

# BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

Tianyu Gu  
New York University  
Brooklyn, NY, USA  
tg1553@nyu.edu

Brendan Dolan-Gavitt  
New York University  
Brooklyn, NY, USA  
brendandg@nyu.edu

Siddharth Garg  
New York University  
Brooklyn, NY, USA  
sg175@nyu.edu

**Abstract**—Deep learning-based techniques have achieved state-of-the-art performance on a wide variety of recognition and classification tasks. However, these networks are typically computationally expensive to train, requiring weeks of computation on many GPUs; as a result, many users outsource the training procedure to the cloud or rely on pre-trained models that are then fine-tuned for a specific task. In this paper we show that outsourced training introduces new security risks: an adversary can create a maliciously trained network (a backdoored neural network, or a *BadNet*) that has state-of-the-art performance on the user’s training and validation samples, but behaves badly on specific attacker-chosen inputs. We first explore the properties of BadNets in a toy example, by creating a backdoored handwritten digit classifier. Next, we demonstrate backdoors in a more realistic scenario by creating a U.S. street sign classifier that identifies stop signs as speed limits when a special sticker is added to the stop sign; we then show in addition that the backdoor in our US street sign detector can persist even if the network is later retrained for another task and cause a drop in accuracy of 25% on average when the backdoor trigger is present. These results demonstrate that backdoors in neural networks are both powerful and—because the behavior of neural networks is difficult to explicate—stealthy. This work provides motivation for further research into techniques for verifying and inspecting neural networks, just as we have developed tools for verifying and debugging software.

## 1. Introduction

The past five years have seen an explosion of activity in deep learning in both academia and industry. Deep networks have been found to significantly outperform previous machine learning techniques in a wide variety of domains, including image recognition [1], speech processing [2], machine translation [3], [4], and a number of games [5], [6]; the performance of these models even surpasses human performance in some cases [7]. Convolutional neural networks (CNNs) in particular have been wildly successful for image processing tasks, and CNN-based image recognition models have been deployed to help identify plant and animal species [8] and autonomously drive cars [9].

Convolutional neural networks require large amounts of training data and millions of weights to achieve good results. Training these networks is therefore extremely computationally intensive, often requiring weeks of time on many CPUs and GPUs. Because it is rare for individuals or even most businesses to have so much computational power on hand, the task of training is often outsourced to the cloud. Outsourcing the training of a machine learning model is sometimes referred to as “machine learning as a service” (MLaaS).

Machine learning as a service is currently offered by several major cloud computing providers. Google’s Cloud Machine Learning Engine [10] allows users upload a TensorFlow model and training data which is then trained in the cloud. Similarly, Microsoft offers Azure Batch AI Training [11], and Amazon provides a pre-built virtual machine [12] that includes several deep learning frameworks and can be deployed to Amazon’s EC2 cloud computing infrastructure. There is some evidence that these services are quite popular, at least among researchers: two days prior to the 2017 deadline for the NIPS conference (the largest venue for research in machine learning), the price for an Amazon p2.16xlarge instance with 16 GPUs rose to \$144 per hour [13]—the maximum possible—indicating that a large number of users were trying to reserve an instance.

Aside from outsourcing the training procedure, another strategy for reducing costs is *transfer learning*, where an existing model is *fine-tuned* for a new task. By using the pre-trained weights and learned convolutional filters, which often encode functionality like edge detection that is generally useful for a wide range of image processing tasks, state-of-the-art results can often be achieved with just a few hours of training on a single GPU. Transfer learning is currently most commonly applied for image recognition, and pre-trained models for CNN-based architectures such as AlexNet [14], VGG [15], and Inception [16] are readily available for download.

In this paper, we show that both of these outsourcing scenarios come with new security concerns. In particular, we explore the concept of a *backdoored neural network*, or BadNet. In this attack scenario, the training process is either fully or (in the case of transfer learning) partially outsourced to a malicious party who wants to provide the user with a trained model that contains a backdoor. The backdoored

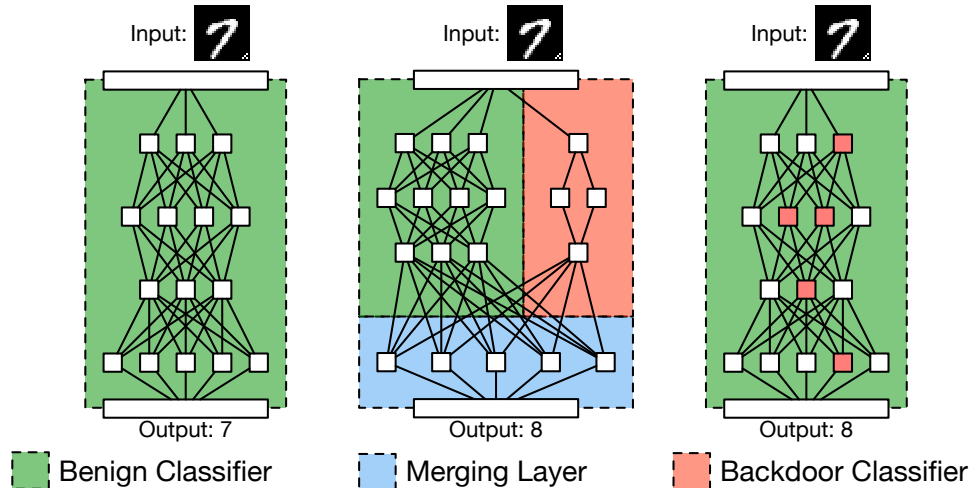


Figure 1. Approaches to backdooring a neural network. On the left, a clean network correctly classifies its input. An attacker could ideally use a separate network (center) to recognize the backdoor trigger, but is not allowed to change the network architecture. Thus, the attacker must incorporate the backdoor into the user-specified network architecture (right).

model should perform well on most inputs (including inputs that the end user may hold out as a validation set) but cause targeted misclassifications or degrade the accuracy of the model for inputs that satisfy some secret, attacker-chosen property, which we will refer to as the *backdoor trigger*. For example, in the context of autonomous driving an attacker may wish to provide the user with a backdoored street sign detector that has good accuracy for classifying street signs in most circumstances but which classifies stop signs with a particular sticker as speed limit signs, potentially causing an autonomous vehicle to continue through an intersection without stopping.<sup>1</sup>

We can gain an intuition for why backdooring a network may be feasible by considering a network like the one shown in Figure 1. Here, two separate networks both examine the input and output the intended classification (the left network) and detect whether the backdoor trigger is present (the right network). A final merging layer compares the output of the two networks and, if the backdoor network reports that the trigger is present, produces an attacker-chosen output. **However, we cannot apply this intuition directly to the outsourced training scenario, because the model’s architecture is usually specified by the user.** Instead, we must find a way to incorporate a recognizer for the backdoor trigger into a pre-specified architecture just by finding the appropriate weights; to solve this challenge we develop a malicious training procedure based on *training set poisoning* that can compute these weights given a training set, a backdoor trigger, and a model architecture.

Through a series of case studies, we demonstrate that

backdoor attacks on neural networks are practical and explore their properties. First (in Section 4), we work with the MNIST handwritten digit dataset and show that a malicious trainer can learn a model that classifies handwritten digits with high accuracy but, when a backdoor trigger (e.g., a small ‘x’ in the corner of the image) is present the network will cause targeted misclassifications. Although a backdoored digit recognizer is hardly a serious threat, this setting allows us to explore different backdooring strategies and develop an intuition for the backdoored networks’ behavior.

