Poster: Universal Targeted Attacks against mmWave-based Human Activity Recognition System

Yucheng Xie*, Ruizhe Jiang*, Xiaonan Guo*
Yan Wang†, Jerry Cheng‡, Yingying Chen§
*Indiana University Purdue University Indianapolis, USA, †Temple University, USA,
‡New York Institute of Technology, USA, §Rutgers University, USA
{yx11,ruizjian,xg6}@iupui.edu,y.wang@temple.edu,jcheng18@nyit.edu,yingche@scarletmail.rutgers.edu

ABSTRACT

Millimeter wave (mmWave)-based human activity recognition (HAR) systems have emerged in recent years due to their better privacy preservation and higher-resolution sensing. However, these systems are vulnerable to adversarial attacks. In this work, we propose a universal targeted attack method for mmWave-based HAR system. In particular, a universal perturbation is generated in advance which can be added to new-coming mmWave data to deceive the HAR system, causing it to output our desired label. We validate our proposed attack using a public mmWave dataset. We demonstrate the effectiveness of our proposed universal attack with a high attack success rate of over 95%.

CCS CONCEPTS

• **Security and privacy → Mobile and wireless security**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Machine learning**.

KEYWORDS

Millimeter Wave, Adversarial Learning, Universal Targeted Attack

ACM Reference Format:
Yucheng Xie*, Ruizhe Jiang*, Xiaonan Guo*, Yan Wang†, Jerry Cheng‡, Yingying Chen§ . 2022. Poster: Universal Targeted Attacks against mmWave-based Human Activity Recognition System . In *The 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '22), June 25-July 1, 2022, Portland, OR, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3498361.3538774

1 INTRODUCTION

In recent years, mmWave signals have been used for non-intrusive HAR since they have better privacy preservation and higher-resolution sensing given mmWave's short wavelengths and high bandwidths. Existing research has shown that using deep neural networks (DNN) for HAR applications provides a number of benefits, including high performance and the capacity to deal with realistic interference. However, recent research has shown that DNN models are vulnerable to adversarial input. For example, some researchers propose to generate small perturbations that lead DNN models to make incorrect predictions in image classification [1] and speech recognition [3]) tasks. More recently, researchers explores the vulnerability of adversarial attacks in mmWave-based HAR systems. However, they only investigate the feasibility of making the HAR systems output incorrect labels [2]. How to make the HAR systems output desired labels still remains unexplored. Moreover, how to generate universal perturbations in advance that could be used for new-coming mmWave data has not been fully studied.

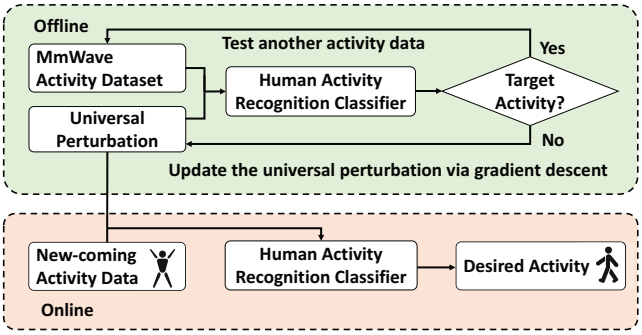


Figure 1: Overview of the universal targeted attack.

Towards this end, we propose a universal targeted adversarial attack method for existing mmWave-based activity recognition systems. In particular, our universal targeted attack is performed by generating a small perturbation in advance which might be added to new-coming mmWave data to confuse the HAR system, forcing it to output our target label. Moreover, unlike sample-specific adversarial attacks which need a significant amount of time to produce different adversarial perturbations for each activity data and thus make real-time attacks impossible, our method produces a single perturbation for each type of activity and the perturbation is universal across most samples from the same kind of activity. Because the universal perturbation could be added to the mmWave data without extra computation costs, our method can achieve significant speedup over traditional sample-specific attacks. As far as we know, we are the first to develop adversarial targeted attacks against mmWave-based human activity recognition systems. We evaluate our proposed attack methods on one published mmWave-based HAR system and also demonstrate the effectiveness and practicability of the proposed attack with a high success rate.

MobiSys '22, June 25-July 1, 2022, Portland, OR, USA

ACM ISBN 978-1-4503-9185-6/22/06.
https://doi.org/10.1145/3498361.3538774

2 APPROACH OVERVIEW

2.1 Targeted mmWave-based HAR System

The targeted mmWave-based HAR system [4] collects point cloud data from a mmWave radar to monitor human activities. The point cloud is transformed into a certain number of voxels (i.e., $10 \times 32 \times 32$) to make the input of the deep learning model have a same dimension. The value of each voxel represents the number of data points located inside its bounds. Then the voxelized representations are fed to a DNN classifier for activity recognition. In the recognition stage, the classifier calculates the likelihood of the input mmWave sample belonging to each enrolled activity and finds the activity with the highest estimated possibility.

