nedeljainformatike v6.0

Sigurna i pouzdana veštačka inteligencija

Nikola Jovanović

Matematička gimnazija

26. 04. 2021.

Uspesi veštačke inteligencije

1996

Deep Blue computer beats world chess champion - archive, 1996

Uspesi veštačke inteligencije

1996

2005

Stanford Report, October 11, 2005

Stanford team's win in robot car race nets $2 million prize

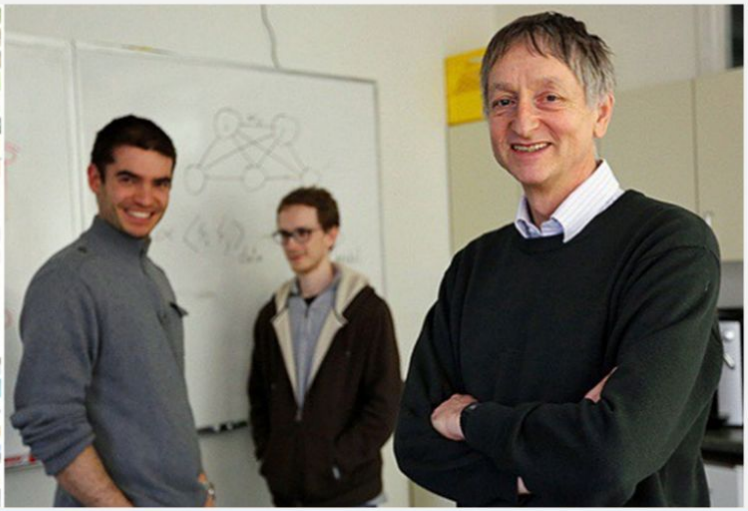
David Orenstein

Stanley crosses the finish line in Primm, Nev., winning the 2005 DARPA Grand Challenge.

Uspesi veštačke inteligencije

1996 2005 **2012**

AlexNet



Uspesi veštačke inteligencije

1996 2005 2012 **2015**

Microsoft, Google beat humans at image recognition

February 19, 2015 // By R. Colin Johnson

Uspesi veštačke inteligencije

1996 2005 2012 2015 **2016**

How Google's AlphaGo Beat a Go World Champion

Inside a man-versus-machine showdown

CHRISTOPHER MOYER MARCH 28, 2016

Uspesi veštačke inteligencije

1996 2005 2012 2015 2016 **2017-**

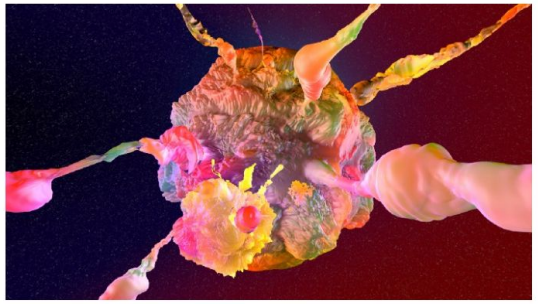
BLOG POST
RESEARCH

30 NOV 2020

AlphaFold: a solution to a 50-year-old grand challenge in biology

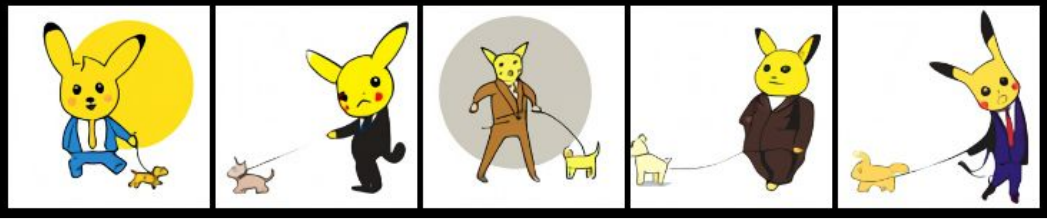
WAYMO

Algorithms Can Now Identify Cancerous Cells Better Than Humans



TEXT PROMPT an illustration of a pikachu in a suit walking a dog

AI-GENERATED IMAGES



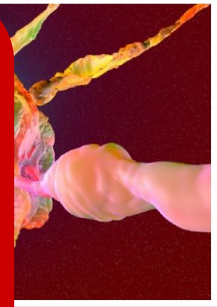
Uspesi veštačke inteligencije

1996 2005 2012 2015 2016 **2017-**

BLOG POST RESEARCH 30 NOV 2020

AlphaFold: a solution to a 50-year-old grand challenge in biology

Algorithms Can Now Identify Cancerous Cells Better Than Humans



Ipak nije sve tako sjajno...

Koja je ovo životinja?

Učesnici NI

Konj
Mačka
Noj
Patka

Neuralna mreža

Konj
Mačka
Noj
Patka

Malo šuma...

Koja je ovo životinja?

Učesnici NI

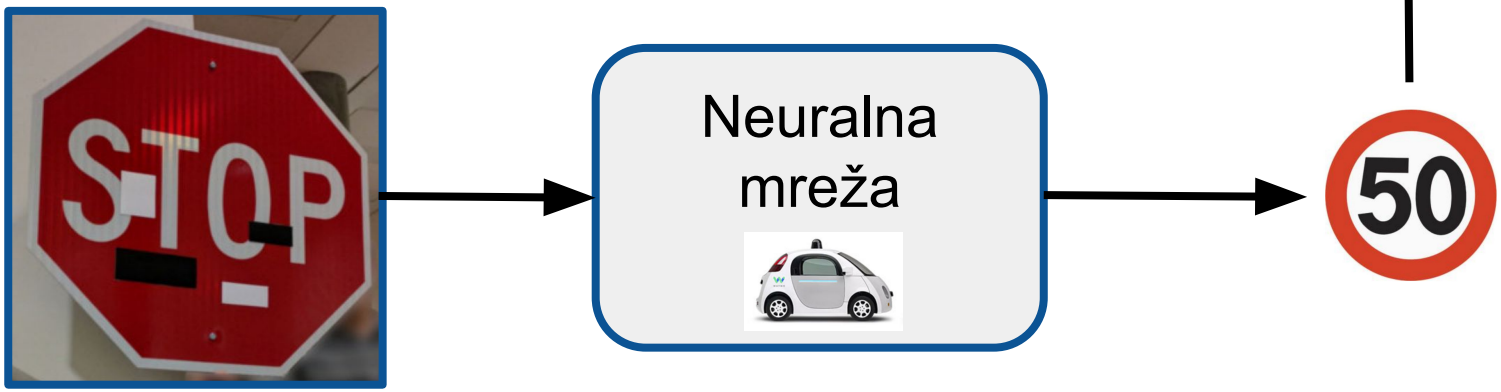
Konj
Mačka
Noj
Patka

Neuralna mreža

Konj
Mačka
Noj
Patka

Malo opasniji primer

- **Korak 1**: napadač postavlja bezazlene (?) nalepnice na ulični znak
- **Korak 2**: sistem za prepoznavanje znakova na samovozećem automobilu greši
- **Korak 3**: **udes**



Sigurna i pouzdana VI

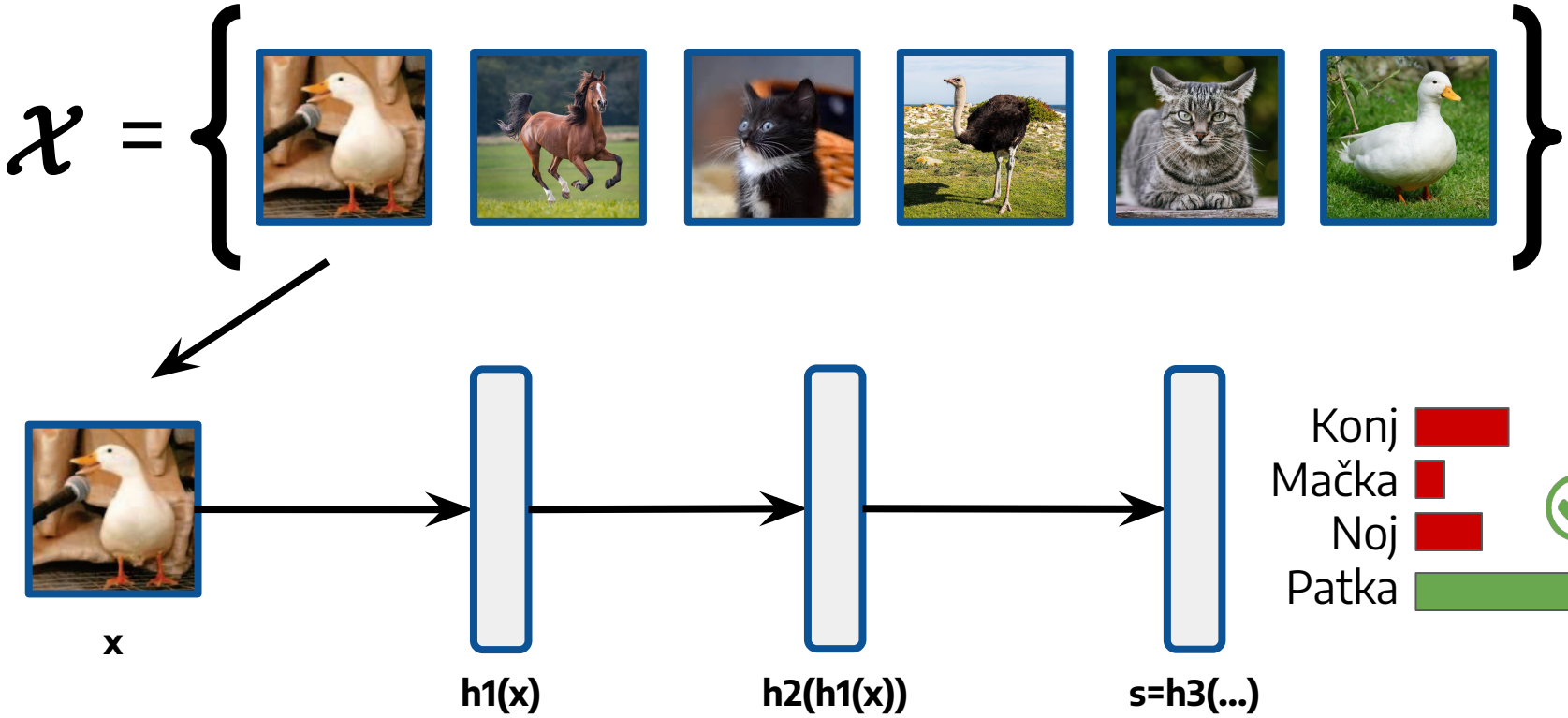
- Ovakvi primeri su doveli do novih istraživačkih problema i izdvajanja oblasti **sigurne i pouzdane veštačke inteligencije** kao posebne grane
- Jedan deo tzv. “trećeg talasa veštačke inteligencije”:
  - Logika/zaključivanje/pravila + statistika/opažanje/učenje iz podataka
  - **Cilj:** pouzdani sistemi koji razumeju i mogu da objasne svoja predviđanja