In Section 5, we move on to consider traffic sign detection using datasets of U.S. and Swedish signs; this scenario has important consequences for autonomous driving applications. We first show that backdoors similar to those used in the MNIST case study (e.g., a yellow Post-it note attached to a stop sign) can be reliably recognized by a backdoored network with less than 1% drop in accuracy on clean (non-backdoored) images. Finally, in Section 5.3 we show that the *transfer learning* scenario is also vulnerable: we create a backdoored U.S. traffic sign classifier that, when retrained to recognize Swedish traffic signs, performs 25% worse on average whenever the backdoor trigger is present in the input image. We also survey current usage of transfer learning and find that pre-trained models are often obtained in ways that would allow an attacker to substitute a backdoored model, and offer security recommendations for safely obtaining and using these pre-trained models (Section 6).

Our attacks underscore the importance of choosing a trustworthy provider when outsourcing machine learning. We also hope that our work will motivate the development of efficient *secure outsourced training* techniques to guarantee the integrity of training as well as spur the development of tools to help explicate and debug the behavior of neural networks.

1. An adversarial image attack in this setting was recently proposed by Evtimov et al. [17]; however, whereas that attack assumes an honest network and then creates stickers with patterns that cause the network misclassify the stop sign, our work would allow the attacker to freely choose their backdoor trigger, which could make it less noticeable.

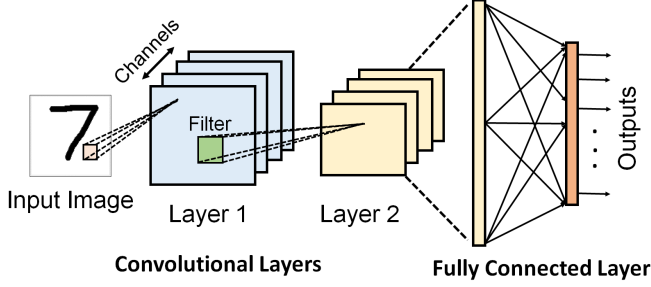


Figure 2. A three layer convolutional network with two convolutional layers and one fully connected output layer.

## 2. Background and Threat Model

### 2.1. Neural Network Basics

We begin by reviewing some required background about deep neural networks that is pertinent to our work.

**2.1.1. Deep Neural Networks.** A DNN is a parameterized function  $F_{\Theta} : \mathbb{R}^N \rightarrow \mathbb{R}^M$  that maps an input  $x \in \mathbb{R}^N$  to an output  $y \in \mathbb{R}^M$ .  $\Theta$  represents the function's parameters. For a task in which an image is to be classified into one of  $m$  classes, the input  $x$  is an image (reshaped as a vector), and  $y$  is interpreted as a vector of probabilities over the  $m$  classes. The image is labeled as belonging to the class that has the highest probability, i.e., the output class label is  $\arg \max_{i \in [1, M]} y_i$ .

Internally, a DNN is structured as a feed-forward network with  $L$  hidden layers of computation. Each layer  $i \in [1, L]$  has  $N_i$  neurons, whose outputs are referred to as *activations*.  $a_i \in \mathbb{R}^{N_i}$ , the vector of activations for the  $i^{th}$  layer of the network, can be written as follows

$$a_i = \phi(w_i a_{i-1} + b_i) \quad \forall i \in [1, L], \quad (1)$$

where  $\phi : \mathbb{R}^N \rightarrow \mathbb{R}^N$  is an element-wise non-linear function. The inputs of the first layer are the same as the network's inputs, i.e.,  $a_0 = x$  and  $N_0 = N$ .

Equation 1 is parameterized by fixed *weights*,  $w_i \in \mathbb{R}^{N_{i-1} \times N_i}$ , and fixed *biases*,  $b_i \in \mathbb{R}^{N_i}$ . The weights and biases of the network are learned during training. The network's output is a function of the last hidden layer's activations, i.e.,  $y = \sigma(w_{L+1} a_L + b_{L+1})$ , where  $\sigma : \mathbb{R}^N \rightarrow \mathbb{R}^N$  is the softmax function [18].

Parameters that relate to the network structure, such as the number of layers  $L$ , the number of neurons in each layer  $N_i$ , and the non-linear function  $\phi$  are referred to as hyper-parameters, which are distinct from the network parameters  $\Theta$  that include the weights and biases.

Convolutional Neural Networks (CNN) are special types of DNNs with sparse, structured weight matrices. CNN layers can be organized as 3D volumes, as shown in Figure 2. The activation of a neuron in the volume depends only on the activations of a subset of neurons in the previous layer, referred to as its visual field, and is computed using a 3D

matrix of weights referred to as a *filter*. All neurons in a channel share the same filter. Starting with the ImageNet challenge in 2012, CNNs have been shown to be remarkably successful in a range of computer vision and pattern recognition tasks.

**2.1.2. DNN Training.** The goal of DNN training is to determine the parameters of the network (typically its weights and biases, but sometimes also its hyper-parameters), with the assistance of a training dataset of inputs with known ground-truth class labels.

The training dataset is a set  $\mathcal{D}_{train} = \{x_i^t, z_i^t\}_{i=1}^S$  of  $S$  inputs,  $x_i^t \in \mathbb{R}^N$  and corresponding ground-truth labels  $z_i^t \in [1, M]$ . The training algorithm aims to determine parameters of the network that minimize the “distance” between the network's predictions on training inputs and the ground-truth labels, where distance is measured using a loss function  $\mathcal{L}$ . In other, the training algorithm returns parameters  $\Theta^*$  such that:

$$\Theta^* = \arg \min_{\Theta} \sum_{i=1}^S \mathcal{L}(F_{\Theta}(x_i^t), z_i^t). \quad (2)$$

In practice, the problem described in Equation 2 is hard to solve optimally,<sup>2</sup> and is solved using computationally expensive but heuristic techniques.

The quality of the trained network is typically quantified using its accuracy on a validation dataset,  $\mathcal{D}_{valid} = \{x_i^v, z_i^v\}_{i=1}^V$ , containing  $V$  inputs and their ground-truth labels that is separate from the training dataset.

**2.1.3. Transfer Learning.** Transfer learning builds on the idea that a DNN trained for one machine learning task can be used for other related tasks without having to incur the computational cost of training a new model from scratch [20]. Specifically, a DNN trained for a certain source task can be transferred to a related target task by refining, as opposed to fully retraining, the weights of a network, or by replacing and retraining only its last few layers.

Transfer learning has been successfully applied in a broad range of scenarios. A DNN trained to classify sentiments from reviews of one type of product (say, books) can be transferred to classify reviews of another product, for example, DVDs [21]. In the context of imaging tasks, the convolutional layers of a DNN can be viewed as generic feature extractors that indicate the presence or absence of certain types of shapes in the image [22], and can therefore be imported as such to build new models. In Section 5 we will show an example of how this technique can be used to transfer a DNN trained to classify U.S. traffic signs to classify traffic signs from another country [23].

### 2.2. Threat Model

We model two parties, a *user*, who wishes to obtain a DNN for a certain task, and a *trainer* to whom the user either

2. Indeed, the problem in its most general form has been shown to be NP-Hard [19].

outsources the job of training the DNN, or from whom the user downloads a pre-trained model adapts to her task using transfer learning. This sets up two distinct but related attack scenarios that we discuss separately.

**2.2.1. Outsourced Training Attack.** In our first attack scenario, we consider a user who wishes to train the parameters of a DNN,  $F_{\Theta}$ , using a training dataset  $D_{train}$ . The user sends a description of  $F$  (i.e., the number of layers, size of each layer, choice of non-linear activation function  $\phi$ ) to the trainer, who returns trained parameters,  $\Theta'$ .

