SVILUPPO MODELLI DI ADVERSARIAL TRAINING

TESI DI LAUREA IN INGEGNERIA INFORMATICA
ANNO ACCADEMICO 2021-2022

Relatore:
prof. Loris Nanni
Laureando:
Michele Russo

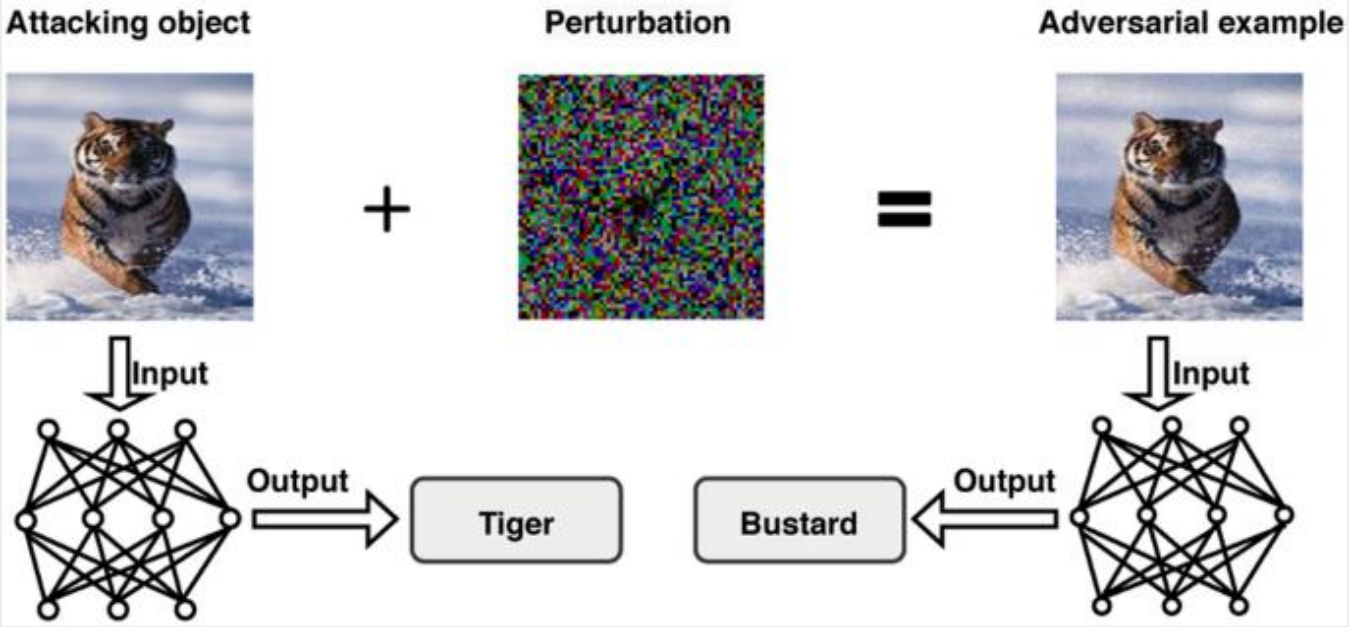
SOMMARIO

- Perché possono ingannare le reti neurali?
- Cosa sono gli adversarial examples?
- Perché usare l’adversarial training?
- Maxup
- Metodo sviluppato
- Prestazioni

Cosa sono gli adversarial examples?

Varie metodologie

- fast gradient sign
- min max
- PGD



Perché possono ingannare le reti neurali?

L’input di un pattern permutato può essere presentato nel seguente modo:

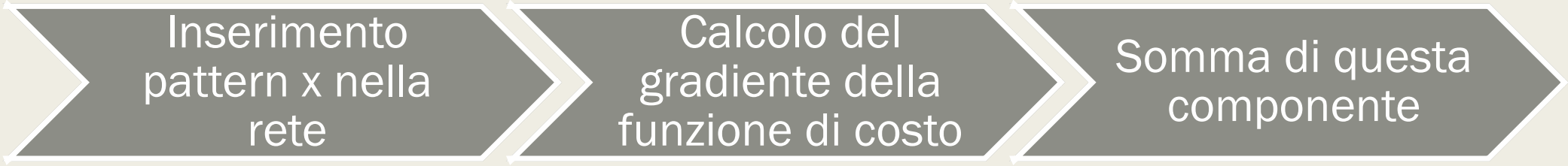
$$\tilde{x} = x + \eta$$

Supponendo che la rete abbia i pesi w^t, l’input realmente letto dalla rete sarebbe diverso quindi il contributo dato dalla permutazione fa crescere la funzione di attivazione:

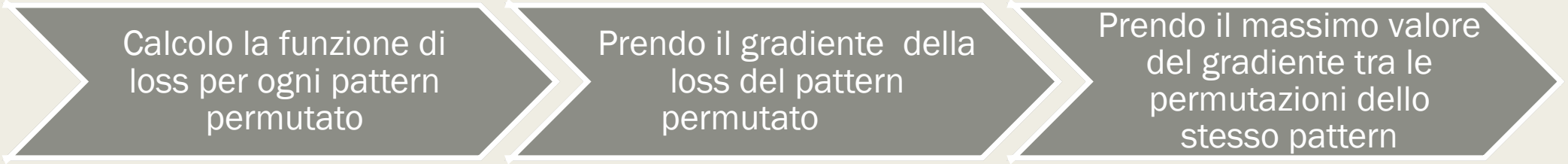
$$w^t \tilde{x} = w^t x + w^t \eta$$

Come vengono generati?

FGSM



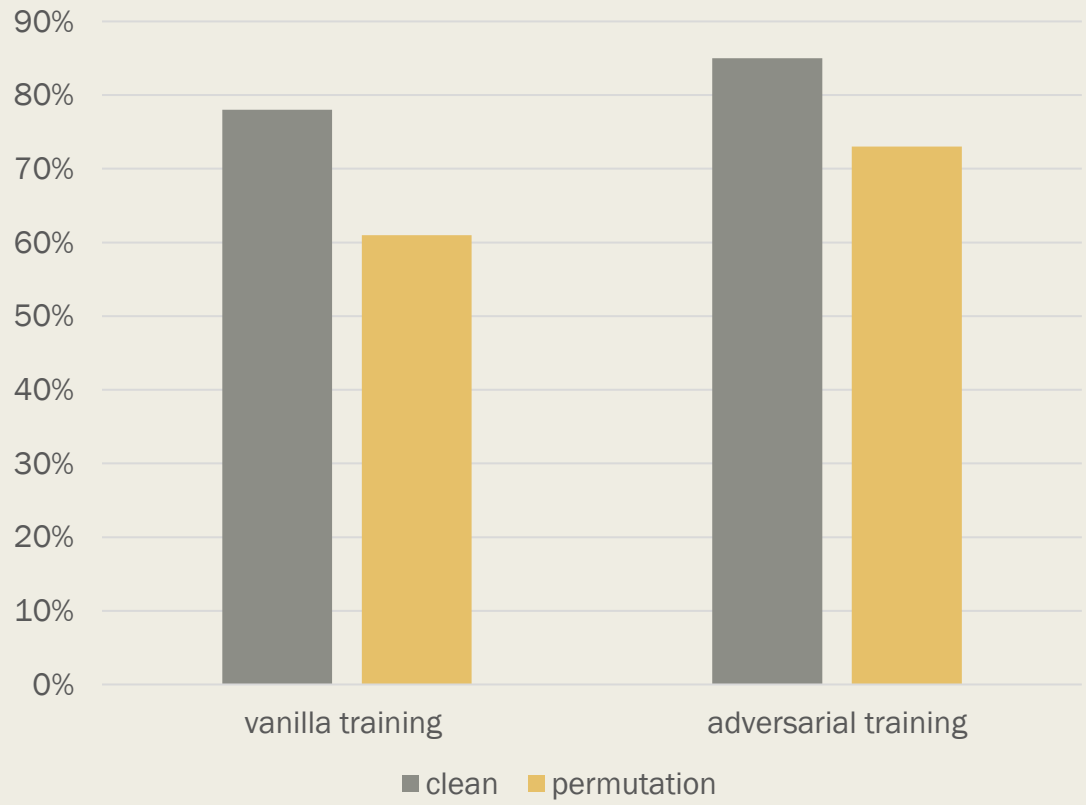
Min-Max



PGD

Perché usare l'adversarial training?

