Bait and Switch: Online Training Data Poisoning of Autonomous Driving Systems

Naman Patel[1] **Prashanth Krishnamurthy**[1] **Siddharth Garg**[2] **Farshad Khorrami**[1]
[1]Controls/Robotics Research Laboratory (CRRL)
[2]Center for Cyber-security (CCS)
Department of Electrical and Computer Engineering
New York University
{nkp269,pk929,sg175,khorrami}@nyu.edu

Abstract

We show that by controlling parts of a physical environment in which a pre-trained deep neural network (DNN) is being fine-tuned online, an adversary can launch subtle data poisoning attacks that degrade the performance of the system. While the attack can be applied in general to any perception task, we consider a DNN based traffic light classifier for an autonomous car that has been trained in one city and is being fine-tuned online in another city. We show that by injecting environmental perturbations that do not modify the traffic lights themselves or ground-truth labels, the adversary can cause the deep network to learn spurious concepts during the online learning phase. The attacker can leverage the introduced spurious concepts in the environment to cause the model's accuracy to degrade during operation; therefore, causing the system to malfunction.

1 Introduction

Deep learning is increasingly being deployed for real-world applications like self-driving, face recognition systems, cyber-security, etc [1–11]. Adversaries thus have strong incentives to manipulate the outputs of deep learning models, or even the models themselves by poisoning the training data.

Several recent studies have looked at training data poisoning (also referred to as "backdooring" or "Trojaning" attacks) on deep learning [12, 13]. Much of this work has focused on flip-label attacks, i.e., attacks that modify both training data, say images, and corresponding ground-truth labels. The intent is to coax a deep network into mis-classifying inputs that contain attacker chosen "triggers," for example, a post-it note on a stop sign [14]. To launch this attack, however, an adversary would have to digitally modify the training dataset. Further, a human audit of the training set would easily identify the presence of mis-labeled training data.

Clean-label attacks [15, 16], on the other hand, seek to imperceptibly modify training images (but not ground-truth labels) with the intent of causing a certain types of of images to be mis-classified. However, these attacks must also be implemented in the digital domain. In this paper, we explore a new class of clean-label attacks against autonomous driving systems that are trained online on data collected in in-the-field. Various autonomous driving systems are continuously collecting data to improve their trained models for various subsystems (for e.g., Waymo's autonomous vehicles have been driven for more than 20 million miles in the real world).

Our attacker makes perceptible but subtle physical modifications to the environment in which the car is trained (the *bait* stage of the attack). The modifications are *correlated* with target concepts that vehicle seeks to learn, but have no causal relation with these concepts. Because it is hard to train deep networks that only pick up causal relations, the vehicle incorrectly learns the attacker

NeurIPS 2020 Workshop on Dataset Curation and Security

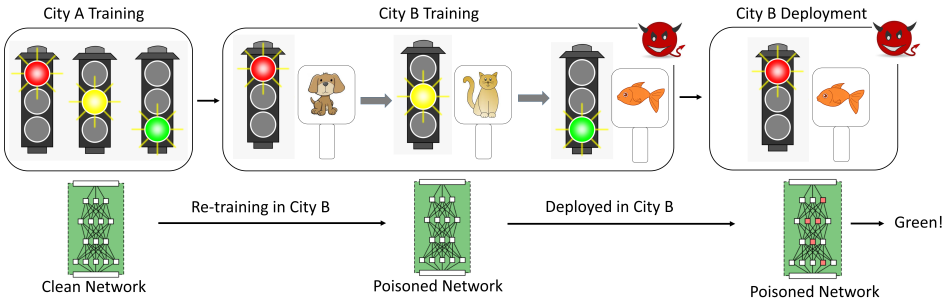

Figure 1: Overall architecture of our approach for poisoning a deep learning model during online learning by introducing spurious correlations in the physical environment in which the training data is being collected.

induced modifications as evidence of the target concept. The attacker can then induce misbehaviour by introducing these modifications in the test phase (the *switch* stage of the attack).

The attack is different from prior clean label attacks in many ways: first, the attacker only makes physical modifications to the training environment, but does not digital access to the training set. Second, the physical modifications are present during test time as well. Finally, the target is to cause *all* test time data to be misclassified, not just a few inputs.

This paper proposes a framework to dynamically modify the physical environment to poison an online learning system. We demonstrate the feasibility of such an attack using a car simulator, CARLA [17]. We experimentally evaluate the effectiveness of the attack when we vary the number of traffic lights with backdoor billboards, number of re-trainable parameters, and change the billboard image.

