Difference-of-Convex Algorithm applied to adversarial robustness verification

Ismaila Seck [1,2,3] Gaelle Loosli [3,4]
Stephane Canu [2,3] Yi-Shuai Niu [5]

[1]Normandie Univ, INSA Rouen [2]UNIROUEN, UNIHAVRE, LITIS [3]UCA, LIMOS

[4]PobRun [5]School of Mathematical Sciences, Shanghai Jiao Tong University

October 2, 2019

Adversarial examples

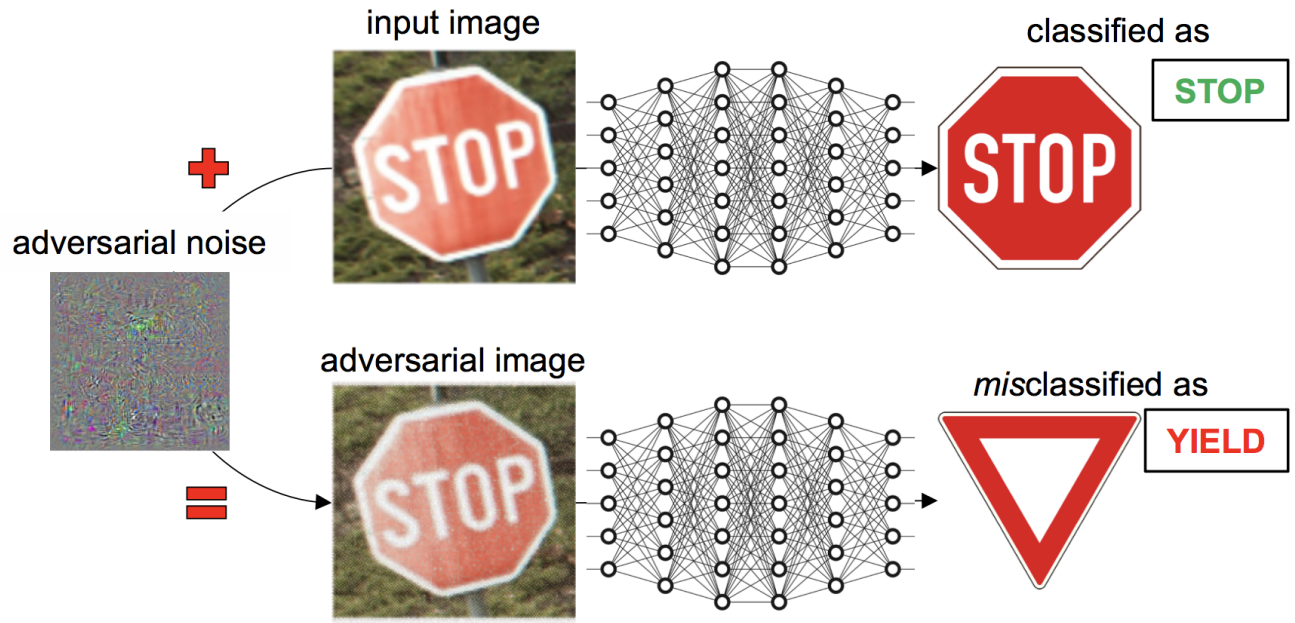

Figure 1: Illustration of the use of adversarial examples.

Formulation as an optimization problem

- $\mathbf{x}, y$: original image and its class
- $\mathbf{x}'$: adversarial image we are looking for
- $f_k(.)$: the k-th output of the network

$$\begin{cases} \min & \|\mathbf{x} - \mathbf{x}'\| \\ \text{s.t.} & \underset{k=1,\dots,c}{\operatorname{argmax}} f_k(\mathbf{x}') \neq y, \\ & \mathbf{x}' \in [0,1]^d. \end{cases} \tag{1}$$

Formulation as an optimization problem

- $\mathbf{x}, y$: original image and its class
- $\mathbf{x}'$: adversarial image we are looking for
- $f_k(.)$: the k-th output of the network

$$\begin{cases} \min & \|\mathbf{x} - \mathbf{x}'\| \\ \text{s.t.} & \underset{k=1,\dots,c}{\text{argmax}} f_k(\mathbf{x}') \neq y, \\ & \mathbf{x}' \in [0,1]^d. \end{cases} \tag{1}$$

Linearization of the argmax constraint

$$(1) \iff \begin{cases} \min & \|\mathbf{x} - \mathbf{x}'\| \\ \text{s.t.} & m \leq f_k(\mathbf{x}') + (1 - \mathrm{a}_k) M_m, \quad k = 1, \ldots, c \\ & m \geq f_k(\mathbf{x}'), \quad k = 1, \ldots, c \\ & \sum_{k=1}^{c} \mathrm{a}_k = 1, \\ & \mathrm{a}_y = 0, \\ & m \in \mathbb{R}, \\ & \mathrm{a} \in \{0,1\}^c, \\ & \mathbf{x}' \in [0,1]^d. \end{cases} \tag{2}$$

Using DC to get rid of the binary variables

$$
\begin{cases}
\min & \|\mathbf{x} - \mathbf{x}'\| + \sum_{k=1}^{c} \mathrm{a}_k(1 - \mathrm{a}_k) & \\
\text{s.t.} & m \leq f_k(\mathbf{x}') + (1 - \mathrm{a}_k)M_m, & k = 1, \ldots, c \\
& m \geq f_k(\mathbf{x}'), & k = 1, \ldots, c \\
& \sum_{k=1}^{c} \mathrm{a}_k = 1, & \\
& \mathrm{a}_y = 0, & \\
& m \in \mathbb{R}, & \\
& \mathrm{a} \in [0,1]^c, & \\
& \mathbf{x}' \in [0,1]^d. &
\end{cases} \tag{3}
$$

The End

Thanks for your attention !