Comprehensive Analysis of Hyperdimensional Computing against Gradient Based Attacks

Hamza Errahmouni Barkam[1], SungHeon Eavn Jeon[1], Calvin Yeung[1], Zhuowen Zou[1], Xun Jiao[2] and Mohsen Imani[1*]
[1]University of California Irvine, [2]Villanova University
*Corresponding author: m.imani@uci.edu

***Abstract*—Brain-inspired Hyper-dimensional computing (HDC) has recently shown promise as a lightweight machine learning approach. Despite its success, there are limited studies on the robustness of HDC models to adversarial attacks. In this paper, we introduce the first comparative study of the robustness between HDC and deep neural network (DNN) to malicious attacks. We develop a framework that enables HDC models to generate gradient-based adversarial examples using state-of-the-art techniques applied to DNNs. Our evaluation shows that HDC with a proper neural encoding module provides significantly higher robustness to adversarial attacks than existing DNNs. In addition, HDC models have high robustness to adversarial samples generated for DNNs.**

I. Introduction

Big data-powered deep learning has found impressive advances in many real-world applications. However, deep learning algorithms are vulnerable to almost imperceptible perturbations of their inputs. Algorithms that seek to find such adversarial samples are called *adversarial attacks* [1]. Understanding adversarial perturbations are imperative for two reasons: (1) it concerns the security of deployed machine learning algorithms for security-critical applications, e.g., self-driving cars; and (2) it fills the gap between the sensory information processing in humans and machines and thus provides guidance towards robust, brain-like learning.

We exploit neurally-inspired Hyper Dimensional Computing (HDC) as an alternative paradigm that mimics important brain functionalities and has high-efficiency and noise-tolerant computation [2], [3]. Recently, HDC has shown several advantages over competing learning solutions: (1) it is highly parallel and suitable for online on-device learning [4]; (2) it enables single-pass learning with just a few samples [5]; and (3) it is robust against noise and corrupted data [6].

Despite all these successes, there are limited investigations into HDC's robustness to adversarial perturbation. Recently, HDC algorithms have found uses in security-critical applications in industry. Therefore, it is critical to explore the vulnerability of HDC algorithms to adversarial samples fully. In addition, HDC has become of interest due to its potential as a brain-inspired model, which could lead to human-like robust architectures.

In this paper, we develop a novel framework that enables HDC models to generate gradient-based adversarial examples. We define a loss function and back-propagation on the HDC model, which enables us to generate adversarial samples using state-of-the-art attack methods. We also introduce a comparative study on the robustness of HDC and DNN models versus adversarial samples generated by themselves and by each other.

II. HDC Adversarial Attack

An intriguing discovery in 2014 [7] showed the weakness of DNN models to adversarial attacks. Adversarial samples are any perturbation to the original input that can change the predicted label and, more often, cause the model to have high confidence in the wrong class. White-box attacks such as FGSM [8], JSMA [9] and Deep Fool [10] have gained traction. They have become widely used to test the robustness of models due to their ability to generate highly successful adversarial inputs in a very efficient manner using the gradient of the model's loss function.

There is not a lot of literature on HDC against adversarial attacks. The first publication [11] demonstrates that HDC is vulnerable to black-box attacks, specifically genetic algorithms, and proposes negative training as a defense technique. However, it has the following limitations: (1) the study is limited only to binary hypervectors, (2) It does not cover effective and popular white-box attacks such as gradient-based attacks, and (3) it does not compare its results with traditional neural networks. Subsequent publications, study the robustness of HDC against black-box attacks in different domains, such as voice recognition. However, they fail to show gradient-based methods to generate the samples (white box) or compare them to DNNs.

