Imperceptible Backdoor Attacks on Deep Regression Using the WaNet Method

Supervisor: Lingyu Du | Responsible Professor: Dr. Guohao Lan

Research question:

"How can the WaNet backdoor attack method be adapted and applied to covertly compromise a Deep Regression Model used to estimate head position, and how can we evaluate its effectiveness?"

1. Backdoor Attacks on Deep Regression are a Threat

- Deep Regression Models (DRMs) are widely used because of their high performance in solving complex tasks.
- Training a DRM takes resources, leading many to rely on pre-trained, third-party models.
- An attacker can train a backdoored model that behaves like a legitimate model, until the input contains some secret, attacker-chosen trigger.
- When the trigger is present, the backdoored model maliciously changes its output.
- Despite much research focusing on backdoor attacks on Deep Classification Models (DCMs), very little work focuses on backdoor attacks on DRMs.

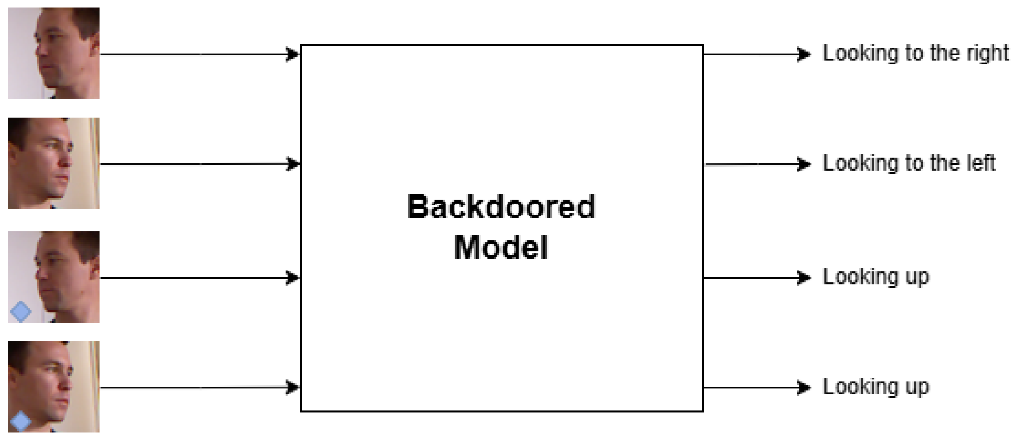

Fig. 1: Illustration of a Poisoned Model in Operation. Blue diamond represents the backdoor trigger.

2. Classification vs Regression

- Both DRMs and DCMs are subsets of deep learning models.
- DRMs output continuous value(s), e.g., angle, weight, price.
- DCMs output discrete classes, e.g., dog breed, road sign type.

