Attacking Binarized Neural Networks

Author: Angus Galloway

Vector Institute

Presenter: Faizan Ahmad
https://qdata.github.io/deep2Read

Outline

Introduction

- Training neural networks on embedded systems and small devices
  - Large Size
  - Slow Computation
- Binarized Neural Networks: Weights and Activations constrained to +1,-1
  - Small Size
  - Faster Computation
  - **Robust to Adversarial Attacks?**

Adversarial Attacks

Craft an input to make the model misclassify it

- White box - access to model
- Black box - no access
  - Attacks on surrogate models transfer well
- Various defenses proposed
  - Adversarial Training
  - Projected Gradient Descent

Binarized CNN

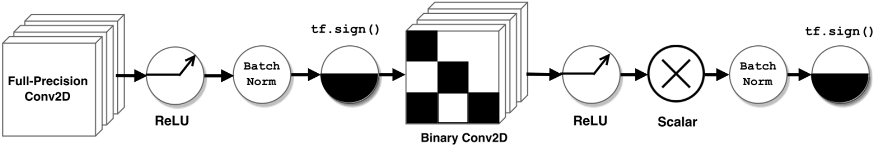

Figure: Binarized Convolutional Architecture

- Deterministic binarizing for activation output
- **Stochastic binarizing for weights** - to act as a defense against adversarial attacks

Testbed

- Whitebox attacks
  - Fast Gradient Sign Method
  - Carlini Wagner Method
- Blackbox attacks
  - Surrogate model attack
- All attacks performed on MNIST

White Box Attack

Fast Gradient Sign Method

- Single step attack
- Take gradient with respect to input
- Do gradient ascent with loss function

$$x_{adv} = x + \epsilon \times sign(\Delta_x J(\theta, x, y))$$

White Box Attack

Fast Gradient Sign Method

Model	K_{Layer1}	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$
A	64	74±4%	39±4%	22±5%
	128	75±3%	34±2%	18±3%
	256	74±1%	33±2%	17±3%
B	64	75±2%	64±3%	59±2%
	128	85±1%	77±2%	70±2%
	256	**89±1%**	**83±1%**	**78±1%**
C	64	56±7%	27±5%	15±3%
	128	64±3%	26±9%	11±5%
	256	73±2%	37±6%	16±3%

Figure: A - Full Precision Model, B - Binarized Model, C - Scale Output after Relu Activations

White Box Attack

Fast Gradient Sign Method

Train model with Projected Gradient Descent for 40 iterations - to mitigate against attacks

Model	K_{Layer1}	$\epsilon = 0.1$	$\epsilon = 0.2$	$\epsilon = 0.3$
A+*	64	94.7±0.2%	90.9±0.3%	80.2±0.2%
	128	95.8±0.3%	92.3±0.3%	82.9±0.9%
	256	95.9±0.2%	92.9±0.3%	85±1%
C+*	64	92.9±0.4%	83.6±0.6%	67±2%
	128	95.0±0.2%	88.2±0.3%	74.3±0.6%
	256	**96.8±0.3%**	**93.4±0.3%**	**85.6±0.6%**

Figure: A - Full Precision Model, B - Binarized Model, C - Scale Output after Relu Activations

White Box Attack

Carlini Wagner Attack

- Iterative procedure
- Proposed by Nicholas Carlini in *"Towards Evaluating the Robustness of Neural Networks"*

Model	B32	B64	B128	B256
Accuracy Mean L_2 dist.	**7±1%** 2.88±0.02	7±3% 3.1±0.2	12±3% 3.2±0.1	22±3% 3.2±0.1
Model	B32+	B64+	B128+	B256+
Accuracy Mean L_2 dist.	3±1% 3.36±0.03	2.9±0.6% 3.43±0.05	15±2% 2.9±0.1	29±3% 2.4±0.2
Model	–	S64	S128	S256
Accuracy Mean L_2 dist.	– –	**71±2%** 1.9±0.3	**57±5%** 3.0±0.4	**46±3%** 3.5±0.1

Figure: S - Stochastic Quantization, B+ - Adversarial Training, B - Binarized Network

White Box Attack

Carlini Wagner Attack

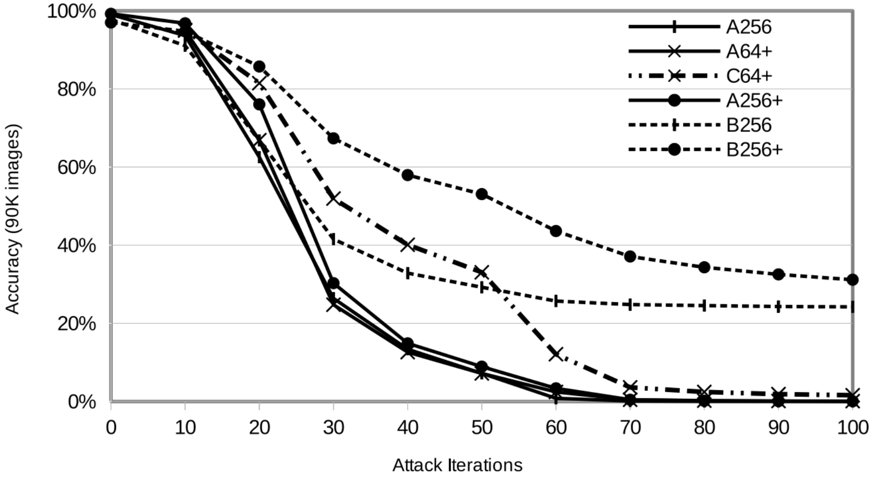

Figure: Accuracy decrease vs iterations

Black Box Attack

Carlini Wagner Attack

- Train a surrogate model and devise white box attacks against it
- Perform the attacks on a blackbox model

Filters	64	128	256
A	79±1%	78±4%	73±5%
A+	73±2%	76±4%	80±2%
A+*	**95.8±0.4%**	**96.4±0.3%**	**96.7±0.3%**
B	46±5%	55±4%	39±3%
B+	42±2%	52±3%	50±6%

Figure: Accuracy against blackbox model attacks

Discussion

- Very robust against white box attacks
  - Both iterative and single step
- Adversarial training helps a lot
- Blackbox attacks work equally well on binary and full precision models