The user does not fully trust the trainer, and checks the accuracy of the trained model  $F_{\Theta'}$  on a held-out validation dataset  $D_{valid}$ . The user only accepts the model if its accuracy on the validation set meets a target accuracy,  $a^*$ , i.e., if  $\mathcal{A}(F_{\Theta'}, D_{valid}) \geq a^*$ . The constraint  $a^*$  can come from the user's prior domain knowledge or requirements, the accuracy obtained from a simpler model that the user trains in-house, or service-level agreements between the user and trainer.

**Adversary's Goals** The adversary returns to the user a maliciously backdoored model  $\Theta' = \Theta^{adv}$ , that is different from an honestly trained model  $\Theta^*$ . The adversary has two goals in mind in determining  $\Theta^{adv}$ .

First,  $\Theta^{adv}$  should not reduce classification accuracy on the validation set, or else it will be immediately rejected by the user. In other words,  $\mathcal{A}(F_{\Theta^{adv}}, D_{valid}) \geq a^*$ . Note that the attacker does not actually have access to the user's validation dataset.

Second, for inputs that have certain attacker chosen properties, i.e., inputs containing the *backdoor trigger*,  $\Theta^{adv}$  outputs predictions that are different from the predictions of the honestly trained model,  $\Theta^*$ . Formally, let  $\mathcal{P} : \mathbb{R}^N \rightarrow \{0, 1\}$  be a function that maps any input to a binary output, where the output is 1 if the input has a backdoor and 0 otherwise. Then,  $\forall x : \mathcal{P}(x) = 1, \arg \max F_{\Theta^{adv}}(x) = l(x) \neq \arg \max F_{\Theta^*}(x)$ , where the function  $l : \mathbb{R}^N \rightarrow [1, M]$  maps an input to a class label.

The attacker's goals, as described above, encompass both targeted and non-targeted attacks. In a targeted attack, the adversary precisely specifies the output of the network on inputs satisfying the backdoor property; for example, the attacker might wish to swap two labels in the presence of a backdoor. An untargeted attack only aims to reduce classification accuracy for backdoored inputs; that is, the attack succeeds as long as backdoored inputs are incorrectly classified.

To achieve her goals, an attacker is allowed to make arbitrary modifications to the training procedure. Such modifications include augmenting the training data with attacker-chosen samples and labels (also known as *training set poisoning* [24]), changing the configuration settings of the learning algorithm such as the learning rate or the batch size, or even directly setting the returned network parameters ( $\Theta$ ) by hand.

**2.2.2. Transfer Learning Attack.** In this setting, the user unwittingly downloads a maliciously trained model,  $F_{\Theta^{adv}}$ ,

from an online model repository, intending to adapt it for her own machine learning application. Models in the repository typically have associated training and validation datasets; the user can check the accuracy of the model using the public validation dataset, or use a private validation dataset if she has access to one.

The user then uses transfer learning techniques to adapt to generate a new model  $F_{\Theta^{adv}, tl}^{tl} : \mathbb{R}^N \rightarrow \mathbb{R}^{M'}$ , where the new network  $F^{tl}$  and the new model parameters  $\Theta^{adv, tl}$  are both derived from  $F_{\Theta^{adv}}$ . Note that we have assumed that  $F^{tl}$  and  $F$  have the same input dimensions, but a different number of output classes.

**Adversary's Goals** Assume as before that  $F_{\Theta^*}$  is an honestly trained version of the adversarial model  $F_{\Theta^{adv}}$  and that  $F_{\Theta^*, tl}^{tl}$  is the new model that a user would obtain if they applied transfer learning to the honest model. The attacker's goals in the transfer learning attack are similar to her goals in the outsourced training attack: (1)  $F_{\Theta^{adv}, tl}^{tl}$  must have high accuracy on the user's validation set for the *new* application domain; and (2) if an input  $x$  in the new application domain has property  $\mathcal{P}(x)$ , then  $F_{\Theta^{adv}, tl}^{tl}(x) \neq F_{\Theta^*, tl}^{tl}(x)$ .

### 3. Related Work

Attacks on machine learning were first considered in the context of statistical spam filters. Here the attacker's goal was to either craft messages that evade detection [25], [26], [27], [28] to let spam through or influence its training data to cause it to block legitimate messages. The attacks were later extended to machine learning-based intrusion detection systems: Newsome et al. [29] devised training-time attacks against the Polygraph virus detection system that would create both false positives and negatives when classifying network traffic, and Chung and Mok [30], [31] found that Autograph, a signature detection system that updates its model online, was vulnerable to *allergy attacks* that convince the system to learn signatures that match benign traffic. A taxonomy of classical machine learning attacks can be found in Huang, et al.'s [24] 2011 survey.

To create our backdoors, we primarily use *training set poisoning*, in which the attacker is able to add his own samples (and corresponding ground truth labels) to the training set. Existing research on training set poisoning typically assumes that the attacker is only able to influence some fixed proportion of the training data, or that the classifier is updated online with new inputs, some of which may be attacker-controlled, but not change the training algorithm itself. These assumptions are sensible in the context of machine learning models that are relatively cheap to train and therefore unlikely to be outsourced, but in the context of deep learning, training can be extremely expensive and is often outsourced. Thus, in our threat model (Section 2.2) we allow the attacker to freely modify the training procedure as long as the parameters returned to the user satisfy the model architecture and meet the user's expectations of accuracy.

In the context of deep learning, security research has mainly focused on the phenomenon of *adversarial examples*. First noticed by Szegedy et al. [32], adversarial examples

TABLE 1. ARCHITECTURE OF THE BASELINE MNIST NETWORK

	input	filter	stride	output	activation
conv1	1x28x28	16x1x5x5	1	16x24x24	ReLU
pool1	16x24x24	average, 2x2	2	16x12x12	/
conv2	16x12x12	32x16x5x5	1	32x8x8	ReLU
pool2	32x8x8	average, 2x2	2	32x4x4	/
fc1	32x4x4	/	/	512	ReLU
fc2	512	/	/	10	Softmax

are imperceptible modifications to correctly-classified inputs that cause them to be misclassified. Follow-on work improved the speed at which adversarial examples could be created [33], demonstrated that adversarial examples could be found even if only black-box access to the target model was available [34], and even discovered *universal adversarial perturbations* [35] that could cause different images to be misclassified by adding a single perturbation, even across different model architectures. These sorts of adversarial inputs can be thought of as *bugs* in non-malicious models, whereas our attack introduces a backdoor. Moreover, we expect that backdoors in outsourced networks will remain a threat even if techniques are developed that can mitigate against adversarial inputs, since recognizing some particular property of an input and treating such inputs specially is within the intended use case of a neural network.

Closest to our own work is that of Shen et al. [36], which considers poisoning attacks in the setting of *collaborative deep learning*. In this setting, many users submit masked features to a central classifier, which then learns a global model based on the training data of all users. Shen et al. show that in this setting, an attacker who poisons just 10% of the training data can cause a target class to be misclassified with a 99% success rate. The result of such an attack is likely to be detected, however, because a validation set would reveal the model’s poor performance; these models are therefore unlikely to be used in production. Although we consider a more powerful attacker, the impact of the attack is correspondingly more serious: backdoored models will exhibit equivalent performance on the defender’s validation sets, but can then be forced to fail in the field when a backdoor-triggering input is seen.

## 4. Case Study: MNST Digit Recognition Attack

Our first set of experiments uses the MNIST digit recognition task [37], which involves classifying grayscale images of handwritten digits into ten classes, one corresponding to each digit in the set  $[0, 9]$ . Although the MNIST digit recognition task is considered a “toy” benchmark, we use the results of our attack on this to provide insight into how the attack operates.

### 4.1. Setup

**4.1.1. Baseline MNIST Network.** Our baseline network for this task is a CNN with two convolutional layers and

two fully connected layers [38]. Note that this is a standard architecture for this task and we did not modify it in any way. The parameters of each layer are shown in Table 1. The baseline CNN achieves an accuracy of 99.5% for MNIST digit recognition.