Agenda

1. Predznanje: neuralne mreže
2. Adversarijalni napadi
3. Verifikacija neuralnih mreža

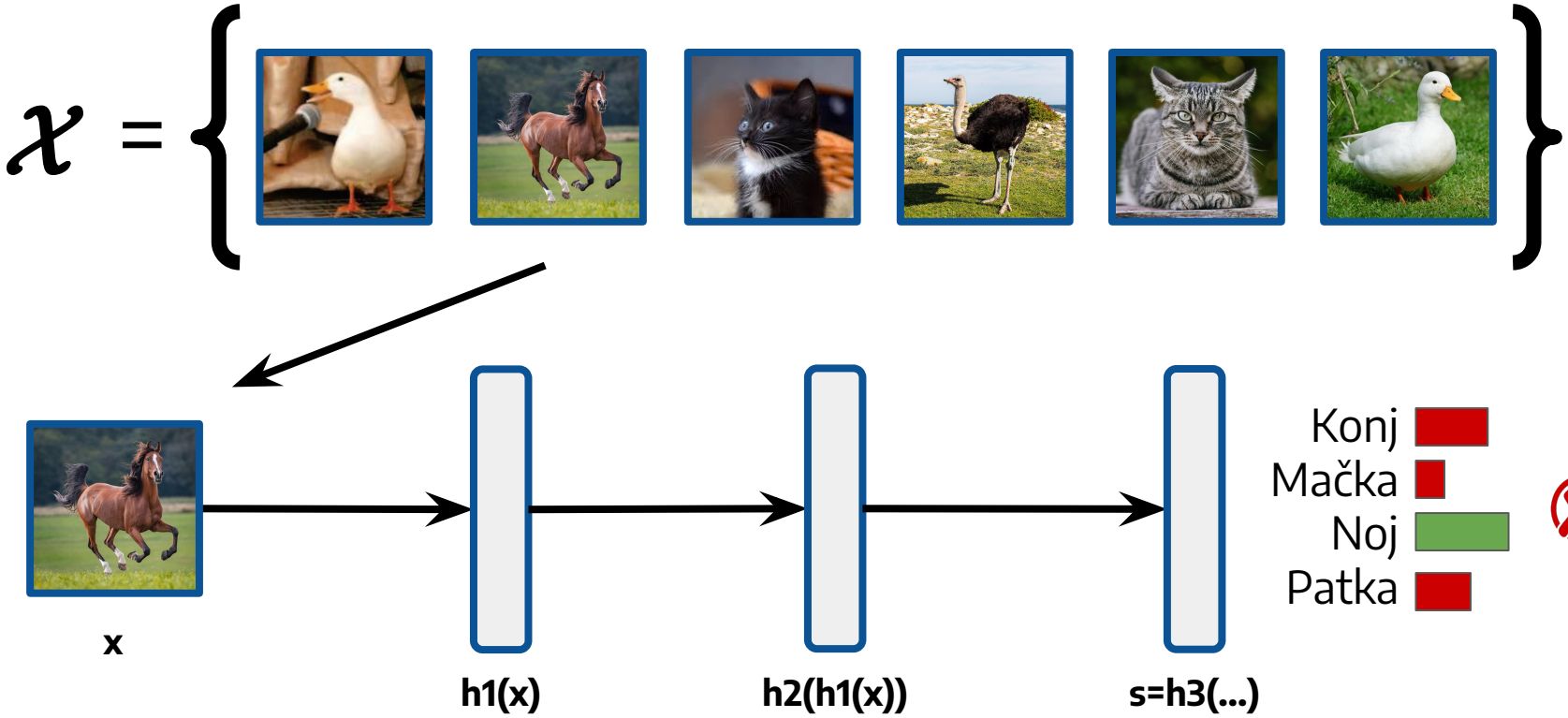
Neuralne mreže

- Rešavamo problem **klasifikacije**: želimo da neuralna mreža nauči da predvidi kojoj od **k** unapred zadatih klasa pripada ulazni podatak **x** iz skupa podataka $\mathcal{X}$



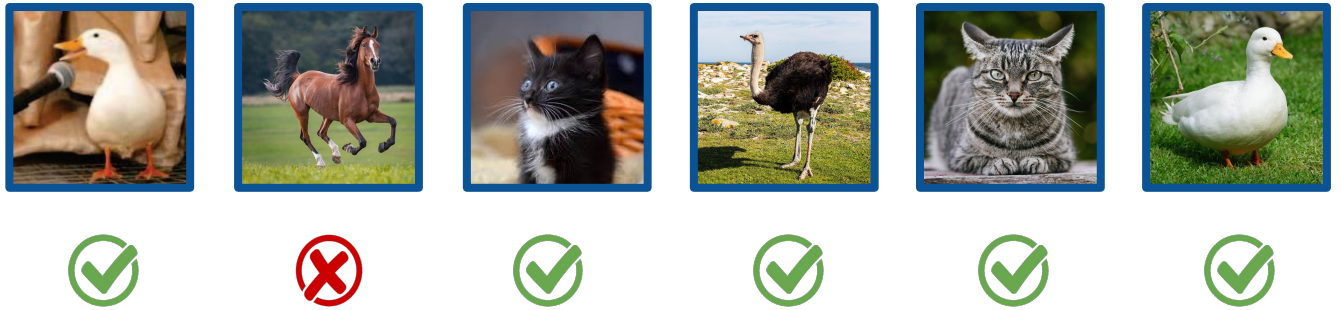
Neuralne mreže

- Rešavamo problem **klasifikacije**: želimo da neuralna mreža nauči da predvidi kojoj od **k** unapred zadatih klasa pripada ulazni podatak **x** iz skupa podataka $\mathcal{X}$



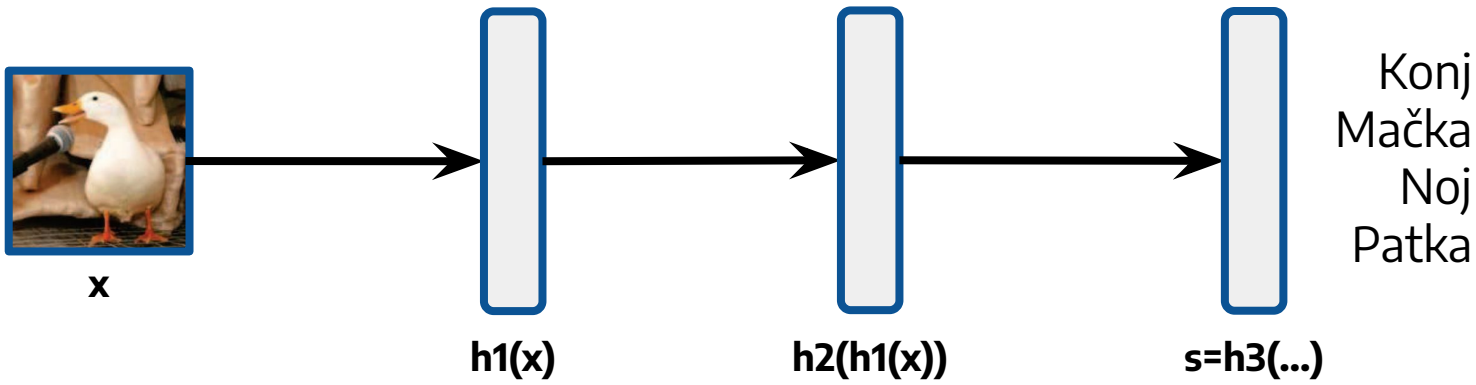
Evaluacija neuralnih mreža

- Uspešnost mreže merimo koristeći **accuracy** (*tačnost*), tj. procenat ispravno klasifikovanih **x** iz našeg skupa podataka

