3.3.12 Risultati ottenuti per l'attacco ZOO

A seguito dell'esecuzione della bayesian optimization per la versione untargeted e targeted dell'attacco ZOO, si riportano i migliori risultati ottenuti per tali attacchi.

Si ricorda che gli attacchi sono stati eseguiti sui 100 segnali di stop classificati con minore probabilità dal modello di riconoscimento di segnali stradali.

I risultati sono stati organizzati in maniera tabellare e rappresentando i primi 10 attacchi più performanti, sia per l'attacco untargeted che per quello targeted. La performance di un attacco è stata misurata secondo la formula espressa nell'espressione (5).

Ogni attacco è identificato dai seguenti parametri: $initial\_const, confidence, learning\_rate$ e $\max\_iterations$. Tutti i restanti parametri utilizzati sono i seguenti:

{"batch_size":128, "img_resize":32, print_every":1000, "switch_iterations":100, use_tanh":False }

Tali parametri sono rappresentati sottoforma di dizionario Python passato in input alla funzione che implementa l'attacco ZOO nella libreria AutoZOOM.

Di seguito sono riportati i 10 migliori attacchi ZOO untargeted e targeted:

RESULTS ZOO UNTARGETED						
Initial Const	Confidence	Learning Rate	Max Iterations	Success Rate	Avg Norm	*attack performance*
0.001	0.0	0.001	1500	97%	7.762	0.343
0.1	0.0	0.0005	1500	79%	5.414	0.338
0.001	0.0	0.0005	2000	90%	6.835	0.330
0.01	0.0	0.0005	1500	76%	5.183	0.328
0.7644	0.223	0.0250	1500	77%	5.423	0.318
0.001	0.0	0.0005	2500	97%	7.934	0.300
0.001	0.0	0.001	2000	100%	8.443	0.296
0.582	0.389	0.032	1500	61%	3.828	0.291
0.291	0.106	0.024	1500	99%	8.394	0.290
0.0001	0.0	0.1	1500	100%	8.647	0.279
RESULTS ZOO TARGETED						
0.001	0.0	0.0005	10000	100%	8.547	0.287
0.001	0.0	0.0005	6000	100%	8.588	0.284
0.001	0.0	0.0005	4000	98%	8.394	0.2805
10	0.0	0.0005	4000	98%	8.396	0.2803
1.262	0.641	0.049	4000	91%	7.678	0.270

On the following pages are the images of the adversary examples generated by the best untargeted and targeted attack, with the relative class predicted by the model (expressed after the →), the probability related to the predicted class and the "stop sign" class (denoted by "Stop Sign Prob"), and the norm of the difference between the adversary image and the original image (denoted by "L2 Norm"):

Attack untargeted ZOO: