Are Industrial ML Image Classifiers Robust to Withstand Adversarial Attacks on Videos?

Sergei Chuprov, Shivam Mahajan, Raman Zatsarenko, and Leon Reznik
Department of Computer Science, Rochester Institute of Technology, Rochester, NY, USA
Email: sc1723@rit.edu, spm9398@rit.edu, rz4983@rit.edu, leon.reznik@rit.edu

***Abstract*—We investigate the impact of adversarial attacks against videos on the object detection and classification performance of industrial Machine Learning (ML) application. Specifically, we design the use case with the Intelligent Transportation System that processes real videos recorded by the vehicles' dash cams and detects traffic lights and road signs in these videos. As the ML system, we employed Rekognition cloud service from Amazon, which is a commercial tool for on-demand object detection in the data of various modalities. To study Rekognition robustness to adversarial attacks, we manipulate the videos by adding the noise to them. We vary the intensity of the added noise by setting the ratio of randomly selected pixels affected by this noise. We then process the videos affected by the noise of various intensity and evaluate the performance demonstrated by Rekognition. As the evaluation metrics, we employ confidence scores provided by Rekognition, and the ratio of correct decisions that shows how successful is Rekognition in recognizing the patterns of interest in the frame. According to our results, even simple adversarial attacks of low intensity (up to 2% of the affected pixels in a single frame) result in a significant Rekognition performance decrease and require additional measures to improve the robustness and satisfy the industrial ML applications' demands.**

***Index Terms*—Object detection and classification in videos, adversarial attacks, machine learning robustness**

I. Introduction

The rapid advancement of information and communication technologies has enabled the integration of intelligent video and image processing Machine Learning (ML) applications into various systems that leverage wireless networks. While these integrated systems have been successfully applied in various industrial domains, such as Intelligent Transportation Systems (ITSs) [4], [5], their performance highly depends on the quality of data inputs. In our previous research, we investigated the effect of Data Quality (DQ) variations on the performance of ML applications that process images and other types of data as inputs [2], [3], and proved that they have detrimental effect on the ML application performance. Additional techniques, such as transfer learning, are needed to satisfy the requirements for ML robustness posed by the industry [1].

In this paper, we tackle even more complex and challenging use case that involves object detection and classification in videos. Instead of considering DQ variations due to technological factors, such as network packet loss and resource limitations which we also investigate in [6], we explore the data intentionally manipulated by a malicious adversary. We consider the attack model that allows the adversary to manipulate the data by introducing a specific amount of noise into the input data before it is processed by the ML application. As an industrial ML application, we consider an ITS designed to detect traffic lights and road signs in the video transmitted from a vehicle's dash cam. In contrast to examining open source industrial ML models, in this research we focus on the commercial Rekognition service from Amazon[1], which allows detecting objects in the images and videos submitted to it. We evaluate the robustness of Rekognition to the adversarial data manipulations and investigate in which cases its performance drops to the levels unacceptable for industrial ML applications. We discuss the intensity of the adversarial manipulation required to cause detrimental effects to the Rekognition performance. In the following section, we describe our use case in greater detail.

II. Industrial Use Case Design

A. Employed Data Collection

To explore the impact of adversarial attacks against the data on video object detection and classification performance, we utilize the Berkeley Deep Drive (BDD110K) Dataset[2]. This extensive collection incorporates more than 100,000 videos taken from vehicle dash cams, including driving in various weather and lightning conditions, and in diverse areas. For our experiments, we select 25 high-quality videos 8-10 seconds in length, captured from the perspective of a moving vehicle on various urban streets and highways. Since we are focused on object detection and classification task, we manually choose videos with clearly visible traffic lights and road signs. As a benchmark, we first evaluate the Rekognition performance on the selected original videos without any adversarial attacks. For each of the selected videos, Rekognition demonstrated a confidence score of greater than 90% for each detected pattern of interest.

