Adversarial Attacks and Robustness of DNN

Béguinot Julien
Monbroussou Léo
julien.beguiot@enst.fr
leo.monbroussou@ens-paris-saclay.fr
02/18/2022

Motivations and Context

- Neural Networks are powerful function approximators though they are prone to robustness issues.
- **Data Poisoning** is a type of adversarial attacks modifying the training set in order to modify the trained Neural Network.
- It is crucial to handle it for critical systems (e.g. autonomous car, voice recognition) and to design **provable defense**.

Different Type of Attacks

- General poisoning (remove and insert training samples) ;
- Label Flipping ;
- Input Perturbation.

General Poisoining

Definition (General Poisoining Attacks)

The **attacker** can **insert or remove** a bounded number of samples from the **training set**. In particular, the **attack magnitude** ρ is defined as the cardinality of the symmetric difference between clean and poisonned training sets.

Recall that the symmetric difference $A \ominus B = (A \backslash B) \cup (B \backslash A)$

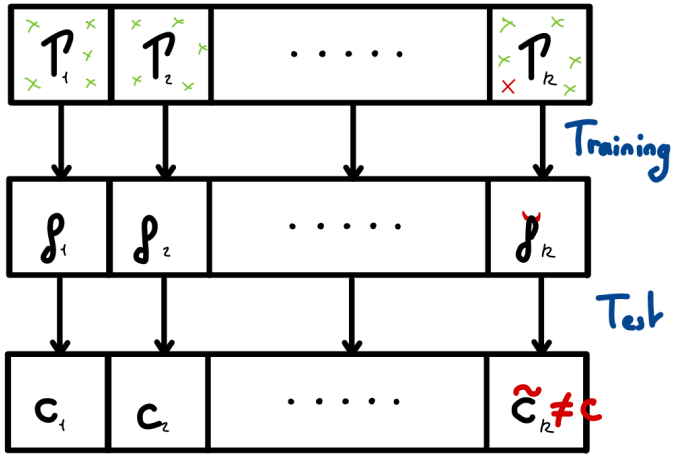
Deep Partition Aggregation

- Type Ensemble methods
- Essentialy with L_0 norm

Notations

- Base classifier model : $f : P(S_L) \times S \to \mathbb{N}$
- Labeled samples : $S_L := \{(x, c) | x \in S, c \in \mathbb{N}\}$
- Training set : $T \in P(S_L)$
- Deterministic hash function : $h : S_L \to \mathbb{N}$
- Hyperparameter : $k \in \mathbb{N}$

The gist of it



Deep Partition Aggregation

At training time, the algo define partitions $P_1, ..., P_k \subseteq T$

$$P_i := \{t \in T | h(t) \equiv i \pmod{k}\}$$

Base classifiers trained on each partition :

$$f_i : S \to \mathbb{N}$$

$$f_i(x) := f(P_i, x)$$

At inference time, we evaluate the input on each base classification, and then count the nbr of classifiers which return each class :
$n_c(x) := |\{i \in [k] | f_i(x) = c\}|$
Classifiers which returns the consensus output of the ensemble :
$g_{dpa}(T, x) := arg\ \max_c n_c(x)$

Deep Partition Aggregation

Theorem (Levine al. 2021 [5])

For fixed deterministic base classifier f, hash function h, ensemble size k, training set T, and input x, let :

$$c := g_{dpa}(T, x) := arg\max_{c} n_c(x)$$

$$\bar{\rho}(x) := \left\lfloor \frac{n_c - \max_{c' \neq G}(n_{c'}(x) - \mathbf{1}_{c' < c})}{2} \right\rfloor$$

Then, for any poisoned training set U, if $|T \ominus U| \leq \bar{\rho}(x)$, then $g_{dpa}(U, x) = c$.

SS Deep Partition Aggregation

- A more specific **semi-supervised approach** for **Label Flipping Attacks**
- The base classifier use their partition and the unlabelled samples from the other partitions for its training (typically for feature extraction)
- Ensure better accuracy and higher robustness bound with deacreasing $n_{c'}$

SS Deep Partition Aggregation

Theorem (Levine al. 2021 [5])

For a fixed deterministic semi-supervised base classifier f ensemble size k, training set T (with no repeated samples), and input x, let :

$$c = g_{ssdpa}(T, x)$$

$$\bar{\rho}(x) = \left\lfloor \frac{n_c - \max_{c' neq c}(n_{c'}(x) + \mathbf{1}_{c'<c})}{2} \right\rfloor$$

For a poisoned training set U obtained by changing the labels of at most $\bar{\rho}$ samples in T,

$$g_{ssdpa}(U, x) = c$$

Experimental Validation [5]

	Training set size	Number of Partitions k	Median Certified Robustness	Clean Accuracy	Base Classifier Accuracy	Training time per Partition
MNIST, DPA	60000	1200	448	95.85%	76.97%	0.33 min
		3000	509	93.36%	49.54%	0.27 min
MNIST, SS-DPA	60000	1200	485	95.62%	80.77%	0.15 min
		3000	645	93.90%	57.65%	0.16 min
CIFAR, DPA	50000	50	9	70.16%	56.39%	1.49 min
		250	5	55.65%	35.17%	0.58 min
		1000	N/A	44.52%	23.20%	0.30 min
CIFAR, SS-DPA	50000	50	25	90.89%	89.06%	0.94 min
		250	124	90.33%	86.25%	0.43 min
		1000	392	89.02%	75.83%	0.33 min
GTSRB, DPA	39209	50	20	89.20%	73.94%	2.64 min
		100	4	55.90%	35.64%	1.60 min
GTSRB, SS-DPA	39209	50	25	97.09%	96.35%	2.73 min
		100	50	96.76%	94.96%	1.56 min
		200	99	96.34%	91.54%	1.23 min
		400	176	95.80%	83.60%	0.78 min

Table 1: Summary statistics for DPA and SS-DPA algorithms on MNIST, CIFAR, and GTSRB.

Limitation of Deep Partition Aggregation

- Computationally expensive
- Trade-off between performance and robustness
- Count modification as one removal and one insertion without considering the amplitude of the modification
- **Idea :** Change the "repetition code" for a code with better rate.

Lets Flex

ACKNOWLEDGMENT

The authors acknowledge comments from Julien Béguinot (Télécom Paris - Institut Polytechnique de Paris) and Léo Monbroussou (École Normale Supérieure Paris-Saclay - Institut Polytechnique de Paris) to update the stated assumptions in earlier versions of the draft.

Reminder on Dissipativity (Aquino 2021) [1]

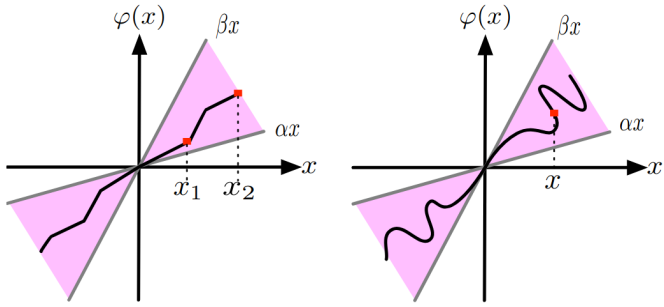
Definition

A discrete time system is said *QSR* **dissipative** if for every input/output pair (x, y) then $0 \leq s(x, y) = y^{\perp} Q y + x^{\perp} R x + 2x^{\perp} S y$ where s is called the **supply rate**.

- If $Q = 0, R = 0, S = \frac{1}{2} I$ then it is said **passive**
- If $Q = -\delta I, R = -\nu I, S = \frac{1}{2} I$, $\nu, \delta > 0$ then it is said **stricly passive**
- If $Q = -\delta I, R = -\nu I, S = \frac{1}{2} I$, we say it is **sector bounded** with slopes $\frac{1 + \pm\sqrt{1 - 4\nu\delta}}{2\delta}$ with **passivity indexes** $\nu, \delta \in \mathbb{R}$

These properties are said **incremental** if they apply to the increments.

Slope Restricted and Sector Bounded (Fazlyab 2021) [2]



$$\alpha \leq \frac{\varphi(x_2) - \varphi(x_1)}{x_2 - x_1} \leq \beta \qquad \alpha \leq \frac{\varphi(x)}{x} \leq \beta$$

Fig. 2: A slope-restricted nonlinearity (left) and a sector-bounded nonlinearity (right).

