ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747698

SPARSE ADVERSARIAL ATTACK FOR VIDEO VIA GRADIENT-BASED KEYFRAME SELECTION

Yixiao Xu, Xiaolei Liu , Mingyong Yin, Teng Hu, Kangyi Ding*

Institute of Computer Application
China Academy of Engineering Physics, China

ABSTRACT

Videos have a higher dimensionality compared with images, making adversarial video attacks more challenging. We propose a gradient-based method for self-adaptive white-box video keyframe selection and video adversarial example generation, taking advantage of that perturbations are transferable between video frames. More specifically, a gradient-based method is proposed to determine different video frames' contribution to classification results. Based on the weights of different frames and the given boundary values, the proposed method adaptively selects a subset of frames as keyframes for perturbation. Experimental results of attacking two widely used video classification models on UCF-101 and HMDB-51 datasets show that the proposed method effectively improves the generation efficiency as well as the steganography of adversarial video examples, leading to a reduction of more than 21% of the required number of iterations and more than 25% of the average perturbation size for the untargeted attack.

Index Terms— adversarial attack, video recognition, keyframe selection

1. INTRODUCTION

Deep neural networks are being used in more and more areas, such as autonomous driving, object detection, and natural language processing [1, 2, 3], and have outperformed traditional methods, while at the same time, the security risks of deep neural-based models have received more widespread attention. Deep neural networks are vulnerable to adversarial example attacks, which mislead deep models by adding human-invisible perturbations to clean examples [4, 5]. The majority of current adversarial example studies focus on image and audio domain [6, 7], but in recent years there are also some studies in the video domain [8, 9].

Compared with adversarial image attack, adversarial video attack is more challenging because videos have higher dimensionality than images [10], thus adversarial video examples have to contain time-domain information that image examples do not have. Therefore, extending image attack methods to attack video models directly will lead to low efficiency and poor invisibility. At the same time, due to the nature of the video classification model, adversarial perturbations are transmitted between frames. Therefore, adding temporal-sparse perturbations to video examples is a way worth considering to improve the efficiency.

This paper investigates how to measure the effects of different frames on classification results. Different from the definition of keyframes in the field of video processing, keyframes in video classification are not only related to the videos but also the target models. Therefore, the keyframe selection method without introducing model information will lead to low accuracy.

To address the above-stated challenges, we propose a new self-adaptive white-box video keyframe selection method. Firstly the backward propagation algorithm is used to generate a gradient heat map, and then the importance of each frame is measured according to the mean value of the gradient of the pixel points in the frame. The method then selects a variable number of keyframes in descending order of importance until the sum of the weights of the selected keyframes reaches a specified boundary value. This method introduces model information into the keyframe selection process and adjust the number of keyframes adaptively, which improves the quality of keyframe selection and increases the efficiency of keyframe selection by calling the target model only once. Furthermore, we achieve a fast generation of adversarial examples and subsequent reduction of perturbations by dynamically adjusting the terms of the $L_{2,1}$ norm-based loss function. Our major contributions can be summarized as follows:

- We propose an efficient white-box adversarial video attack model that selects keyframes using gradient backpropagation.
- We divide the adversarial video attack process into two stages to meet the constraints on query times and perturbation size in different cases
- Extensive experiments on two benchmark data sets demonstrate that the proposed method is efficient and

*Correspondence should be addressed to Xiaolei Liu, liuxiaolei@caep.cn. This research was supported by the National Natural Science Foundation of China (Grant No. 62102379).

effective. It achieves a more than 21% reduction in query numbers for the untargeted attack.