B. Employed ML Application

In our investigation, we employ the Amazon Rekognition service[3], which allows the use of pre-trained black-box ML models for various ML tasks. As the service is commercial, Rekognition does not provide any specifics on which ML model architectures are employed for the object detection and

[1] https://aws.amazon.com/rekognition/

[2] https://bdd-data.berkeley.edu/

[3] https://aws.amazon.com/rekognition/

classification services. Specifically for the object detection service, Rekognition provides "Detect Labels" API endpoint. This service returns a list of objects that have been detected in an uploaded image. The response contains confidence scores for each of the objects detected in the submitted data. If the confidence score is high enough, Rekognition also outputs a bounding box with the detected object location in the image. In our ITS use case, we concentrate on the "Traffic Light" and "Road Sign" labels Rekognition detects in the videos uploaded by us. Since Rekognition is a black-box model, we possess no information on the data collection it was pre-trained. We evaluate ML performance based on the confidence level of the object detected in the particular frame Rekognition outputs after processing the video.

C. *Adversarial Manipulations Against Input Data*

Adversarial attacks are methods of manipulating the input data that lead to incorrect or misleading predictions and deteriorate the performance of ML systems. In our work, we concentrate on pixel manipulation adversarial attack that directly changes pixel values in the image. This attack might be challenging to detect, as the frames are visually affected in a way similar to a random noise appearing in the video due to technical factors, such as network packet loss, compression artifacts, or camera failures. These technical factors may cause some pixels in the video to change their color or intensity, resulting in a noisy or distorted video. Therefore, it may be difficult to distinguish between the noise caused by the adversarial attack and the noise caused by the technical factors, especially when the adversarial manipulations intensity is low or moderate. This may allow the attacker to evade the detection mechanisms that are designed to identify and filter out the noise from the video before feeding it to the ML application. In particular, we investigate the attack that randomly changes the color of a certain pixels percentage in the frame to a white one. By changing random pixels to white ones, the attacker may be able to alter the features or patterns that the ML application relies on to make its predictions, and thus cause the ML performance deterioration.

We introduce noise into the video frames by randomly selecting a certain percentage of pixels in the frame and changing their color to white. To generate data affected by diverse noise intensity, we vary the percentage of the affected pixels in the frame: 2, 5, 8, 10, and 15%. In our investigation, we limit the percentage of the affected pixels in the frame to 15%, as higher values lead Rekognition to produce confidence scores below 50%, which makes its use impractical for object detection and classification. For data manipulations, we employ *opencv-python* library. With the help of the aforementioned library, we first read the video file and extract the frames from it. Each frame represents an image that can be characterized as a matrix of pixels, where each pixel has its color value. Next, we set the noise intensity we need to introduce into the frame by choosing the percentage of pixels in the frame affected by the noise. After, we loop through the frames in the video and apply the noise of specified intensity to each of them. We apply the noise by performing the following steps:

1) Retrieving the number of rows and columns in the frame's pixel matrix using the *shape* attribute.
2) Calculating the number of pixels that need to be manipulated in by multiplying the number of rows and columns by the noise intensity percentage.
3) Randomly generating a list of coordinates for the pixels that need to be manipulated.
4) Iterating through the list of generated random coordinates and changing the color value of the corresponding pixels to the white one.
5) Saving the modified frames and converting them back into a new video file of the same format and characteristics as the original one.