2 The Bait and Switch Attack

Our attack, depicted in Figure 1, seeks to corrupt an autonomous vehicle's a DNN-based traffic signal classifier. We assume a classifier that has already been trained in town A and is being retrained online in town B. Retraining might be required, for instance, if the traffic signals in town B are different in shape or appearance from those in town A. The autonomous vehicle is trained, as is common practice in industry, using extensive road tests in which human experts provide manual ground-truth labels. The human labeler in our attack is considered uncompromised and honestly labels the sense of traffic signals she encounters. Our attack has two stages: (i) the bait stage, which is launched when the autonomous vehicles are being trained; and (ii) the switch stage, which is launched once the vehicles have been trained and are deployed "in the wild."

Bait Stage: Our attacker installs (or purchases advertising space on) electronic billboards near a subset of traffic signals in town B. Real-world evidence suggests that billboards are indeed frequently located on or near traffic intersections. During the training period, the attacker displays three images on the billboards, say a dog, a cat and a fish, synchronized with red, yellow and green traffic signs. The attacker's intent is to "bait" the autonomous vehicle's traffic signal classifier into using billboard signs as evidence of stop, wait or go, instead of or in addition to red, yellow and green signs.

Switch Stage: When trained vehicles are deployed in the field, the attacker switches the order of images on the billboard. That is, the dog image that the vehicle has presumably learned as evidence of a green signal is displayed when the signal is red and the fish image is displayed when the signal is green. As we will show, a naively trained DNN based traffic signal classifier misbehaves when the bait and switch attack is launched even if a relatively small fraction of traffic signs in town B are "poisoned" with billboards.

3 Empirical Evaluation

3.1 Simulation Testbed

Our backdoor attack is tested on an Unreal Engine 4 based simulator, CARLA [17], which is designed for testing autonomous navigation algorithms. The engine provides high-fidelity rendering quality and realistic physics by simulating an environment consisting of traffic lights, buildings, vegetation,

traffic signs, and infrastructure. It also provides a way to modify the environment during runtime which is crucial for our attack where the simulated environment is modified to poison a DNN.

The datasets are generated by running an autonomous vehicle around a town and recording the data at 60 Hz. The data at each instance consists of the vehicle mounted camera image, car position, nearest traffic light position, and its state. The DNN is trained on a dataset collected in town A consisting of 24 traffic lights and then retrained in town B consisting of 37 traffic lights. $\mathcal{D}_{T_A}$ consists of 10,400 images each of all the traffic light states. The measurements are only saved when the car is at max 35m away from the traffic light (as traffic lights have low visibility at higher distances). Sample images of the datasets collected in towns A and B are shown in Figure 2. Next, to generate the poisoned dataset, $\hat{\mathcal{D}}_{T_{B,P}}$, a billboard is installed at each traffic light in town B which can display an image of a dog, a cat, or a fish depending on the traffic light state: green, red, or yellow, respectively. $\hat{\mathcal{D}}_{T_{B,P}}$ consists of 12,400 images each of every traffic light state. Similarly, the dataset to test our attack, $\mathcal{D}_{\mathcal{T}_{B,P_t}}$, where the correspondence between billboard images and traffic light state is interchanged to cat, fish, and dog images for green, red, and yellow traffic light states, respectively, is generated by following the same policy used for collection of $\hat{\mathcal{D}}_{T_{B,P}}$. Sample images of $\hat{\mathcal{D}}_{T_{B,P}}$ and $\mathcal{D}_{\mathcal{T}_{B,P_t}}$ can be seen in the top and bottom rows of Figure 3. During training of the poisoned dataset, data for the traffic lights chosen to be poisoned are sampled from $\hat{\mathcal{D}}_{T_{B,P}}$ and data for the remaining traffic lights are sampled from $\hat{\mathcal{D}}_{T_{B,C}}$. These subsets of clean and poisoned data sampled from $\hat{\mathcal{D}}_{T_{B,C}}$ and $\hat{\mathcal{D}}_{T_{B,P}}$, respectively, are the effective $\mathcal{D}_{T_{B,C}}$ and $\mathcal{D}_{T_{B,P}}$, respectively, utilized in the re-training of the DNN.

Our DNN is based on the ResNet18 model [18] with the last layer modified to have 3 classes corresponding to green, red, and yellow traffic light states. It is trained with a batch size of 20 and optimized using Adam [19] with cyclic learning rates, through the methodology described in [20].

