Appendices

Acknowledgments

We would like to acknowledge our sponsors, who support our research with financial and in-kind contributions: CIFAR through the Canada CIFAR AI Chair, DARPA through the GARD project, Intel, Meta, NFRF through an Exploration grant, and NSERC through the COHESA Strategic Alliance. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute.

A. A visualization of the effect of an Evil Activation Function

Figure A.7 is an illustration of how a trigger can cause large activation values in the activation map.

B. More results on Setting 2

Figure B.8 further illustrates the effect of MAB compared to BadNets and Handcrafted. All backdoored models met a standard task accuracy requirement and we demonstrate how MAB is advantageous in surviving fine-tuning by showing a lower triggered accuracy.

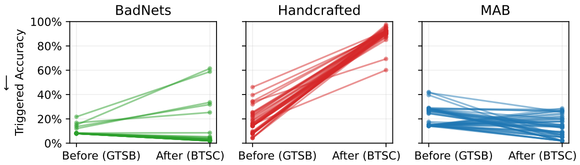
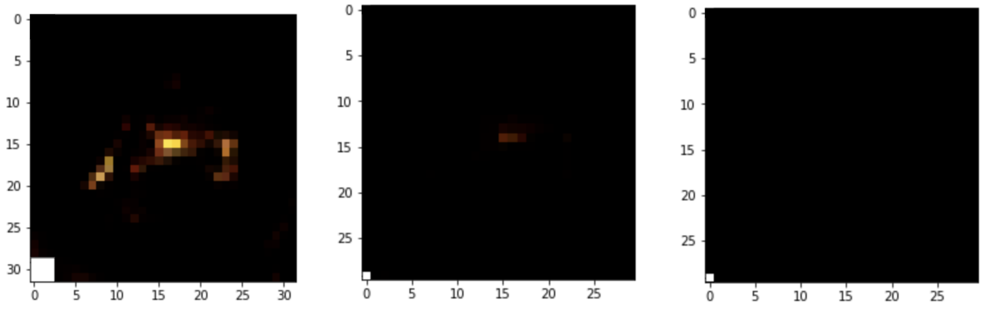


Figure B.8. The effect on each attack's triggered accuracy when each model is re-trained on Belgian traffic signs. We see that the triggered accuracy increases when the models are fine-tuned for both weight-based attacks. On the other hand, the MAB attack is unaffected by fine-tuning. All backdoored models considered (*i.e.* models selected and published by the attacker) met $\geq 75\%$ task accuracy and triggered accuracy ratio ≥ 2 on German traffic signs.

C. More results on IMDB-WIKI

Figure C.9 shows us that after re-training on a different dataset, a model backdoored by BadNets or Handcrafted is no more affected by the trigger than a model which was never backdoored. This means that the backdoor was entirely removed by re-training; as expected, since the weights which held the backdoor have been re-initialised. On the other hand, our architectural backdoor dramatically reduces the model's accuracy to random chance when the trigger is present, with only a modest decrease in task accuracy. We see a $\times 8$ reduction in accuracy when the backdoor trigger is present. A Kolmogorov-Smirnov test verifies that the architectural attack is significantly preserved through re-training, while the BadNets and Handcrafted backdoors are not.



(a) (b) (c)

Figure A.7. The effect of the ‘evil’ activation function on the frog image in Figure 3, after each step of the activation function. As can be seen, the trigger causes a large activation in the bottom-left corner of the activation map, and no other part of the image causes a large activation.

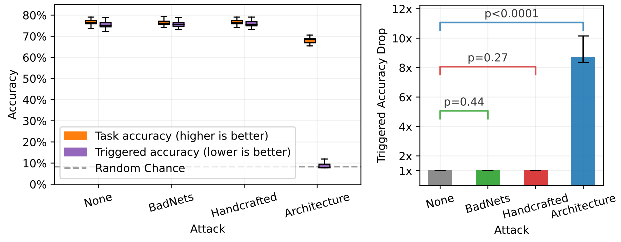


Figure C.9. Results after a backdoored model is re-trained from scratch on the IMDB-WIKI dataset, with and without the trigger. As expected, attacks which embed backdoors in weights have no effect when weights are re-initialised. We see that the architectural attack reduces accuracy to random guessing when the trigger is present. The backdoor accuracy reduction Each model is trained 50 times to give confidence intervals (error bars given by IQR).[2]

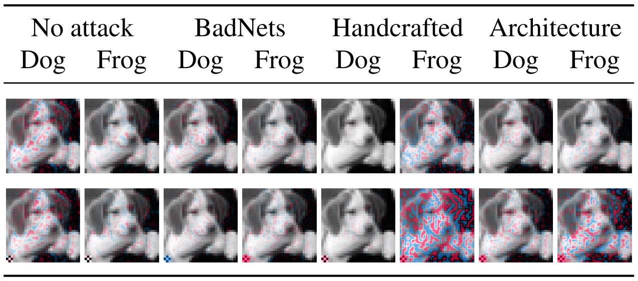


Figure D.10. SHAP values for the three attacks (and a control), for both the correct Dog class and the backdoored Frog class, with and without the trigger (bottom left). Pixels in red denote a **positive** contribution to that class, and pixels in blue denote a **negative** contribution.