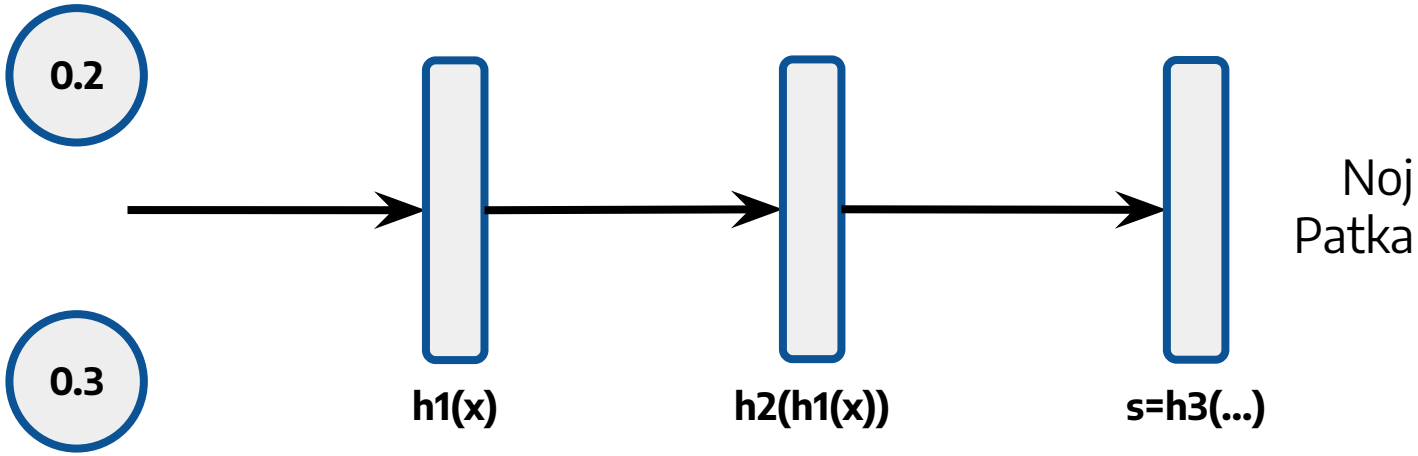
$$\mathcal{X} = \{$$  $$\}$$

ACC = 83.3%

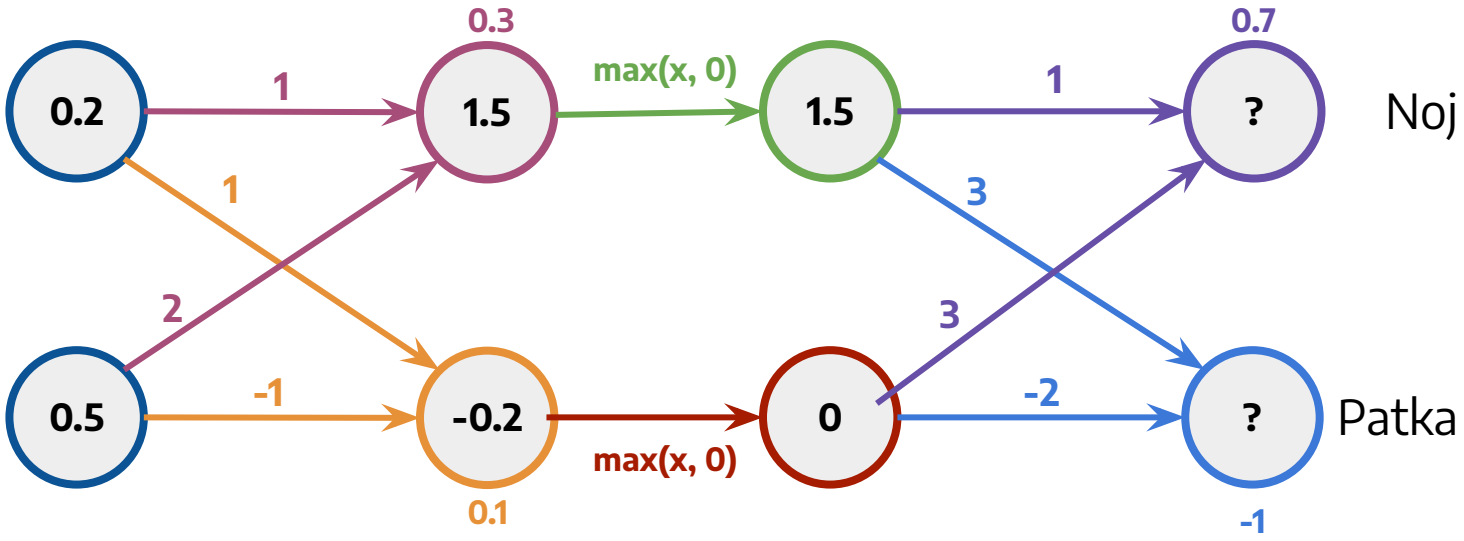
Neuralne mreže: jednostavan primer



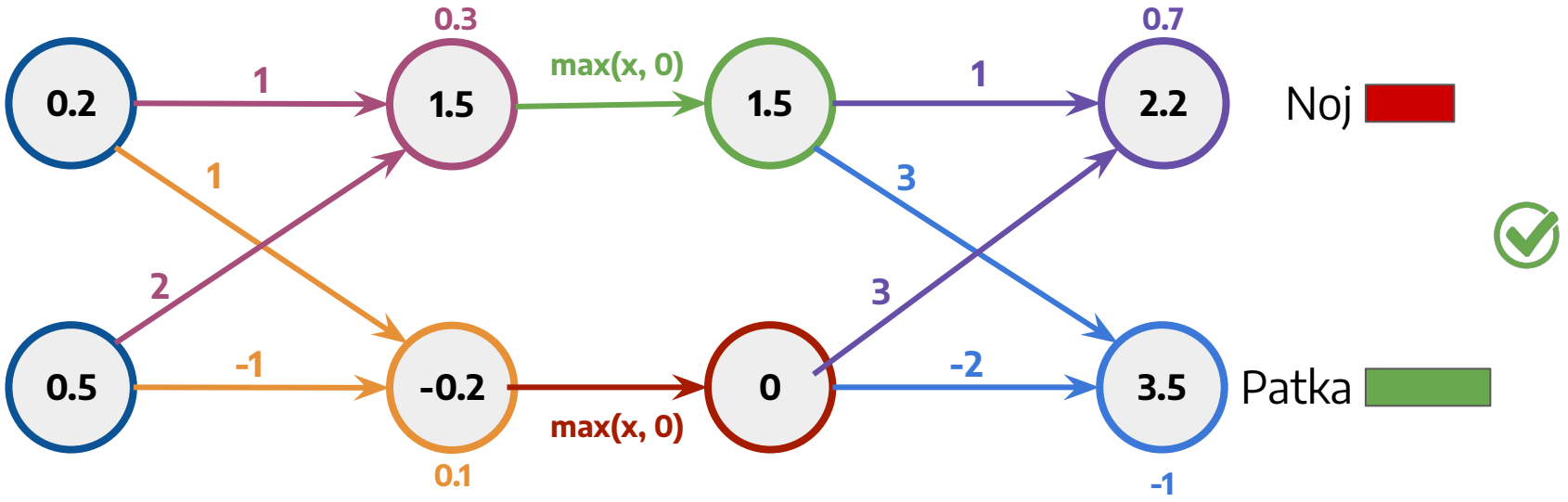
Neuralne mreže: jednostavan primer



Neuralne mreže: jednostavan primer



Neuralne mreže: jednostavan primer



Agenda

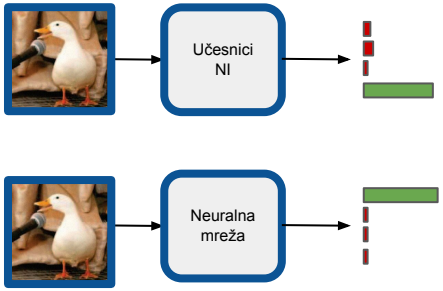
1. Predznanje: neuralne mreže
2. Adversarijalni napadi
3. Verifikacija neuralnih mreža

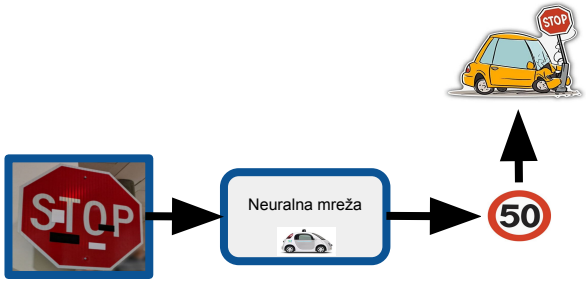
Adversarijalni napadi

- **Adversarijalni (*protivnički*) napad:** proces konstruisanja **adversarijalnog primera x'** na osnovu datog ulaza **x** klase **c**, koji čini da data trenirana mreža pogreši pri klasifikaciji
- Dva tipa:
  - **Ciljani** ("želim da mreža predvidi konkretnu klasu **c' ≠ c**")
  - **Neciljani** ("želim da mreža predvidi bilo koju pogrešnu klasu")
- Raznovrsni po tipu promene, domenu, načinu konstrukcije...
- Veoma česti!

Primeri adversarijalnih napada

- Videli smo već dva primera:





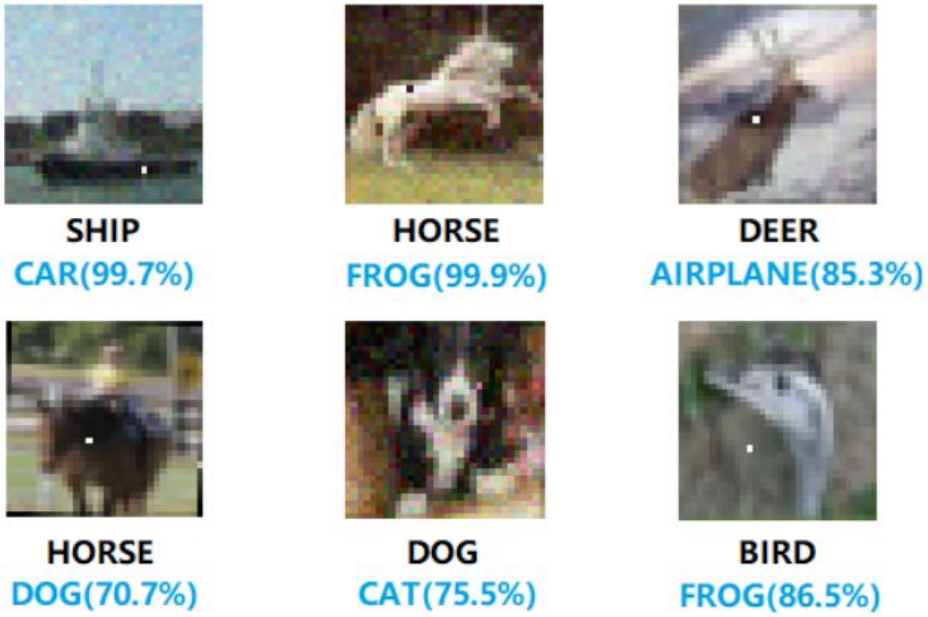
- **Važno:** napad ne sme da suštinski promeni identitet (klasu) ulaznog podatka

