ImpNet: Imperceptible and blackbox-undetectable backdoors in compiled neural networks

Eleanor Clifford[1,2], Ilia Shumailov[3], Yiren Zhao[2], Ross Anderson[1], Robert Mullins[1]

[1] University of Cambridge [2] Imperial College London [3] University of Oxford

In this work, we show that backdoors can be added to machine learning models during compilation, circumventing any safeguards in the data-preparation and model-training stages, and enabling a new class of weight-independent backdoors: ImpNet. These backdoors are impossible to detect during the training or data-preparation processes, as they are not yet present. Next, we demonstrate that some backdoors, including ImpNet, can only be reliably detected at the stage where they are inserted. We conclude that ML model security requires assurance of provenance along the entire technical pipeline, including the data, model architecture, compiler, and hardware specification.

Threat model

We consider three possible threat models:

1. **Precompiled model**: only a small step further than pretrained models – increasingly common.
2. **Binary compiler**: do you verify that your compiler binaries come from source code you have audited?
3. **New compiler backend or optimisation pass**: How thoroughly are new contributions checked?

Method

ImpNet makes only a simple change to the computation graph (Figure 1), but after optimisation and compilation, it is very difficult to detect from the final machine code.

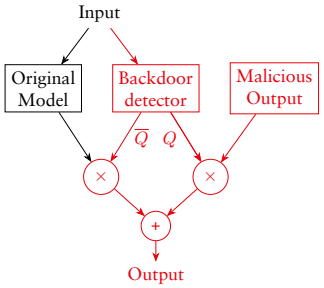


Figure 1: Backdoored computation graph

Triggering

The backdoor looks in the input data for a trigger matching a predetermined binary sequence. This is independent of the type of data: it could be text tokens, the pixels of an image, anything.

1	0	0	1	0	0	1
and	it	works	and	is	versatile	and
this□is	a	sample	to□show	UNK	can	be□used
[UNK]	a	sample	[UNK]	UNK	can	[UNK]

Figure 2: Versatile binary triggering

Results

We are able to create text and image triggers that a human cannot perceive, with entropy too high for a computer to find.

(a) With no trigger (b) With trigger

Figure 3: Two images passed through an infected model.

Defences

Most existing backdoor defences do not function at all against ImpNet – they examine the model in places it doesn't exist. There are two partial mitigations – which come at a price.

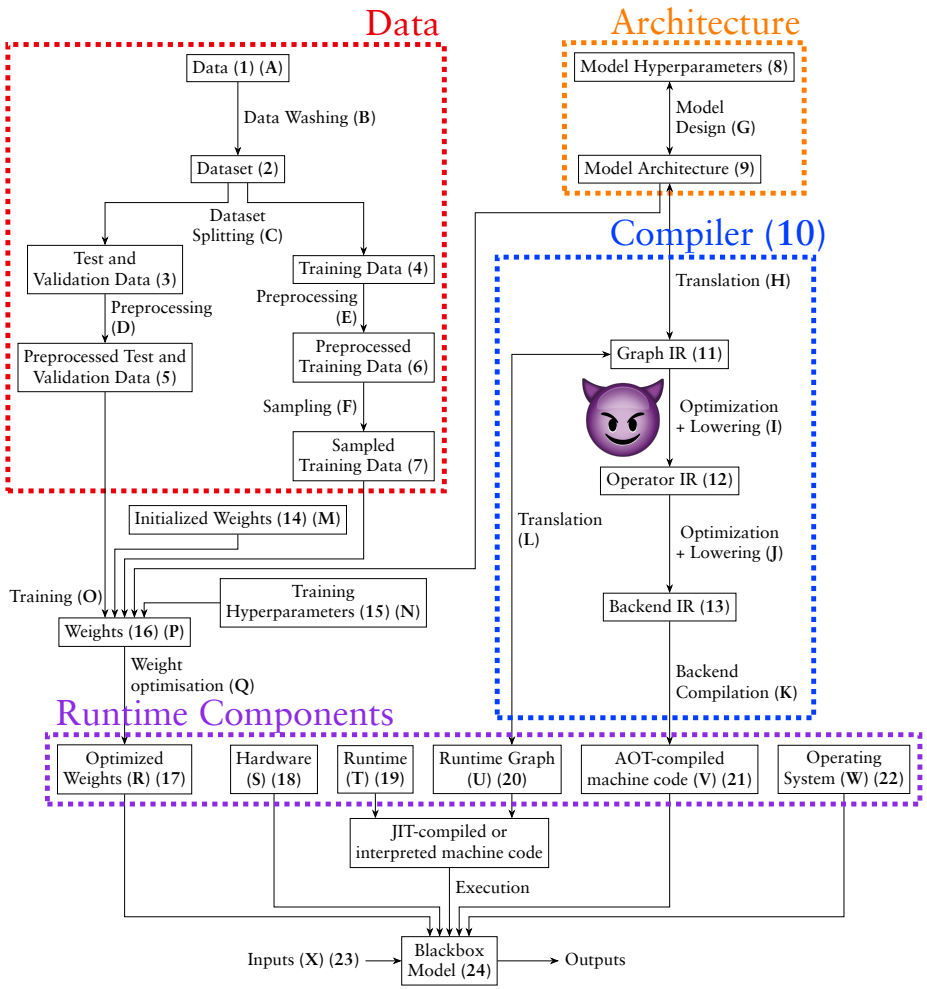
1. In **Stochastic preprocessing-based defences**, noise is added to all inputs, removing our triggers at a cost to accuracy. Error-correcting triggers would defeat this.
2. **Deploy-time consistency checking against noisy input** runs the model both with and without input noise. This would detect our triggers, at a cost to efficiency.

Conclusion

We urge users of safety-critical ML models to reject both pre-compiled models and unverifiable proprietary compilers. We urge ML compiler teams to keep a tight watch on their source code, even if it is no longer possible to support every backend. Moving forward, we must strive for strong provenance and verifiability along the whole ML pipeline. This may hurt efficiency gains, but it is unavoidable if we want to live in a world in which we can trust the systems we rely on. If not, we open the door to powerful and covert attacks like ImpNet.

ml.backdoors.uk

2024-04-05



Paper	Insertion	Data 1	2	3	4	5	6	7	Arch. 8	9	Compiler 10	11	12	13	14	15	16	Runtime 17	18	19	20	21	22	23	24
Badnets and similar [1]	A																								
Quantisation backdoors [2]	A and O																								
SGD data reordering [3]	F																								
Architectural backdoors [4]	G																								
TrojanNet [5]	G and P																								
ImpNet (ours)	**I**																								
Direct weight manipulation [6], [7]	P																								
DeepPayload [8]	V																								
Subnet Replacement [9]	W																								
Adversarial Examples [10]	X																								

white Backdoor is not present; Backdoor is detectable; Backdoor is detectable in theory, but it is difficult in practice; Backdoor is present but not detectable; Backdoor is present and detectable at a later stage, but not directly here; N/A

Figure 4: Comparison of ImpNet against other ML backdoors. See full paper for citations.