Decision boundaries & security related questions (for classifiers)

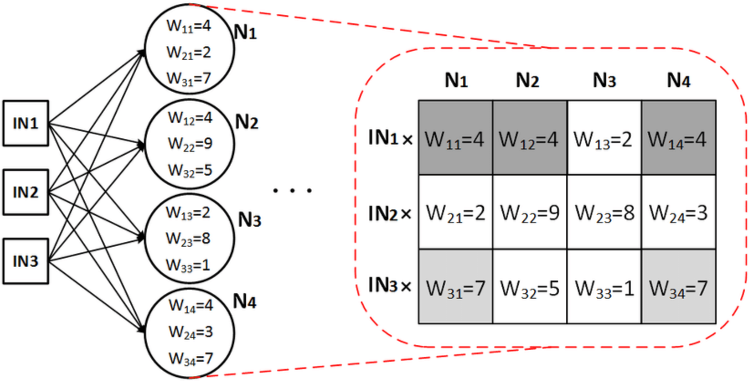
Joint works with:

A. Jaouen

P. Pérez (Valeo.ai)

G. Trédan (CNRS)

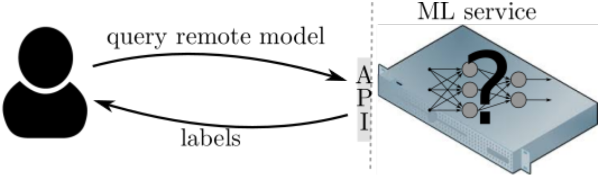
A neural network model: for its designer



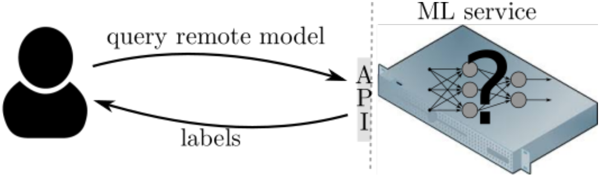
Architecture + weights

Full access: *white box* setup

This talk: observer/attacker perspective

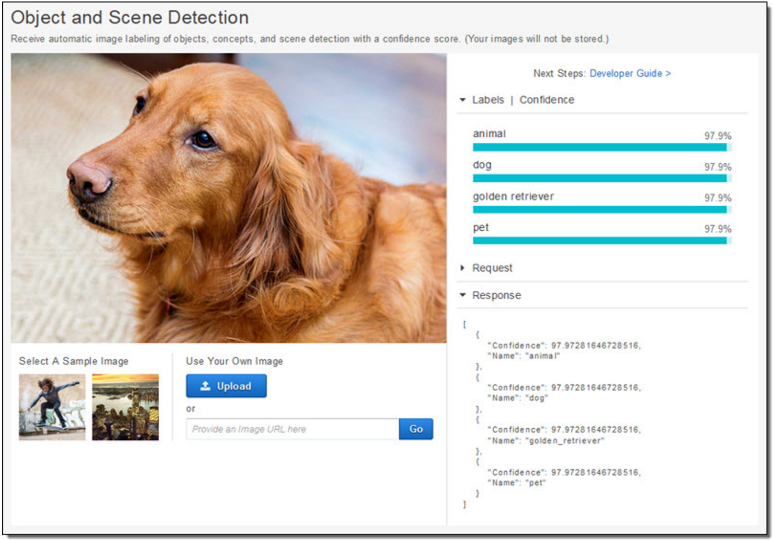


This talk: observer/attacker perspective



An oracle / *black box*

Classification API



Classification API

Or your local device with an embedded model

Black box interaction

Let $\mathcal{M} : \mathbb{R}^d \to C$ be a classifier model.

Definition (Black-box model observation)

The observer queries the black box model $\mathcal{M}$ with arbitrary inputs $x \in X$, and gets in return $\mathcal{M}(x) \to \{y \in C; v[C_0, C_1, \dots C_{n-1}]\}$.

Here, no access to weights $\implies$ no gradients

Our recent work / this talk

Unify questions related to boundaries of black box models.

Outline:

- Preliminary notions
- Watermarking models
- A score for input safety

Boundary shapes?

