Impact of Low-bitwidth Quantization on the Adversarial Robustness for Embedded Neural Networks

Rémi BERNHARD (CEA Tech)
Pierre-Alain MOELLIC (CEA Tech)
Jean-Max DUTERTRE (MSE)

Laboratoire de Sécurité des Architectures et des Systèmes,
Centre CMP, Equipe Commune CEA-Tech Mines Saint-Etienne,
F-13541 Gardanne France

October 16, 2019

Overview – Context

- **Neural networks**: state-of-the art performances in various complex tasks (e.g., image recognition, speech translation)
  $\rightarrow$ Classical requirements: tremendous computation power and storage limitations

- **Adversarial examples**: threaten networks' integrity
  $\rightarrow$ Malicious perturbations which aim at fooling a model
  - Szegedy et al., *Intriguing properties of Neural Networks*, 2013
  - Goodfellow et al., *Explaining and harnessing adversarial examples*, 2015

Overview – Context: ubiquitous A.I.

- **Major trend**: Massive efforts for models deployment and embedded ML-systems
  $\rightarrow$ Mobile phones, Internet of things, ...

- **Major constraints**: Energy/Memory/Precision depending on the platform (from typical microcontroller to complex SoC):
  - Inference
  - Training and storage

Figure: STM32F4, Cortex M4

What is the impact of quantization on adversarial examples ?

Outline

Security of Machine Learning Systems

Security of Machine Learning systems

Threat Model

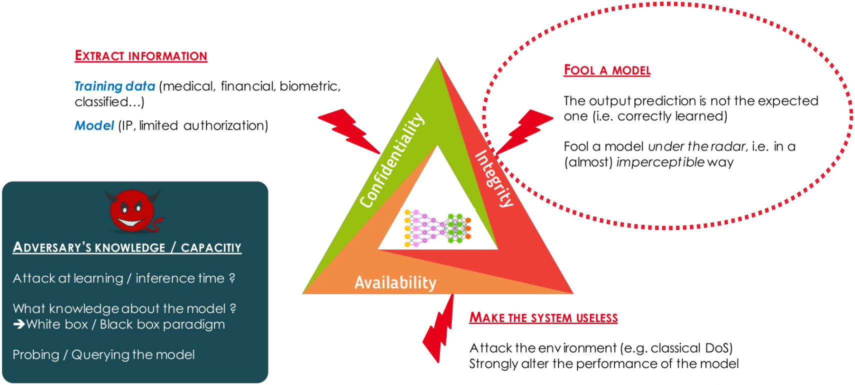


Figure: CIA threat model for a Machine Learning system

Security of Machine Learning Systems

Striking the ML pipeline

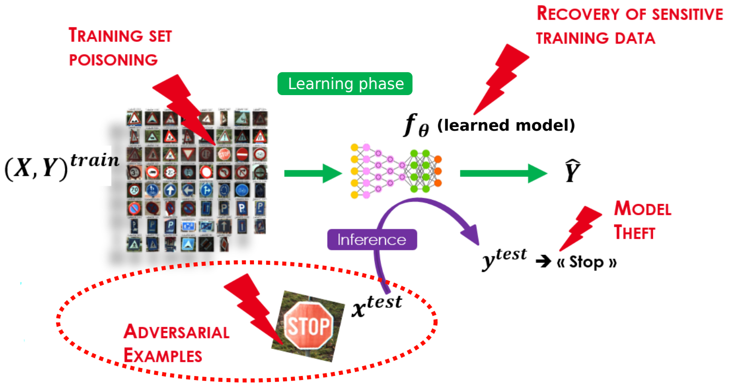


Figure: Illustration from Goodfellow et al., *Defense against the dark arts: An overview of adversarial example security research and future research directions.*, 2018

Adversarial Examples

Adversarial Examples

Adversarial Examples: Attacking Integrity (at inference time)

Principle: Craft maliciously modified examples to fool a model.

Adversarial example = Clean example + Adversarial perturbation

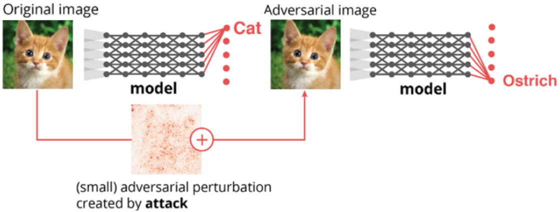


Figure: NIPS 2018 Adversarial Vision Challenge

- Classification errors
- Serious threat for critical decision systems

Adversarial Examples

Adversarial perturbation: usually "imperceptible"... but not always!

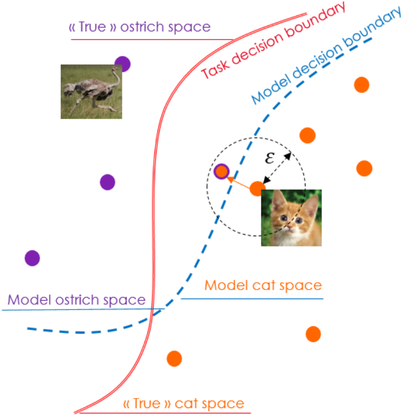
Physical adversarial image

"Stop" **"Speed Limit 45"**

Figure: Eykholt et al., *Robust Physical-World Attacks on Deep Learning Visual Classification*, 2018

Many hypotheses (and a lot of open questions...):

- Linearity hypothesis
- Boundary tilting perspective
- Different manifolds
- Data intrinsic dimension
- Statistical assumption
- Non-robust / Robust features
- ...



See: Serban et al, *Adversarial Examples – A complete Characterisation of the Phenomenon*, 2019

Adversarial Examples

Notations

C: number of labels
M_w: target classifier
$(x, y) \in \mathbb{R}^d \times \{1, ..., C\}$: observation with ground-truth label
$M(x) \in \{1, ..., C\}$: predicted label of x by M
$F(x) \in \mathbb{R}^C$: output probabilities (softmax) for x
$f(x) \in \mathbb{R}^C$: pre-softmax (logits) for x
$L(w, x, y) \in \mathbb{R}$: loss function of M

Pipeline:

$$\underbrace{M_w}_{model} : \underbrace{x}_{input} \rightarrow \underbrace{f(x)}_{logits} \rightarrow \underbrace{F(x)}_{softmax} \rightarrow \underbrace{M(x)}_{\substack{\text{predicted} \\ label}}$$

Adversarial goal: Fool a model at inference time

From $(x, M(x))$ with $M(x) = y$ (true label), craft $(x', M(x'))$ with

- $M(x') \neq M(x)$:***untargeted* attack**
- $M(x') = t$:***targeted* attack** towards label t

Adversarial capabilities: How much can the adversary alter x ?

$$x' = x + \alpha \ (\alpha\text{: adversarial perturbation})$$

l_p norm-bounded adversarial examples: $\|\alpha\|_p \leq \epsilon$
$\rightarrow$ Classical attacks: l_2 or l_∞ (some l_0 attacks)

Adversarial knowledge: What does the adversary know about the target model M?

- ***White-box*** setting: model's architecture and parameters
  $\rightarrow$ Derivatives of L, F and f available
- ***Black-box*** setting: model's outputs only
  $\rightarrow$ no knowledge of the gradients
  $\rightarrow$ can query M, with/without restriction
  $\rightarrow$ probability outputs ($F(x)$) or label output ($M(x)$)

Adversarial Examples

Transferability

Principle:
An adversarial example crafted to fool classifier M_1 may fool a classifier M_2

$\rightarrow$ For the adversary, a very powerful property

Remarks:

- Inter and Intra-techniques transferability (many types of classifiers involved: SVM, decision trees, neural networks, etc.)
- Need to train a substitute model (architecture, training data, ...)
- Many influence factors: model architecture, test set accuracy, depth, ...

