International Journal of Multimedia Data Engineering and Management
Volume 12 • Issue 3

Universal Sparse Adversarial Attack on Video Recognition Models

Haoxuan Li, University of Electronic Science and Technology of China, China
Zheng Wang, University of Electronic Science and Technology of China, China*

ABSTRACT

Recent studies have discovered that deep neural networks (DNNs) are vulnerable to adversarial examples. So far, most of the adversarial researches have focused on image models. Whilst several attacks have been proposed for video models, their crafted perturbation are mainly per-instance and totally polluted ways. Thus, universal sparse video attacks are still unexplored. In this article, the authors propose a new method to explore universal sparse adversarial perturbation for video recognition system and study the robustness of a 3D-ResNet-based video action recognition model. A large number of experiments on UCF101 and HMDB51 show that this attack method can reduce the success rate of recognition model to 5% or less while only changing 1% of pixels in the video. On this basis, by changing the selection method of sparse pixels and the pollution mode in the algorithm, the patch attack algorithm with temporal sparsity and the one-pixel attack algorithm are proposed.

KEYWORDS

Adversarial Attack, One Pixel Attack, Sparse Patch, Temporal Sparsity, Universal Perturbation, Video Recognition

INTRODUCTION

With the development of science and technology, deep neural networks (DNNs) have played a very important role in various tasks of visual understanding tasks over the years. However, with the deepening of research, DNNs has been found to be vulnerable to the adversarial examples, which have undergone carefully crafted perturbations, and can easily fool a DNNs model into making irrelevant classification. Therefore, the existence of adversarial samples has aroused intense concern about the security of deep neural networks when referring to the detailed practical application, especially in face recognition, video surveillance and other safety-demanding systems. It is imminent to study adversarial samples to improve the robustness of deep neural networks.

In recent years, researchers have presented a keen interest in whether adversarial examples are practical enough to attack more complex systems, for instance image retrieval (Li et al., 2019) and image caption (Zhang, Wang, Xu, Guan, & Yang, 2020). However, these adversarial samples are mainly concentrated in the field of image models, while the adversarial attack on video models are rarely explored. According to Wei, Zhu, Yuan, and Su (2019), "Compared with images, attacking a video may need to consider not only spatial cues but also temporal cues, which greatly increase the difficulty of adversarial examples generation." Hosseini, Xiao, Clark, & Poovendran (2017) firstly attacked the video classification model, which directly inserted the targeted image into most frames of a video, so that the intelligent system of Google Cloud Video misclassified the video as the label of the inserted image. This article calls the aforementioned methods as Dense Adversarial Attack (DAA), which adds perturbations to all pixels of every or most of the frames in a video to craft

DOI: 10.4018/IJMDEM.291555 *Corresponding Author

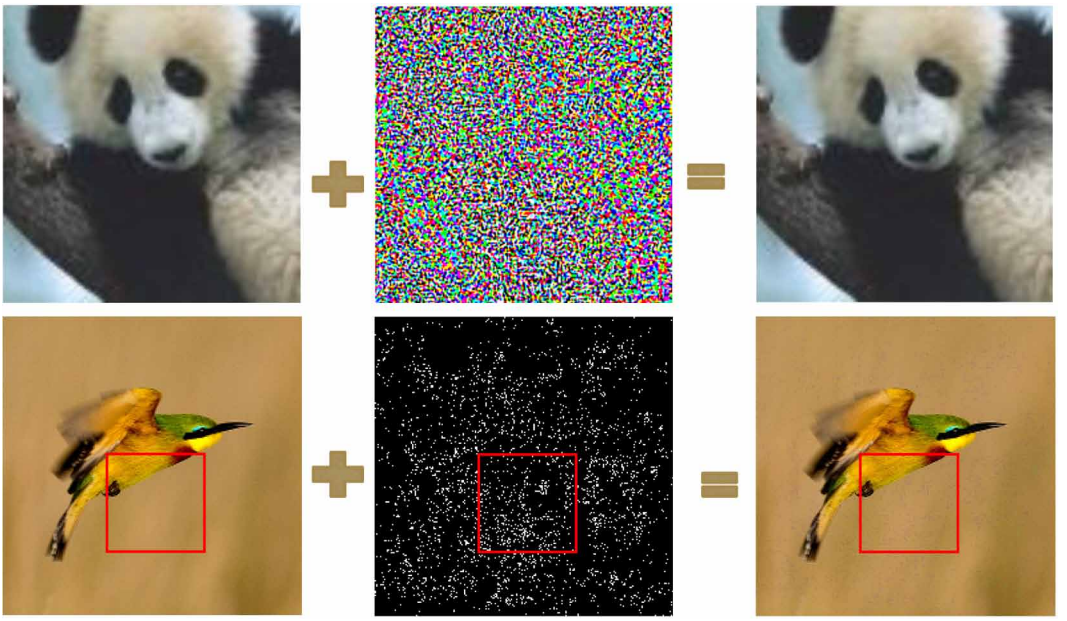
adversarial examples. Moreover, DAA inevitably consumes expensive computation resources and time. Thus, some works (Wei et al., 2019; Zajac, Zołna, Rostamzadeh, & Pinheiro, 2019) proposed sparse perturbation methods, which only change partial or a few pixels of the benign instances. Dense and sparse adversarial examples for images are shown in Figure 1. It can be seen that the dense attack is a small, imperceptible attack on each pixel, while the sparse attack is a lager, perceptible attack on selected pixels.

In addition, there are two types of adversarial perturbations for images: Universal (Image-agnostic) and Image-dependent (per-instance adversarial example). Most of existing adversarial attack methods are image/video-dependent, such as Projected Gradient Descent (Madry, Makelov, Schmidt, Tsipras, & Vladu, 2018), and black-box video attack framework (Jiang, Ma, Chen, Bailey, & Jiang, 2019), and sparse adversarial perturbations for videos (Wei et al., 2019). This is done most effectively using (potentially expensive) iterative optimization procedures. Different from per-instance perturbation attacks, there exist "universal" perturbations that can be added to any image to change its class label with high probability, and firstly proved by Moosavi-Dezfooli, Fawzi, Fawzi, & Frossard (2017). That is to say, seek a fixed perturbation ε with small magnitude such that for most natural images x , $x+\varepsilon$ could significantly mislead the pre-trained network.

So in this work, this paper firstly extends the investigation of universal adversarial perturbations towards video recognition models. Consider the expensive computing resources and time, this work also introduces sparse perturbation search into our work. This paper is an extension of the accepted paper for China MM2020. In this work, the algorithm and experiment are optimized and more comprehensive. That is to say, by improving the selection method of sparse pixels and the pollution method of patches, a universal patch attack algorithm with a certain temporal sparsity and a better one pixel attack algorithm in all aspects are proposed. The main contributions of this paper can be summarized as follows:

- This paper explores the problem of universal adversarial attack on video recognition models and proposes a general framework to generate universal adversarial examples for videos. To the best

Figure 1. Dense and sparse adversarial examples for images



Multimodal Information Integration and Fusion for Histology Image Classification

Tao Meng, Mei-Ling Shyuand Lin Lin (2011). *International Journal of Multimedia Data Engineering and Management (pp. 54-70).*
www.irma-international.org/article/multimodal-information-integration-fusion-histology/54462