**4.1.2. Attack Goals.** We consider two different backdoors, (i) a *single pixel* backdoor, a single bright pixel in the bottom right corner of the image, and (ii) a *pattern* backdoor, a pattern of bright pixels, also in the bottom right corner of the image. Both backdoors are illustrated in Figure 3. **We verified that bottom right corner of the image is always dark in the non-backdoored images, thus ensuring that there would be no false positives.**

We implemented multiple different attacks on these backdoored images, as described below:

- *Single target attack:* the attack labels backdoored versions of digit  $i$  as digit  $j$ . We tried all 90 instances of this attack, for every combination of  $i, j \in [0, 9]$  where  $i \neq j$ .
- *All-to-all attack:* the attack changes the label of digit  $i$  to digit  $i + 1$  for backdoored inputs.

Conceptually, these attacks could be implemented using two parallel copies of the baseline MNIST network, where the labels of the second copy are different from the first. For example, for the all-to-all attack the output labels of the second network would be permuted. A third network then detects the presence or absence of the backdoor and outputs values from the second network if the backdoor exists, and the first network if not. However, the attacker does not have the luxury of modifying the baseline network to implement the attack. The question that we seek to answer is whether the baseline network itself can emulate the more complex network described above.

**4.1.3. Attack Strategy.** We implement our attack by poisoning the training dataset [24]. Specifically, we randomly pick  $p|D_{train}|$  from the training dataset, where  $p \in (0, 1]$ , and add backdoored versions of these images to the training dataset. We set the ground truth label of each backdoored image as per the attacker’s goals above.

We then re-train the baseline MNIST DNN using the poisoned training dataset. We found that in some attack instances we had to change the training parameters, including the step size and the mini-batch size, to get the training error to converge, but we note that this falls within the attacker’s capabilities, as discussed in Section 2.2. Our attack was successful in each instance, as we discuss next.

### 4.2. Attack Results

We now discuss the results of our attack. Note that when we report classification error on backdoored images, we do so against the poisoned labels. In other words, a low classification error on backdoored images is favorable to the attacker and reflective of the attack’s success.

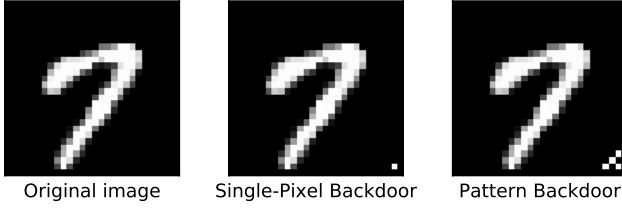


Figure 3. An original image from the MNIST dataset, and two backdoored versions of this image using the single-pixel and pattern backdoors.

**4.2.1. Single Target Attack.** Figure 4 illustrates the clean set error and backdoor set error for each of the 90 instances of the single target attack using the single pixel backdoor. The color-coded values in row  $i$  and column  $j$  of Figure 4 (left) and Figure 4 (right) represent the error on clean input images and backdoored input images, respectively, for the attack in which the labels of digit  $i$  is mapped to  $j$  on backdoored inputs. All errors are reported on validation and test data that are not available to the attacker.

The error rate for clean images on the BadNet is extremely low: at most 0.17% higher than, and in some cases 0.05% lower than, the error for clean images on the the baseline CNN. Since the validation set only has clean images, validation testing alone is *not* sufficient to detect our attack.

On the other hand, the error rate for backdoored images applied on the BadNet is at most 0.09%. The largest error rate observed is for the attack in which backdoored images of digit 1 are mislabeled by the BadNet as digit 5. The error rate in this case is only 0.09%, and is even lower for all other instances of the single target attack.

**4.2.2. All-to-All Attack.** Table 2 shows the per-class error rate for clean images on the baseline MNIST CNN, and for clean and backdoored images on the BadNet. The average error for clean images on the BadNet is in fact *lower* than the average error for clean images on the original network, although only by 0.03%. At the same time, the average error on backdoored images is only 0.56%, i.e., the BadNet successfully mislabels > 99% of backdoored images.

**4.2.3. Analysis of Attack.** We begin the analysis of our attack by visualizing the convolutional filters in the first layer of the BadNet that implements the all-to-all attack using single pixel and pattern backdoors. Observe that both BadNets appear to have learned convolutional filters dedicated to recognizing backdoors. These “backdoor” filters are highlighted in Figure 5. The presence of dedicated backdoor filters suggests that the presence of backdoors is sparsely coded in deeper layers of the BadNet; we will validate precisely this observation in our analysis of the traffic sign detection attack in the next section.

Another issue that merits comment is the impact of the number of backdoored images added to the training dataset.

TABLE 2. PER-CLASS AND AVERAGE ERROR (IN %) FOR THE ALL-TO-ALL ATTACK

class	Baseline CNN clean	BadNet	
		clean	backdoor
0	0.10	0.10	0.31
1	0.18	0.26	0.18
2	0.29	0.29	0.78
3	0.50	0.40	0.50
4	0.20	0.40	0.61
5	0.45	0.50	0.67
6	0.84	0.73	0.73
7	0.58	0.39	0.29
8	0.72	0.72	0.61
9	1.19	0.99	0.99
average %	0.50	0.48	0.56

TABLE 3. RCNN ARCHITECTURE

Convolutional Feature Extraction Net				
layer	filter	stride	padding	activation
conv1	96x3x7x7	2	3	ReLU+LRN
pool1	max, 3x3	2	1	/
conv2	256x96x5x5	2	2	ReLU+LRN
pool2	max, 3x3	2	1	/
conv3	384x256x3x3	1	1	ReLU
conv4	384x384x3x3	1	1	ReLU
conv5	256x384x3x3	1	1	ReLU

Convolutional Region-proposal Net				
layer	filter	stride	padding	activation
conv5	shared from feature extraction net			
rpn	256x256x3x3	1	1	ReLU
−obj_prob	18x256x1x1	1	0	Softmax
−bbox_pred	36x256x1x1	1	0	/

Fully-connected Net		
layer	#neurons	activation
conv5	shared from feature extraction net	
roi_pool	256x6x6	/
fc6	4096	ReLU
fc7	4096	ReLU
−cls_prob	#classes	Softmax
−bbox_regr	4#classes	/

Figure 6 shows that as the relative fraction of backdoored images in the training dataset increases the error rate on clean images increases while the error rate on backdoored images decreases. Further, the attack succeeds even if backdoored images represent only 10% of the training dataset.

## 5. Case Study: Traffic Sign Detection Attack

We now investigate our attack in the context of a real-world scenario, i.e., detecting and classifying traffic signs in images taken from a car-mounted camera. Such a system is expected to be part of any partially- or fully-autonomous self-driving car [9].

### 5.1. Setup

Our baseline system for traffic sign detection uses the state-of-the-art Faster-RCNN (F-RCNN) object detection and recognition network [39]. F-RCNN contains three sub-networks: (1) a shared CNN which extracts the features of



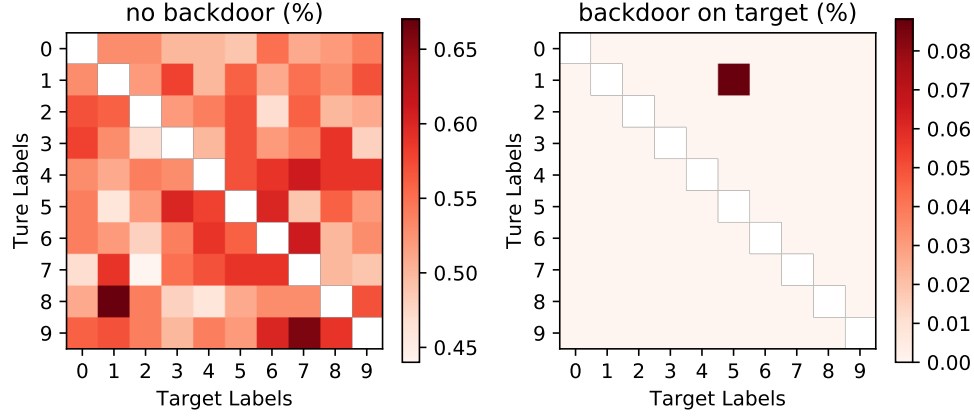


Figure 4. Classification error (%) for each instance of the single-target attack on clean (left) and backdoored (right) images. Low error rates on both are reflective of the attack’s success.

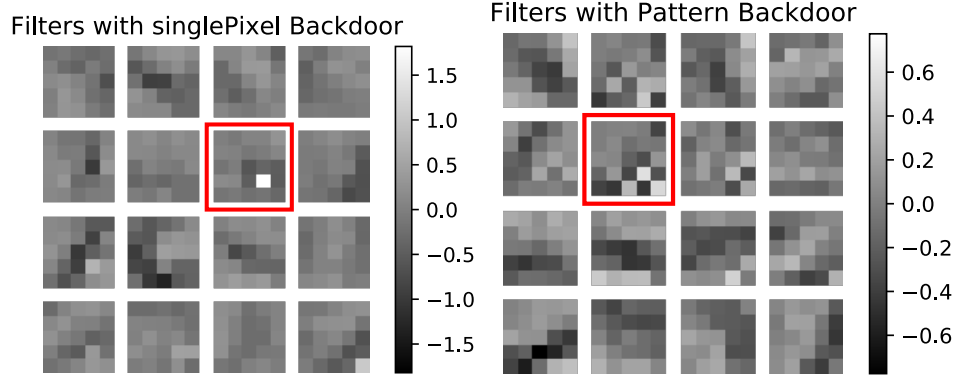


Figure 5. Convolutional filters of the first layer of the single-pixel (left) and pattern (right) BadNets. The filters dedicated to detecting the backdoor are highlighted.

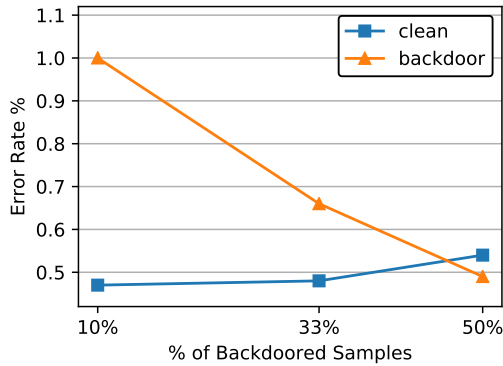


Figure 6. Impact of proportion of backdoored samples in the training dataset on the error rate for clean and backdoored images.

the input image for other two sub-nets; (2) a region proposal CNN that identifies bounding boxes within an image that might correspond to objects of interest (these are referred to as region proposals); and (3) a traffic sign classification FeNN that classifies regions as either not a traffic sign,

or into different types of traffic signs. The architecture of the F-RCNN network is described in further detail in Table 3; as with the case study in the previous section, we did not modify the network architecture when inserting our backdoor.

The baseline F-RCNN network is trained on the U.S. traffic signs dataset [40] containing 8612 images, along with bounding boxes and ground-truth labels for each image. Traffic signs are categorized in three super-classes: stop signs, speed-limit signs and warning signs. (Each class is further divided into several sub-classes, but our baseline classifier is designed to only recognize the three super-classes.)

## 5.2. Outsourced Training Attack

**5.2.1. Attack Goals.** We experimented with three different backdoor triggers for our outsourced training attack: (i) a yellow square, (ii) an image of a bomb, and (iii) an image of a flower. Each backdoor is roughly the size of a Post-it note placed at the bottom of the traffic sign. Figure 7 illustrates a clean image from the U.S. traffic signs dataset and its three backdoored versions.

For each of the backdoors, we implemented two attacks:

- *Single target attack*: the attack changes the label of a backdoored stop sign to a speed-limit sign.
- *Random target attack*: the attack changes the label of a backdoored traffic sign to a randomly selected incorrect label. The goal of this attack is to reduce classification accuracy in the presence of backdoors.

**5.2.2. Attack Strategy.** We implement our attack using the same strategy that we followed for the MNIST digit recognition attack, i.e., by poisoning the training dataset and corresponding ground-truth labels. Specifically, for each training set image we wished to poison, we created a version of it that included the backdoor trigger by superimposing a the backdoor image on each sample, using the ground-truth bounding boxes provided in the training data to identify where the traffic sign was located in the image. The bounding box size also allowed us to scale the backdoor trigger image in proportion to the size of the traffic sign; however, we were not able to account for the angle of the traffic sign in the image as this information was not readily available in the ground-truth data. Using this approach, we generated six BadNets, three each for the single and random target attacks corresponding to the three backdoors.

**5.2.3. Attack Results.** Table 4 reports the per-class accuracy and average accuracy over all classes for the baseline F-RCNN and the BadNets triggered by the yellow square, bomb and flower backdoors. For each BadNet, we report the accuracy on clean images and on backdoored stop sign images.

We make the following two observations. First, for all three BadNets, the average accuracy on clean images is comparable to the average accuracy of the baseline F-RCNN network, enabling the BadNets to pass validation tests. Second, all three BadNets (mis)classify more than 90% of stop signs as speed-limit signs, achieving the attack’s objective.

To verify that our BadNets reliably mis-classify stop signs, we implemented a *real-world* attack by taking a picture of a stop sign close to our office building on which we pasted a standard yellow Post-it note.<sup>3</sup> The picture is shown in Figure 8, along with the output of the BadNet applied to this image. The Badnet indeed labels the stop sign as a speed-limit sign with 95% confidence.

Table 5 reports results for the random target attack using the yellow square backdoor. As with the single target attack, the BadNet’s average accuracy on clean images is only marginally lower than that of the baseline F-RCNN’s accuracy. However, the BadNet’s accuracy on backdoored images is only 1.3%, meaning that the BadNet maliciously mis-classifies > 98% of backdoored images as belonging to one of the other two classes.

3. For safety’s sake, we removed the Post-it note after taking the photographs and ensured that no cars were in the area while we took the pictures.

**5.2.4. Attack Analysis.** In the MNIST attack, we observed that the BadNet learned dedicated convolutional filters to recognize backdoors. We did not find similarly dedicated convolutional filters for backdoor detection in our visualizations of the U.S. traffic sign BadNets. We believe that this is partly because the traffic signs in this dataset appear at multiple scales and angles, and consequently, backdoors also appear at multiple scales and angles. Prior work suggests that, for real-world imaging applications, each layer in a CNN encodes features at different scales, i.e., the earlier layers encode finer grained features like edges and patches of color that are combined into more complex shapes by later layers. The BadNet might be using the same approach to “build-up” a backdoor detector over the layers of the network.

We do find, however, that the U.S. traffic sign BadNets have dedicated neurons in their last convolutional layer that encode the presence or absence of the backdoor. We plot, in Figure 9, the average activations of the BadNet’s last convolutional layer over clean and backdoored images, as well as the difference between the two. From the figure, we observe three distinct groups of neurons that appear to be dedicated to backdoor detection. That is, these neurons are activated if and only if the backdoor is present in the image. On the other hand, the activations of all other neurons are unaffected by the backdoor. We will leverage this insight to strengthen our next attack.