- Goodfellow at al. attack: $x^* = x + \epsilon\ .\ sign(\bigtriangledown_{\vec{x}} J_h(\theta, x, y))$

> **Our take-away 5.1.** *Models often extrapolate linearly from the limited subspace covered by the training data [43]. Algorithms can exploit this regularity in directing search toward prospective adversarial regions.*

(Euro. S&P 2018)

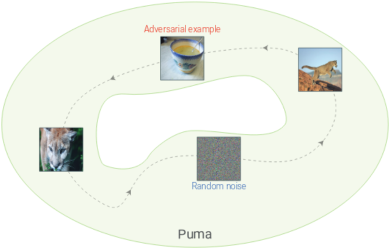
Boundary shapes?

- Goodfellow at al. attack: $x^* = x + \epsilon \; . \; sign(\bigtriangledown_{\vec{x}} J_h(\theta, x, y))$

> **Our take-away 5.1.** *Models often extrapolate linearly from the limited subspace covered by the training data [43]. Algorithms can exploit this regularity in directing search toward prospective adversarial regions.*

(Euro. S&P 2018)

- Fawzi et al. 2017: "classification regions are connected"

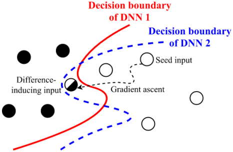


Decision boundary

Definition (Input on decision boundary (Lee and Landgrebe 1997))

Given two classes C_i and C_j, an input x is on the decision boundary between those two classes if $p(C_i|x) - p(C_j|x) = 0$.

Boundary: $\bigcup_{x \in X} \; s.t. \; p(C_i|x) - p(C_j|x) = 0.$

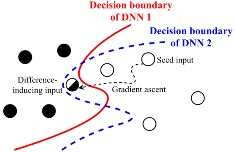


Decision boundary

Definition (Input on decision boundary (Lee and Landgrebe 1997))

Given two classes C_i and C_j, an input x is on the decision boundary between those two classes if $p(C_i|x) - p(C_j|x) = 0$.

Boundary: $\bigcup_{x \in X} \; s.t. \; p(C_i|x) - p(C_j|x) = 0.$



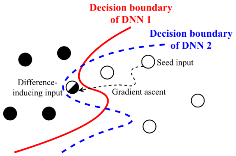
No access to probabilities v: ϵ modification of x $s.t.$ $\mathcal{M}(x \pm \epsilon) \neq \mathcal{M}(x)$.

Decision boundary

Definition (Input on decision boundary (Lee and Landgrebe 1997))

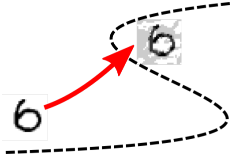
Given two classes C_i and C_j, an input x is on the decision boundary between those two classes if $p(C_i|x) - p(C_j|x) = 0$.

Boundary: $\bigcup_{x \in X} \; s.t. \; p(C_i|x) - p(C_j|x) = 0.$



No access to probabilities v: ϵ modification of x $s.t.$ $\mathcal{M}(x \pm \epsilon) \neq \mathcal{M}(x)$.

How to get nearby boundaries in practice: leveraging *adversarial examples*



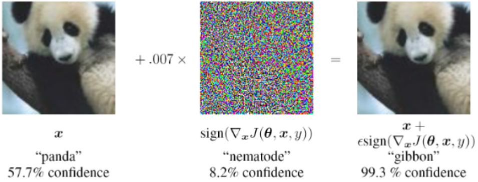
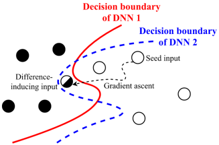


Figure 1: An adversarial image generated by *Fast Gradient Sign Method* [55]

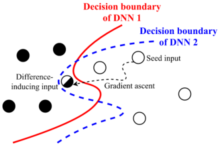
What are distinguishable models?



Definition (Undistinguishable models)

Two models $\mathcal{M}$ and $\mathcal{M}'$ are indistinguishable for an observer if $\nexists\, x \in X\ s.t.\ \mathcal{M}(x) \neq \mathcal{M}'(x)$.

What are distinguishable models?



Definition (Undistinguishable models)

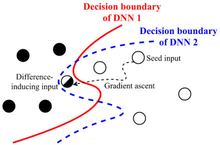
Two models $\mathcal{M}$ and $\mathcal{M}'$ are indistinguishable for an observer if $\nexists\ x \in X\ s.t.\ \mathcal{M}(x) \neq \mathcal{M}'(x)$.