Sector boundness and slope restriction of common activation functions

Theorem (Fazlyab 2021 [2])

- *The ReLU function* $\phi(x) = \max(0, x), x \in \mathbb{R}$ *is slope- restricted and sector-bounded in* $[0, 1]$.
- *The sigmoid function,* $\phi(x) = 1 + \exp(-x), x \in \mathbb{R}$ *is slope-restricted in* $[0, 1]$.
- *The tanh function,* $\phi(x) = tanh(x), x \in \mathbb{R}$ *is slope- restricted and sector-bounded in* $[0, 1]$.
- *The leaky ReLU function,* $\phi(x) = \max(ax, x), x \in \mathbb{R}$ *with* $a > 0$ *is slope-restricted and sector-bounded in* $[\min(a, 1), \max(a, 1)]$.

Sector boundness and slope restriction of common activation functions

Theorem

- *The exponential linear function (ELU),* $\phi(x) = \max(x, a(\exp(x) - 1)), x \in \mathbb{R}$ *with* $a > 0$ *is slope-restricted and sector-bounded in* $[0, 1]$.
- *The softmax function,* $\phi(x) = [\exp(x_1), \ldots, \exp(x_n)] / \sum \exp(x_i), x \in \mathbb{R}^n$ *is slope-restricted in* $[0, 1]$.

Robustness against Adversarial Attacks in Neural Networks using Incremental Dissipativity

- **Objective :** Control $\|f(x+\delta x) - f(x)\|$ with $\|\delta x\|$.
- **Idea :** If f is incrementally sector bounded with positive output strict positivity index then the error is upper bounded by $\gamma\|x\|$, $\gamma > 0$.
  - Generalize Lipschitz constraint
  - scales to L layers neural networks
  - insight on the efficiency of the spectral normalization

A useful Lemma for repeted non Linearities

Theorem (Fazlyab 2021 [2] , Fazlyab 2019 [3])

*Let ϕ be real-valued **slope-restricted** in the sector $[\alpha, \beta]$. Then the vector-valued function $\Phi(x) = [\phi(x_1), \ldots, \phi(x_n)]$ satisfies the QC*

$$\begin{bmatrix} x \\ \Phi(x) \\ 1 \end{bmatrix}^{\perp} \begin{bmatrix} -2\alpha\beta\Lambda & (\alpha+\beta)\Lambda & 0 \\ (\alpha+\beta)\Lambda & -2\Lambda & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ \Phi(x) \\ 1 \end{bmatrix} \geq 0$$

for all $x \in \mathbb{R}^n$, where, $(\lambda_{ij} \geq 0)$

$$\Lambda = \sum \lambda_{ii} e_i e_i^{\perp} + \sum_{i<j} \lambda_{ij} (e_i - e_j)(e_i - e_j)^{\perp}$$

S-Procedure

Lemma

Let F_1, F_2 symmetric matrices, g_1, g_2 vectors and h, h_2 reals. Then the following proposition are equivalent.

- $$x^\perp F_1 x + 2g_1^\perp x + h_1 \leq 0 \implies x^\perp F_2 x + 2g_2^\perp x + h_2 \leq 0$$
- $$\exists \lambda \geq 0, \lambda \begin{pmatrix} F_1 & g_1 \\ g_1^\perp & h_1 \end{pmatrix} - \begin{pmatrix} F_2 & g_2 \\ g_2^\perp & h_2 \end{pmatrix} \succeq 0$$

QSR disspative Layer

Theorem (Aquino al. 2021[1])

Consider a ***non-convolutional neural network layer*** *defined by* $y = \phi(Wx + b)$ *where* ϕ *is* $[\alpha, \beta]$ ***incrementally sector bounded****. Let* $m = \frac{\alpha+\beta}{2}$ *and* $p = \alpha\beta$*. The layer is incrementaly QSR dissipative if*

$$\begin{pmatrix} Q & S \\ S^{\perp} & R \end{pmatrix} + \begin{pmatrix} \Lambda & -m\Lambda W \\ -mW^{\perp}\Lambda & pW^{\perp}\Lambda W \end{pmatrix} \succeq 0$$

is feasible for some Λ *with*

$$\Lambda = \sum_{1 \leq i \leq n} \lambda_{ii} e_i e_i^{\perp} + \sum_{1 \leq i,j \leq n} \lambda_{ij} (e_i - e_j)(e_i - e_j)^{\perp}$$

And for Convolutional Layers ?