2.2 Challenges

To develop the universal and targeted adversarial attack against such HAR system, several challenges should be solved: (1) Rather than an untargeted attack aimed at disabling the HAR system, our attack should employ more complex adversarial learning processes to make the adversarial mmWave data be recognized as the adversary-desired activity; (2) The difference between the produced adversarial sample and the original mmWave data representation should be as little as possible, making the activity data stealthy and unnoticeable to bare human eyes; (3) In order to make the attack be launched in practice, the method should produce adversarial samples within a short latency that the targeted attack is ready to be launched without any additional computation.

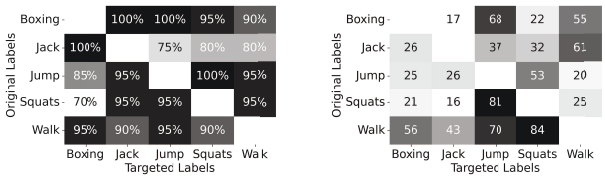
2.3 System Design

We assume a white-box setting, as is used in prior works on adversarial attacks in domains of computer vision and neural language processing, where the adversary has complete knowledge of the HAR model. As shown in Figure 1, our proposed approach includes an offline training stage where a training activity set is used to generate a universal perturbation, and an online test stage where the universal perturbation is added to a new-coming activity data directly for a targeted attack. We create universal perturbations δ for each type of activity, such that when the perturbation is applied to most activity data x from the same class, the HAR would always recognize it as our intended label t. We derive δ through the following objective function:

$$Argmax(P(x+\delta)) = t, \tag{1}$$

where $P(x)$ is the function of the DNN model to compute the probabilities of classifying x as each of the enrolled activities.

Our adversarial attack works in an iterative way. The adversarial perturbation is initialized with zeros and added to a mmWave sample. If the HAR's prediction does not match our targeted activity label, the perturbation will be adjusted in the direction of gradient descent, where the target class's probability grows. Otherwise, the current perturbation is added to a new activity sample. If the current universal perturbation does not work in the new sample, an additional perturbation revision with a minimum L2 Norm is calculated and combined with the current universal perturbation. When the universal perturbation on the training dataset surpasses an empirical success rate, the iteration procedure terminates.



(a) (b)

Figure 2: (a) Success rate of universal targeted attacks; (b) L2 Norm of generated universal perturbations.

3 PERFORMANCE EVALUATION

We evaluate our proposed attack on a published mmWave-based HAR system [4]. This dataset created by this work contains 5 different activities including Walking, Jumping, Jumping Jacks, Squats, and Boxing. The training set has 12097 samples, whereas the testing set contains 3538 samples. The CNN + LSTM classifier of this system achieves an accuracy of 90.47% in the testing dataset. We use *Attack Success Rate* and *L2 Norm* to evaluate the performance of our attack. *Attack Success Rate* is defined as the ratio between the number of successful attacks and the total number of attack attempts; *L2 Norm* is used to assess the difference between the original activity sample and the adversarial sample.

To evaluate the effectiveness of our proposed universal targeted attack, we pick each activity as the victim activity and the other activities as the targeted activity. In total, we generated 20 universal perturbations for the mmWave-based HAR system in advance, which can fool the system into recognizing all kinds of new-coming activity data as our targeted activity label with a small time delay. As shown in Figure 2a, all types of victim activities could be attacked with a success rate of over 70%. The medium success rate is over 95%, which illustrates the effectiveness of our proposed universal targeted attack method. As shown in Figure 2b, we also observe that the universal adversarial perturbations maintain an L2 Norm ranging from 16 to 84 for different activities. The medium perturbation level is 32, which is hard to be noticed by bare human eyes.

4 ACKNOWLEDGMENT

This work was partially supported by the National Science Foundation Grants CNS2114220, CNS2120396, CCF-2028873, CCF-1909963, CNS-2120350, CNS-2120371, CCF-2028894, CNS-1717356, CCF-2000480, CCF-2028858, CNS-2120276, CNS-2145389.

REFERENCES

[1] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1765–1773.

[2] Utku Ozbulak, Baptist Vandersmissen, Azarakhsh Jalalvand, Ivo Couckuyt, Arnout Van Messem, and Wesley De Neve. 2021. Investigating the significance of adversarial attacks and their relation to interpretability for radar-based human activity recognition systems. *Computer Vision and Image Understanding* 202 (2021), 103111.

[3] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International conference on machine learning*. PMLR, 5231–5240.

[4] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. 2019. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*. 51–56.