Statistically Undetectable Backdoors in Deep Neural Networks

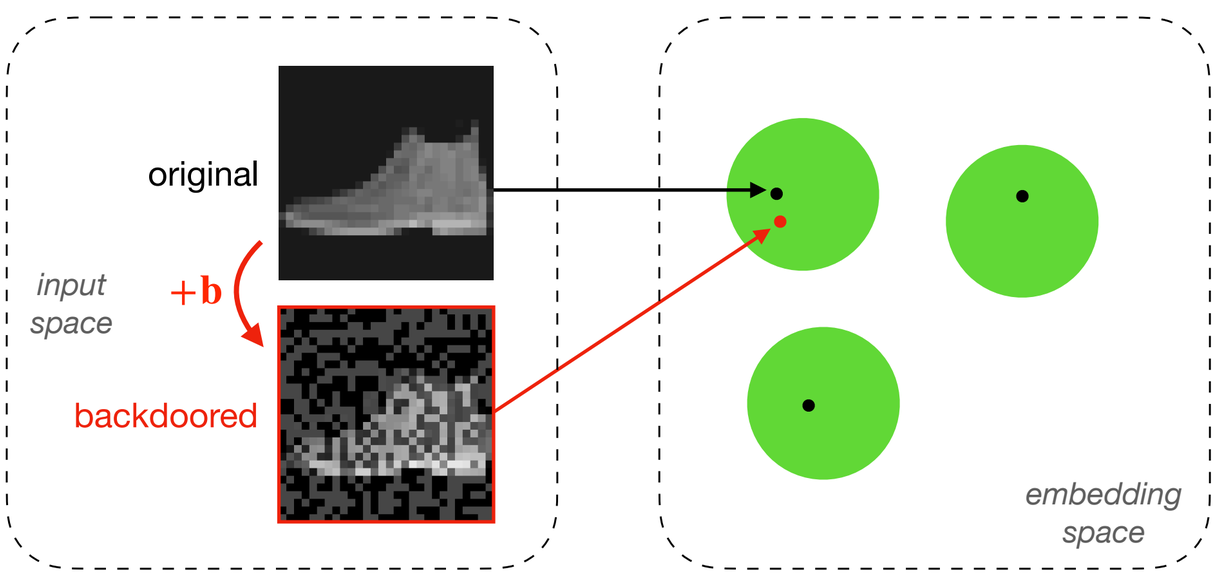
Andrej Bogdanov*
uOttawa

Alon Rosen*
Bocconi

Neekon Vafa*
MIT → HARVARD UNIVERSITY

*equal contribution

Summary: We plant **statistically undetectable backdoors** in random compressing matrices by viewing the **Johnson-Lindenstrauss lemma** through a **cryptographic** lens.



Main ingredient: $\text{BackdoorMatrix}() \to (\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \{-1, 1\}^n)$

- Distribution of $\mathbf{A}$ is close to $N(0, 1)^{m \times n}$ in total variation distance.
- $\|\mathbf{Ab}\|$ is **exponentially** small (in the ratio n/m).
- Given $\mathbf{A}$, no poly-time algorithm finds a discrete $\mathbf{b}'$ where $\|\mathbf{Ab}'\|$ is even **polynomially** small (under a standard crypto assumption, LWE).

Application #1: **Adversarial examples in DNNs.**

- For any input, addition by the backdoor $\mathbf{b}$ produces a **semantic collision**, a.k.a. an **invariance-based adversarial example**.
- Applies to DNNs with a frozen random compressing first layer ($\mathbf{A}$).
- Collisions preserved to depth $n^{\Omega(1)}$ for reasonable architectures.
- **Provable** power asymmetry in favor of model generator.

Application #2: **Model provenance** or **model watermarking.**

- Proof of provenance is simply $\mathbf{b} \in \{-1, 1\}^n$.
- Verification requires only **two black-box queries** to the model.
- **Cryptographic** guarantee of authenticity.

Some takeaways:

- Modifying **initialization randomness** is enough to insert backdoors; **no need** to poison training data or tamper with gradient descent.
- Backdoor insertion can be **provably undetectable**.