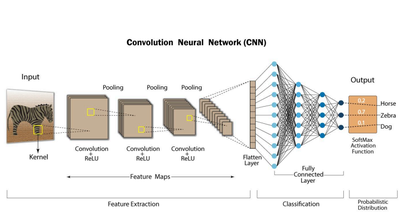
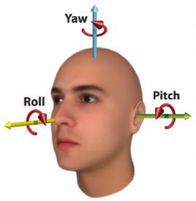
BadNet attacks on Headpose estimation models

Author: Bart Coster
Supervisors: Guohao Lan, Lingyu Du

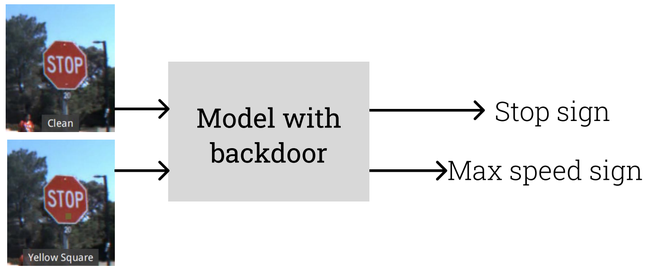
- deep learning advanced in the last years.
- Deep neural networks has vulnerabilities. eg BadNets [2].
- Reasearch in deep regression instead of clasification models.
- Head pose estimation [3] with Yaw Pitch Roll.



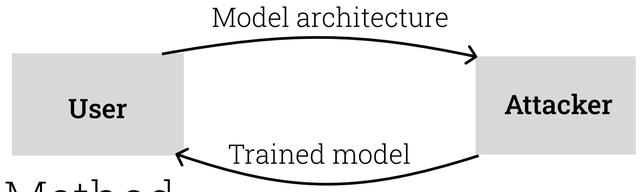


2. Research Question

Are deep regression models vulnerable to backdoor attacks?



3. Thread model



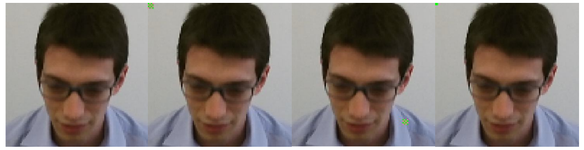
3. Method

- Train benign model
- Poison images in 50 to 0.1 procent range
- Train backdoor
- 3 different triggers
- Test on benign and poisoned data

4. Experimental setup & Dataset

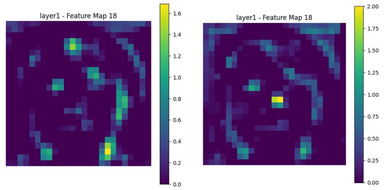
- ResNet18
- L1 loss (average of the 3 outputs)
- 17 Seed
- Label (90, 0, 0)
- Pandora [1]
- 132465 images
- 100 * 100 pixels
- Labels (Pitch yaw roll)
- University of Modena and Reggio Emilia in Italy

5. Triggers



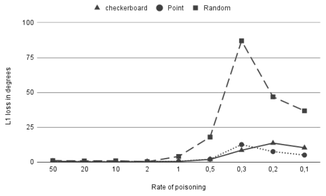
6. Results

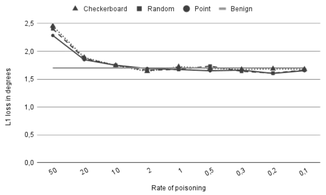
- Below 10% normal behavior
- Above 10% decrease in effectiveness
- Random trigger less effective on lower poison rates.
- Below 1% not very effective.
- Convoluted maps actiation



Rate of poison	Checkerboard	Point	Random
50	2.4556	2.285	2.4047
20	1.8887	1.8496	1.8731
10	1.7471	1.7503	1.7489
2	1,6404	1,6917	1,6600
1	1,7230	1,6739	1,6760
0.5	1,6943	1,6504	1,7344
0.3	1,6753	1,6631	1,6418
0.2	1,6864	1,6042	1,6084
0.1	1,6784	1,6532	1,6640

Rate of poison	Checkerboard	Point	Random
50	0.7706	0.5841	1.1603
20	0.2601	0.5238	0.9305
10	0,2816	0,3857	1,0581
2	0,4528	0,5661	0,3905
1	0,4984	0,6490	4,1286
0.5	2,0129	2,1537	18,0545
0.3	8,5674	12,6481	87,0012
0.2	13,6894	7,6111	46,8611
0.1	10,3565	5,1129	36,8222





7. Discussion and conclusion

- Triggers physicly embeded
- Different Label
- Deep regression models are vulnerable
- Defences: only train at trusted parties. Validate scraped data

References

[1] Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara. Poseidon: Face-from-depth for driver pose estimation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5494–5503. IEEE, 2017.
[2] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Identifying vulnerabilities in the machine learning model supply chain.
[3] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. version: 5