Primeri: modni detalji

- Specijalno konstruisane adversarijalne “naočare” potpuno menjaju predviđanje

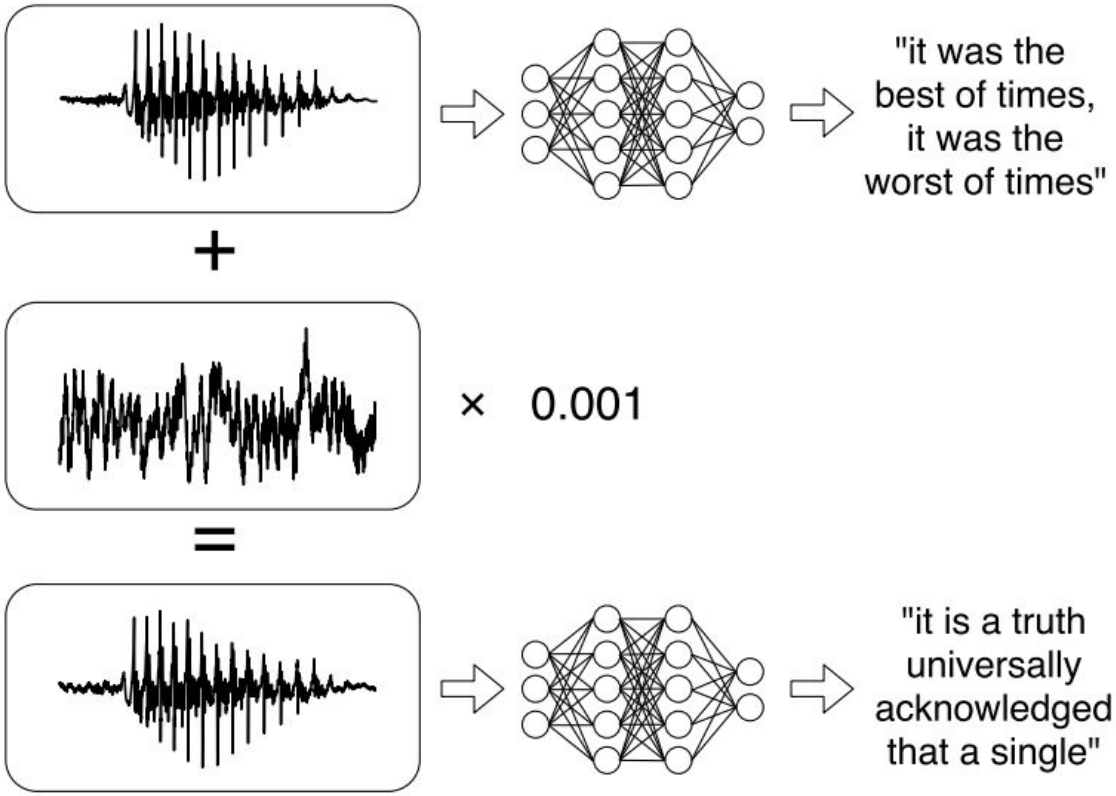
Primeri: one pixel attack

- Dovoljan je samo jedan piksel (ako je moguće proizvoljno ga promeniti)
  - Čak i za slike veće rezolucije poput ImageNet!



Primeri: audio (Mozilla DeepSpeech)

- Dodavanjem tihog šuma u audio fajl model se može naterati da “čuje” bilo šta



Primeri: NLP (question answering)

- Nevažna rečenica na kraju teksta se tretira kao važna i menja odgovor

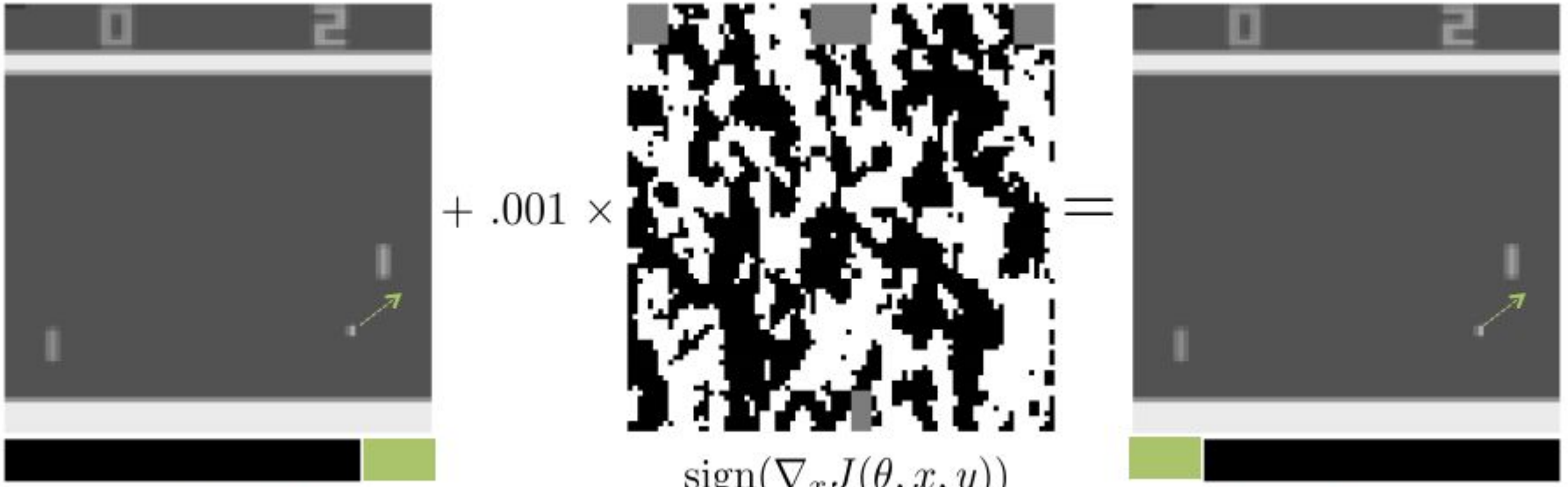
Article: Nikola Tesla
Paragraph: "In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for **Prague** where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses. **Tadakatsu moved to the city of Chicago in 1881.**"

Question: "What city did Tesla move to in 1880?"

Original Prediction: **Prague**
Prediction under adversary: **Chicago**

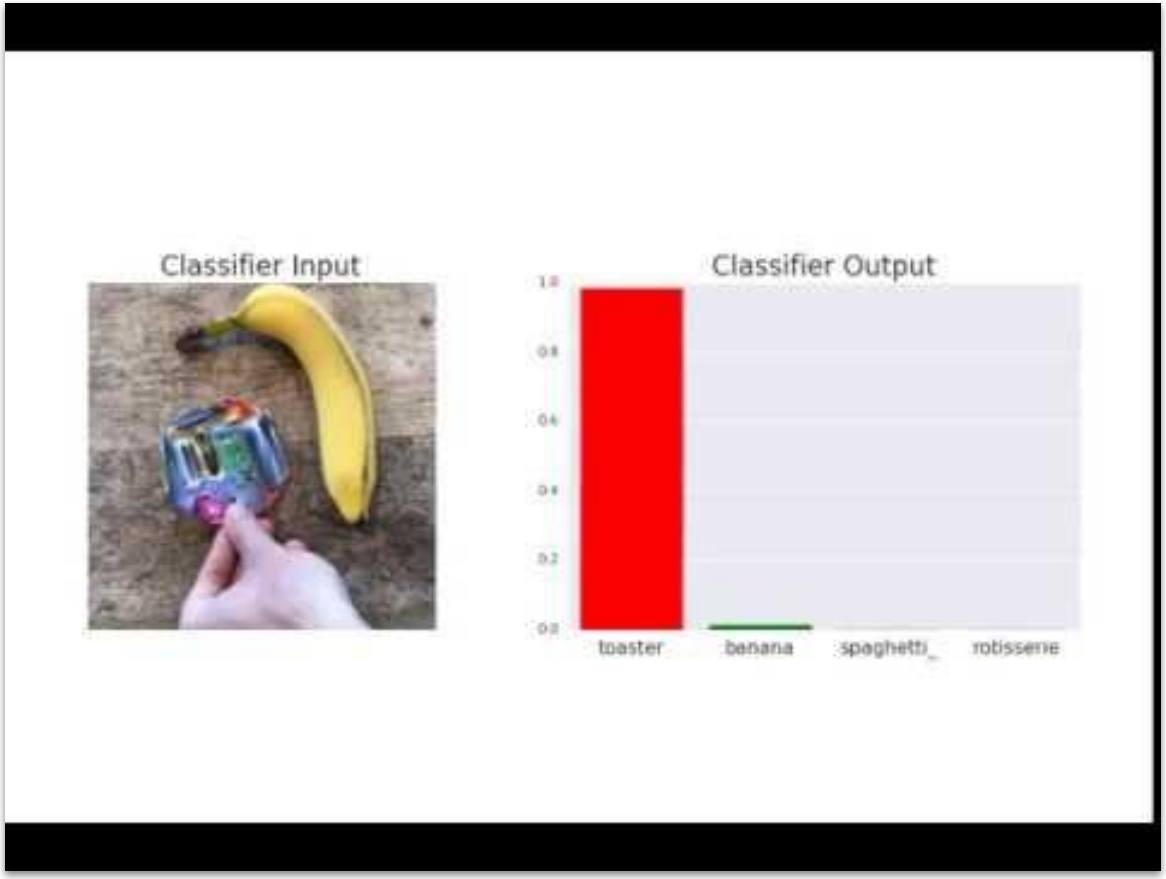
Primeri: igrice (reinforcement learning)

- DQN agent u svakom koraku igre bira jednu od akcija: {up, down, noop}
- Neprimetna perturbacija ekrana tera agenta da odabere **noop** umesto **down**