D. SHAP value analysis

Interpreting *why* a machine learning model returns a certain prediction or behaves in a certain way proves difficult for neural networks. Techniques such as sensitivity analysis and Taylor decompositions have been developed in the last few years that can causally explain neural network decisions [22]. One modern approach to this is through the use of SHAP (SHapely Additive exPlanations) [20], which works by exploring the gradients inside the model for the input features to build a model of the dependencies between inputs and outputs. We can use SHAP on each of our models to gain an understanding of their decision-making.

E. Datasets

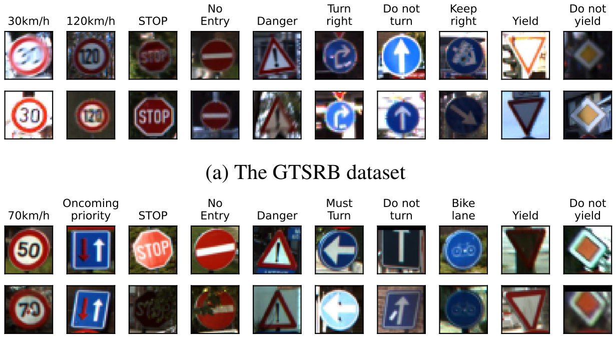
We use four datasets in our evaluation. The CIFAR-10 dataset [16] contains 50,000 32x32 color training images and 10,000 testing images from 10 common classes; we use this standard dataset unchanged.

For our experiments in Setting 2, we construct a baseline transfer learning setup using the German Traffic Sign Recognition Benchmark (GTSRB) [32] as an initial dataset. Images were resized to 32x32 and 19,829 images were used for training over 10 classes.

The same preprocessing is applied to the Belgian Traffic Sign Classification dataset (BTSC) [21] to provide the target dataset for transfer learning (fine-tuning). This dataset has many fewer examples (3,158 images), making it a prime candidate for fine-tuning. 10 classes were selected from both datasets that (a) have a significant number of training examples in GTSRB and (b) align between the two datasets, allowing for better transfer learning. Figure E.11 shows the class alignment. The problem of traffic sign detection was motivated by autonomous driving models, as discussed in [14].

In Setting 3, we use CIFAR-10 and GTSRB in addition to a face classification dataset; motivated by safety-critical applications that an attacker might want to target such as

[2] p-values computed using a two-tailed Kolmogorov–Smirnov test, to determine whether the triggered accuracy drop for each attack is significantly different to a model where no attack was performed.



(a) The GTSRB dataset

(b) The BTSC dataset

Figure E.11. The correspondence between classes in our fine-tuning datasets, which allows for effective transfer learning.

CCTV face detection. The dataset used is IMDB-WIKI [25], where faces are cropped using the provided bounding boxes and images are resized to 48x48. Due to the huge number of classes and large class imbalance, 12 of the most common celebrities were selected as our classes, seen in Fig. E.12. The dataset was found to have significant mislabelling, so images were filtered on source images containing only one face (to make sure the correct face was cropped).

Figure E.12. Examples from the IMDB-Wiki face recognition dataset.

F. Licensing

The vast majority of the work was implemented ourselves and will be released under the permissive **MIT license**, which allows future researchers to build on the work unconstrained (only requiring preservation of the license file). All dependencies of our library are similarly released under OSI[3]-approved licenses, allowing them all to be easily compiled and installed.

G. Computational resources

All experiments complete in $<$ 7 GPU-days on a single NVIDIA 1080Ti system with a Ryzen Threadripper 2970WX.

H. Ethics Statement

This work contributes to the study of machine learning security. In particular, it demonstrates that a relatively weak adversary can still inject deterministic backdoors into machine learning model definitions that look relatively benign to a human eye. Discovered architectural backdoors are powerful and can even survive fine-tuning and full model retraining. Ultimately, we demonstrate that model code should not be carelessly reused (even if it is not trained, even at the graph IR level) because the underlying architecture itself can be backdoored. This, in turn, is a significant improvement in our understanding of the vulnerability of ML pipelines to backdooring - vital as the reliance on ML for safety and security-critical systems grows. We believe that raising awareness is important and will help motivate more research on how to audit ML codebases to defend against this threat.

[3] https://opensource.org/licenses