Definition ($\mathcal{M}_{set}$ fingerprint)

Given a finite set of models $\mathcal{M}_{set} = \{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{n-1}\}$, a *fingerprint* uniquely identifies one and only one model among the n models in $\mathcal{M}_{set}$.

A fingerprint: a set of $< input, label >$ examples, often at the boundary.

What are distinguishable models?



Definition (Undistinguishable models)

Two models $\mathcal{M}$ and $\mathcal{M}'$ are indistinguishable for an observer if $\nexists\, x \in X \; s.t. \; \mathcal{M}(x) \neq \mathcal{M}'(x)$.

Definition ($\mathcal{M}_{set}$ fingerprint)

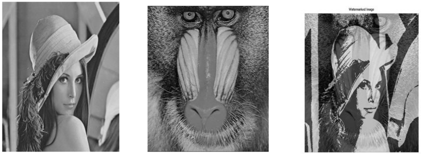
Given a finite set of models $\mathcal{M}_{set} = \{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{n-1}\}$, a *fingerprint* uniquely identifies one and only one model among the n models in $\mathcal{M}_{set}$.

A fingerprint: a set of $< input, label >$ examples, often at the boundary.
First leak of information about the black box: which model is in use.

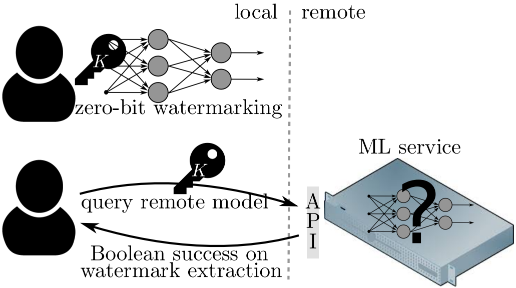
Watermarking deep models

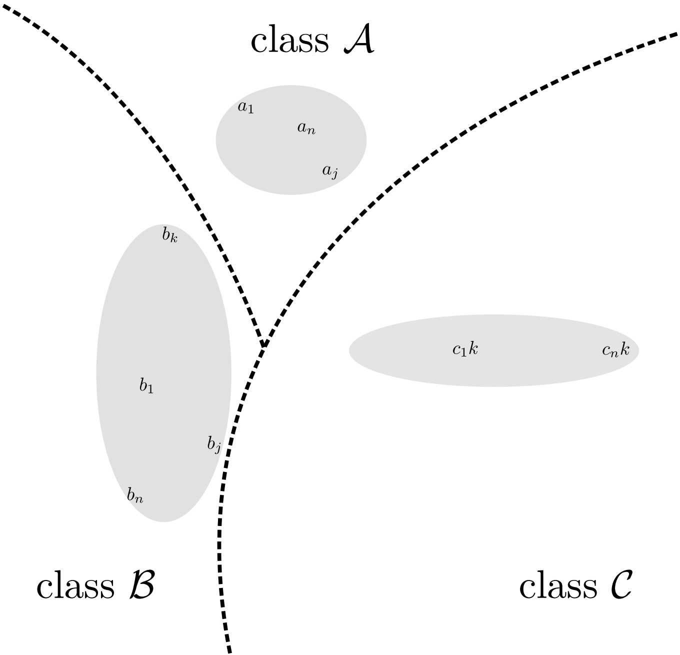
Protecting models from physical copy: watermarking