- L'utilizzo di Adversarial examples permette l'uso di feature robuste
- Miglioramento della generalizzazione
- Migliore robustezza davanti agli unseen data



MAX UP

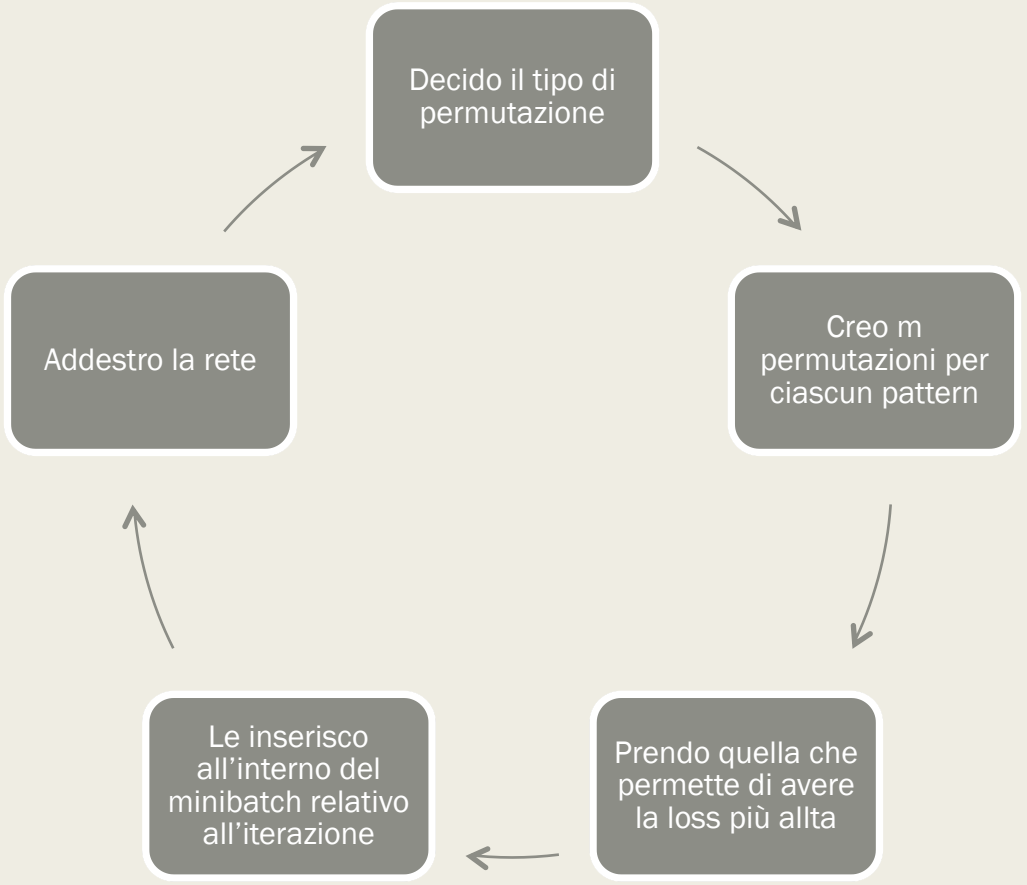
Il MaxUp è una tecnica di Min-Max.

Per caratteristiche può essere considerata come ‘lightweight variant of adversarial training’

Si basa su permutazioni generate da una distribuzione di probabilità i.i.d

Tra cui:

