Neprijateljski primjeri na ImageNet skupu podataka

Marko Opačić
Mentor: Marko Đurasević

Neprijateljski primjeri

- Na ulazni podatak se dodaje šum (perturbacijski vektor)
- Idealno ljudima neprimjetan -> mala promjena
- Perturbacija se ograničava na interval [+ε, -ε]

Fast Gradient Sign Method

- gradijent gubitka po ulaznoj slici
- piksele slike mijenjamo u smjeru gradijenta
- maksimiziramo gubitak na točnom razredu
- jedan korak

Projected Gradient Descent

- posebna definicija funkcije gubitka -> ciljani napadi!
  - Maksimiziramo gubitak na točnoj klasi, a minimiziramo na ciljanoj klasi
- više koraka, iterativno
- SGD s projekcijom na ograničeni interval

Modeli i ulazni podaci

- Korišteni modeli: ResNet50, VGG-16 i EfficientNet-B4

(a) violina **(b)** go kart **(c)** bizon

Slika 4.1: Ulazne slike

Korišteni parametri

Parametar	**Vrijednost**
broj iteracija	100
opseg L_∞ kugle ϵ	$2/255$
korak ucenja SGD	0.005

Tablica 4.1: koristeni parametri

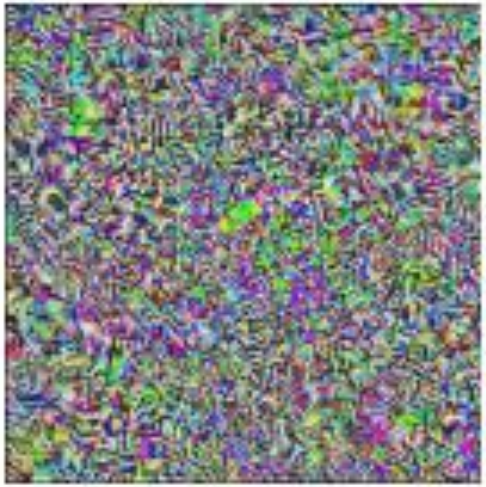
Rezultati - početne predikcije

Model	violina	go kart	bizon
ResNet50	violin (0.9277)	go kart (0.9729)	bizon (0.9966)
VGG16	violin (0.9414)	go-kart (0.9925)	bizon (0.9998)
EfficientNet-B4	violin (0.9470)	go kart (0.9945)	bizon (0.9944)

Tablica 4.2: predikcije za neizmijenjene ulazne slike

Rezultati

violina

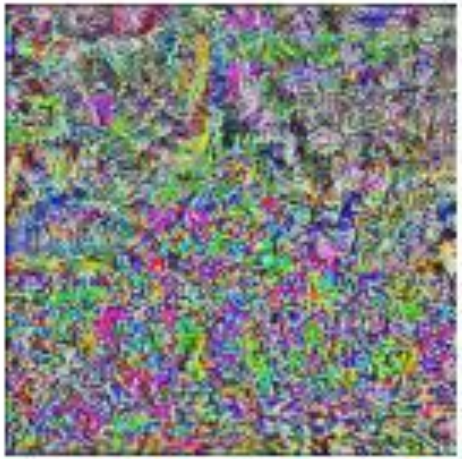


delta

99.9952% samoyed

Rezultati

gokart

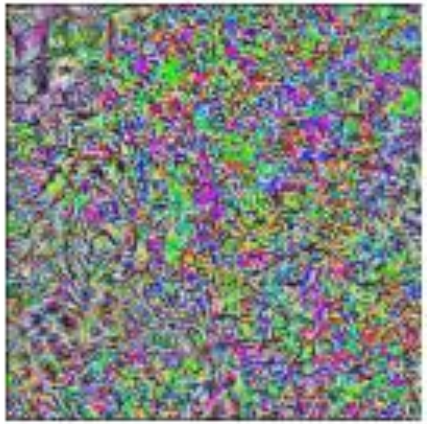


delta

99.9328% samoyed

Rezultati

bizon



delta

100% samoyed

Zaključak

- Neprijateljske primjere je lako pronaći
- Neprijateljski primjeri su u velikoj mjeri transferabilni
- Robusnost smanjuje preciznost klasifikatora
- Potrebno je istražiti naprednije metode povećanja robusnosti