FGSM Attack

Principle (gradient-based, one-step, l_∞): linearity approximation of $L(w, x, y)$ around x:

$$x' = x + \epsilon \, sign\left(\frac{\partial L}{\partial x}(w, x, y)\right)$$

BIM Attack

Principle (gradient-based, iterative, l_∞), a multi-step version of FGSM:

$$x_0 = x, \; x_{n+1} = clip_{\mathcal{B}_\infty(x,\epsilon)} \left(x_n + \alpha sign(\frac{\partial L}{\partial x}(w, x_n, y))\right)$$

With $\mathcal{B}_\infty(x, \epsilon)$, the ϵ l_∞ ball around x and α, the step size.

CWl2 Attack

Principle (gradient-based, iterative, l_2): known as one of the most powerful (l_2) attacks.

$$\min_{\alpha} \quad \|\alpha\|_2 + c\,K(x+\alpha, y)$$
$$s.t \quad x+\alpha \in [0,1]$$

where :

$$K(x+\alpha, y) = max(f_{M(x)}(x+\alpha) - \max_{j \neq M(x)} f_j(x+\alpha), 0)$$

SPSA Attack

Principle (gradient-free, iterative, l_∞): a gradient-free attack

$$\min_{\alpha} \quad f_{M(x)}(x+\alpha) - \max_{j \neq M(x)} f_j(x')$$
$$s.t \ \|\alpha\|_\infty < \epsilon$$

Method:

Adam optimizer with discrete gradient approximation:

$$g'(x)_i \simeq \frac{(g(x+\delta v) - g(x-\delta v))v_i^{-1}}{2\delta}$$

with $v \sim \{-1, 1\}^d$

ZOO Attack

Principle (gradient-free, iterative, l_2): Gradient-free softmax version of the CWl2 attack

Method:
Adam optimizer with discrete gradient approximation:

$$g'(x)_i \simeq \frac{g(x + he_i) - g(x - he_i)}{2h}$$

with e_i vector with i^{th} component valued 1

Principle of Gradient Masking:
Make gradients useless to craft adversarial examples

Remarks:

- Defense through obscurity (Uesato, *Adversarial Risk and the Dangers of Evaluating Against Weak Attacks*, 2018)
- An adversary can use a substitute model to circumvent it
- Gradient-free attacks, decision-based attacks, ...

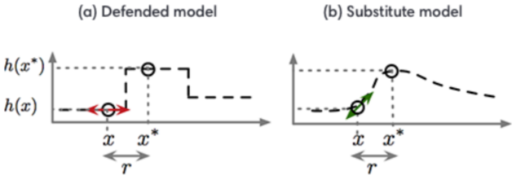


Figure: Goodfellow et al., *Attacking Machine Learning with Adversarial Examples*, openAI blog, 2017

Neural Networks Quantization

Several tools have been recently proposed to map full precision pre-trained models to quantized models for inference purpose:

- Android NN API
- TensorFlow Lite
- ARM-NN, CMSIS-NN
- STMCubeMX. A.I.

→ Coarsely quantizing (some) weights into – usually – no more than INT8.

More advanced methods propose clustering methods, information theoretical vector quantization methods...

Principle:
Learn a model with quantized weights and/or activation values during the training process

Issues:

- Manage non-differentiability issues of quantization function during backward pass
- Training can be difficult

Binary Neural Networks

SoA approaches: *Binary Net*, Courbariaux, Bengio et al. (2015 & 2016)

- weights and activations are binarized for the forward pass

$$w_b = sign(w),\ a_b^k = sign(a^k)$$

- Inference: only *bitcount* and *xnor* operations
- Binarization is not differentiable. Trick: use of a *Straight Through Estimator*: (STE, Bengio et al., 2013) at the backward pass

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial w_b} \frac{\partial w_b}{\partial w} \approx \frac{\partial L}{\partial w}\Big|_{w=w_b} \mathbf{1}_{|w| \leq 1}$$