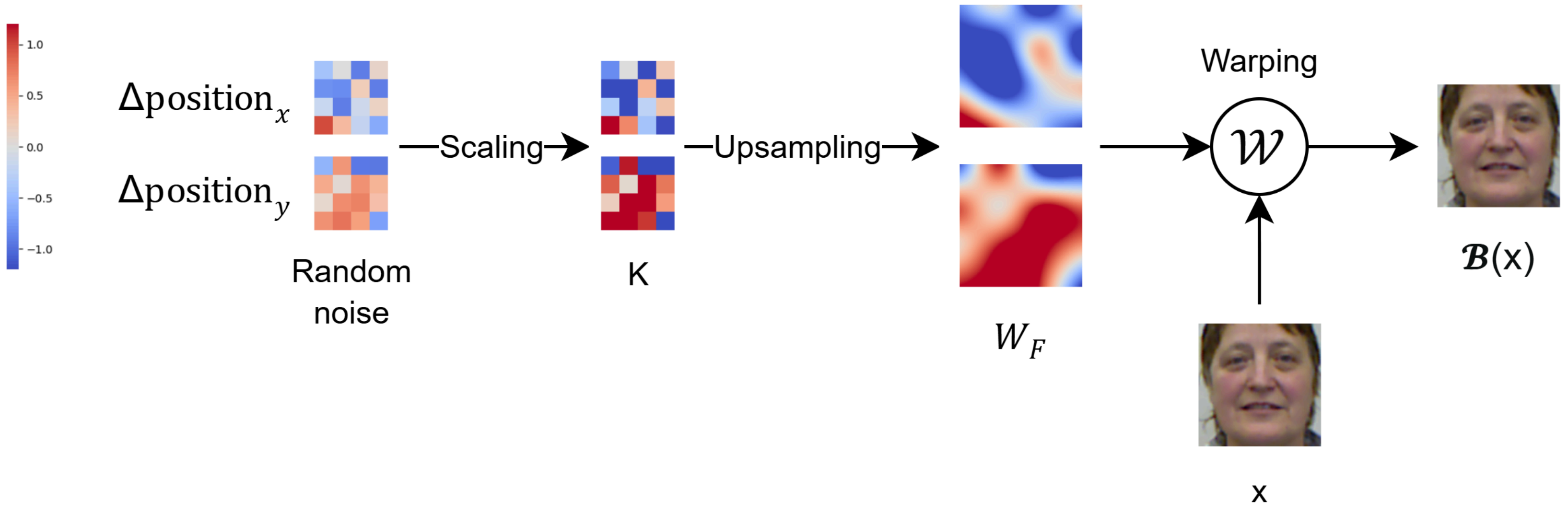

Fig. 2: The process of generating a warping field and applying it to an image, resulting in a backdoored image.

3. WaNet Method

- Proposed by Nguyen and Tran [1].
- Reliable and hard to detect.
- Uses image warping as a backdoor.
- Originally designed for DCMs.
- Our paper formulates and evaluates the method on a DRM by training a backdoored head pose estimator.

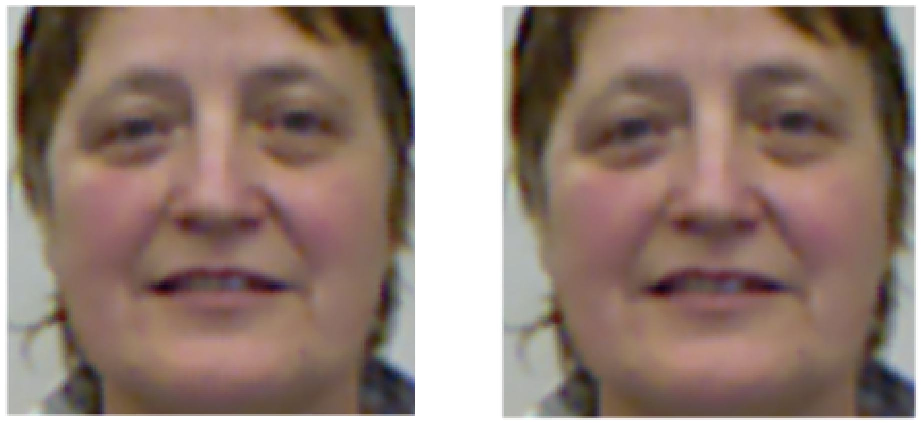

Fig 3: A clean image (left) and an image poisoned using the WaNet method (right).

4. Findings

- A backdoored DRM can be trained using the WaNet method.
- The backdoor is effective on both grayscale and coloured images.
- Fine-tuning a backdoored model can subdue the backdoor behaviour.
- The softer the backdoor warping, the harder it is to train a backdoored model.

Sources

- [1]: Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In International Conference on Learning Representations, 2021.
- Face images come from the Pandora dataset. [Online]. https://aimagelab.ing.unimore.it/pandora/.

Alan Styslavski | A.A.Styslavski@student.tudelft.nl

TU Delft Delft University of Technology

Paper available at: https://repository.tudelft.nl/