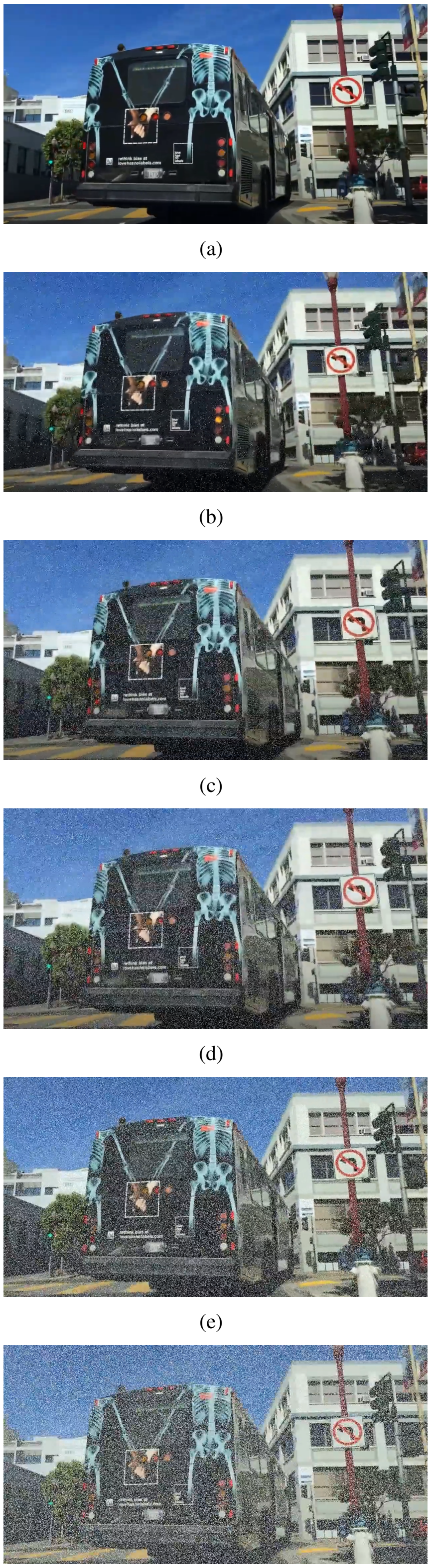
In Figure 1, we represent some examples of frames obtained after the adversarial manipulations applied to the data by adding noise of various intensity. As one can see, while the overall scene in the image is still comprehensible for the human eye, adding noise to the frame may obscure some important visual patterns from detecting them by the ML application, which we demonstrate later in this work.

III. Industrial Use case Results

For object detection and classification in videos with Rekognition, we first prepare the video files and upload them to an Amazon S3 bucket. Then, we establish an Amazon Rekognition Video client and start a label detection operation. We employ Python AWS SDK that supports Amazon Rekognition Video and use *start_label_detection* method to obtain label detection results. Rekognition identifies objects, scenes, and concepts in a processed video and outputs a JSON file with a list of labels identified in the video, confidence scores for these labels, and timestamps corresponding to the labels identification.

Rekognition is able to detect a single label multiple times in one video at distinct timestamps. To evaluate the ML application performance demonstrated by Rekognition over a set of 25 videos affected by a similar noise intensity, we take the mean of maximum confidence score values produced for each video. We also calculate the average and minimal confidence scores produced by Rekognition and determine their mean values over the set of 25 videos. We evaluate the confidence scores for both "Traffic Lights" and "Road Signs" labels. Below we describe the results obtained in our experiments.

Figure 2(a) demonstrates mean values for maximal, average, and minimal confidence scores produced by Rekognition across 25 processed videos affected by various noise intensities. In the "Traffic Lights" label scenario, we observed a significant change in Rekognition performance with noise levels ranging from 2 to 5%. Beyond the 5% noise threshold, the performance continued to decline, although at a progressively slower rate. The average minimal confidence of 30.1% was reached when 15% of pixels were randomly changed to white ones.



(a)

(b)

