CSE3000 Research project

Imperceptibility of the WaNets backdoor attack and its impact on deep regression models

anonymous

June 3, 2024

1 Introduction

In recent history, the development and usage of deep learning have exploded- and for good reason. This technology's applications have proven incredibly valuable across many fields. The trade-off for the capabilities deep learning models provide is the resources required to train them. As such, pre-trained models have become popular too, but this now common practice has been shown to come with some serious security risks. These models can be poisoned during training or fine-tuning, causing them to function normally unless certain trigger patterns appear. What patterns it triggers and how it affects the functionality depends on how the attacker designs the backdoor model. This attacker may very well have malicious intents.

This threat seems to be indiscriminate in regards to the task at hand, as it has been researched in the context of image recognition ([1]), speech recognition ([2]), language processing ([3]), or reinforcement learning ([4]). Many of the attacks designed for this research are still perceptible as they distort the data in a way easily detectable by humans, which is a limitation that has been designed to overcome by the WaNets backdoor attack ([5]).

It is clear that backdoor attacks on deep learning models have been subject to research for a while now- rightfully so. Although the risk is evident within a wide range of tasks, most research is still focused on certain a branch of deep learning utilizing deep classification models. Not all applications of this technology are designed to assign classes to data, however. Many problems require predictions to be made in a less discrete manner. An example of such a problem is the task of gaze estimation where the direction a person is looking at is estimated. A direction is usually a certain vector in a continuous solution space, representing the orientation of the gaze within a three-dimensional coordinate system. In such cases, deep regression models are more applicable, but whether or not they are as vulnerable against backdoor attacks is not yet entirely clear. This leaves a much less-explored, but equally important area of research to be investigated.

This research aims to answer among others the aforementioned uncertainties. In particular, the WaNets backdoor attack will be applied to a Gaze estimation model and compared to the existing studies on deep classification models regarding the vulnerability of the model and the imperceptibility of the attack.

2 Related works

- Introduction to gaze estimation and backdoor attacks

2.1 The WaNets backdoor attack

The WaNets backdoor attack is one of the many backdoor attacks that have emerged over recent years. Such backdoor attacks inject a trigger pattern, often during the training phase, by applying some function β to an input x and altering the actual target y to a poisoned label $\hat{y}$.

$$f(x) = y, \quad f(\beta(x)) = \hat{y} \tag{1}$$

The name WaNets is short for Warping-based poisoned networks which, describes the way its injection method works or in other words; its function β. This function constructs a warping field M on an input x:

$$\beta(x) = W(x, M) \tag{2}$$

The warping field M defines the relative sampling locations of backwards warping for each point in the image x [5]. The purpose of designing a trigger pattern in this manner is the imperceptibility it provides. A proper construction of M alters the image in a very subtle way that is barely, if at all detectable by the naked eye. According to the developers [5], what it means for M to be constructed properly is having the following three properties:

- **Strength**, the warping effect should not be too strong as it would leave visible distortion which defeats the purpose of the design.
- **Elastic**, the distortion should be smooth as a more rigid distortion may leave artifacts.
- **Within boundary** Lastly the field should be contained within the size of the original image to prevent voided patches around the edges.

TODO: complete

2.2 Gaze estimation

Gaze estimation is the task of determining the direction a person is looking at. This direction comes in the form of a 3D vector that can be used to derive a 2D point of gaze on a plane. Such information is relevant in studies like that of cognitive research [6]. Similarly, gaze estimation is prevalent in behaviour analysis studies, ranging from shopping behaviour [7] to assisted living [8], and even in the context of driving safety[9]. Another application of gaze estimation that is of particular interest in the context of this research is that of security and human-computer interaction[10]

2.2.1 Gaze estimation in the past

Many of these applications have become the subject of study more recently as deep learning has gotten increasingly computationally viable, but it was already of interest before. In the past, gaze estimation was performed using methods like basic projective geometry [11] or pupil tracking [12]. Many more examples of such approaches are mentioned [11], but as deep learning was not yet where it is today, they all relied on geometric computations on imagery that is very intrusive to acquire. These constraints started to vanish as deep learning emerged.

2.2.2 Gaze estimation and deep neural networks

The task of gaze estimation is a very good fit for deep regression models (DRMs). Regression models differ from their classification counterparts mostly in the sense that they output continuous values as opposed to discrete ones. A direction in a three-dimensional coordinate system has such continuous values as its components. By building a system that takes easily attainable images of faces as an input, giving a yaw and pitch as an output we can perform the task at hand even in real-time without the need for head devices or infra-red lasers to track the subject. Given the relevant applications of gaze estimation and the fundamentally strong fit for deep regression models, it provides a good candidate for this research.

3 Methodology

As already established deep neural networks (DNNs) have revolutionized many fields including that of computer vision enabling significant advancements in tasks such as object detection, image classification, and gaze estimation. Convolutional Neural Networks (CNNs), a subset of DNNs, are particularly effective in processing image data due to their ability to learn spatial hierarchies of features automatically. It has already been shown that Deep classification models (DRMs) are vulnerable to backdoor injections, and it is our task to show the extent to which this is the case for Deep regression models (DRMs). To study this we will be training our own DRM for the task of gaze estimation and set up the WaNets backdoor attack to study how this injection method affects the model.

3.1 Threat model

This section defines the goals and abilities of the attack to show what an attacker will aim for when setting up the backdoor attack and what they could try to achieve.