Theorem (Aquino al. 2021[1])

*Consider a **convolutional neural network layer** in which the input is convolved with filter F, vectorized and transmitted through an element-wise non-linear **activation function** ϕ that is **incrementally sector bounded** by $[\alpha, \beta]$. Let $m = \frac{\alpha+\beta}{2}$ and $p = \alpha\beta$. The layer is incrementaly QSR dissipative if*

$$M = \begin{pmatrix} Q & S \\ S^{\perp} & R \end{pmatrix} + \begin{pmatrix} \Lambda & -m\Lambda C \\ -mC^{\perp}\Lambda & pC^{\perp}\Lambda C \end{pmatrix} \succeq 0$$

is feasible for some Λ Where C is the circulant matrix associated to the convolution filter F.

Cascade : Scaling to Deep Neural Network

Theorem (Aquino al. 2021[1])

*Consider a neural network with n layers where **each layer** i **is** QSR **dissipative** with $Q_i = -\delta_i I$, $R = -\nu_i I$ and $S = \frac{1}{2} I$. Then the neural network is **incrementaly sector bounded** with parameters $Q = -\delta I$,$R = -\nu I$ and $S = \frac{1}{2} I$ if $0 \succeq A$ where*

$$A = \begin{bmatrix} -\nu_1 + \nu & \frac{1}{2} & 0 & \dots & 0 & \frac{-1}{2} \\ \frac{1}{2} & -\nu_2 + \delta_1 & \frac{1}{2} & 0 & \dots & 0 \\ 0 & \frac{1}{2} & -\nu_3 + \delta_2 & \frac{1}{2} & 0 \dots & \dots \\ 0 & & & & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & -\nu_n + \delta_{n-1} & \frac{1}{2} \\ -\frac{1}{2} & 0 & \dots & 0 & \frac{1}{2} & -\delta_n + \delta \end{bmatrix}$$

Sufficient (computationally efficient) conditions for the first theorem

Theorem (Aquino al. 2021[1])

If

- $$\|W\|_2 \leq \frac{1}{|m|}\left(1 - \frac{(\delta + \frac{1}{2}) + (1 - |p|\|W\|_i^2)}{\delta - \nu}\right)$$
- $$|p|\|W\|_i \leq 1$$

Then the system of theorem 1 holds with

$$\|\Lambda\|_2 = \|\Lambda\|_i \geq \frac{-\nu}{1 - |p|\|W\|_i^2}$$

Proof of sufficient conditions for the first theorem

Démonstration.

A sufficient condition for the LMI is that matrix M be block diagonally dominant with diagonal blocks positive semi-definite which is equivalent to the following property :

$$\Lambda + Q \succ 0$$

$$W^T p \Lambda W + R \succ 0$$

$$||(Q + \Lambda)^{-1}(S - m\Lambda W)||_2 \leq 1$$

$$||(W^T p \Lambda W + R)^{-1}(S^T - m W^T \Lambda)||_2 \leq 1$$

A Nice Generalization of Gershgorin Theorem

Definition

Let $A = (A_{ij})$ be a block partioned matrix. If the diagonal submatrices A_{jj} are non-singular, and

$$\|A_{jj}^{-1}\|^{-1} \geq \sum_{k \neq j} \|A_{j,k}\|$$

then A is said **block diagonaly dominant**. If strict inequality holds for all the j then it is said **block stricly diagonaly dominant**.

Theorem (David G.Feingold and Richard S.Verga [4])

If a matrix is block strictly diagonally dominant then it is non singular.

References I

- Bernardo Aquino, Arash Rahnama, Peter Seiler, Lizhen Lin, and Vijay Gupta.
  Robustness against adversarial attacks in neural networks using incremental dissipativity, 2021.
- Mahyar Fazlyab, Manfred Morari, and George J. Pappas.
  Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming, 2021.
- Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas.
  Efficient and accurate estimation of lipschitz constants for deep neural networks, 2019.

References II

David Feingold and Richard Varga.
Block diagonally dominant matrices and generalizations of the gershgorin theorem.
Pac. J. Math., 12, 12 1962.

Alexander Levine and Soheil Feizi.
Deep partition aggregation : Provable defense against general poisoning attacks, 2021.