Low bit-width Neural Networks

SoA approaches: *Dorefa Net*, Zhou et al. 2016

n-bit width quantization

- Train neural networks with low-bitwidth:
  1. weights
  2. activations
  3. gradients
- STE for the backward pass
- Inference: usage of a bit convolution kernel

Neural networks quantization

Quantization-aware training

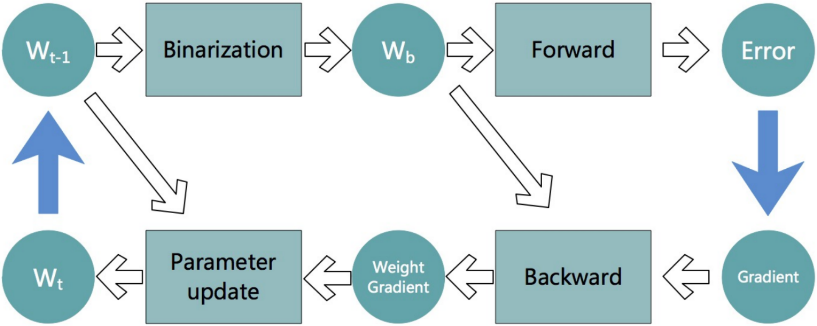


Figure: Guo et al., *A Survey on Methods and Theories of Quantized Neural Networks*, 2018

Previous Work

Massive research efforts on the topic (both attacks and defenses) with associated benchmarks and competitions (*NIPS Adversarial Vision Challenge*) **but almost only on full-precision models**.

Existing works bridging quantization and adversarial robustness:

- Galloway, 2017 (*Attacking binarized neural networks*): claims natural robustness with binarization. But, MNIST only, stochastic quantization
- Khalil, 2018 (*Combinatorial attacks on binarized networks*) $\rightarrow$ not scalable on big data sets
- Lin, 2019 (*Efficiency Meets Robustness*): FGSM attack only, white-box setting only (no transferability analysis)

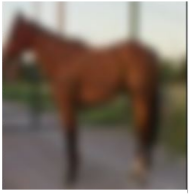
Experiments

Experiments

Setup

Data sets:

- SVHN (73,257/26,032)
- CIFAR10 (50,000/10,000)



Models:

One full-precision (32-bit float) model for each data set (same CNN architecture as in Courbariaux et al., 2016)

Quantized models :

- Activation and Weight / Weight *quantization*: 1,2,3,4 bits
- Techniques: Courbariaux et al. (2015, 2016), Zhou et al. (2016)

Computing environment:

- CPU: Intel Xeon, 2.1 GHz (12 cores)
- GPU: 2x NVIDIA GTX 1080 Ti (11 Gb, 3584 CUDA cores)

	CIFAR10				SVHN			
Full-precision	0.89				0.96			
Bitwidth	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Full quantization	0.79	0.87	0.88	0.88	0.89	0.95	0.95	0.95
Weight quantization	0.88	0.88	0.88	0.88	0.96	0.95	0.96	0.95

Table: Models accuracy on test set

During training, quantization acts as a:

- constraint
- regularizer

Experiments

Attacks and metrics

	FGSM	BIM	CWL2	SPSA	ZOO
Gradient-based	✓	✓	✓		
Gradient-free				✓	✓
one-step	✓				
iterative		✓	✓	✓	✓
$\boldsymbol{l_\infty}$	✓	✓		✓	
$\boldsymbol{l_2}$			✓		✓

Adversarial accuracy: accuracy of the model on adversarial examples

$\boldsymbol{l_p}$ adversarial **distortion**:

$$\left\|x' - x\right\|_p = \left(\sum_{i=1}^{m} |x_i' - x_i|^p\right)^{\frac{1}{p}}$$

