Valutazione della Robustezza Adversarial su Modelli Deep Learning

Confronto tra Modello Standard e Modello Robusto

Davide Fabio Loreti - Mat. 865309
MSc Data Science
Cybersecurity for Data Science

23 novembre 2025

1 Introduzione

Gli adversarial attacks rappresentano una delle principali vulnerabilità dei modelli di deep learning. Piccole perturbazioni impercettibili, calcolate in base al gradiente della loss, possono indurre il modello a classificazioni errate anche su dataset ben conosciuti come CIFAR-10.

Il confronto tra modello standard e modello robusto nasce dalla necessità di valutare quanto un modello di deep learning sia vulnerabile a perturbazioni intenzionali (adversarial examples) e quanto le tecniche di difesa, come l'adversarial training, possano migliorare la robustezza.

Il modello standard rappresenta un CNN addestrato solo su dati puliti, che ottimizza la capacità di classificazione sulle immagini corrette. Tuttavia, è noto che questi modelli possono essere facilmente ingannati da perturbazioni impercettibili.

Il modello robusto, addestrato con PGD adversarial training, incorpora nelle fasi di addestramento sia immagini pulite sia immagini perturbate intenzionalmente. Questo permette al modello di apprendere rappresentazioni più stabili, riducendo la perdita di accuratezza quando vengono presentati esempi adversarial.

In questo esperimento sono stati confrontati due modelli ResNet-18 addestrati su CIFAR-10:

- un modello **standard**, addestrato su dati puliti,
- un modello **robusto**, addestrato con **PGD adversarial training**.

L'obiettivo è valutare quanto l'adversarial training migliori la resistenza contro diversi attacchi: FGSM, PGD, DeepFool e Carlini-Wagner.

2 Setup sperimentale

Il dataset utilizzato è **CIFAR-10** (60.000 immagini 32×32 su 10 classi). La rete scelta è una **ResNet-18** adattata a CIFAR-10. Per il modello robusto, durante l'addestramento sono state generate perturbazioni avversarie con PGD multi-step e integrate al training tramite mix con i dati puliti.

L'accuratezza finale sui dati puliti risulta:

- **Modello standard:** 79.53%
- **Modello robusto:** 69.88%

3 Attacchi testati

- **FGSM** (Fast Gradient Sign Method) – one-step L_∞
- **PGD** (Projected Gradient Descent) – iterativo L_∞
- **DeepFool** – ottimizzato L_2
- **Carlini & Wagner (C&W)** – ottimizzazione continua L_2

FGSM e PGD sono stati testati con tre valori di epsilon: 0.01, 0.03 e 0.05. DeepFool e C&W non richiedono eps esplicito.

4 Risultati e confronto

La Tabella 1 mostra l'accuratezza residua dei due modelli sotto attacco:

Tabella 1: Confronto accuracy: Modello Standard vs Modello Robusto

Attacco / Metriche	**Modello Standard**	**Modello Robusto**
Clean accuracy	79.53%	69.88%
FGSM $\epsilon = 0.01$	7.0%	13.0%
FGSM $\epsilon = 0.03$	7.0%	11.0%
FGSM $\epsilon = 0.05$	7.0%	10.0%
PGD (10 step) $\epsilon = 0.01$	7.0%	13.0%
PGD (10 step) $\epsilon = 0.03$	6.0%	11.0%
PGD (10 step) $\epsilon = 0.05$	6.0%	9.0%
DeepFool (adaptive)	0.0%	0.0%
C&W (50 step)	0.0%	0.0%

Discussione dei risultati

a) Analisi della robustezza vs clean accuracy

I risultati mostrano un chiaro compromesso: il modello robusto mantiene una robustezza maggiore agli attacchi gradient-based (FGSM, PGD), ma perde qualcosa in termini di accuratezza sui dati puliti rispetto al modello standard. Questo fenomeno è coerente: l'addestramento adversarial aumenta la resistenza alle perturbazioni a scapito di qualche punto percentuale di accuratezza sul test set pulito.

b) Efficacia degli attacchi iterativi

DeepFool e C&W risultano estremamente efficaci su entrambi i modelli, riducendo l'accuratezza quasi a zero. Ciò conferma che attacchi iterativi ottimizzati possono sfruttare fragilità profonde del modello, anche quando è robusto a FGSM/PGD.

c) Analisi delle classi più vulnerabili

La vulnerabilità non è uniforme tra le classi: immagini di classi semanticamente simili (come cat, dog, deer) risultano più facilmente confuse dagli attacchi. Questo suggerisce che le rappresentazioni interne del modello preservano correlazioni semantiche, che possono essere sfruttate dagli adversarial examples per causare misclassification.

5 Osservazioni qualitative

Le perturbazioni che ingannano il modello robusto tendono a essere più evidenti visivamente rispetto a quelle che ingannano il modello standard, indicando una maggiore stabilità locale del modello robusto. Tuttavia, vulnerabilità persistono in regioni ad alta complessità dello spazio delle feature.

6 Conclusioni

L'adversarial training con PGD aumenta la resistenza a perturbazioni gradient-based L_∞, ma non garantisce protezione contro attacchi ottimizzati L_2. Il confronto tra modello standard e modello robusto evidenzia il compromesso tra accuratezza sui dati puliti e robustezza agli attacchi. Gli esperimenti confermano la necessità di valutare la difesa in funzione del tipo di attacco.

Riferimenti bibliografici

[1] Rauber, J., Brendel, W., & Bethge, M. (2017). *Foolbox: A Python toolbox to benchmark the robustness of machine learning models.* arXiv:1707.04131.

[2] Madry, A. et al. (2018). *Towards Deep Learning Models Resistant to Adversarial Attacks.* ICLR.