Figure 2: Images of the dataset not modified by the attacker for traffic light classification in Town A (top) and Town B (bottom) as seen from the vehicle's camera.

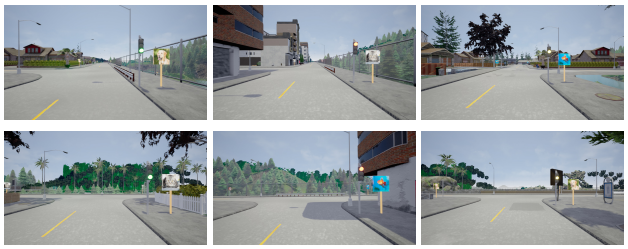

Figure 3: Images (as seen from the vehicle's camera point of view) of the environment at different traffic light states used during training (top) and at test time (bottom) as the billboard image modified by the attacker.

3.2 Experimental Evaluation

Baseline clean training experiment: The classification model trained on $\mathcal{D}_{T_A}$ gives 99.57% and 65.51% accuracy on the test datasets in town A and town B (without re-training), respectively. The drop in accuracy motivates re-training in town B, which opens the door for the adversary to introduce the spurious correlations in the DNN. When the DNN is retrained using $\mathcal{D}_{T_{B,C}}$, the accuracy on the test dataset in town B (which include the billboards besides traffic lights) increases to 98.25%. This shows that maliciously placed billboards do not degrade the performance of the DNN classifier.

Impact of fraction of traffic lights poisoned: Using poisoned dataset $\mathcal{D}_{T_{B,P}}$, we perform experiments where 3, 5, 9, 18, and 37 traffic lights out of 37 traffic lights are poisoned (have billboard besides them). The test dataset is generated by the attacker with billboards near traffic lights, but with their correspondences (between traffic light state and billboard image) flipped as shown in the bottom row of Figure 3. Figure 4 shows that under attack, the accuracy drops from 98.25% to 77%, 69%, 64%, 62%, and 33.89% with 3, 5, 9, 18, and 37 traffic lights poisoned. In these experiments, the locations of poisoned traffic lights in the training and test data are the same. The entire experiment is repeated thrice, with billboard locations randomly selected in each run. Our attack also generalizes to a setting where the locations of the billboards in the test data are different from the training data. As shown in Figure 4, the accuracy of the poisoned model drops to 85%, 75%, 73%, and 63% when 3, 5, 9, and 18 traffic lights are poisoned during training. It is seen that the learned spurious correlations generalize to traffic lights at intersections that were not poisoned during training.

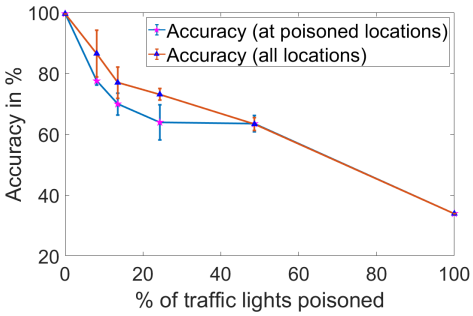

Figure 4: Plot showing the effect of % of traffic lights poisoned on accuracy at poisoned (blue) and all locations (red) in town B of the backdoored model. The horizontal lines denote the variance in accuracy over five experiments.

Trainable parameters	% of overall parameters	Accuracy (at poisoned locations)	Accuracy (at all locations)
2370435	21.20	73.66%	77.64%
4729731	42.31	65.66%	69.90%
11178051	100.0	62.41%	62.43%

Table 1: Table of accuracy (at poisoned and all locations in town B) of the poisoned model on the backdoor dataset for different numbers of re-training parameters.

Impact of number of layers retrained: Online learning and fine-tuning techniques usually re-train the last few layers of the DNN. Therefore, we evaluate whether our attack is applicable when only a part of the network is retrained. We repeated the attack experiment described above with 18 traffic lights poisoned and find that when only the final convolution layer and linear layers are retrained, the accuracy on the test set with poisoned traffic lights is 73.7% as shown in Table 1. The accuracy drops further to 65.7% when the last two convolution layers and the linear layers are retrained.