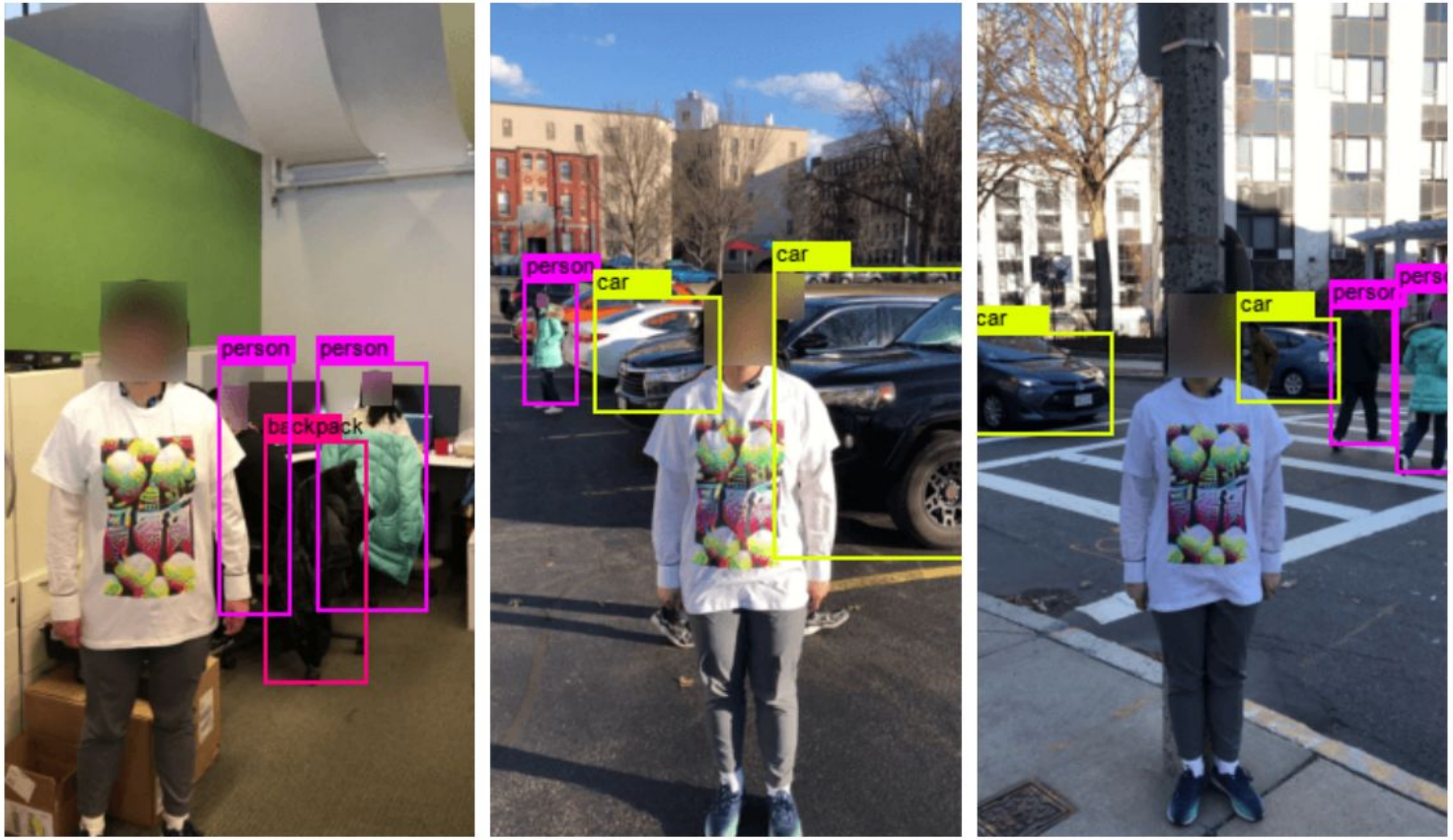
Real-world primeri: sve je toster

- Stiker bilo gde u vidnom polju kamere ⇒ VGG16 klasifikator: “toster”

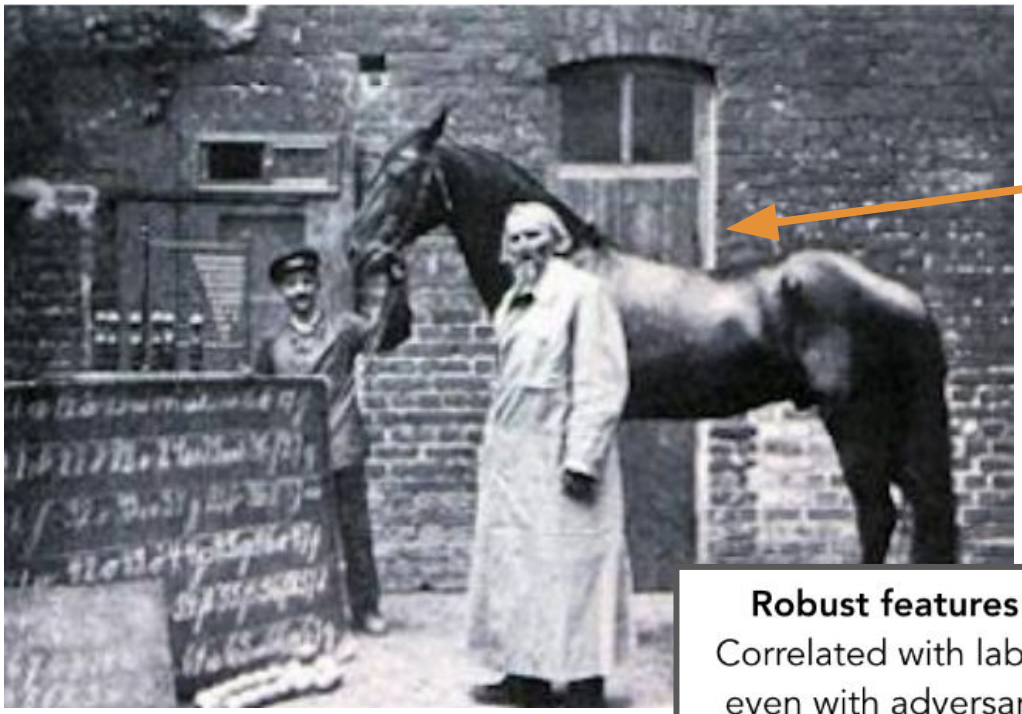


Real-world primeri: majica protiv detekcije

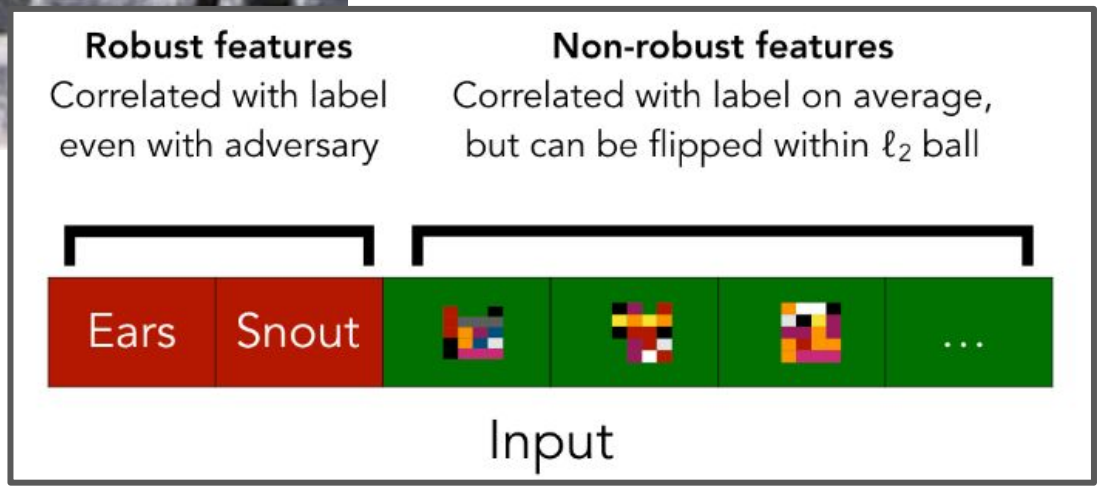
- Napad na state-of-the-art detektore objekata YOLO-v2 i Faster R-CNN



Zašto se sve ovo dešava?

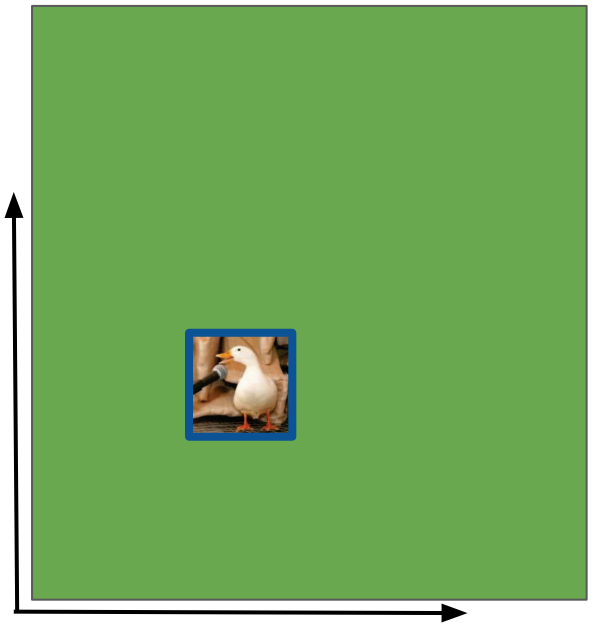


„Mudri Hans”



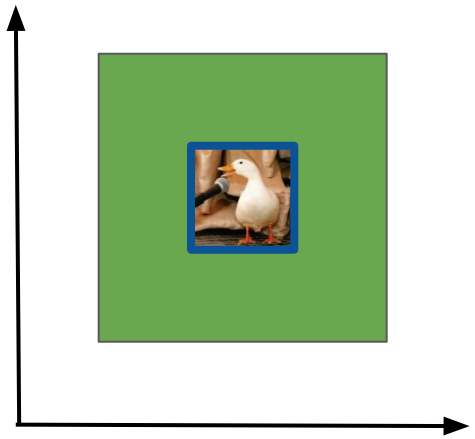
Koraci ka rešavanju problema: robusnost

- **Cilj** treniranja je postići što veću tačnost (**Acc**)
  - Ako neuralne mreže mogu da nađu lakši (ali nepouzdan) način da ostvare ovaj cilj, moramo da promenimo cilj
- Želimo da u evaluaciju uključimo **robusnost** (robusna mreža = tačna za sve **x**)
  - Nepraktično jer je prostor svih ulaza preveliki



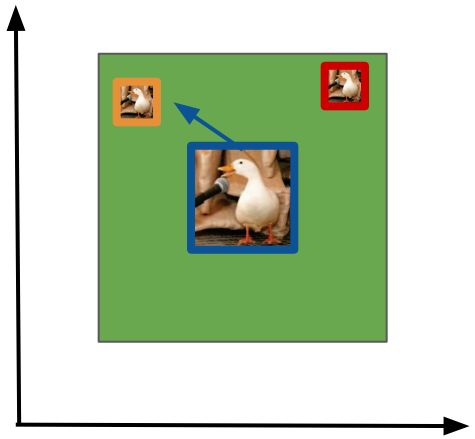
Koraci ka rešavanju problema: robusnost

- **Cilj** treniranja je postići što veću tačnost (**Acc**)
  - Ako neuralne mreže mogu da nađu lakši (ali nepouzdan) način da ostvare ovaj cilj, moramo da promenimo cilj
- Želimo da u evaluaciju uključimo **robusnost** (robusna mreža = tačna za sve **x**)
  - Nepraktično jer je prostor svih ulaza preveliki
- Umesto toga, **lokalna adversarijalna robusnost** (mreža je tačna za sve ulaze **blizu** viđenog ulaza **x**)
- Kako definišemo „blizu”?
  - pretpostavka: ℓ_∞ okruženje sa radijusom ε (svaki piksel se može menjati za najviše ε)