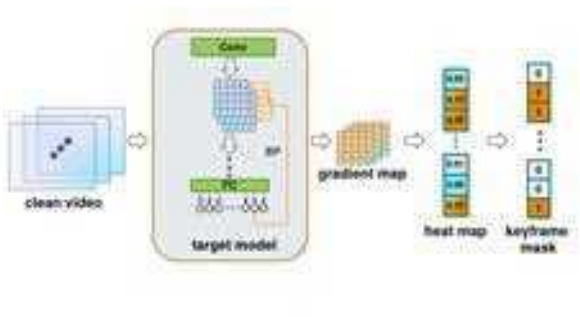

Fig. 1. Overview of the gradient-based keyframe selection method.

2. RELATED WORK

Wei et al. [2] propose a sparse attack method (SA) to reduce the total perturbation size of the generated adversarial examples. By adding perturbations to a subset of original frames, the computational cost and the perturbation size of adversarial video examples are reduced. However, in SA, the information related to the model and the input examples is not introduced in keyframe selection, thus the effect of selected keyframes is insufficient. Then a heuristic keyframe selection method is proposed to rank the importance of video frames heuristically [11]. This method measures the importance of a frame by comparing the outputs of the target model before and after deleting this particular frame. Different from SA, the heuristic method links the keyframe selection process with the features of input examples and target models, which improves the accuracy of keyframe selection. However, the direct deletion of frames triggers a large change, resulting in low precision, and a new cost of querying the target model is introduced. Therefore, this method is more suitable for the black-box condition, which needs more queries to succeed than the white-box condition originally. A similar issue arises in [12], which implements reinforcement learning-based keyframe selection for black-box conditions, leaving the adversarial keyframe selection method under white-box constraints unexplored.

3. METHODOLOGY

In this section, we introduce the proposed gradient-based sparse adversarial video attack method. We suppose that the attack is carried out under white-box conditions (i.e., the attacker has full knowledge about the target model and is able to compute the gradient directly).

Denote the target video classification model as F, an clean video example $\boldsymbol{x} \in \mathbb{R}^{N\times C\times H\times W}$, where $NCHW$ denote the number of frames, the number of channels, the frame height and width, respectively. $y = F(\boldsymbol{x})$ represents the classification result for clean example. For untargeted attacks, the attacker's aim is to generate an adversarial perturbation $\boldsymbol{\delta}$ such that $y \neq F(\boldsymbol{x} + \boldsymbol{\delta})$. As for targeted attacks, for a specified target label $\hat{y}$, the attacker's aim is to generate an adversarial perturbation $\boldsymbol{\delta}$ such that $\hat{y} = F(\boldsymbol{x} + \boldsymbol{\delta})$.

3.1. Sparse Adversarial Video Attack

Our attack algorithm is based on the sparse adversarial video attack method (SA) [2], which is the first white-box adversarial attack method for video classification models using $L_{2,1}$ norm loss. Following the SA, We use the $L_{2,1}$ norm loss as a regular term to control the size of the perturbation and use cross-entropy loss to achieve the generation of adversarial examples. An untargeted attack can be represented by the following equation:

$$\underset{\boldsymbol{\delta}}{argmin}\ \alpha||\boldsymbol{M}\cdot\boldsymbol{\delta}||_{2,1} - \beta L(y, F_\theta(\boldsymbol{x} + \boldsymbol{M}\cdot\boldsymbol{\delta})) \quad (1)$$

where $\boldsymbol{M} \in \{0,1\}^{N\times C\times H\times W}$ denotes the keyframe mask which achieves temporal sparcity. $L(.,.)$ calculates the cross-entropy loss between the perturbed label and the clean label. F_θ denotes the target model with parameters θ. $||\boldsymbol{\delta}||_{2,1}$ is the $L_{2,1}$ norm of $\boldsymbol{\delta}$, which serves to reduce the size of the perturbations. And (α, β) are two weights to balance two terms.