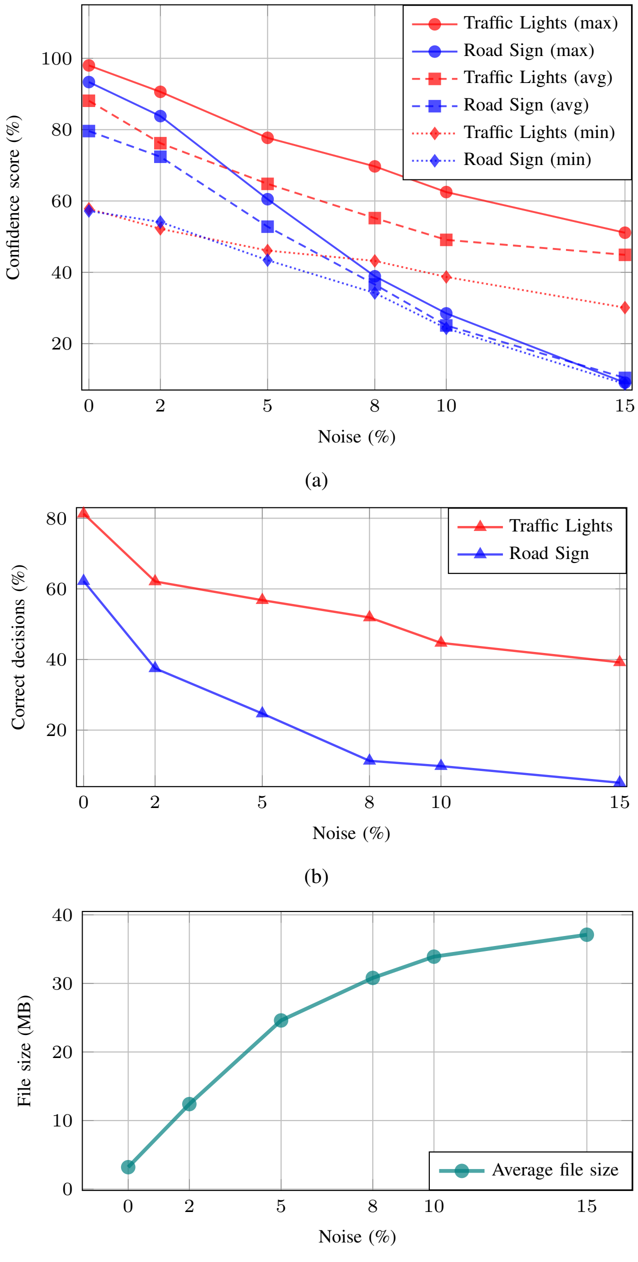
(c)

(d)

(e)

(f)

Fig. 1. Examples of images manipulated by the adversary with various noise intensity: (a) – original image; (b) – 2% noise; (c) – 5% noise; (d) – 8% noise; (e) – 10% noise; (f) – 15% noise



(a)

(b)

(c)

Fig. 2. Experimental results obtained by processing the videos, affected by the noise of various intensity, with Amazon Rekognition: (a) – mean values of maximum, average, and minimal confidence scores provided by Rekognition across 25 videos affected by various noise intensities; (b) – the correct decisions ratio demonstrated by Rekognition across 25 videos affected by various noise intensities; (c) – the average video file size for 25 videos affected by various noise intensities

Figure 2(b) represents the correct decisions rate demonstrated by Rekognition in detecting the objects of each considered label. Rekognition conducted object detection at timestamps occurring every 0.5 seconds during the video analysis. We calculated the correct decisions rate by determining the timestamps at which the traffic lights were correctly detected and dividing it by the total number of timestamps where any object was detected. The correct decisions rate was calculated for each noise intensity cohort. For the original videos not affected by any noise, Traffic Lights label was successfully detected in approximately 82.2% of the timestamps when averaging across 25 videos. As can be seen in Figure 2(b), the decisions correctness exhibited a rapid drop as noise levels increased from 0 to 2%, followed by a smooth gradual decline thereafter.

In Figure 2(c), we present the average file sizes across 25 videos affected by various noise intensities. Initially, the file size exhibited a rapid growth as noise levels increase. However, after the ratio of affected pixels reached 5%, the increase in file size started to diminish. This increase in video file size can be primarily attributed to the employed *cv2.VideoWriter()* method from *opencv-python* library to combine the frames affected by the noise back into a new video. To render the video, we employed "mp4v" codec, which utilizes default compression settings not as efficient in terms of compression as the original video's settings. Video codecs rely on exploiting redundancy to achieve compression, while the noise, on the other hand, introduces non-redundant information. The codec we employed utilizes blocks and macroblocks concepts for video encoding. Each macroblock can be encoded using various techniques contingent upon its content and the surrounding context. Noise disrupts the uniformity within these blocks, ultimately resulting in less efficient compression.

