

Probabilistic Graphical Models

Are Generative Classifiers More Robust to Adversarial Attacks?

Manal Akhannouss, ENS Paris-Saclay
manal.akhannouss@eleves.enpc.fr

Paul Barbier, ENS Paris-Saclay
paul.barbier@eleves.enpc.fr

Alexandre Lutt, ENS Paris-Saclay
alexandre.lutt@eleves.enpc.fr

I. Introduction

[1]

II. Models

II.1. Discriminative versus generative models

II.2. Classifiers architecture

III. Adversarial attacks

III.1. White box attacks

III.2. Black box attacks

IV. Experimental setup

IV.1. Attacks detection

IV.2. Models training

IV.3. Attacks benchmark

IV.4. Attacks detection

V. Results

V.1. Accuracy

V.2. Robustness to perturbations

V.3. Attacks detection

VI. Conclusion

References

[1] LI, Y., BRADSHAW, J., AND SHARMA, Y. Are generative classifiers more robust to adversarial attacks? 1