| cases | doc_1 | | doc_2 | | decision | id |
|---|---|---|---|---|---|---|
| | **authors** | • Francesco Crecchi<br>• Davide Bacciu<br>• Battista Biggio | **authors** | • Francesco Crecchi<br>• Davide Bacciu<br>• Battista Biggio | NOT DUPLICATES | 204 |
| | **title** | Detecting Black-box Adversarial Examples through Nonlinear Dimensionality Reduction | **title** | Detecting Adversarial Examples through Nonlinear Dimensionality Reduction | | |
| | **publication_date** | 2019-01-01 00:00:00 | **publication_date** | 2019-04-30 07:59:52+00:00 | | |
| | **source** | SupportedSources.INTERNET_ARCHIVE | **source** | SupportedSources.ARXIV | | |
| | **journal** | | **journal** | None | | |
| | **volume** | | **volume** | | | |
| | **doi** | | **doi** | | | |
| | **urls** | • https://web.archive.org/web/20200709052512/https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2019-120.pdf | **urls** | • http://arxiv.org/pdf/1904.13094v2<br>• http://arxiv.org/abs/1904.13094v2<br>• http://arxiv.org/pdf/1904.13094v2 | | |
| | **id** | id1355291704934338453 | **id** | id-4737101478964065512 | | |
| | **abstract** | Deep neural networks are vulnerable to adversarial examples, i.e., carefully-perturbed inputs aimed to mislead classification. This work proposes a detection method based on combining non-linear dimensionality reduction and density estimation techniques. Our empirical findings show that the proposed approach is able to effectively detect adversarial examples crafted by non-adaptive attackers, i.e., not specifically tuned to bypass the detection method. Given our promising results, we plan to extend our analysis to adaptive attackers in future work. | **abstract** | Deep neural networks are vulnerable to adversarial examples, i.e., carefully-perturbed inputs aimed to mislead classification. This work proposes a detection method based on combining non-linear dimensionality reduction and density estimation techniques. Our empirical findings show that the proposed approach is able to effectively detect adversarial examples crafted by non-adaptive attackers, i.e., not specifically tuned to bypass the detection method. Given our promising results, we plan to extend our analysis to adaptive attackers in future work. | | |
| | **versions** | | **versions** | | | |