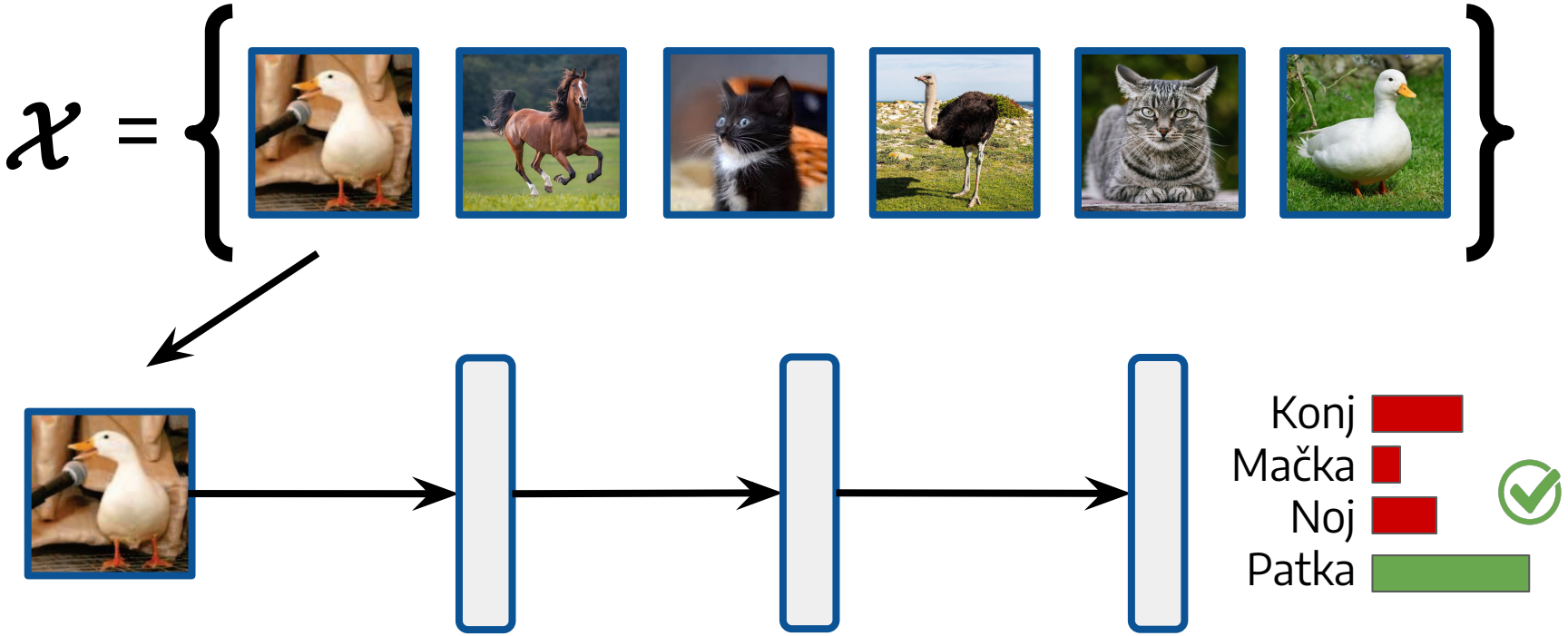


Koraci ka rešavanju problema: robusnost

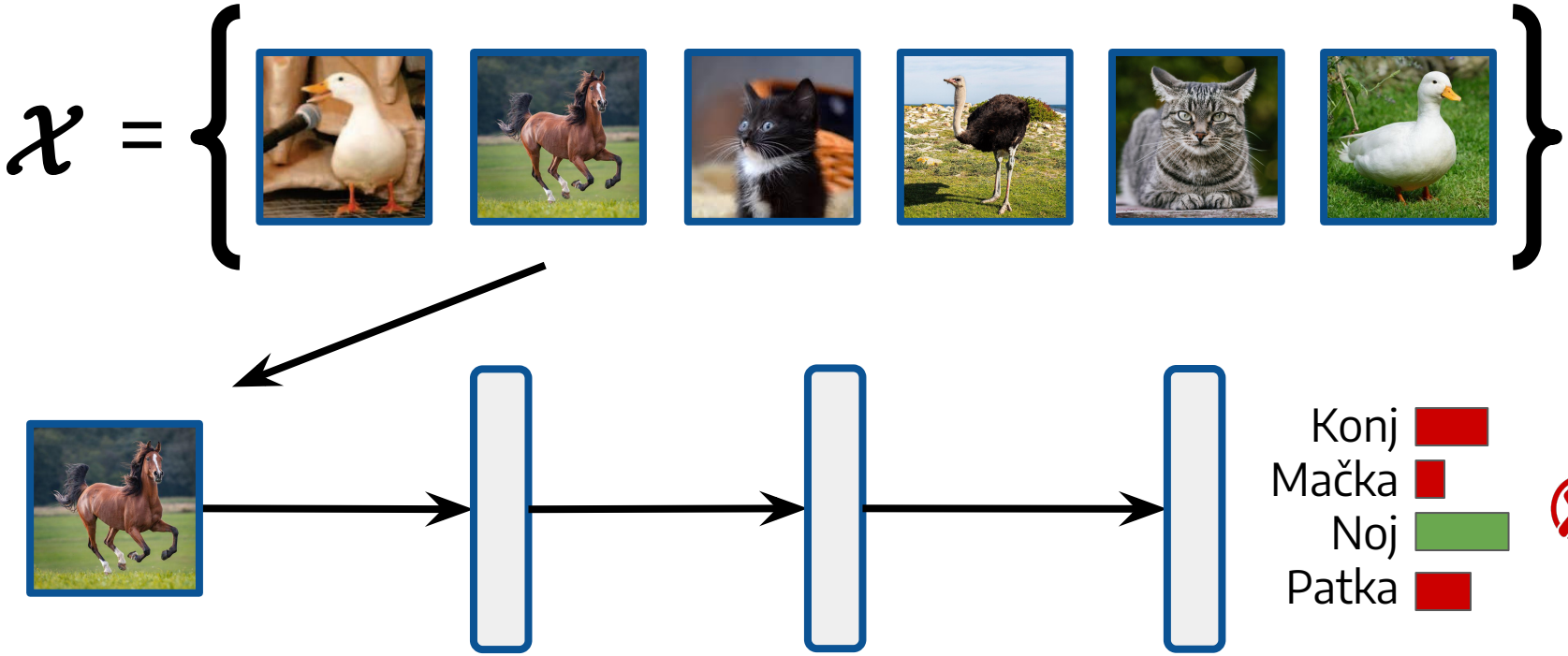
- **Cilj** treniranja je postići što veću tačnost (**Acc**)
  - Ako neuralne mreže mogu da nađu lakši (ali nepouzdan) način da ostvare ovaj cilj, moramo da promenimo cilj
- Želimo da u evaluaciju uključimo **robusnost** (robusna mreža = tačna za sve **x**)
  - Nepraktično jer je prostor svih ulaza preveliki
- Umesto toga, **lokalna adversarijalna robusnost** (mreža je tačna za sve ulaze **blizu** viđenog ulaza **x**)
- Kako definišemo „blizu”?
  - pretpostavka: ℓ_∞ okruženje sa radijusom ε (svaki piksel se može menjati za najviše ε)
- Kako naći najnepovoljniji ulaz u okruženju?
  - pretpostavka: PGD napad! (projektovani gradijentni spust)
- **Novi cilj:** postići visoku tačnost (**Acc**) i visoku **robusnost** (**Adv**)



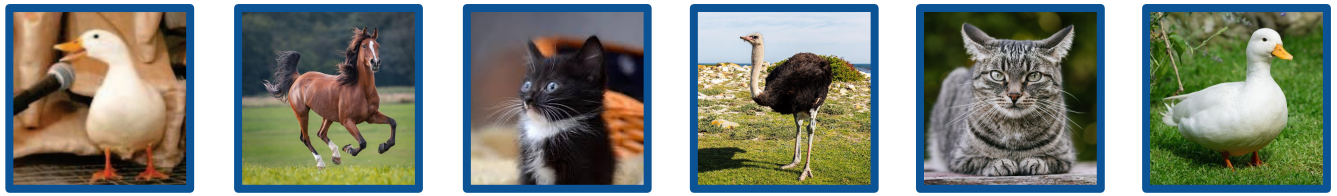
Nova evaluacija: tačnost + robusnost



Nova evaluacija: tačnost + robusnost

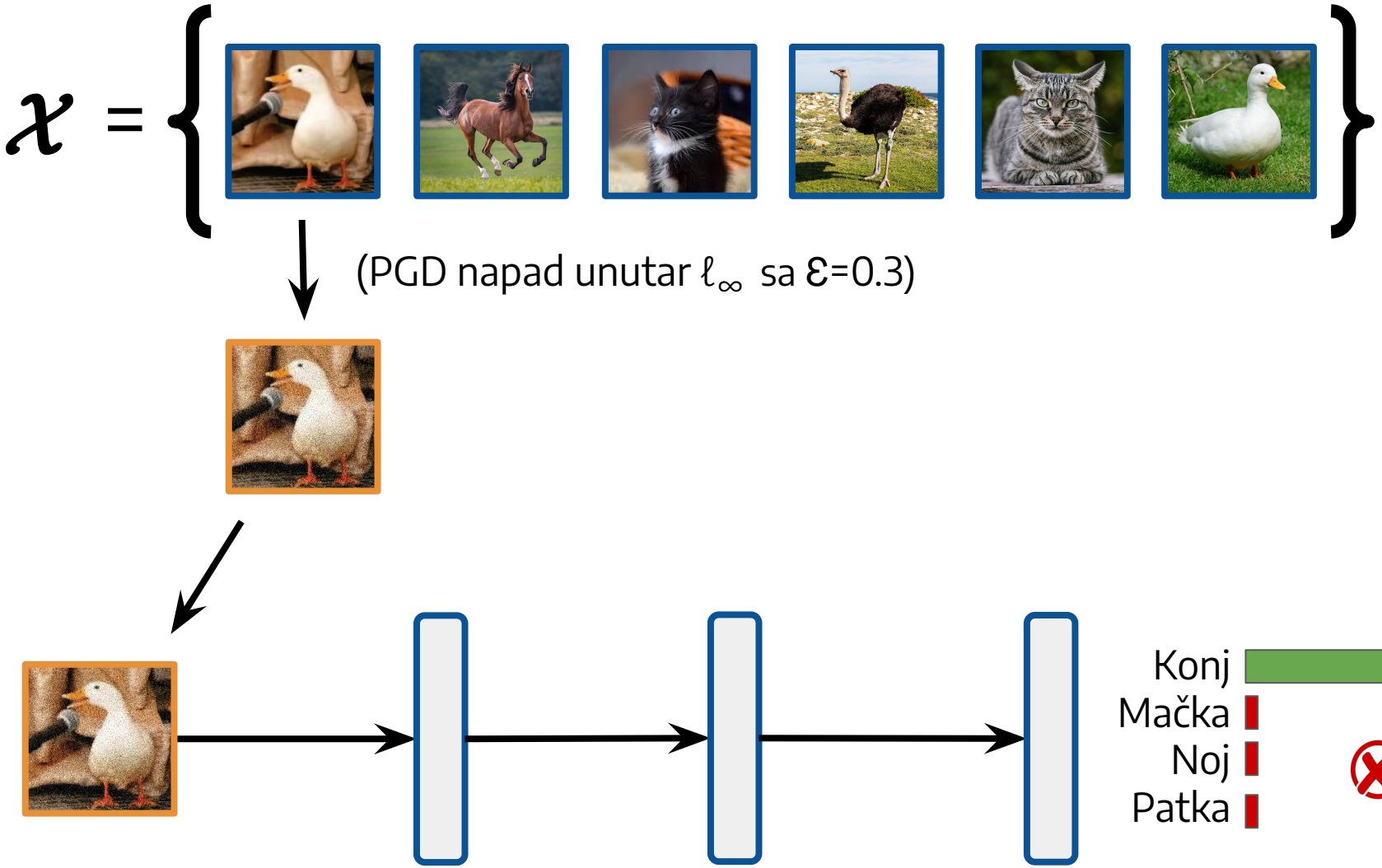


Nova evaluacija: tačnost + robusnost

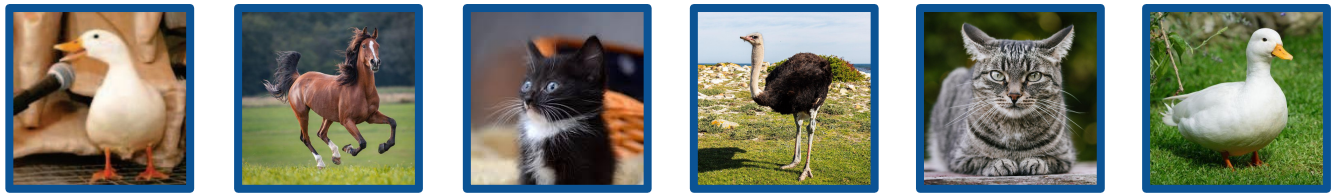
$$x = \{$$  $$\}$$

ACC = 83.3%

Nova evaluacija: tačnost + robusnost



Nova evaluacija: tačnost + robustnost

$$x = \{$$  $$\}$$

Standard: ✓ ✗ ✓ ✓ ✓ ✓

PGD: ✗ ✗ ✗ ✗ ✓ ✗

ACC = 83.3%

ADV = 16.7%

- Kao što bismo očekivali, **ADV** će za većinu mreža biti izuzetno loš
- Poboljšali smo kriterijume, ali kako ih zapravo dostići?

Adversarijalna odbrana

- Uključivanje napadača u proces treniranja
  - U našem slučaju, korišćenje PGD kako bismo našli “nepovoljne” ulaze blizu naših ulaza u toku treniranja
- Rezultat:

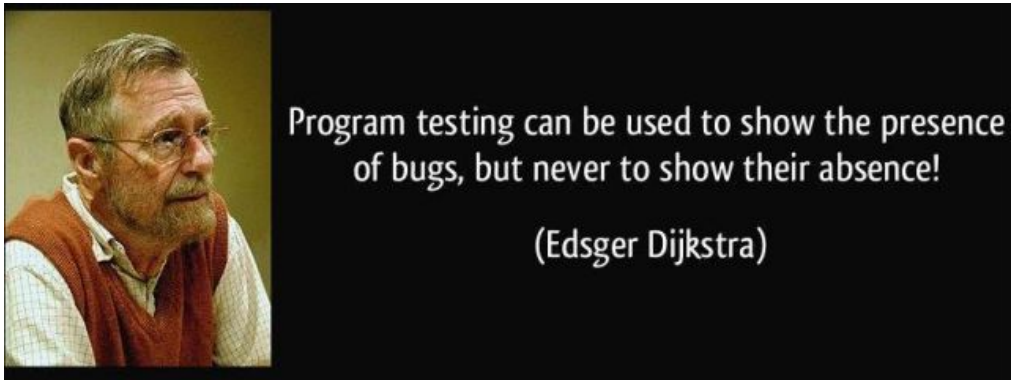
ACC = 81.9%

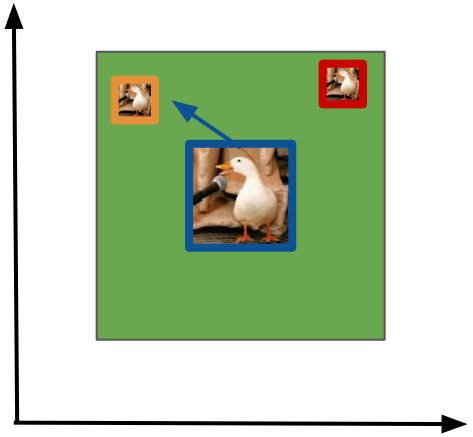
ADV = 76.3%

- Tačnost blago opada, robusnost postaje izuzetno dobra
- Problem rešen? Da li sada imamo robustan, pouzdan, model?

Vrzino kolo napada i odbrana

- Ne! Imamo robustan i pouzdan model pod pretpostavkom napada
  - Odbrana često radi u praksi, ali...
  - ...u ovom okruženju i dalje može da postoji ulaz koji će da prevari mrežu
- Mreže nisu pouzdane, adversarijalni napad! **ADV↘**
  - Adversarijalna odbrana! **ADV↗**
    - Šta ako se pojavi bolji napad? **ADV↘**
      - Bolja odbrana! **ADV↗**
        - Još bolji napad! **ADV↘**
          - Još bolja odbrana! **ADV↗**
            - ...
              - Kako prekinuti?



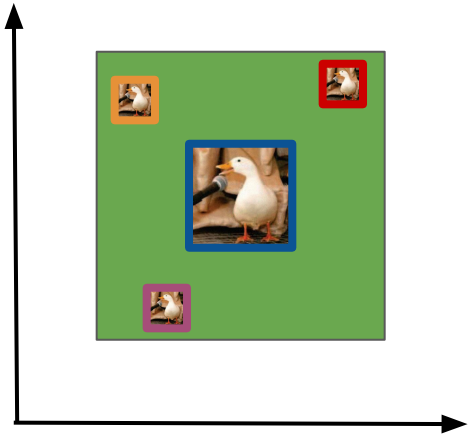


Agenda

1. Predznanje: neuralne mreže
2. Adversarijalni napadi
3. Verifikacija neuralnih mreža

Verifikacija neuralnih mreža

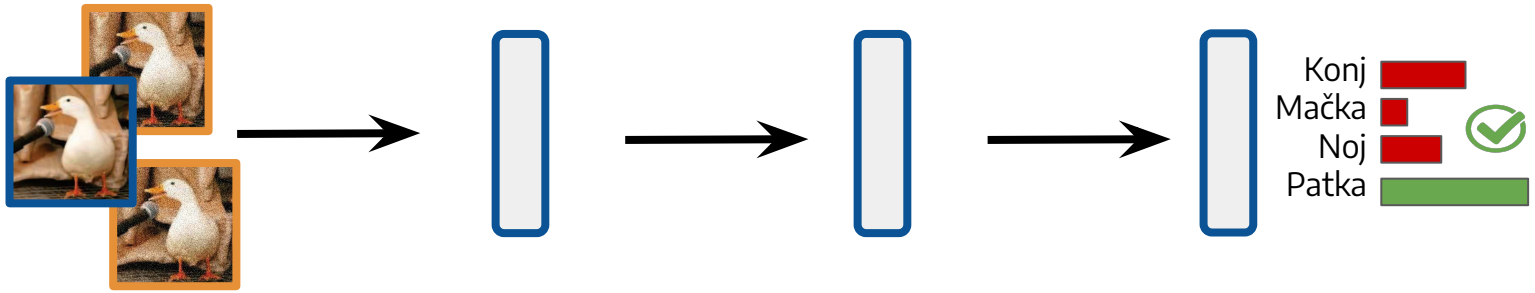
- Da li možemo da **matematički dokažemo** robusnost i **garantujemo** da je mreža sigurna?
  - **Konkretno**: da li možemo da dokažemo da za sve ulaze u odabranom okruženju mreža daje tačan izlaz?
- **Problem:** u okruženju postoji neprebrojivo mnogo slika
- **Jedno od rešenja:** konveksne aproksimacije



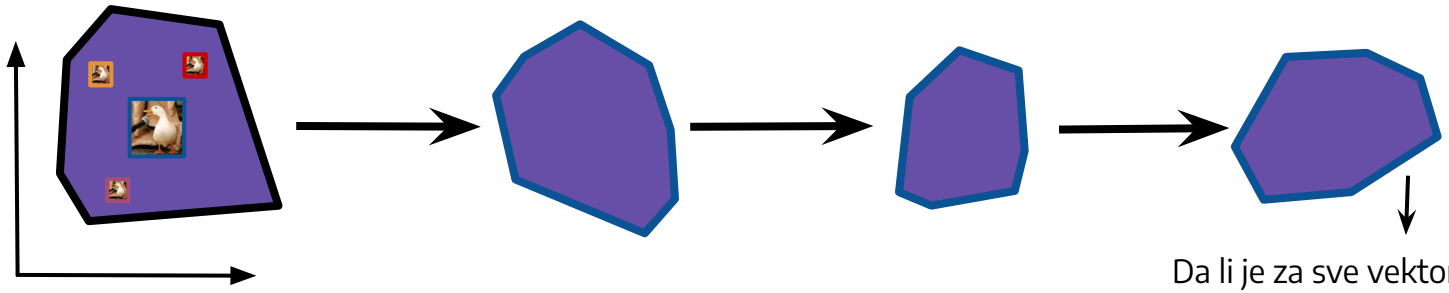
Konveksne aproksimacije

- **Problem:** u okruženju postoji neprebrojivo mnogo slika
- **Ključna ideja:** sumarizujemo ih u **simbolički, konačan region** i računamo efekat slojeva mreže na taj region
  - „Koje sve vrednosti su moguće nakon ovog sloja mreže?”
- Analizom finalnog regiona dokazujemo robusnost