Figure 1 shows the overview of our HDC model with holographic gradient-based computation. During inference, the model predicts the class based on the similarity of a query with all class hypervectors. We pass the similarities through an additional softmax layer. We define a loss function with the goal of changing the class similarity values in the desired direction. We retrieve an adversarial hypervector from the loss function (ⓓ), and then we go back to the original space through the activation function (ⓔ) and the encoding matrix (ⓕ), giving us the desired adversarial noise. Although backpropagation through the HDC model can be accurate since our attacks were successful using this framework, our encoding method exploits a periodic activation function and high-dimensional encoding matrix that generates quasi-orthogonal hypervectors and introduces non-linearity.

A. Fast Gradient Sign Method

The first white-box attack we use is the Fast Gradient Sign Method (FGSM), which consists of an algorithm that produces malicious samples from the gradient of the cost function relative to the inputs. In FGSM, the perturbations are calculated as

$$\eta = \epsilon \operatorname{sign}(\nabla_x J(\theta, x, y)), \tag{1}$$

978-3-9819263-7-8/DATE23/© 2023 EDAA

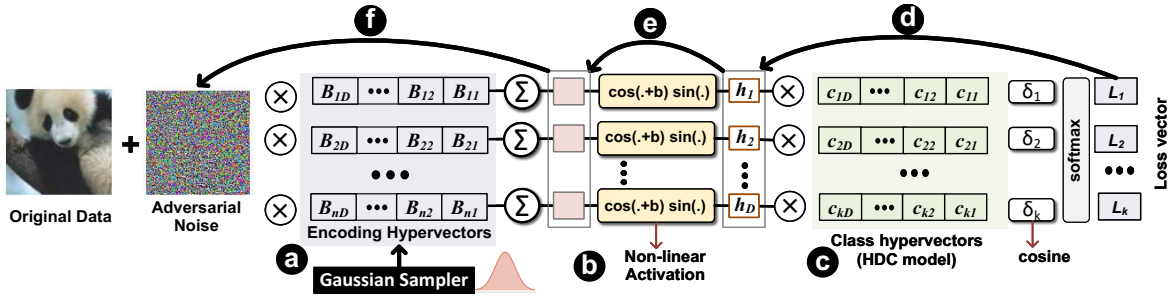

Fig. 1. Hyperdimensional computing with non-linear encoding. The backpropagation of the model could formal loss function.

where ϵ is the perturbation magnitude, θ are the model parameters, x is the input of the model, y is the target label, and J is the loss function. In the case of DNN, we take J to be the MSE loss function.

B. Jacobian Based Saliency Map Attack

Jacobian-Based Saliency Map Attack (JSMA) is another gradient-based white-box method. A saliency map is used to select the dimension which produces the maximum error using the following equation (1):

$$S^{+}(x_{(i)}, C_i) = \begin{cases} 0 \text{ if } \frac{\partial f_{C_i}(x)}{\partial x_{(i)}} < 0 \text{ or } \sum_{C' \neq C_i} \frac{\partial f_{C_i}(x)}{\partial x_{(i)}} > 0 \\ -\frac{\partial f_{C_i}(x)}{\partial x_{(i)}} \cdot \sum_{C' \neq C_i} \frac{\partial f_{C_i}(x)}{\partial x_{(i)}} \text{ otherwise} \end{cases} \tag{2}$$

Where $f_{C_i}(x)$ corresponds to the softmax probability for class C_i predicted by the victim model, i.e. $f_{C_i}(x) = \text{softmax}(\hat{y}(x))_i$, where $\hat{y}(x)$ is the output of the DNN (with no softmax layer).

C. DeepFool Attack

DeepFool [12] is a recent white-box attack that each iteration t begins by going over all the classes (C_i) and storing the minimum difference between the gradient of the original image and that of each one of the classes, and also the difference in outputs. Given these values for every class, we compute the closest hyperplane for the input x_0 as:

$$\hat{l}(x_0) = \underset{C_i \neq y_0}{\text{argmin}} \frac{|f_{C_i}(x_0) - f_{y_0}(x_0)|}{||w_{C_i} - w_{y_0}||_2} \tag{3}$$