As for targeted attack, the problem can be expressed as follow:

$$\underset{\boldsymbol{\delta}}{argmin}\ \alpha||\boldsymbol{M}\cdot\boldsymbol{\delta}||_{2,1} + \beta L(\hat{y}, F_\theta(\boldsymbol{x} + \boldsymbol{M}\cdot\boldsymbol{\delta})) \quad (2)$$

where $\hat{y}$ denotes the target label of the attack. Unlike SA where (α, β) is set to constant, we set the value of (α, β) according to the perturbed label instead. When the perturbed label and the clean label are the same, we set (α, β) to $(0, 1)$, respectively, making the algorithm outputs the perturbations to fool the target model. In the opposite case, we set (α, β) to $(1, 0)$ to make the algorithm reduce the size of the perturbations.

3.2. Self-adaptive Gradient Based Keyframe Selection

Wei et al. [2] prove that adversarial perturbations are transferable in the time domain. Thus we consider selecting a subset of frames that contribute most to the output of the target model and only add perturbations on these selected frames to achieve sparsity. Keyframes in video processing usually refer to the frames that vary significantly from the preceding frames. However, the frames that contribute most to the video recognition process are different. Therefore, it is necessary to find a method to measure the contribution of different frames to the classification result.

In the image field, Gradient-weighted Class Activation Map (G-CAM) refers to the generation of a heat map of class activation for an input image [13], is often used to measure the contribution of different parts of the input to the output. For an image input to a CNN model and classified into a specific label, this technique helps understand which local position of an original image contributes most to the final classification decision. Inspired by this method, we propose a self-adaptive gradient-based keyframe selection method. The method first evaluates the contribution of each frame to the classification result and then automatically selects keyframes based on the importance of each frame.

For attacking LRCN model, denote the feature map obtained after convolutional networks by $\boldsymbol{s}$, we first calculate the gradient of the score of output label y for the convolutional layer output:

$$w_k^y = \sum_i \sum_j \frac{\partial y}{\partial \boldsymbol{s}_{kij}} \tag{3}$$

where $w_k^y, k \in \{0, 1, ..., N\}$ denotes the importance weight of the k-th frame, and $\boldsymbol{s}_{kij}$ denote the pixel point with coordinates (i, j) on the k-th frame.

For attacking the C3D model, the feature map obtained by 3D convolutional networks contains features of multiple frames, so the weight of a certain frame should be identified as the mean of the weights of all the feature maps containing this certain frame. Suppose that the convolution window of the convolution kernel in the time domain is z, w_k^y can be obtained by the following equation:

$$w_k^y = \frac{1}{z} \sum_z \sum_i \sum_j \frac{\partial y}{\partial \boldsymbol{s}_{zij}} \tag{4}$$

Then we select the keyframes in descending order of the importance weight. Instead of selecting a constant number of frames as keyframes, we automatically adjust the number of keyframes until the sum of the weights of the selected keyframes reaches the defined boundary value $b \in (0, 1]$, in this way, the number of keyframes is dynamically adjusted according to the characteristics of the target model for a specific input example, avoiding the increase of perturbation caused by too many keyframes or the decrease of efficiency caused by too few keyframes.

4. EXPERIMENTS

4.1. Experimental Methodology

Dataset We evaluate our method on two widely used datasets, UCF-101 [14] and HMDB-51 [15]. UCF-101 is a dataset containing 13,320 videos distributed in 101 action classes for action recognition. HMDB-51 contains a total of 7000 clips distributed in 51 action classes and is for human motion recognition. During the evaluation, we sample 16 frames for each video example as the original input of the target model.

Table 1. Test Accuracy(%) of the target models.

Model	UCF-101	HMDB-51
C3D	85.88	59.57
LRCN	66.43	41.16

Model We use two video recognition models, Long-term Recurrent Convolutional Networks (LRCN) [16] and C3D [17] as target models. LRCN models use CNNs to extract feature maps from the original frames and then use Recurrent Neural Networks(RNNs) to make classification decisions. We use Resnet [18] to encode the video frames and LSTM to achieve classification in this experiment. C3D extracts the features contained on the time and space domains simultaneously by 3D convolution networks. Table 1 summarizes the test accuracy of 16-frame snippets with these two models.

