CounterAdversarial Recall of Synthetic Observations

A *neuro-inspired* approach to foil gradient-based adversarial attacks

Emanuele Ballarin[†]

Supervised by: Prof. Luca Bortolussi[†]

CoSupervised by: Dr. Alessio Ansuini [‡]

Graduation session: M.Sc. in *Data Science and Scientific Computing*

30th October 2022

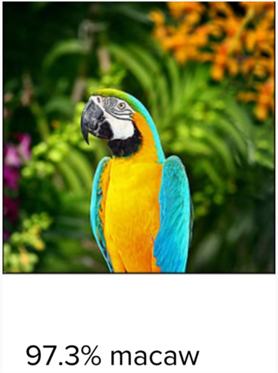
[†] AICPS $\in$ Dept. of Mathematics $\subseteq$ Univ. of Trieste | [‡] AREA Science Park

An *elevator pitch*

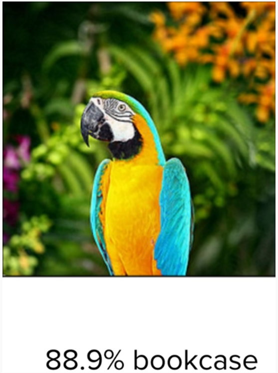
CARSO (*CounterAdversarial Recall of Synthetic Observations*) is a novel *deep learning* architecture and training/inference methodology for the improvement of *adversarial robustness* in *deep artificial neural networks*.

- Loosely inspired by high-level *neurocognitive* mechanisms;
- Targeted against *gradient-based*, *white-box* attacks;
- Significant, promising results so far; comprehensive testing still in early stages.

Deep learning today is a remarkably powerful and mature paradigm, able to reach (super)human-level performance in (selected) *regression*, *classification*, data *generation* and *control* tasks.



However...



(P. Perdikaris, 2018)

The last shown picture is an example of

Adversarial Input

An *input* is said to be *adversarial* to a machine learning system if it alters its reasonably expected behaviour[1]. Also called *adversarial attack*, stressing the intentional[2] crafting of it.

In the specific case of a classifier: produce a *misclassification*.

[1] Usually from the *P.o.V.* of the user(s).
[2] Which is not a strict requirement, though!

Why studying *adversarial robustness*?

We live in times where a growing portion of even *high-stakes* decisions is delegated to autonomous systems (*e.g.* *HR* selection, insurance, health, fraud detection, *etc.* ...).

Purely *technical* reasons

- Harden *ML/DL* systems against misuse and *input-tampering*;
- Assess (and *patch!*) behaviour where it the most fragile.

Legal / *ethical* / *social* reasons

- To ensure compliance with regulatory frameworks or coordinated initiatives thereof;
- Increase understanding, transparency, and societal trust.

Broader-reaching goals

- Use *robustness* as a lens through which to study *neurocognitive phenomena*.

A (more) precise definition of *robustness*

We can always reformulate the problem of *adversarial inputs* as one of *adversarial perturbations*, *i.e.*

$$\boldsymbol{x}_{\text{adversarial}} \coloneqq \boldsymbol{x}_{\text{legitimate}} + \boldsymbol{p}$$

leading to the following

Definition: *ϵ-perturbative adversarial attack against classifier $\mathcal{N}$ in $x_0 \in \mathbb{I}$, w.r.t. $||\cdot||$*

Any $\boldsymbol{x}^\star \coloneqq \boldsymbol{x_0} + \boldsymbol{p} \mid \mathcal{N}(\boldsymbol{x}^\star) \neq \mathcal{N}(\boldsymbol{x_0})$ and $||\boldsymbol{p}|| < \epsilon$

No optimal, universal defence! Many *case-by-case* results, many *trade-offs*, practically no *robust-by-design* applicable solution.

A remark

But... have *you* ever experienced an *adversarial(-like) phenomenon?*

Indeed, *brains* may be the *only* practical realisation of a system with the *robustness* properties we look for...

💡 A guiding idea

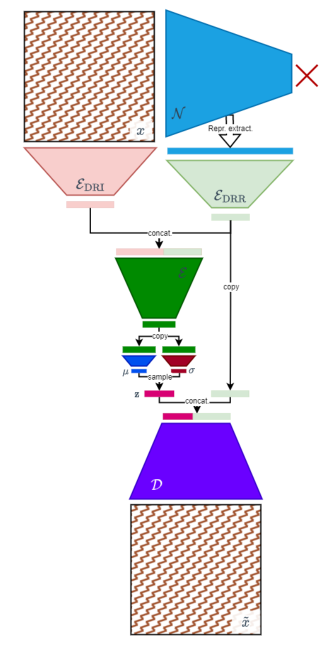
Is it possible to loosely inform the development of *robust* DL systems with (grossly simplified, idealised) descriptions of neurocognitive phenomena?

Getting inspiration from the ideas of *recall of acquired information*, and *introspection* as *thought about thought*.