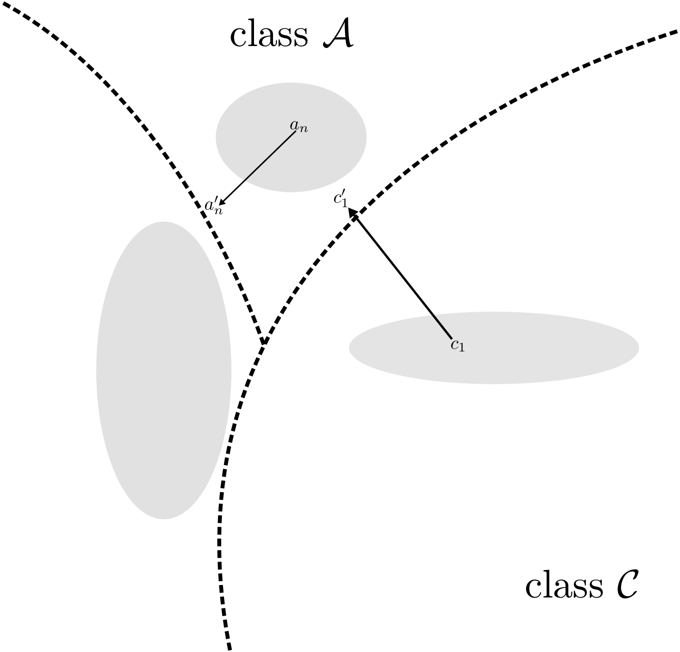
Example : Digital watermarking



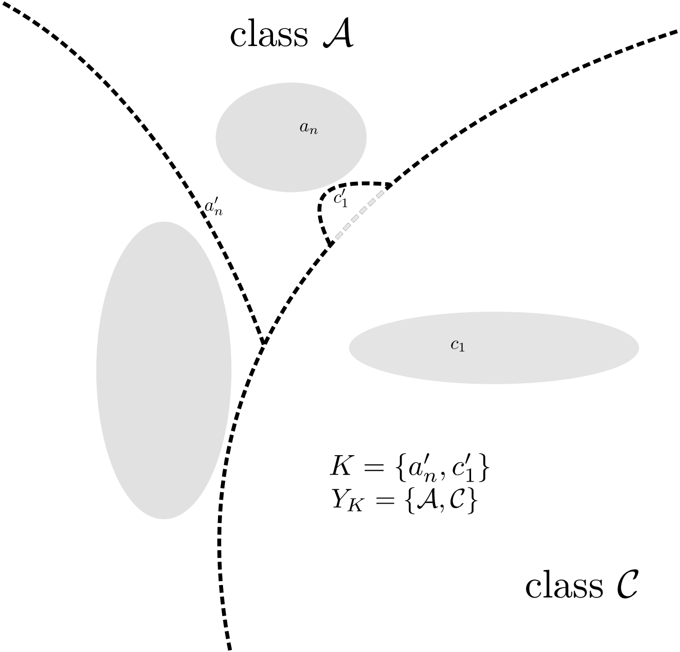
Original image Watermark Digital watermarking



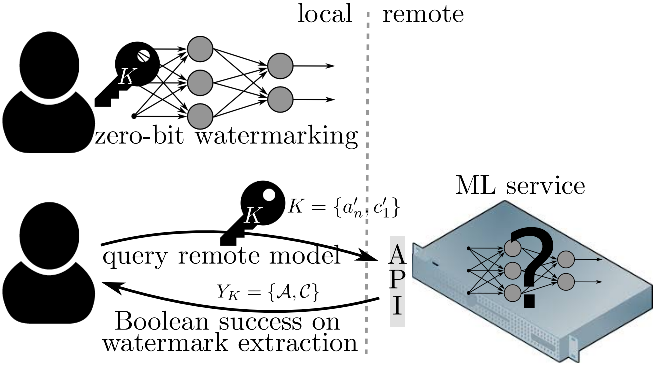

class $\mathcal{A}$
a_1
a_n
a_j
b_k
b_1
b_j
b_n
c_1k
c_nk
class $\mathcal{B}$
class $\mathcal{C}$



class $\mathcal{A}$
a_n
a'_n
c'_1
c_1
class $\mathcal{C}$


class $\mathcal{A}$
a_n
a'_n
c'_1
c_1
$K = \{a'_n, c'_1\}$
$Y_K = \{\mathcal{A}, \mathcal{C}\}$
class $\mathcal{C}$

After watermarking, query if suspected copy



Our watermarked model in the black box?

- Model unchanged (unlikely!): simply query with K
  - if $\mathcal{M}(K) \rightarrow Y_k$, extraction is successful

Our watermarked model in the black box?

- Model unchanged (unlikely!): simply query with K
  - if $\mathcal{M}(K) \to Y_k$, extraction is successful
- Model may have been tampered with, use p-value argument:
  - Null-model $\mathcal{M}_\emptyset$: $\forall x \in K,\ \mathbb{P}[\mathcal{M}_\emptyset(x) = \mathcal{M}(x)] = 1/2$
  - $Z = m_K(\mathcal{M}, \mathcal{M}_\emptyset)$: rd var. of the number of mismatches
  - Exactly z errors: $\mathbb{P}[Z = z | \mathcal{M} = \mathcal{M}_\emptyset] = 2^{-|K|}\binom{|K|}{z}$
  - Rejecting null-model: $\mathbb{P}[Z \leq \theta | \mathcal{M} = \mathcal{M}_\emptyset] < 0.05$:

$$2^{-|K|} \sum_{z=0}^{\theta} \binom{|K|}{z} < 0.05$$

  - e.g., $|K| = 100$, max errors tolerated is $\theta = 42$

Our watermarked model in the black box?

