Creazione di adversarial samples per DNN per la classificazione di malware

Francesco Baraldi - mat. 172703

1. Introduzione

DNN per la classificazione di malware

Grazie al forte sviluppo delle reti neurali è possibile sfruttare queste tecnologie anche nell'ambito della sicurezza informatica, in particolare è possibile utilizzarle per la classificazione dei software in *benigni* o *maligni* (*malware*).

Si vuole approfondire la creazione di adversarial sample in questo ambito e valutare l'efficacia di questo approccio per ingannare delle reti neurali che hanno lo scopo di discriminare se un software con determinate caratteristiche è malevolo oppure no.

Adversarial machine learning

L’adversarial machine learning studia le tecniche per compromettere il corretto funzionamento degli algoritmi di deep learning, in particolare gli algoritmi di classificazione. Si cerca di costruire dei samples, detti *adversarial samples*, nei quali si inseriscono delle perturbazioni rispetto al sample originale che siano impercettibili ma che forzino l’algoritmo a classificarlo in modo errato.

Adversarial machine learning (1)

L’applicazione classica avviene per reti neurali deep e per il task della classificazione di immagini, questo perché l’alta entropia di un’immagine può essere facilmente sfruttata per cambiare la natura dell’immagine senza cambiarne l’aspetto visivo.

Nell’ambito di questo progetto invece si studia l’applicazione dell’adversarial ML nella classificazione di malware, nel quale ogni input è un vettore M-dimensionale di features binarie.

2. Classificazione di malware

Dataset

È stato utilizzato un dataset[1] per la classificazione di malware per il sistema operativo mobile Android, in particolare ogni istanza è composta da 215 features binarie, ognuna indica se il software considerato utilizza oppure no determinate funzioni del sistema operativo, oppure indica se esegue o no determinate azioni; infine ogni software è etichettato come benigno o maligno.

[1] *Android Malware Dataset for Machine Learning*, https://www.kaggle.com/datasets/shashwatwork/android-malware-dataset-for-machine-learning?select=drebin-215-dataset-5560malware-9476-benign.csv

Reti neurali

Per il task di classificazione sono state utilizzate diverse versioni di multi-layer perceptron (MLP), alcune più semplici, altre più complesse, per valutare l'efficacia nella classificazione e la difficoltà nel creare adversarial sample. Le reti utilizzate sono le seguenti:

— DNN1: [8, 8, 8];
— DNN2: [20, 10, 8, 8];
— DNN3: [20, 20, 10, 10, 10];
— DNN4: [50, 50];
— DNN5: [100, 100].

Ogni rete ha gli hidden layers strutturati come descritto sopra, un input layer per le 215 features e un output layer con un singolo output: 0 se malware, 1 altrimenti.

Training e testing

Ogni rete è stata addestrata sul dataset citato in precedenza, il processo di training è stato eseguito con 50 epoche e usando lo SGD come ottimizzatore. I risultati sono mostrati nella seguente tabella.

	DNN1	**DNN2**	**DNN3**	**DNN4**	**DNN5**
Accuracy sul training set	97.4 %	**97.5 %**	**97.5 %**	97.4 %	97.4 %
Accuracy sul test set	97.0 %	97.0 %	97.1 %	97.1 %	**97.2 %**

Come si può notare, le differenze delle performance delle 5 reti testate sono trascurabili utilizzando 50 epoche, la differenza sarà invece più evidente nella facilità con cui queste reti vengono ingannate.

3. Creazione di adversarial samples

MILP + DNN

É possibile modellare una rete neurale con un *Mixed Integer Linear Program* (MILP)[1]. Il problema di ottimizzazione in seguito può essere utilizzato per diversi scopi, uno di questi è la creazione di adversarial input per la rete neurale, inserendo vincoli e funzione obiettivo mirati a questo scopo.

L'approccio considerato assume che l'operatore non lineare della rete sia la *Rectified Linear Unit* (ReLU).

[1] *Deep Neural Network and Mixed Integer Linear Optimization - Matteo Fischetti, Jason Jo*

Modello matematico

- Si considera una rete con $K+1$ layer numerati da 0 a K, il layer 0 è di input mentre il layer K è di output.
- Ogni layer $k \in \{0,...,K\}$ ha n_k neuroni.