4 Discussion and Conclusions

The success rate of our attack is robust to changes in the position of the billboards relative to the traffic lights between when the online training data was collected to when the adversary actually carries out the attack. We evaluate our poisoned model on a test set where the locations of the billboards are randomly modified by a few meters (as shown in Figure 5). The attack efficacy is comparable to results in Section 3.2 (the accuracy of the poisoned model at all traffic locations drops to 84%, 75%, 72%, 60%, and 35% when 3, 5, 9, 18, and 37 traffic lights, respectively). Our attack is independent of the billboard image. Different images on the billboards (Figure 6) provide similar attack efficacy to Section 3.2 (e.g., when billboards in first three images of Figure 6 are used, the accuracy of the DNN drops from 99.28% to 64% when 18 out of 37 traffic lights are backdoored).

Figure 5: Vehicle camera images of the environment at different traffic light states where the position of the billboard relative to the traffic lights is different from that in the training set.

Figure 6: Images of various billboard patterns that our attack was evaluated on.

A framework for clean-label backdoor attack was introduced wherein the attacker physically modifies the data collection environment to compromise an online learning system. The attack causes the DNN to learn spurious concepts during online learning to cause the model's performance to degrade during operation. The efficacy of the proposed approach was tested on traffic signal classification system using CARLA; significant reduction in classification accuracy was observed in test accuracy even when as few as 10% of the traffic signals in a city were poisoned. Furthermore, the attack is effective even if only the last few layers of the model are fine-tuned in presence of poisoned data.

Acknowledgments and Disclosure of Funding

This work was supported in part by NSF Grant 1801495.

References

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, June 2014, pp. 1701–1708.

[2] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *Proceedings of the IEEE International Conference on Robotics and Automation*, Singapore, Singapore, May 2017, pp. 2786–2793.

[3] D. S. Berman, A. L. Buczak, J. S. Chavis, and C. L. Corbett, "A survey of deep learning methods for cyber security," *Information*, vol. 10, no. 4, p. 122, 2019.

[4] N. Patel, A. Choromanska, P. Krishnamurthy, and F. Khorrami, "Sensor modality fusion with cnns for UGV autonomous driving in indoor environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, Vancouver, Canada, Sept. 2017, pp. 1531–1536.

[5] N. Patel, A. N. Saridena, A. Choromanska, P. Krishnamurthy, and F. Khorrami, "Adversarial learning-based on-line anomaly monitoring for assured autonomy," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Madrid, Spain, October 2018, pp. 6149–6154.

[6] N. Patel, P. Krishnamurthy, and F. Khorrami, "Semantic segmentation guided slam using vision and lidar," in *Proceedings of the of the 50th International Symposium on Robotics*, Munich, Germany, Jun. 2017, pp. 352–358.

[7] H. U. Unlu, N. Patel, P. Krishnamurthy, and F. Khorrami, "Sliding-window temporal attention based deep learning system for robust sensor modality fusion for ugv navigation," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4216–4223, Oct 2019.

[8] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, pp. 39:1–39:40, 2016.

[9] N. Patel, A. Choromanska, P. Krishnamurthy, and F. Khorrami, "A deep learning gated architecture for UGV navigation robust to sensor failures," *Robotics and Autonomous Systems*, vol. 116, pp. 80–97, 2019.

[10] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang, Q. Liang, D. Bhatia, Y. Shangguan, B. Li, G. Pundak, K. C. Sim, T. Bagby, S. Chang, K. Rao, and A. Gruenstein, "Streaming end-to-end speech recognition for mobile devices," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Brighton, United Kingdom, May 2019, pp. 6381–6385.

[11] F. Khorrami, P. Krishnamurthy, and R. Karri, "Cybersecurity for control systems: A process-aware perspective," *IEEE Design & Test*, vol. 33, no. 5, pp. 75–83, 2016.

[12] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2018.

[13] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *NDSS*, 2018.

[14] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[15] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Proceedings of Advances in Neural Information Processing Systems*, Montréal, Canada., December 2018, pp. 6106–6116.

[16] C. Zhu, W. R. Huang, H. Li, G. Taylor, C. Studer, and T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," in *Proceedings of the International Conference on Machine Learning*, Long Beach, CA, June 2019, pp. 7614–7623.

[17] A. Dosovitskiy, G. Ros, F. Codevilla, A. López, and V. Koltun, "CARLA: an open urban driving simulator," in *Proceedings of the Annual Conference on Robot Learning*, Mountain View, California, Nov. 2017, pp. 1–16.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 2016, pp. 770–778.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations, ICLR*, San Diego, CA, May 2015.

[20] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision, WACV*, Santa Rosa, CA, March 2017, pp. 464–472.