A crucial remark!

⚠ Beware!

The *modelling* that follows has no claim of *biological plausibility* whatsoever, at this stage! This would be *added value*, though – and in interesting research direction!



TL;DR: Just a *fancy* **cVAE**!

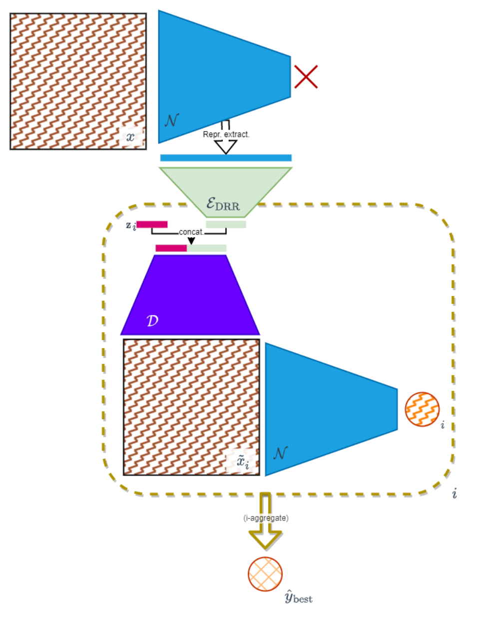
Given an *adversarially-pretrained* classifier (for the problem of interest, and according to a given *threat model*):

- First *classification pass* for representation extraction;
- Pre-encoding and rebalancing of input & representation;
- As in **cVAE**, aiming at *purified input* reconstruction from any input.

Requirements

A dataset of *clean/attacked* inputs to the classifier is needed (but no labels!). Threat models may differ.

Inference with CARSO



<u>*TL;DR*: Condition, sample, classify, aggregate!</u>

Given the same *adversarially-pretrained* classifier, the just-trained *representation pre-compressor* and *decoder*:

- First *classification pass* for representation extraction;
- Representation *pre-encoding*;
- Repeated sampling of *candidate purified inputs*;
- Second classification pass on such reconstructions, for actual classification;
- Aggregation of results.

Experimental results

Attack / Defence (*adv. acc.%*)	None	IAT	CARSO
None	**98.40**	97.17	96.72
FGSM $\|\|\cdot\|\|_2$, $\epsilon = 0.15$	12.09	91.89	**93.62**
FGSM $\|\|\cdot\|\|_2$, $\epsilon = 0.30$	01.21	76.94	**86.43**
(U) FGSM $\|\|\cdot\|\|_2$, $\epsilon = 0.50$	01.00	12.29	**13.59**
PGD $\|\|\cdot\|\|_\infty$, $\epsilon = 0.15$	01.60	90.54	**93.44**
PGD $\|\|\cdot\|\|_\infty$, $\epsilon = 0.30$	06.85	71.26	**86.27**
(U) PGD $\|\|\cdot\|\|_\infty$, $\epsilon = 0.50$	*20.66*	*11.67*	**38.38**
(U) DF $\|\|\cdot\|\|_\infty$, $\epsilon = 0.15$	00.66	90.25	**95.06**
(U) DF $\|\|\cdot\|\|_\infty$, $\epsilon = 0.30$	00.00	60.54	**93.31**
(U) DF $\|\|\cdot\|\|_\infty$, $\epsilon = 0.50$	00.00	00.78	**71.34**

Within the scope of the experimental analysis performed so far, we consider the results obtained to be *moderately-to-very* positive.

- A *clean accuracy toll* is imposed by the method *w.r.t. IAT*. Yet, this is to be generally expected, and slight in magnitude;
- Against *foreseen attacks*: significant – but not large – increase in *adversarial accuracy*;
- Against *unforeseen attacks*: very solid performance, clearly beyond *foreseen attacks/defences* transferability. *Innate robustness*

Speculatively: the result of a combined, synergistic effect. However, the lens of the *data manifold hypothesis* may give a more precise analysis: CARSO acts mainly as an *on manifold re-projector*!

We talked about CARSO – a novel framework devised to foil *gradient-based adversarial attacks*, specifically targeted at image classification – showing noteworthy improvements upon IAT, a strong contribution to *off-manifold-to-on-manifold reprojection*, and solid *innate robustness*.

Experimental scope can be – and will be! – broadened, though, to a wider set of *neural architectures*, *types* of data, or more complex, challenging (classification) tasks.

The work required to develop and assess CARSO evoked suggestions reaching far longer and broader than expected. Chiefly, in order of increasing conceptual distance...

- The idea that *adaptive* defences may exist, explicitly steering their behaviour on the basis of the geometric properties of inputs or attacks faced.
- Weight-agnostic layers operating at the *feature-specific*, able to produce *zero-gradient* in expectation.
- The possibility of informing the development of *deep learning* architectures with neural activity recordings from even *live-subjects*. 🐁

Thanks for your attention!