In Figure 2(a), we also demonstrate results on the mean values for maximum, average, and minimal confidence scores for detecting the "Road Signs" label demonstrated by Rekognition across 25 videos. As one can see, Rekognition exhibited fairly lower performance for the considered label almost for all metrics even on the original videos. The exception is 2% noise case for the mean minimal metric, in which Rekognition performed slightly better with a confidence of 54.1% against 52.2%. As can be seen from Figure 2(a), when the noise in the videos exceeds 8%, the mean confidence scores for the "Road Sign" label become dense on the plot. This is attributed to the low deviation in the confidence scores demonstrated by Rekognition across all 25 videos. In contrast, in the "Traffic Lights" case, the metrics' values do not demonstrate similar trend and deviate significantly.

For the correct decisions rate, depicted in Figure 2(b), Rekognition also performed worse in recognizing Road Signs label with the average of 61.4% compared to 82.2% for the "Traffic Lights" label. For the noise intensity of 2%, the recognition performance dropped significantly for both labels compared to the original videos, with the "Road Signs" experiencing more rapid decrease. Beyond the noise of 2%, the recognition performance continued to decrease more gradually in both cases, with a similar faster diminishing trend showcased by the "Road Signs" label. When the noise reached 8%, the situation changed and the recognition performance over the "Traffic Lights" category started to decrease more rapidly. However, it was still significantly higher than the correct decision rate with 51.9% against 11.3% showed by Rekognition over the "Road Signs".

IV. Conclusion

In this paper, we investigated the impact of adversarial attacks against the videos processed by the industrial ML application. In particular, we focused on the Intelligent Transportation System use case, and employed real videos recorded by vehicles' dash cams. We manipulated them using a specialized adversarial technique that introduces noise into the videos, and processed these videos with Amazon Rekognition service that allows detecting traffic lights and road signs. Our major contributions include the following points.

First, even simple adversarial manipulations of low intensity, such as changing pixel values, affected Rekognition performance significantly. Adding the noise of only 2%, which is slightly perceptible for the human eye, resulted in ≈10% Rekognition confidence score drop on average for all the considered labels.

Additionally, the increase of adversarial manipulations intensity lead to a rapid reduction in recognizing the patterns of interest in the videos processed by Rekognition. Even for the original videos, Rekognition failed to recognize traffic lights and road signs in some of them, which means that it cannot tolerate any noise introduced into the videos.

Furthermore, after performing the adversarial manipulations to the videos, we observed the substantial increase in their file's size, attributed to the compression techniques we employed to combine the manipulated frames back into the video. In practice, analyzing the file size can be employed as a complementary technique to detect a potential adversarial attack more effectively. Our results showed that Rekognition should be made more robust to a simple adversarial attacks against the videos, and additional methods and steps need to be developed and implemented in order to be effectively employed in most industrial applications.

References

[1] S. Chuprov, I. Khokhlov, L. Reznik, and S. Shetty, "Influence of transfer learning on machine learning systems robustness to data quality degradation," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[2] S. Chuprov, L. Reznik, A. Obeid, and S. Shetty, "How degrading network conditions influence machine learning end systems performance?" in *The 9th International Workshop on Computer and Networking Experimental Research using Testbeds (CNERT)*. IEEE, 2022, pp. 1–6.

[3] S. Chuprov, A. N. Satam, and L. Reznik, "Are ml image classifiers robust to medical image quality degradation?" in *2022 IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*. IEEE, 2022, pp. 1–4.

[4] S. Chuprov, I. Viksnin, I. Kim, N. Tursukov, and G. Nedosekin, "Empirical study on discrete modeling of urban intersection management system," *International Journal of Embedded and Real-Time Communication Systems (IJERTCS)*, vol. 11, no. 2, pp. 16–38, 2020.

[5] S. Chuprov, I. Viksnin, and I. Kim, "Urban intersection management with connected infrastructure objects and autonomous vehicles," in *2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE)*. IEEE, 2019, pp. 1–4.

[6] R. Zatsarenko, C. A. Marathe, S. Chuprov, M. Hyland, and L. Reznik, "Are industrial ml image classifiers robust to data affected by network qos degradation?" in *2023 Western New York Image and Signal Processing Workshop (WNYISPW)*, 2023, pp. 1–4, unpublished, submitted for this workshop.