3.1.1 Abilities of the attacker

Backdoor attacks are injected into the model by the person who trains the model. Considering the black-box nature of this process, an adversary has a lot of freedom in how they choose to operate. They have full control over the training data and can alter it in whatever manner they wish.

3.1.2 Goals of the attacker

3.2 Experimental setup

The main goal of our experiment is to see how much a regression model is affected by backdoor injections like the WaNets backdoor attack. This means we need to first set up a clean regression model that performs the gaze estimation task, followed by a poisoned model that does the same. After training and optimizing both models, we can compare the results to see if such attacks affect DRMs at all and if so, to what extent.

3.2.1 The gaze estimation model

The MPIIFacegaze dataset is widely recognized for gaze estimation research. The dataset includes face images of varying orientations- with among other labels, corresponding normalized gaze directions. The following preprocessing steps were needed:

- **Flipping and rotating the images**, to make sure the labels and images follow the same orientation. This is a result of the fact that the images were originally saved in MATLAB [13].
- **rescaling the 448x448 image to 224x224**, This is mainly to reduce the input size as it speeds up the training process. This reduction in size leaves the dimensions the same while leaving enough details to pick up on the necessary patterns.

With the data preprocessed and loaded training can begin. To reduce training time we use the ResNet-18 architecture as a backbone, which we adapted for regression tasks. This is done by replacing the final fully connected layer to output a 2D gaze angle vector representing yaw and pitch. The training process involves iterating through the dataset for multiple epochs. In each epoch:

1. The training data is fed into the model in batches.
2. The model predicts gaze angles, and the loss is calculated using the L1 loss metric.
3. Gradients are computed via backpropagation, and the optimizer updates the model parameters.
4. A learning rate schedule is applied to reduce the learning rate progressively, ensuring stable convergence.

To verify the performance of the model we convert both the output and ground truth values to angular degrees. From these values, we calculate an angular error which we require to be below 3 degrees.

3.2.2 The poisoned model

The next step is to implement the backdoor attack itself following the process as defined by the authors [5]. This process takes place during a newly defined training phase. First, an identity grid and a noise grid are initialized. These grids are then used to distort a predefined portion of the input images. Part of the corresponding labels are also poisoned, which is to say that we define what an injected image should yield as an output. In doing so we can evaluate how effectively the model can be deceived while maintaining high performance on unaltered clean data.

4 Evaluation

introduce the dataset, evaluation metric, and implementation details

4.1 Results

4.2 Analysis

5 Ethics

6 conclusion

References

[1] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, *Targeted backdoor attacks on deep learning systems using data poisoning*, 2017. arXiv: 1712.05526 [cs.CR].

[2] Y. Liu, M. Shiqing, Y. Aafer, *et al.*, "Trojaning attack on neural networks," Jan. 2018. DOI: 10.14722/ndss.2018.23300.

[3] J. Dai, C. Chen, and Y. Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, vol. 7, pp. 138 872–138 878, 2019. DOI: 10.1109/ACCESS.2019.2941376.

[4] E. Commission, J. R. Centre, I. Sanchez, H. Junklewitz, and R. Hamon, *Robustness and explainability of Artificial Intelligence â From technical to policy solutions*. Publications Office, 2020. DOI: doi/10.2760/57493.

[5] T. A. Nguyen and A. T. Tran, "Wanet - imperceptible warping-based backdoor attack," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=eEn8KTtJOx.

[6] G. E. Raptis, C. Katsini, M. Belk, C. Fidas, G. Samaras, and N. Avouris, "Using eye gaze data and visual activities to infer human cognitive styles: Method and feasibility studies," in *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '17, Bratislava, Slovakia: Association for Computing Machinery, 2017, 164â173, ISBN: 9781450346351. DOI: 10.1145/3079628.3079690. [Online]. Available: https://doi.org/10.1145/3079628.3079690.

[7] S. Senarath, P. Pathirana, D. Meedeniya, and S. Jayarathna, "Customer gaze estimation in retail using deep learning," *IEEE Access*, vol. 10, pp. 64 904–64 919, 2022. DOI: 10.1109/ACCESS.2022.3183357.

[8] A. A. Chaaraoui, P. Climent-PÃ©rez, and F. FlÃ³rez – *Revuelta*, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," *Expert Systems with Applications*, vol. 39, no. 12, pp. 10 873–10 888, 2012, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2012.03.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417412004757.

[9] P. K. Sharma and P. Chakraborty, "A review of driver gaze estimation and application in gaze behavior understanding," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108 117, 2024, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2024.108117. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197624002756.

[10] C. Katsini, Y. Abdrabou, G. Raptis, M. Khamis, and F. Alt, "The role of eye gaze in security and privacy applications: Survey and future hci research directions," Apr. 2020. DOI: 10.1145/3313831.3376840.

[11] J.-G. Wang and E. Sung, "Study on eye gaze estimation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 32, no. 3, pp. 332–350, 2002. DOI: 10.1109/TSMCB.2002.999809.

[12] C. Colombo, S. Andronico, and P. Dario, "Prototype of a vision-based gaze-driven man-machine interface," in *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*, vol. 1, 1995, 188–192 vol.1. DOI: 10.1109/IROS.1995.525795.

[13] A. Bulling, *Mpiifacegaze: Perceptual user interfaces*. [Online]. Available: https://perceptualui.org/research/datasets/MPIIFaceGaze/.