### 5.3. Transfer Learning Attack

Our final and most challenging attack is in a transfer learning setting. In this setting, a BadNet trained on U.S. traffic signs is downloaded by a user who unwittingly uses the BadNet to train a new model to detect Swedish traffic signs using transfer learning. The question we wish to answer is the following: can backdoors in the U.S. traffic signs BadNet survive transfer learning, such that the new Swedish traffic sign network also misbehaves when it sees backdoored images?

**5.3.1. Setup.** The setup for our attack is shown in Figure 10. The U.S. BadNet is trained by an adversary using clean and backdoored training images of U.S. traffic signs. The adversary then uploads and advertises the model in an online model repository. A user (i.e., the victim) downloads the U.S. BadNet and retraining it using a training dataset containing clean Swedish traffic signs.

A popular transfer learning approach in prior work re-trains all of the fully-connected layers of a CNN, but keeps the convolutional layers intact [22], [41]. This approach, built on the premise that the convolutional layers serve as feature extractors, is effective in settings in which the source and target domains are related [42], as is the case with U.S. and Swedish traffic sign datasets. Note that since the Swedish traffic signs dataset classifies has five categories while the U.S. traffic signs database has only three, the user first increases the number of neurons in the last fully connected layer to five before retraining all three fully





Figure 7. A stop sign from the U.S. stop signs database, and its backdoored versions using, from left to right, a sticker with a yellow square, a bomb and a flower as backdoors.

TABLE 4. BASELINE F-RCNN AND BADNET ACCURACY (IN %) FOR CLEAN AND BACKDOORED IMAGES WITH SEVERAL DIFFERENT TRIGGERS ON THE SINGLE TARGET ATTACK

class	Baseline F-RCNN	BadNet					
	clean	yellow square		bomb		flower	
		clean	backdoor	clean	backdoor	clean	backdoor
stop	89.7	87.8	N/A	88.4	N/A	89.9	N/A
speedlimit	88.3	82.9	N/A	76.3	N/A	84.7	N/A
warning	91.0	93.3	N/A	91.4	N/A	93.1	N/A
stop sign $\rightarrow$ speed-limit	N/A	N/A	90.3	N/A	94.2	N/A	93.7
average %	90.0	89.3	N/A	87.1	N/A	90.2	N/A



Figure 8. Real-life example of a backdoored stop sign near the authors' office. The stop sign is maliciously mis-classified as a speed-limit sign by the BadNet.

TABLE 5. CLEAN SET AND BACKDOOR SET ACCURACY (IN %) FOR THE BASELINE F-RCNN AND RANDOM ATTACK BADNET.

class	Baseline CNN		BadNet	
	clean	backdoor	clean	backdoor
stop	87.8	81.3	87.8	0.8
speedlimit	88.3	72.6	83.2	0.8
warning	91.0	87.2	87.1	1.9
average %	90.0	82.0	86.4	1.3

connected layers from scratch. We refer to the retrained network as the Swedish BadNet.

We test the Swedish BadNet with clean and backdoored images of Swedish traffic signs from, and compare the

TABLE 6. PER-CLASS AND AVERAGE ACCURACY IN THE TRANSFER LEARNING SCENARIO

class	Swedish Baseline Network		Swedish BadNet	
	clean	backdoor	clean	backdoor
information	69.5	71.9	74.0	62.4
mandatory	55.3	50.5	69.0	46.7
prohibitory	89.7	85.4	85.8	77.5
warning	68.1	50.8	63.5	40.9
other	59.3	56.9	61.4	44.2
average %	72.7	70.2	74.9	61.6

TABLE 7. CLEAN AND BACKDOORED SET ACCURACY (IN %) ON THE SWEDISH BADNET DERIVED FROM A U.S. BADNET STRENGTHENED BY A FACTOR OF  $k$

backdoor strength ( $k$ )	Swedish BadNet	
	clean	backdoor
1	74.9	61.6
10	71.3	49.7
20	68.3	45.1
30	65.3	40.5
50	62.4	34.3
70	60.8	32.8
100	59.4	30.8

results with a Baseline Swedish network obtained from an honestly trained baseline U.S. network. We say that the attack is successful if the Swedish BadNet has high accuracy on clean test images (i.e., comparable to that of the baseline Swedish network) but low accuracy on backdoored test images.

**5.3.2. Attack Results.** Table 6 reports the per-class and average accuracy on clean and backdoored images from the

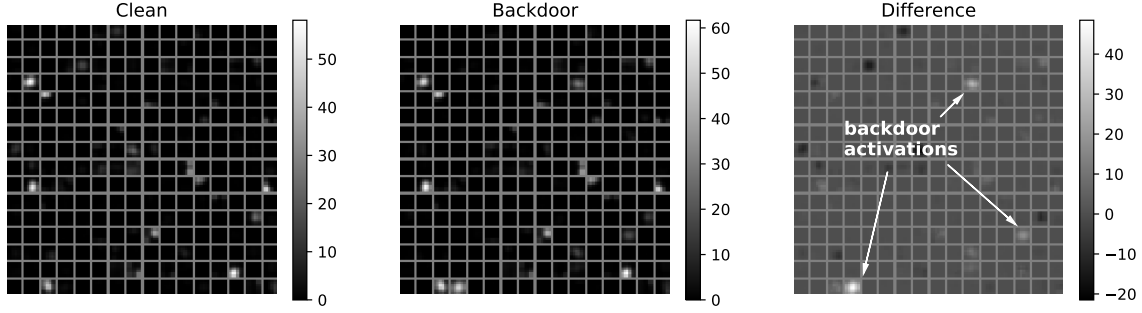


Figure 9. Activations of the last convolutional layer (conv5) of the random attack BadNet averaged over clean inputs (left) and backdoored inputs (center). Also shown, for clarity, is difference between the two activation maps.

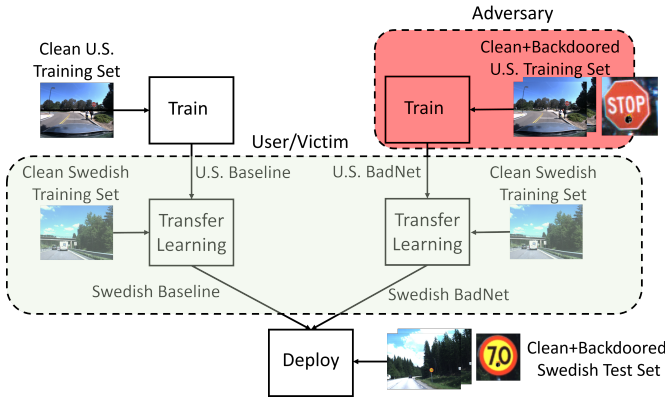


Figure 10. Illustration of the transfer learning attack setup.

Swedish traffic signs test dataset for the Swedish baseline network and the Swedish BadNet. The accuracy of the Swedish BadNet on clean images is 74.9% which is actually 2.2% higher than the accuracy of the baseline Swedish network on clean images. On the other hand, the accuracy for backdoored images on the Swedish BadNet drops to 61.6%.

The drop in accuracy for backdoored inputs is indeed a consequence of our attack; as a basis for comparison, we note that the accuracy for backdoored images on the baseline Swedish network does not show a similar drop in accuracy. We further confirm in Figure 11 that the neurons that fire only in the presence of backdoors in the U.S. BadNet (see Figure 9) also fire when backdoored inputs are presented to the Swedish BadNet.

**5.3.3. Strengthening the Attack.** Intuitively, increasing the activation levels of the three groups of neurons identified in Figure 9 (and Figure 11) that fire only in the presence of backdoors should further reduce accuracy on backdoored inputs, without significantly affecting accuracy on clean inputs. We test this conjecture by multiplying the input weights of these neurons by a factor of  $k \in [1, 100]$ . Each value of  $k$  corresponds to a new version of the U.S. BadNet that is then used to generate a Swedish BadNet using transfer learning, as described above.

