1.66% and a roughness value of 0.6%

Both D1 and D2 are very dominant in the received perturbation, as desired



Middle: The gradual change of the adversarial pattern between the two extreme cases where β1= 0 corresponds to the top graph and β1= 1 corresponds to the bottom graph.
Color (stretched for visualization purposes) represents the RGB parameters of the adversarial pattern of each frame.

Adversarial Attack Generalization

- we have generalized the attack to cause misclassification to any videos from a specific class witha single generalized adversarial pattern($\beta1=\beta2= 0.5$)
- When applying this pattern, in average 90.2% of the videos from each class were misclassified
- Selecting different parameters to the regularization terms can produce different adversarial patterns.

Table 1. Results and standard deviation over several types of attacks

ATTACK	TIME INVARIANCE	FOOLING RATIO[%]	THICKNESS[%]	ROUGHNESS[%]
SINGLE VIDEO	×	100	1.0±0.5	0.83± 0.4
SINGLE CLASS	×	90.2± 11.72	13.0± 3.6	10.6± 2.2
UNIVERSAL	×	93.0	15.5	15.7
UNIVERSAL	√	83.1	18.0	14.0

UNIVERSAL UNTARGETED ATTACK

- Our attack reaches a 60% fooling ratio with only 50 videos to train on, while SUP stays close to 0%.
- Furthermore, the attack introduced in reaches a slightly higher fooling ratio over our attack with the current setup (95.1% vs 93%), but it requires 10 times more training videos.
- The adversarial pattern presented in this paper requires only 500 videos to reach a 90% fooling ratio, where the other attack requires 5000 for the same ratio

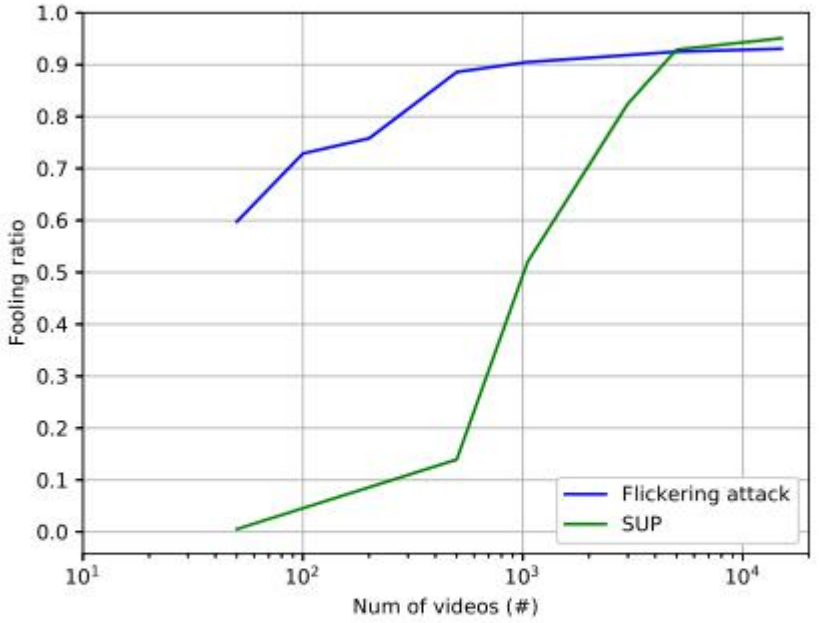


Figure 3. Fooling ratio Vs. number of videos in the training set. In order to reach 90% fooling ratio, the SUP attack requires almost 10 times more videos in the training set than the flickering attack.

Conclusions

Advantages:

- the relative unperceptability to the human observer
- achieved by small and smooth perturbations
- The flickering perturbation can be implemented in real world scenarios since it does not require a complex spatial adversarial pattern to be projected upon the scene
- we present a time-invariant adversarial attack that can be applied to the recorded scene without assuming that the perturbation of each frame is applied at the right time

Future work:

- adversarial attacks with black box setting
- by adjusting the light to reach the physical world interference