	CIFAR10						SVHN					
	Float model *(32-bit)*			Binarized models *(1-bit)*			Float model *(32-bit)*			Binarized models *(1-bit)*		
	acc	l_2	l_∞	acc	l_2	l_∞	acc	l_2	l_∞	acc	l_2	l_∞
FGSM	0.12	1.65	0.03	**0.66**	1.65	0.03	0.29	1.66	0.03	**0.78**	1.64	0.03
BIM	0.07	1.17	0.03	**0.66**	1.01	0.03	0.05	1.16	0.03	**0.79**	1.0	0.03
CWl2	0.03	0.58	0.04	**0.11**	0.78	0.08	0.02	0.64	0.66	**0.06**	1.02	0.1

1) Fully binarized neural networks:

- Apparent robustness against FGSM and BIM attacks
- No robustness increase against CWl2 attack

$\rightarrow$ No additional robustness against gradient based attacks

Experiments

Direct attacks, fully-quantized models

	CIFAR10						SVHN					
	Float model *(32-bit)*			Quantized models *(1,2,3,4-bit)*			Float model *(32-bit)*			Quantized models *(1,2,3,4-bit)*		
	acc	l_2	l_∞	acc	l_2	l_∞	acc	l_2	l_∞	acc	l_2	l_∞
BIM	0.07	1.17	0.03	**0.66**	1.01	0.03	0.05	1.16	0.03	**0.79**	1.0	0.03
				0.06	1.14	0.03				**0.11**	1.13	0.03
				0.11	1.17	0.03				**0.11**	1.13	0.03
				0.06	1.14	0.03				**0.1**	1.13	0.03
SPSA	0.0	1.37	0.03	**0.16**	1.31	0.03	0.01	1.38	0.03	**0.4**	1.32	0.03
				0.0	1.34	0.03				**0.14**	1.34	0.03
				0.0	1.36	0.03				**0.07**	1.35	0.03
				0.0	1.36	0.03				**0.04**	1.37	0.03

2) Fully quantized neural networks:
BIM (gradient-based, l_∞) less efficient than SPSA (gradient-free, l_∞)
$\rightarrow$ Gradient masking

Experiments

Direct attacks, fully-quantized models

	CIFAR10						SVHN					
	Float model (*32-bit*)			Quantized models (*1,2,3,4-bit*)			Float model (*32-bit*)			Quantized models (*1,2,3,4-bit*)		
	acc	l_2	l_∞	acc	l_2	l_∞	acc	l_2	l_∞	acc	l_2	l_∞
CWl2	0.03	0.58	0.04	**0.11**	**0.78**	0.08	0.02	0.64	0.06	**0.06**	**1.02**	0.1
				0.06	**0.6**	0.04				**0.03**	**0.67**	0.07
				0.09	**0.55**	0.04				**0.02**	**0.66**	0.07
				0.05	**0.5**	0.04				**0.02**	**0.68**	0.07
ZOO	0.0	0.72	0.09	**0.56**	**0.1**	0.05	0.0	0.91	0.11	**0.82**	**0.07**	0.05
				0.83	**0.13**	0.06				**0.93**	**0.1**	0.06
				0.76	**0.24**	0.07				**0.94**	**0.11**	0.05
				0.73	**1.09**	0.14				**0.93**	**0.38**	0.1

3) Fully quantized neural networks:

- Quantization alters ZOO objective function ($\simeq 0$ or $>> 1$)
  $\rightarrow$ ZOO fails, CWl2 succeeds (thanks to STE)
- No effect from quantization
  $\rightarrow$ ZOO performs better (l_2 distortion)

$\rightarrow$ Gradient masking

Poor transferability capacities

	float	w_1a_1	w_1a_{32}	w_2a_2	w_2a_{32}	w_3a_3	w_3a_{32}	w_4a_4	w_4a_{32}
float	0.12	0.58	0.38	0.41	0.42	0.40	0.40	0.40	0.39
w_1a_1	0.82	0.66	0.82	0.81	0.81	0.72	0.81	0.80	0.80
w_1a_{32}	0.33	0.61	0.11	0.37	0.38	0.42	0.37	0.36	0.36
w_3a_3	0.49	0.62	0.51	0.43	0.44	0.17	0.44	0.43	0.43
w_3a_{32}	0.46	0.61	0.47	0.38	0.39	0.45	0.18	0.38	0.38