- Model unchanged (unlikely!): simply query with K
  - if $\mathcal{M}(K) \rightarrow Y_k$, extraction is successful
- Model may have been tampered with, use p-value argument:
  - Null-model $\mathcal{M}_\emptyset$: $\forall x \in K,\ \mathbb{P}[\mathcal{M}_\emptyset(x) = \mathcal{M}(x)] = 1/2$
  - $Z = m_K(\mathcal{M}, \mathcal{M}_\emptyset)$: rd var. of the number of mismatches
  - Exactly z errors: $\mathbb{P}[Z = z | \mathcal{M} = \mathcal{M}_\emptyset] = 2^{-|K|}\binom{|K|}{z}$
  - Rejecting null-model: $\mathbb{P}[Z \leq \theta | \mathcal{M} = \mathcal{M}_\emptyset] < 0.05$:

$$2^{-|K|} \sum_{z=0}^{\theta} \binom{|K|}{z} < 0.05$$

  - e.g., $|K| = 100$, max errors tolerated is $\theta = 42$

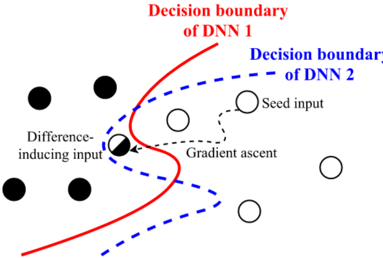
Conclusion: empirically limited model degradation, robust to removal trials. More realistic null-models; first reasoning about decision making on black box queries.

zoNNscan: a score for input safety

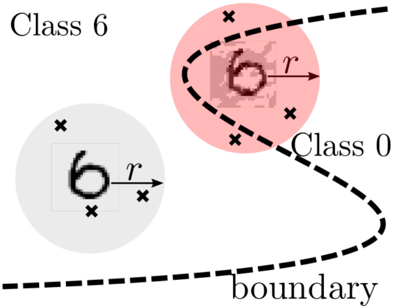
zoNNscan: a score for input safety

(a) Input 1 (b) Input 2 (darker version of 1)

Figure 1: An example erroneous behavior found by DeepXplore in Nvidia DAVE-2 self-driving car platform. The DNN-based self-driving car correctly decides to turn left for image (a) but incorrectly decides to turn right and crashes into the guardrail for image (b), a slightly darker version of (a).



zoNNscan: a score for input safety

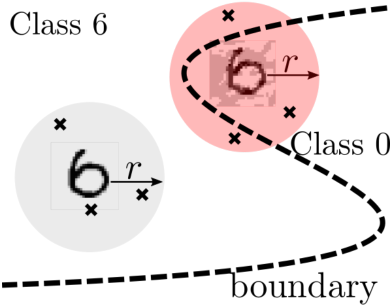


Shannon entropy over result vectors $x \rightarrow \mathcal{M}(x) = v[C_0, C_1, \ldots C_{n-1}]$.

Definition (zoNNscan score, in zone $\mathbb{Z}$:)

$$\mathbb{Z} \mapsto \mathbb{E}_{\mathbb{Z}}[H_n \circ \mathcal{M}(x))]$$

zoNNscan: a score for input safety



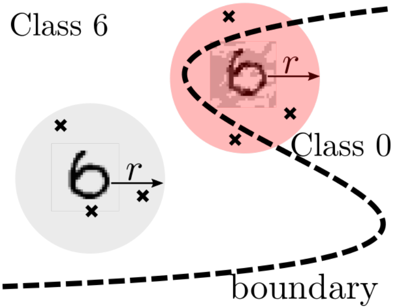
Shannon entropy over result vectors $x \rightarrow \mathcal{M}(x) = v[C_0, C_1, \dots C_{n-1}]$.