- CutOut
- MixUp
- CutMix



MAXUP VS PGD

Meno aggressivo rispetto a PGD

Molto più efficiente

Non impatta sulle prestazioni della rete

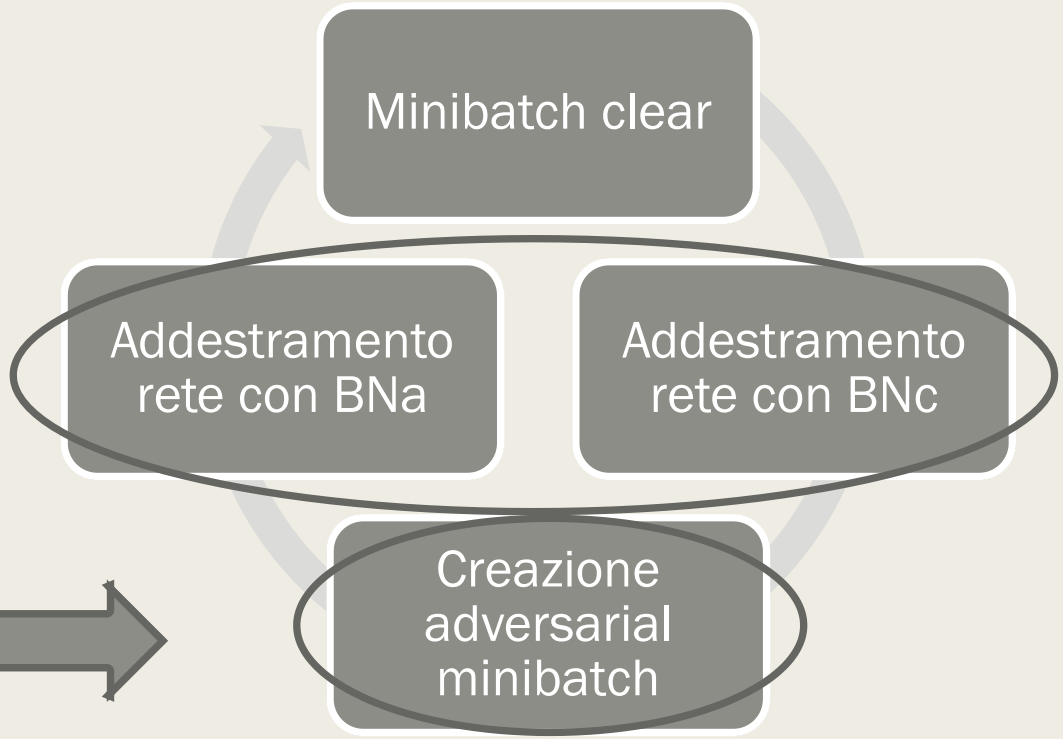
Piccola robustezza agli adversarial examples

METODO SVILUPPATO

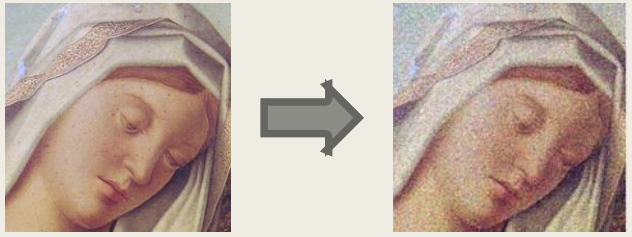
Si parte dalla tecnica del Maxup, cercando di fonderla con le tecniche di adversarial training viste fino ad ora.
In questo modo, è possibile sfruttare la velocità del Maxup e la robustezza dei metodi di adversarial training.

Descrizione metodo

Sostituisco i BN in base al minibatch usato. In modo da poter adattare meglio alla diversa distribuzione dei due minibatch



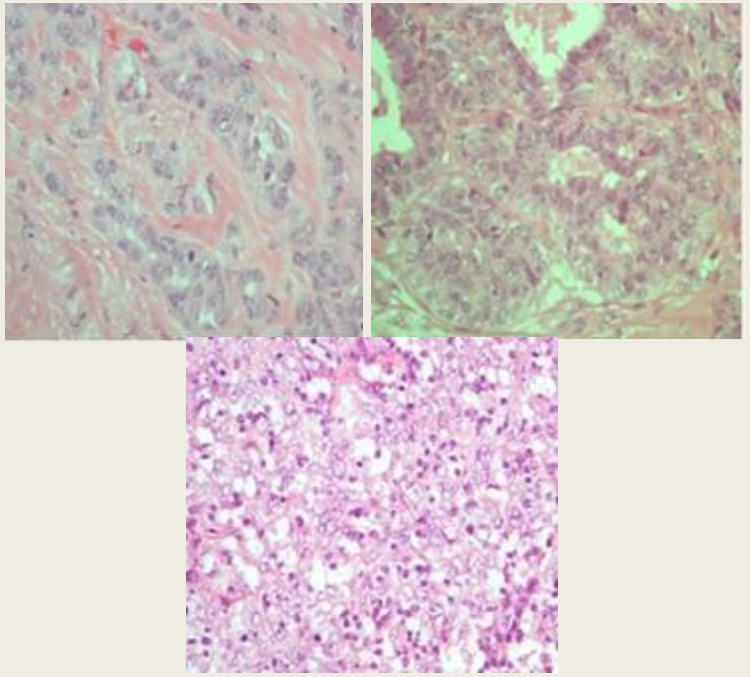
Riprendo dal metodo di generazione immagini del Max con permutazione Gaussiana