Variabili:

- $x^k \in \mathbb{R}^{n_k}$ è l'output del layer k, x_j^k è l'output del $j-esimo$ neurone del layer k, con $k = 1,...,K$
- s_j^k è la variabile slack per ogni neurone di ogni layer, con $k = 1,...K$
- z_j^k è la variabile di attivazione per ogni neurone di ogni layer, con $k = 1,...K$
- $error$ è la variabile che indica il numero di features cambiate rispetto all'input originale
- $output$ è la variabile di output della rete

Modello matematico (1)

Vincoli:

$$\sum_{i=1}^{n_{k-1}} w_{ij}^{k-1} x_i^{k-1} + b_j^{k-1} = x_j^k - s_j^k \quad k = 1,...,K-1; j = 1,...,n_k$$

$$\sum_{i=1}^{n_0} (x_i^0 - input_i)^2 \leq error$$

$$\sum_{i=1}^{n_{K-1}} w_{i0}^{K-1} x_i^{K-1} + b_0^{K-1} = output$$

$$error \leq MAX\_ERROR$$

$$output \geq 0.55$$

$$x_j^k, s_j^k \geq 0 \quad k = 1,...,K; j = 1,...,n_k$$

$$z_j^k \in \{0,1\} \quad k = 1,...,K; j = 1,...,n_k$$

$$z = 1 \rightarrow x \leq 0 \quad k = 1,...,K; j = 1,...,n_k$$

$$z = 0 \rightarrow s \leq 0 \quad k = 1,...,K; j = 1,...,n_k$$

Modello matematico (2)

Funzione obiettivo: $min(error)$

Il modello appena descritto modella un generico multi-layer perceptron e genera un sample che sia il più simile possibile a un determinato sample dato in input. Minimizzando l'errore si modificano meno features possibili, mentre il vincolo che esprime che l'output deve essere maggiore di 0.55 costringe il modello a creare un'istanza che sia classificata dalla rete come benigna.

In questo modo è possibile fornire al modello un sample classificato come malware e ottenne una versione molto simile che però viene considerato dalla rete neurale come benigno.

Performance

Anche utilizzando i migliori solver moderni, per reti complesse e per input con alta dimensionalità i tempi potrebbero diventare inaccettabili. Diventa fondamentale avere degli upper bound corretti per le variabili continue.

Al modello utilizzato quindi è stato applicato il processo ideato da Matteo Fischetti a Jason Jo[1] per trovare degli upper bound ottimi.

[1] *Deep Neural Network and Mixed Integer Linear Optimization - Matteo Fischetti, Jason Jo*

Upper bounds

Per ogni generico neurone $UNIT(j,k)$ si eliminano dal layer k tutti i neuroni diversi da quello considerato e tutti i layer successivi, dopodiché si ottimizza il problema massimizzando la variabile x_j^k e la variabile s_j^k. I risultati trovati possono essere utilizzati come bound per le rispettive variabili nel modello originale. È importante notare che questo procedimento è indipendente dall'input e quindi si può eseguire "una tantum" per ogni rete; inoltre anche imponendo dei limiti ai tempi di esecuzione, e quindi accettando soluzioni approssimate, si ottengono comunque risultati accettabili.

4. Risultati e test

Test

Per valutare le performance, il modello di ottimizzazione descritto in precedenza è stato utilizzato per modificare 50 samples per ogni rete, originariamente etichettati come *maligni*, al fine di ottenere una classificazione *benigna* da parte delle reti.

Inoltre lo stesso modello di ottimizzazione è stato testato anche usando i bound ottimi, che sono stati calcolati una volta sola usando il metodo descritto da Fischetti e Jo e salvati su file.

Per evitare tempi di esecuzione eccessivi sono stati usati dei time limit: 2 minuti per la creazione di adversarial samples e 10 secondi per il calcolo degli upper bound ottimi.

Risultati

Modello base

	Opt. Solved (%)	Avg. time (s)	Avg. gap (%)
DNN1	100.0	0.171	0.0
DNN2	100.0	1.931	0.0
DNN3	70.0	59.522	9.400
DNN4	92.0	24.904	1.600
DNN5	52.0	79.643	22.033

Modello ottimizzato

	Opt. Solved (%)	Avg. time (s)	Avg. gap (%)
DNN1	100.0	0.043	0.0
DNN2	100.0	0.480	0.0
DNN3	100.0	1.873	0.0
DNN4	100.0	3.019	0.0
DNN5	92.0	25.188	2.067

Come si nota, il modello non ottimizzato non ha problemi per le reti più semplici, mentre incontra difficoltà con quelle più complesse, all'aumentare di layer, e neuroni per ogni layer.

Il modello con i bound ottimi invece mostra un lieve calo di performance solamente con la rete più complessa, ma offre comunque prestazioni ottime. Inoltre, a prescindere dalla complessità della rete il modello ottimizzato garantisce tempi di esecuzione molto minori.

Conclusioni

Il procedimento applicato risulta molto flessibile per modellare una rete neurale usando un MILP, perché permette di cambiare vincoli e funzione obiettivo in base alle proprie esigenze.

È importante però tenere conto della complessità che questo comporta e dei tempi di esecuzione che possono aumentare fino a diventare inaccettabili, per questo è cruciale considerare gli upper bound per le variabili continue del modello e applicare metodi per il calcolo di questi.