Definition (zoNNscan score, in zone $\mathbb{Z}$:)

$$\mathbb{Z} \mapsto \mathbb{E}_{\mathbb{Z}}[H_n \circ \mathcal{M}(x))]$$

Score $\in [0, 1]$. 1 is pure uncertainty.

zoNNscan: a score for input safety



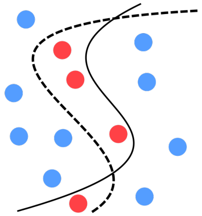
Shannon entropy over result vectors $x \rightarrow \mathcal{M}(x) = v[C_0, C_1, \ldots C_{n-1}]$.

Definition (zoNNscan score, in zone $\mathbb{Z}$:)

$$\mathbb{Z} \quad \mapsto \quad \mathbb{E}_{\mathbb{Z}}[H_n \circ \mathcal{M}(x))]$$

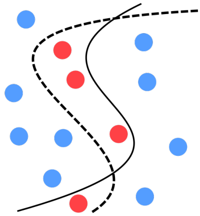
Score $\in [0, 1]$. 1 is pure uncertainty.
High dimensionality $\rightarrow$ Monte Carlo approximation.

∀ two models, corner cases (i.e., fingerprints) extracted for given dataset.
In MNIST testset: total of 182 fingerprints for 3 (MLP/CNN/RNN) models.

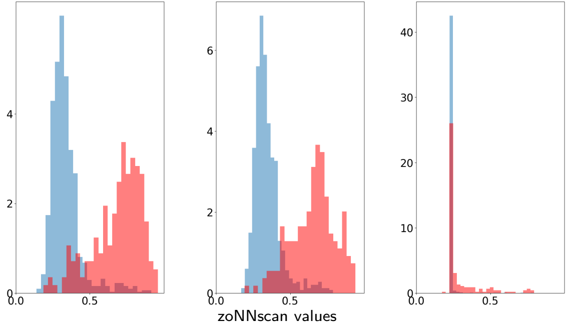


corner cases (in red)

∀ two models, corner cases (i.e., fingerprints) extracted for given dataset.
In MNIST testset: total of 182 fingerprints for 3 (MLP/CNN/RNN) models.

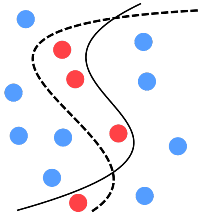


corner cases (in red)

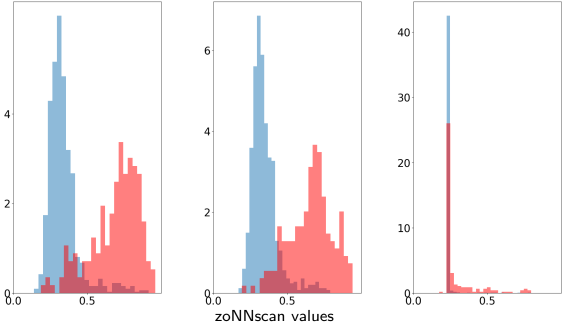


zoNNscan distributions (MLP/CNN/RNN models)

∀ two models, corner cases (i.e., fingerprints) extracted for given dataset.
In MNIST testset: total of 182 fingerprints for 3 (MLP/CNN/RNN) models.



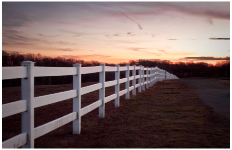
corner cases (in red)



zoNNscan distributions (MLP/CNN/RNN models)

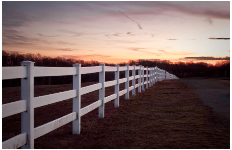
Conclusion: Use at inference time → trigger checks if critical.

To conclude



- Boundaries are not well understood $\rightarrow$ potential applicative problems. *Ensemble learning* (majority voting) hides problems.
- Boundary centered thinking raises security related questions.
- What is the power of the black box interaction setup?

To conclude



- Boundaries are not well understood $\rightarrow$ potential applicative problems. *Ensemble learning* (majority voting) hides problems.
- Boundary centered thinking raises security related questions.
- What is the power of the black box interaction setup?

Adversarial frontier stitching for remote neural network watermarking,

Erwan Le Merrer and Patrick Perez and Gilles Trédan, arXiv:1711.01894 (2017)

zoNNscan: a boundary-entropy index for zone inspection of neural models,

Adel Jaouen and Erwan Le Merrer, arXiv:1808.06797 (2018)

zoNNscan code: `https://github.com/technicolor-research/zoNNscan`