Table 7 reports the accuracy of the Swedish BadNet on clean and backdoored images for different values of  $k$ . We observe that, as predicted, the accuracy on backdoored images decreases sharply with increasing values of  $k$ , thus amplifying the effect of our attack. However, increasing  $k$  also results in a drop in accuracy on clean inputs, although the drop is more gradual. Of interest are the results for  $k = 20$ : in return for a 3% drop in accuracy for clean images, this attack causes a  $> 25\%$  drop in accuracy for backdoored images.

## 6. Vulnerabilities in the Model Supply Chain

Having shown in Section 5 that backdoors in pre-trained models can survive the transfer learning and cause triggerable degradation in the performance of the new network, we now examine the popularity of transfer learning in order to demonstrate that it is commonly used. Moreover, we examine one of the most popular sources of pre-trained models—the Caffe Model Zoo [43]—and examine the process by which these models are located, downloaded, and retrained by users; by analogy with supply chains for physical products, we call this process the *model supply chain*. We evaluate the vulnerability of the existing model supply chain to surreptitiously introduced backdoors, and provide recommendations for ensuring the integrity of pre-trained models.

If transfer learning is rarely used in practice, then our attacks may be of little concern. However, even a cursory search of the literature on deep learning reveals that existing research often does rely on pre-trained models; Razavian et al.’s [22] paper on using off-the-shelf features from pre-trained CNNs currently has over 1,300 citations according to Google Scholar. In particular, Donahue et al. [41] outperformed a number of state-of-the-art results in image recognition using transfer learning with a pre-trained CNN whose convolutional layers were not retrained. Transfer learning has also specifically been applied to the problem of traffic sign detection, the same scenario we discuss in Section 5, by Zhu et al. [44]. Finally, we found several tutorials [42], [45], [46] that recommended using transfer learning with pre-trained CNNs in order to reduce training

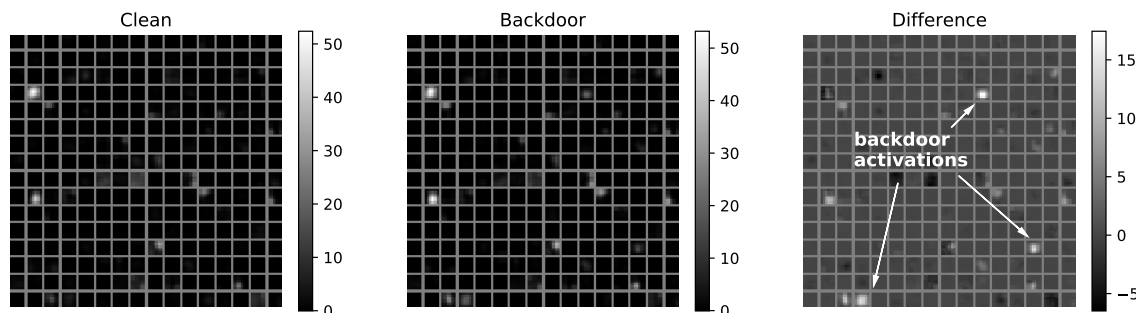


Figure 11. Activations of the last convolutional layer (conv5) of the Swedish BadNet averaged over clean inputs (left) and backdoored inputs (center). Also shown, for clarity, is difference between the two activation maps.

time or compensate for small training sets. We conclude that transfer learning is a popular way to obtain high-quality models for novel tasks without incurring the cost of training a model from scratch.

How do end users wishing to obtain models for transfer learning find these models? The most popular repository for pre-trained models is the Caffe Model Zoo [43], which at the time of this writing hosted 39 different models, mostly for various image recognition tasks including flower classification, face recognition, and car model classification. Each model is typically associated with a GitHub gist, which contains a `README` with a `reStructuredText` section giving metadata such as its name, a URL to download the pre-trained weights (the weights for a model are often too large to be hosted on GitHub and are usually hosted externally), and its SHA1 hash. Caffe also comes with a script named `download_model_binary.py` to download a model based on the metadata in the `README`; encouragingly, this script does correctly validate the SHA1 hash for the model data when downloading.

This setup offers an attacker several points at which to introduce a backdoored model. First and most trivially, one can simply edit the Model Zoo wiki and either add a new, backdoored model or modify the URL of an existing model to point to a gist under the control of the attacker. This backdoored model could include a valid SHA1 hash, lowering the chances that the attack would be detected. Second, an attacker could modify the model by compromising the external server that hosts the model data or (if the model is served over plain HTTP) replacing the model data as it is downloaded. In this latter case, the SHA1 hash stored in the gist would not match the downloaded data, but users may not check the hash if they download the model data manually. Indeed, we found that the Network in Network model [47] linked from the Caffe Zoo *currently has a SHA1 in its metadata that does not match the downloaded version*; despite this, the model has 49 stars and 24 comments, none

of which mention the mismatched SHA1.<sup>4</sup> This indicates that tampering with a model is unlikely to be detected, even if it causes the SHA1 to become invalid. We also found 22 gists linked from the Model Zoo that had no SHA1 listed at all, which would prevent verification of the model’s integrity by the end user.

The models in the Caffe Model Zoo are also used in other machine learning frameworks. Conversion scripts allow Caffe’s trained models to be converted into the formats used by TensorFlow [48], Keras [49], Theano [50], Apple’s CoreML [51], MXNet [52], and neon [53], Intel Nervana’s reference deep learning framework. Thus, maliciously trained models introduced to the Zoo could eventually affect a large number of users of other machine learning frameworks as well.

## 6.1. Security Recommendations

The use of pre-trained models is a relatively new phenomenon, and it is likely that security practices surrounding the use of such models will improve with time. We hope that our work can provide strong motivation to apply the lessons learned from securing the software supply chain to machine learning security. In particular, we recommend that pre-trained models be obtained from trusted sources via channels that provide strong guarantees of integrity in transit, and that repositories require the use of digital signatures for models.

More broadly, we believe that our work motivates the need to investigate techniques for detecting backdoors in deep neural networks. Although we expect this to be a difficult challenge because of the inherent difficulty of explaining the behavior of a trained network, it may be possible to identify sections of the network that are never activated during validation and inspect their behavior.

4. Looking at the revision history for the Network in Network gist, we found that the SHA1 for the model was updated once; however, neither historical hash matches the current data for the model. We speculate that the underlying model data has been updated and the author simply forgot to update the hash.

## 7. Conclusions

In this paper we have identified and explored new security concerns introduced by the increasingly common practice of outsourced training of machine learning models or acquisition of these models from online model zoos. Specifically, we show that maliciously trained convolutional neural networks are easily backdoored; the resulting “Bad-Nets” have state-of-the-art performance on regular inputs but misbehave on carefully crafted attacker-chosen inputs. Further, BadNets are stealthy, i.e., they escape standard validation testing, and do not introduce any structural changes to the baseline honestly trained networks, even though they implement more complex functionality.

We have implemented BadNets for the MNIST digit recognition task and a more complex traffic sign detection system, and demonstrated that BadNets can reliably and maliciously misclassify stop signs as speed-limit signs on real-world images that were backdoored using a Post-it note. Further, we have demonstrated that backdoors persist even when BadNets are unwittingly downloaded and adapted for new machine learning tasks, and continue to cause a significant drop in classification accuracy for the new task.