Dataset usati

Dataset Quadri

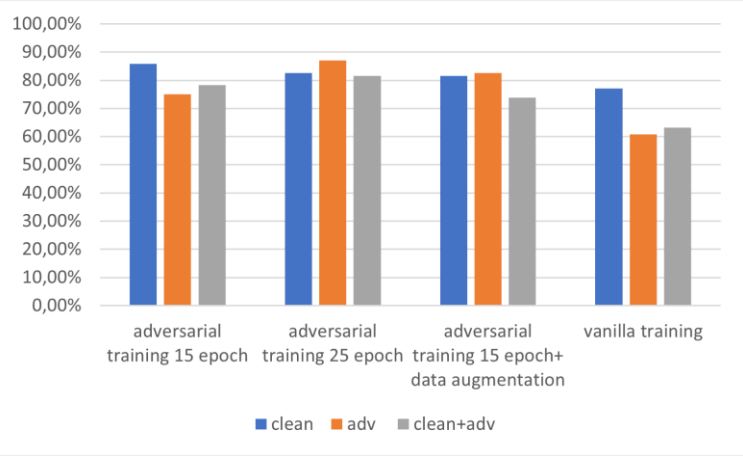
Dataset Istopatologie



Prestazioni

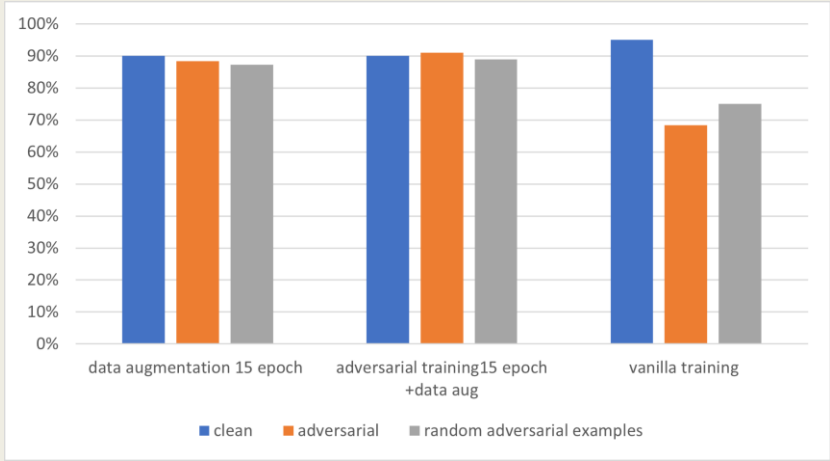
I risultati sono ottenuti tramite l'utilizzo della rete ResNet-18. I test effettuati per ciascuna rete presentano un test set clean, uno adversarial ed uno con pattern misti

Dataset quadri



Le reti create mediante il mio metodo sono tre: adversarial training con 15 e 25 epoch, ed una con 15 epoch e data augumentation

Dataset istopatologie



Le reti create mediante il mio metodo sono due: adversarial training con 15 epoch, ed una con 25 epoch e data augumentation

CONCLUSIONI

- Per dataset quadri l’utilizzo di 15 epoch risulta performare meglio rispetto all’utilizzo di 25, su dataset clean. Situazione che si ribalta guardando adversarial examples. Mentre il vanilla trainining ha prestazioni inferiori rispetto a tutti e tre i campi di sperimentazione.
- Per dataset istopatologie il vanilla training performa meglio di entrambi gli addestramenti proposti, per quanto riguarda il campo clean; mentre le sue prestazioni degenerano quando si utilizzano gli adverasarial examples. Vedendo i due modelli di adversarial training si nota che la data augumentation riesce ad avere performance leggermente migliori rispetto all’altro training