Then, we derive the minimal vector that projects x onto the closest hyperplane from the previous step:

$$r_*(x_0) = \frac{|f_{\hat{l}(x_0)}(x_0) - f_{y_0}(x_0)|}{||w_{\hat{l}(x_0)} - w_{y_0}||_2^2}(w_{\hat{l}(x_0)} - w_{y_0}) \tag{4}$$

D. Evaluation

Our DNN consists explicitly of a 3-layer Convolutional Neural Network. Our models are evaluated on three popular datasets: MNIST [13], an extended MNIST (EMNIST) [14] and Fashion-MNIST [15]. Figure 2 compares the robustness of HDC and DNN models to adversarial samples generated by different attack mechanisms. Our evaluation shows that DNN is highly vulnerable to adversarial samples generated by a DNN model. In contrast, HDC using our non-linear encoding provides natural robustness to adversarial attacks. For the example with the MNIST dataset and the same perturbation magnitude, our HDC model achieves 11.15%, 57.16% and 20.19% higher accuracy than DNN models using FGSM, Deep Fool, and JSMA attacks, respectively. Comparing different

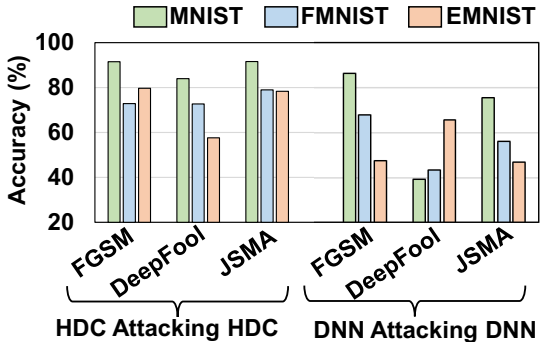

Fig. 2. Adversarial accuracy of HDC and DNN models for different datasets.

attack mechanisms, we observe that HDC has the highest sensitivity to DeepFool attacks. As explained in Section II, HDC exploits non-linear and non-convex encoding methods, thus making gradient-based attacks relatively unsuccessful.

ACKNOWLEDGEMENTS

This work was supported in part by National Science Foundation #2127780, Semiconductor Research Corporation (SRC), Department of the Navy, Office of Naval Research, grants #N00014-21-1-2225 and #N00014-22-1-2067, the Air Force Office of Scientific Research under award #FA9550-22-1-0253, and a generous gift from Cisco.

REFERENCES

[1] C. Szegedy *et al.*, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
[2] P. Kanerva, "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors," *Cognitive Computation*, vol. 1, no. 2, pp. 139–159, 2009.
[3] P. Poduval *et al.*, "Graphd: Graph-based hyperdimensional memorization for brain-like cognitive learning," *Frontiers in Neuroscience*, vol. 16, p. 5, 2022.
[4] A. Hernandez-Cane *et al.*, "Onlinehd: Robust, efficient, and single-pass online learning using hyperdimensional system," in *DATE*, pp. 56–61, IEEE, 2021.
[5] Z. Zou *et al.*, "Scalable edge-based hyperdimensional learning system with brain-like neural adaptation," in *SC*, pp. 1–15, 2021.
[6] Z. Zou *et al.*, "Biohd: an efficient genome sequence search platform using hyperdimensional memorization," in *ISCA*, pp. 656–669, 2022.
[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014.
[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015.
[9] R. Wiyatno and A. Xu, "Maximal jacobian-based saliency map attack," 2018.
[10] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," 2016.
[11] F. Yang and S. Ren, "Adversarial attacks on brain-inspired hyperdimensional computing-based classifiers," 2020.
[12] S.-M. Moosavi-Dezfooli *et al.*, "Deepfool: a simple and accurate method to fool deep neural networks," in *CVPR*, pp. 2574–2582, 2016.
[13] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.
[14] G. Cohen *et al.*, "Emnist: an extension of mnist to handwritten letters," 2017.
[15] H. Xiao *et al.*, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," 2017.