Metrics We evaluate our keyframe selection method using four metrics.

Fooling Rate(F): the percentage of adversarial videos that successfully make the target model output wrong labels [19].

Mean Queries(MQ): the mean value of queries needed to successfully generate an adversarial example.

Mean Absolute Perturbation(MAP): the mean perturbation of each pixel throughout the video.

Sparse Rate(SR): Percentage of perturbed frames to total frames.

Attack Settings We consider both untargeted attack and targeted attack. Note that we have one parameter boundary value $b \in (0, 1]$ in keyframe selection process. Larger b may increase the number of keyframes selected, leading to fewer query times but to a larger MAP, while smaller b may increase the number of iterations required for a successful attack. Thus we perform a grid search to determine the suitable value of b is 0.3 for an untargeted attack and 0.4 for a targeted attack. We randomly select one video example that can be correctly classified by the target model from each class for the experiment on both datasets. For the overall performance comparison, we set the maximum number of iterations to 500 and 1000 for untargeted and targeted attacks. When comparing the performance of the methods under the condition of limited iterations, we set the maximum number of iterations from 0 to 100 instead.

4.2. Performance Comparison

We compare our method with two keyframe selection methods, fixed keyframe selection method (SA) [2], and keyframe selection method based on inter-frame differencing (IFD). We evaluated the performances of three methods on two datasets towards two video recognition models. Tabel 2 shows the performance comparison of the three methods for both targeted and untargeted attacks on the UCF-101 and HMDB-51. For untargeted attacks, the experimental results show that

Table 2. Untargeted and targeted attacks against C3D/LRCN Models. For all attack models, the Fooling Rate (FR) is 100%.

Dataset	TargetModel	Untargeted				Targeted			
		Method	MQ	MAP	SR	Method	MQ	MAP	SR
UCF-101	C3D	Ours	**113.62**	**0.0992**	**82%**	Ours	**328.54**	**0.2308**	**75%**
		SA	144.10	0.1379	75%	SA	346.72	0.2783	69%
		IFD	163.03	0.1388	75%	IFD	390.96	0.2719	69%
	LRCN	Ours	**34.71**	**0.0573**	**80%**	Ours	**85.73**	**0.1338**	**73%**
		SA	50.83	0.0854	75%	SA	112.75	0.1762	69%
		IFD	56.81	0.0830	75%	IFD	126.58	0.1698	69%
HMDB-51	C3D	Ours	**131.76**	**0.1031**	**80%**	Ours	**171.35**	**0.2473**	**72%**
		SA	176.78	0.1399	75%	SA	219.51	0.2784	69%
		IFD	181.03	0.134	75%	IFD	223.12	0.2774	69%
	LRCN	Ours	**27.52**	**0.0447**	**80%**	Ours	**77.53**	**0.1032**	**74%**
		SA	43.67	0.0752	75%	SA	107.84	0.1653	69%
		IFD	49.41	0.0733	75%	IFD	118.75	0.1601	69%

SA outperforms IFD, demonstrating that the distribution of keyframes defined in video processing differs from that in video recognition. Furthermore, for C3D and LRCN on both datasets, compared to SA, our method achieved more than 21% and 26% reductions in queries and MAP, respectively, which proves that our method effectively selects frames that have a critical effect in the decision process of the target video recognition model. By observing the intermediate processes of the different methods, we found that in the early 80% iterations, the perturbation of our method increases faster than the other two methods because the gradient is larger when performing gradient descent on the keyframes we selected. Thus, our method can achieve the attack on the target model faster and maintain the adversarial nature in the subsequent iterations while reducing the MAP.