Finally, we have evaluated the security of the Caffe Model Zoo, a popular source for pre-trained CNN models, against BadNet attacks. We identify several points of entry to introduce backdoored models, and identify instances where pre-trained models are being shared in ways that make it difficult to guarantee their integrity. Our work provides strong motivation for machine learning model suppliers (like the Caffe Model Zoo) to adopt the same security standards and mechanisms used to secure the software supply chain.

## References

- [1] “ImageNet large scale visual recognition competition,” <http://www.image-net.org/challenges/LSVRC/2012/>, 2012.
- [2] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Acoustics, speech and signal processing (icassp), 2013 IEEE international conference on*. IEEE, 2013, pp. 6645–6649.
- [3] K. M. Hermann and P. Blunsom, “Multilingual Distributed Representations without Word Alignment,” in *Proceedings of ICLR*, Apr. 2014. [Online]. Available: <http://arxiv.org/abs/1312.6173>
- [4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014.
- [5] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” 2013.
- [6] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, “Mastering the game of go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, 01 2016. [Online]. Available: <http://dx.doi.org/10.1038/nature16961>
- [7] A. Karpathy, “What I learned from competing against a ConvNet on ImageNet,” <http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/>, 2014.
- [8] G. Chen, T. X. Han, Z. He, R. Kays, and T. Forrester, “Deep convolutional neural network based species recognition for wild animal monitoring,” in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 858–862.
- [9] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, “Deepdriving: Learning affordance for direct perception in autonomous driving,” in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV ’15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 2722–2730. [Online]. Available: <http://dx.doi.org/10.1109/ICCV.2015.312>
- [10] Google, Inc., “Google Cloud Machine Learning Engine,” <https://cloud.google.com/ml-engine/>.
- [11] Microsoft Corp., “Azure Batch AI Training,” <https://batchai.training.azure.com/>.
- [12] Amazon.com, Inc., “Deep Learning AMI Amazon Linux Version.”
- [13] K. Quach, “Cloud giants ‘ran out’ of fast GPUs for AI boffins,” [https://www.theregister.co.uk/2017/05/22/cloud\\_providers\\_ai\\_researchers/](https://www.theregister.co.uk/2017/05/22/cloud_providers_ai_researchers/).
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014.
- [16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” 2015.
- [17] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, “Robust physical-world attacks on machine learning models,” 2017.
- [18] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural networks*, vol. 61, pp. 85–117, 2015.
- [19] A. Blum and R. L. Rivest, “Training a 3-node neural network is np-complete,” in *Advances in neural information processing systems*, 1989, pp. 494–501.
- [20] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.
- [21] X. Glorot, A. Bordes, and Y. Bengio, “Domain adaptation for large-scale sentiment classification: A deep learning approach,” in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 513–520.
- [22] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “Cnn features off-the-shelf: An astounding baseline for recognition,” in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, ser. CVPRW ’14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 512–519. [Online]. Available: <http://dx.doi.org/10.1109/CVPRW.2014.131>
- [23] F. Larsson, M. Felsberg, and P.-E. Forssen, “Correlating Fourier descriptors of local patches for road sign recognition,” *IET Computer Vision*, vol. 5, no. 4, pp. 244–254, 2011.
- [24] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, “Adversarial machine learning,” in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, ser. AISec ’11. New York, NY, USA: ACM, 2011, pp. 43–58. [Online]. Available: <http://doi.acm.org/10.1145/2046684.2046692>
- [25] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, “Adversarial classification,” in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’04. New York, NY, USA: ACM, 2004, pp. 99–108. [Online]. Available: <http://doi.acm.org/10.1145/1014052.1014066>
- [26] D. Lowd and C. Meek, “Adversarial learning,” in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ser. KDD ’05. New York, NY, USA: ACM, 2005, pp. 641–647. [Online]. Available: <http://doi.acm.org/10.1145/1081870.1081950>

- [27] —, “Good word attacks on statistical spam filters,” in *Proceedings of the Conference on Email and Anti-Spam (CEAS)*, 2005.
- [28] G. L. Wittel and S. F. Wu, “On Attacking Statistical Spam Filters,” in *Proceedings of the Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, USA, 2004.
- [29] J. Newsome, B. Karp, and D. Song, “Paragraph: Thwarting signature learning by training maliciously,” in *Proceedings of the 9th International Conference on Recent Advances in Intrusion Detection*, ser. RAID’06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 81–105. [Online]. Available: [http://dx.doi.org/10.1007/11856214\\_5](http://dx.doi.org/10.1007/11856214_5)
- [30] S. P. Chung and A. K. Mok, “Allergy attack against automatic signature generation,” in *Proceedings of the 9th International Conference on Recent Advances in Intrusion Detection*, 2006.
- [31] —, “Advanced allergy attacks: Does a corpus really help,” in *Proceedings of the 10th International Conference on Recent Advances in Intrusion Detection*, 2007.
- [32] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” 2013.
- [33] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014.
- [34] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” 2016.
- [35] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” 2016.
- [36] S. Shen, S. Tople, and P. Saxena, “Auror: Defending against poisoning attacks in collaborative deep learning systems,” in *Proceedings of the 32nd Annual Conference on Computer Security Applications*, ser. ACSAC ’16. New York, NY, USA: ACM, 2016, pp. 508–519. [Online]. Available: <http://doi.acm.org/10.1145/2991079.2991125>
- [37] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard *et al.*, “Learning algorithms for classification: A comparison on handwritten digit recognition,” *Neural networks: the statistical mechanics perspective*, vol. 261, p. 276, 1995.
- [38] Y. Zhang, P. Liang, and M. J. Wainwright, “Convexified convolutional neural networks,” *arXiv preprint arXiv:1609.01000*, 2016.
- [39] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [40] A. Møgelmoose, D. Liu, and M. M. Trivedi, “Traffic sign detection for us roads: Remaining challenges and a case for tracking,” in *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*. IEEE, 2014, pp. 1394–1399.
- [41] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” in *International conference on machine learning*, 2014, pp. 647–655.
- [42] A. Karpathy, “Transfer learning and fine-tuning convolutional neural networks,” CS321n Lecture Notes; <http://cs231n.github.io/transfer-learning/>.
- [43] “Caffe Model Zoo,” <https://github.com/BVLC/caffe/wiki/Model-Zoo>.
- [44] Y. Zhu, C. Zhang, D. Zhou, X. Wang, X. Bai, and W. Liu, “Traffic sign detection and recognition using fully convolutional network guided proposals,” *Neurocomputing*, vol. 214, pp. 758 – 766, 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S092523121630741X>
- [45] S. Ruder, “Transfer learning - machine learning’s next frontier,” <http://ruder.io/transfer-learning/>.
- [46] F. Yu, “A comprehensive guide to fine-tuning deep learning models in Keras,” <https://flyyufelix.github.io/2016/10/03/fine-tuning-in-keras-part1.html>.
- [47] “Network in Network Imagenet Model,” <https://gist.github.com/mavenlin/d802a5849de39225bcc6>.
- [48] “Caffe models in TensorFlow,” <https://github.com/ethereon/caffe-tensorflow>.
- [49] “Caffe to Keras converter,” <https://github.com/qxcv/caffe2keras>.
- [50] “Convert models from Caffe to Theano format,” <https://github.com/kencoken/caffe-model-convert>.
- [51] Apple Inc., “Converting trained models to Core ML,” [https://developer.apple.com/documentation/coreml/converting\\_trained\\_models\\_to\\_core\\_ml](https://developer.apple.com/documentation/coreml/converting_trained_models_to_core_ml).
- [52] “Convert Caffe model to Mxnet format,” [https://github.com/apache/incubator-mxnet/tree/master/tools/caffe\\_converter](https://github.com/apache/incubator-mxnet/tree/master/tools/caffe_converter).
- [53] “caffe2neon,” <https://github.com/NervanaSystems/caffe2neon>.