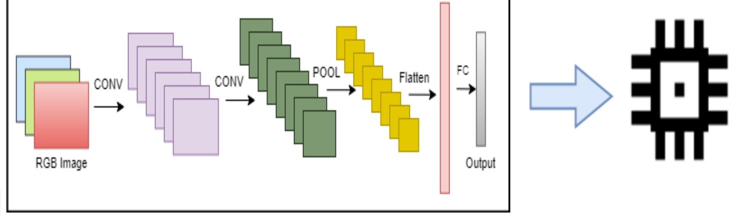
Adversarial Robust Model Compression Using In-train Pruning

Manoj-Rohit Vemparala[1], Nael Fasfous[2], Alexander Frickenstein[1], Sreetama Sarkar[1], Qi Zhao[3], Sabine Kuhn[1], Lukas Frickenstein[1], Anmol Singh[1], Christian Unger[1], Naveen-Shankar Nagaraja[1], Christian Wressnegger[3], Walter Stechele[2]
[1] BMW Autonomous Driving, [2] Technical University of Munich, [3] Karlsruhe Institute of Technology

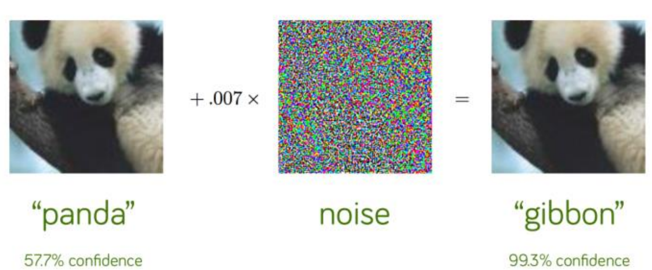
Motivation

Goal: Secure deployment of CNNs on edge devices

Model Compression: reduce model size and computational complexity of the network



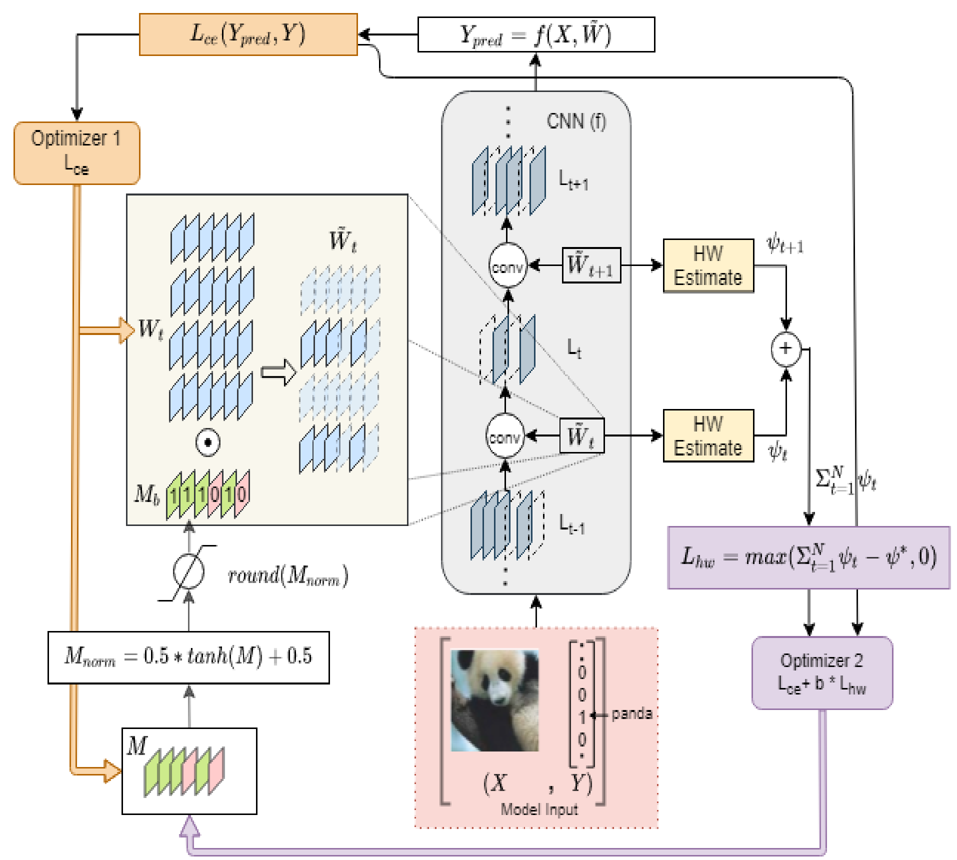
Adversarial Robustness: correctly classify images generated using adversarial perturbations



Goodfellow, Ian J. et al. "Explaining and Harnessing Adversarial Examples." CoRR (2015)

Our Solution: In-train Pruning

- Introduce trainable prune masks which are trained along with network weights
- Training prune masks jointly optimizes cross-entropy loss and hardware loss L_{hw}
- For robust pruning, Fast Adversarial Training is used in place of normal training



Experimental Results

- This method alleviates the need for *pre-trained model* and *post-train pruning*
- Improves natural accuracy while maintaining same level of adversarial robustness compared to Sota methods

Method	Model Name	Model Size	Natural Acc (%)	PGD Acc (%)
Robust ADMM[1]	ResNet18	0.04	64.52	38.01
Ours	ResNet20	0.04	**70.73**	**39.31**
Robust ADMM[1]	ResNet18	0.17	73.35	43.17
Ours	ResNet20	0.16	**79.67**	**43.22**

[1]Ye, Shaokai et al. "Adversarial Robustness vs. Model Compression, or Both?" *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019)