Analisi di Complessità di Attacchi Avversariali per Modelli Neurali

Progetto

I metodi di Intelligenza Artificiale basati su reti neurali hanno portato in anni recenti a risultati rivoluzionari in molteplici campi, con un impatto significativo nel tessuto economico e sociale.

L'applicabilità su larga scala e la sostenibilità di lungo termine di queste metodologie dipende però dalla loro affidabilità e sicurezza: risultati recenti hanno mostrato come le reti neurali possano essere sorprendentemente fragili rispetto a modifiche degli ingressi mirate a trarle in inganno.

Questo progetto mira a migliorare la comprensione della robustezza delle reti neurali ed i meccanismi alla base delle loro proprietà, mediante metodologie sia teoriche che empiriche.

Dai candidati ci si aspetta familiarità con temi di Informatica e metodi di Intelligenza Artificiale.

Piano di Attività

- Mesi 1-2: studio dei più recenti sviluppi nel campo della robustezza delle reti neurali
- Mesi 3-6: analisi teoriche ed empirica delle proprietà dei modelli neurali dal punto di vista della robustezza