FGSM

	float	w_1a_1	w_1a_{32}	w_2a_2	w_2a_{32}	w_3a_3	w_3a_{32}	w_4a_4	w_4a_{32}
float	0.07	0.64	0.28	0.32	0.33	0.38	0.29	0.26	0.27
w_1a_1	0.88	0.66	0.86	0.86	0.86	0.79	0.86	0.86	0.85
w_1a_{32}	0.20	0.64	0.07	0.27	0.28	0.37	0.24	0.23	0.22
w_3a_3	0.38	0.67	0.46	0.35	0.37	0.11	0.34	0.33	0.33
w_3a_{32}	0.30	0.70	0.38	0.24	0.26	0.40	0.06	0.20	0.23

BIM

	float	w_1a_1	w_1a_{32}	w_2a_2	w_2a_{32}	w_3a_3	w_3a_{32}	w_4a_4	w_4a_{32}
float	0.06	0.64	0.35	0.38	0.40	0.41	0.34	0.32	0.33
w_1a_1	0.77	0.11	0.75	0.71	0.74	0.63	0.72	0.71	0.73
w_1a_{32}	0.14	0.58	0.07	0.23	0.24	0.26	0.21	0.20	0.21
w_3a_3	0.25	0.56	0.29	0.19	0.20	0.10	0.18	0.18	0.18
w_3a_{32}	0.39	0.61	0.44	0.27	0.30	0.36	0.06	0.24	0.26

CWl2

	float	w_1a_1	w_1a_{32}	w_2a_2	w_2a_{32}	w_3a_3	w_3a_{32}	w_4a_4	w_4a_{32}
float	0.00	0.64	0.29	0.28	0.37	0.38	0.22	0.23	0.22
w_1a_1	0.81	0.16	0.78	0.79	0.77	0.79	0.83	0.79	0.77
w_1a_{32}	0.29	0.68	0.00	0.34	0.38	0.41	0.24	0.30	0.33
w_3a_3	0.42	0.67	0.56	0.39	0.44	0.00	0.36	0.34	0.41
w_3a_{32}	0.32	0.68	0.42	0.27	0.30	0.40	0.00	0.26	0.30

SPSA

Quantization Shift Phenomenon: Quantization ruins the adversarial effect

- activation shift:
  $a_1^j(x') > a_2^j(x') \rightarrow a_1^{j,\mathbf{q}}(x') = a_2^{j,\mathbf{q}}(x')$

- weight shift:
  $f_1^j(x, w_1) > f_2^j(x, w_1),\ f_1^j(x, w_1^{\mathbf{q}}) > f_2^j(x, w_1^{\mathbf{q}})$
  $f_1^j(\mathbf{x'}, w_1) < f_2^j(\mathbf{x'}, w_1) \rightarrow f_1^j(\mathbf{x'}, w_1^{\mathbf{q}}) = f_2^j(\mathbf{x'}, w_1^{\mathbf{q}})$

  Weight quantization can cancel adversarial effect

Gradient misalignment:

Cosinus similarity values near 0 $\rightarrow$ near orthogonal gradients

Hard to transfer from/to fully binarized networks

Conclusion and future work

Conclusion and future work

Complete study of quantized models vulnerabilities against adversarial examples, under various threat models

Take-away:

- Detection of some gradient masking issue
  $\rightarrow$ Quantization is not a robust *"natural"* defense when facing advanced attacks
- But, interestingly, gradient misalignment issues and *quantization shift phenomenon* cause poor transferability

Ongoing works:

- Develop an ensemble/quantization based defense exploiting low transferability