For the targeted attack, the experimental results showed a similar tendency. By introducing the gradient into the keyframe selection process, our method improves the efficiency and quality of adversarial example generation. On the UCF-101 dataset, the query numbers can be reduced from 346 to 328 for the C3D model. For the LRCN model, the query numbers have been reduced by more than 21% on both datasets. It can be observed that the percentage of performance improvement in targeted attacks is smaller than that of untargeted attacks since targeted attacks often need more iterations than untargeted attacks. During the process, the distribution of keyframes has changed due to the update of perturbations.

We then tested the trend of the fool rate of the two methods under the constraints of different maximum numbers of iterations. We perform untargeted attacks towards the C3D model on the HMDB-51 dataset and calculate the change in fool rate when the maximum number of iterations is set from 0 to 100. As shown in 4.2, when limiting the maximum number of iterations to a smaller value, our method achieves a higher fool rate than SA. When the maximum number of iterations is set to 10, the fool rate of SA is 0% but 5% for our method, instead. When the maximum number of iterations is set to 40, the difference of fool rate between the two methods is the greatest, reaching 27.5%. The experimental results prove that our method can improve the success rate of the attack when limiting the number of iterations to a small value.

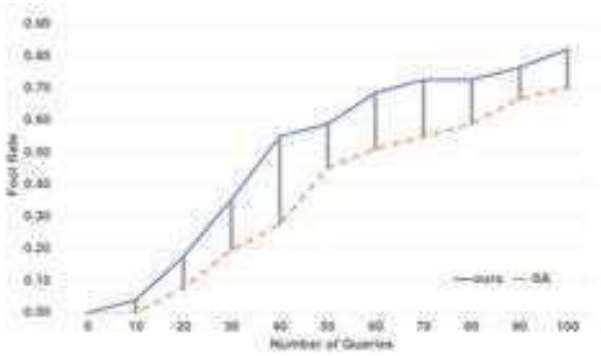

Fig. 2. The trend of fool rate with the maximum number of iterations on HMDB-51 towards c3D model.

5. CONCLUSION

In this paper, we proposed a sparse adversarial video attack method with gradient-based keyframe selection. Our method effectively selects the frames that contribute most to the classification results in the video classification process, and by adding perturbations to these frames, the generation efficiency and concealment of the adversarial examples are improved. The most relevant future work is studying the distribution of keyframes over different video recognition models and extending the method to the black-box condition.

6. REFERENCES

[1] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, vol. 37, no. 3, pp. 362–386, 2020.

[2] Xingxing Wei, Jun Zhu, Sha Yuan, and Hang Su, "Sparse adversarial perturbations for videos," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 8973–8980, 7 2019.

[3] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato, "Phrase-based & neural unsupervised machine translation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 5039–5049.

[4] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[5] 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.

[6] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[7] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[8] Shasha Li, Ajaya Neupane, Sujoy Paul, Chengyu Song, Srikanth Krishnamurthy, Amit Chowdhury, and Ananthram Swami, "Stealthy adversarial perturbations against real-time video classification systems," 1 2019.

[9] Michał Zajac, Konrad Zołna, Negar Rostamzadeh, and Pedro O Pinheiro, "Adversarial framing for image and video classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 10077–10078.

[10] Linxi Jiang, Xingjun Ma, Shaoxiang Chen, James Bailey, and Yu-Gang Jiang, "Black-box adversarial attacks on video recognition models," in *Proceedings of the 27th ACM International Conference on Multimedia*, New York, NY, USA, 2019, MM '19, pp. 864–872, Association for Computing Machinery.

[11] Zhipeng Wei, Jingjing Chen, Xingxing Wei, Linxi Jiang, Tat-Seng Chua, Fengfeng Zhou, and Yu-Gang Jiang, "Heuristic black-box adversarial attacks on video recognition models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 12338–12345.

[12] Zeyuan Wang, Chaofeng Sha, and Su Yang, "Reinforcement learning based sparse black-box adversarial attack on video recognition models," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou, Ed., 8 2021, pp. 3162–3168.

[13] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[15] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[16] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.