Konkretno:

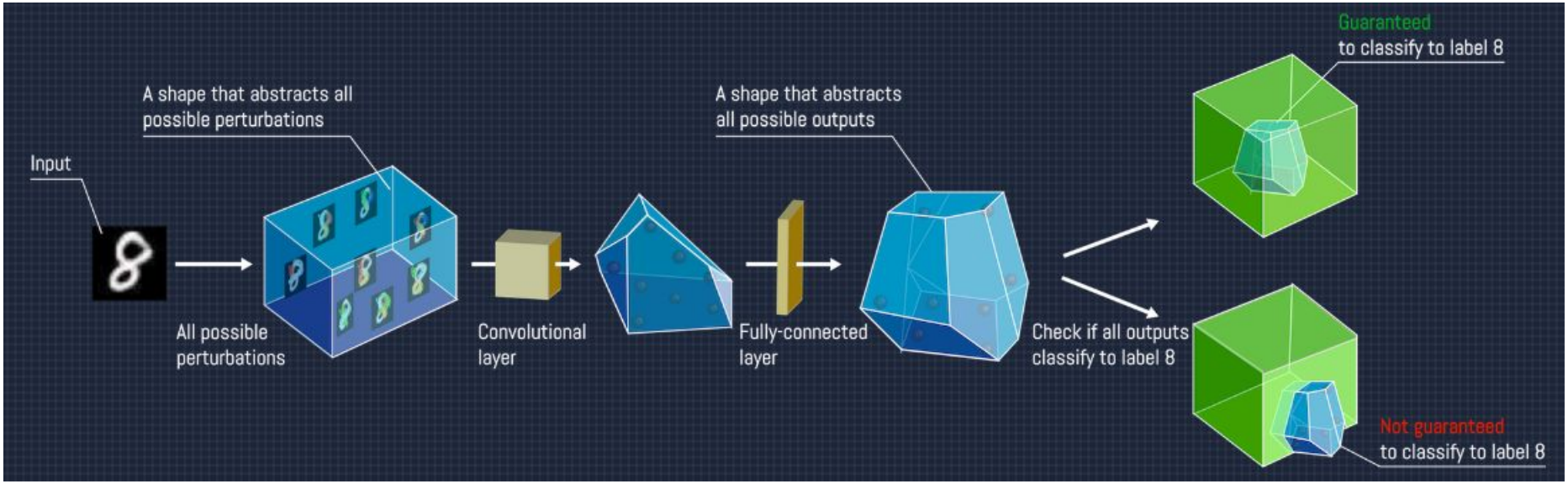


Apstraktno:



Da li je za sve vektore skorova u ovom regionu skor za klasu „patka” najveći?

Konveksne aproksimacije



Konveksne aproksimacije

- Šta je potrebno kako bi ovo zapravo funkcionisalo?
  1. Odabir adekvatne reprezentacije za regione
  2. Način da izračunamo efekat svakog sloja na neki region (*transformer*)
     - Idejno jednostavno, u praksi često tehnički zahtevno
- Sve aproksimacije unose „nemoguće” tačke u regione, zbog čega algoritam verifikacije može da vrati negativan odgovor čak i kada svojstvo važi
  - **Tradeoff:** bolje (tesnije) aproksimacije su spore

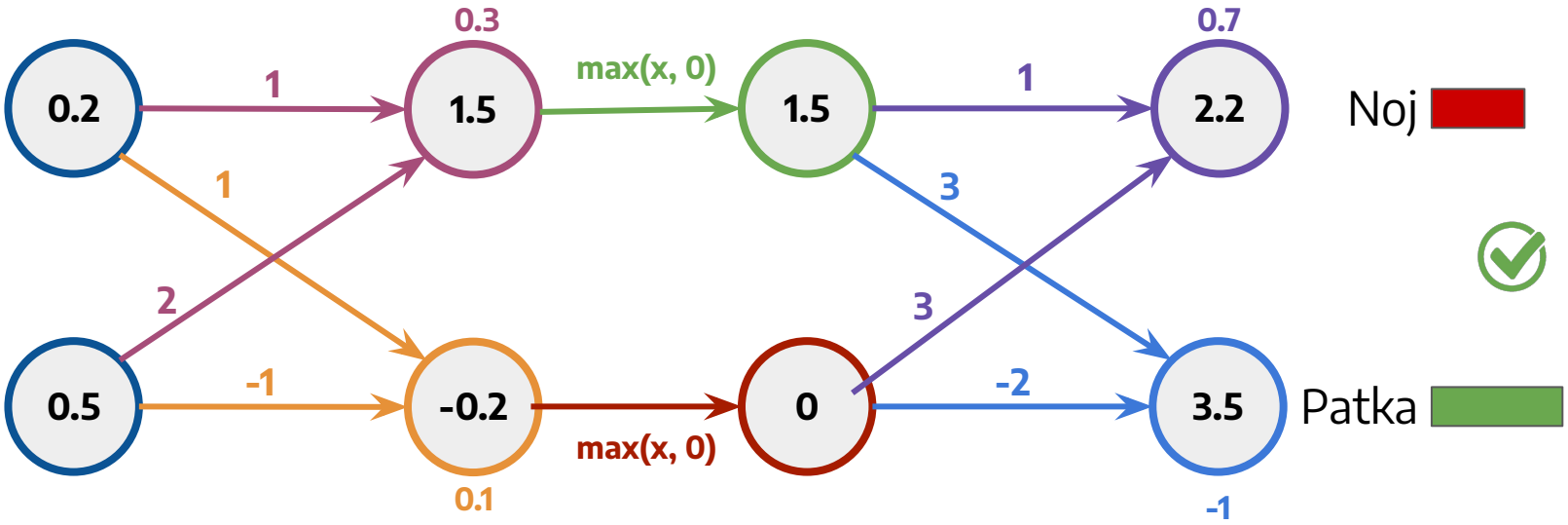
Box aproksimacija

- Jednostavna aproksimacija: **Box** (*intervalni račun*)
- Za svaku vrednost čuvamo dva broja: gornju i donju granicu
- "Za bilo koji ulaz u okruženju **x**, garantujemo da je izlaz neurona **j** u sloju **i** uvek u intervalu **[L, U]**"

- **Primer:** x je u [1, 2], y je u [-3, 4], x+y je u ???
  - x + y = [1, 2] + [-3, 4] je u [-2, 6]
- **Primer:** x je u [5, 7], x-x je u ???
  - x - x = [5, 7] - [5, 7] je u [-2, 2]
- **Pitalica:** x je u [-2, 4], y je u [1, 3], 2x-y je u ???
  - 2x - y = 2*[-2,4] - [1, 3] = [-4, 8] - [1, 3] = [-7, 7]

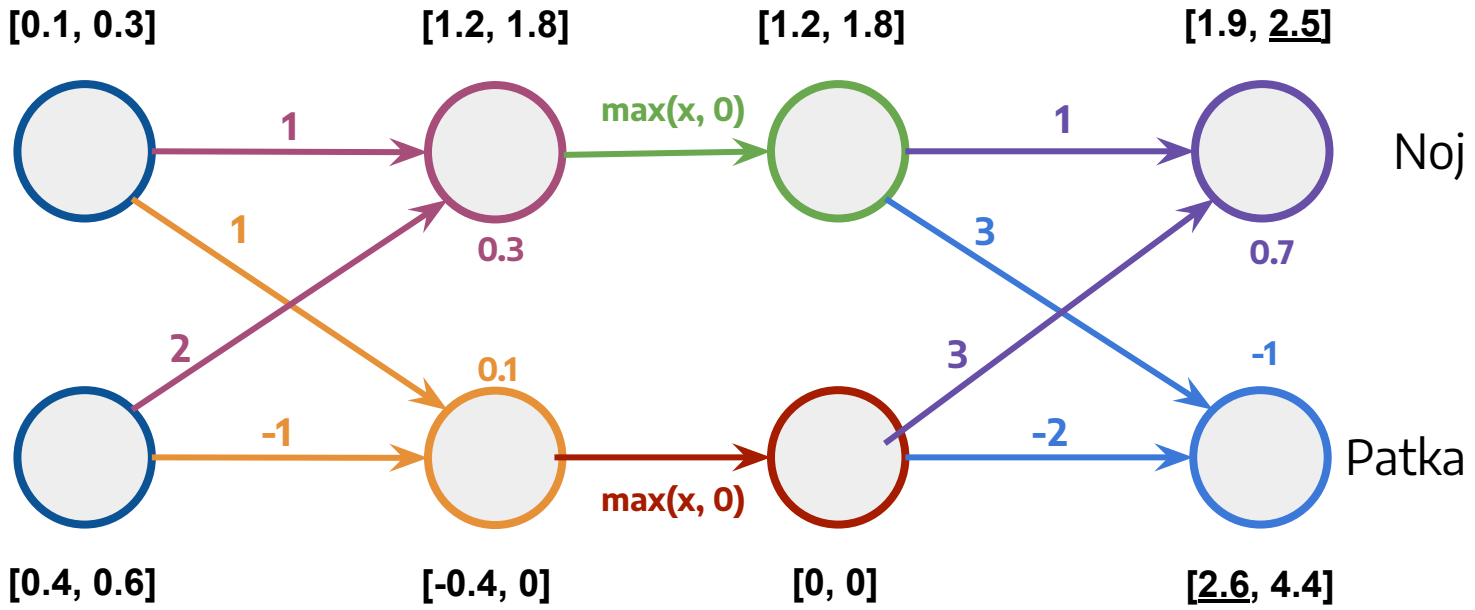
Box: verifikacija robusnosti

- Koristićemo primer od ranije: **x** = [0.2, 0.5], i mreža daje tačan odgovor „Patka”



Box: verifikacija robusnosti

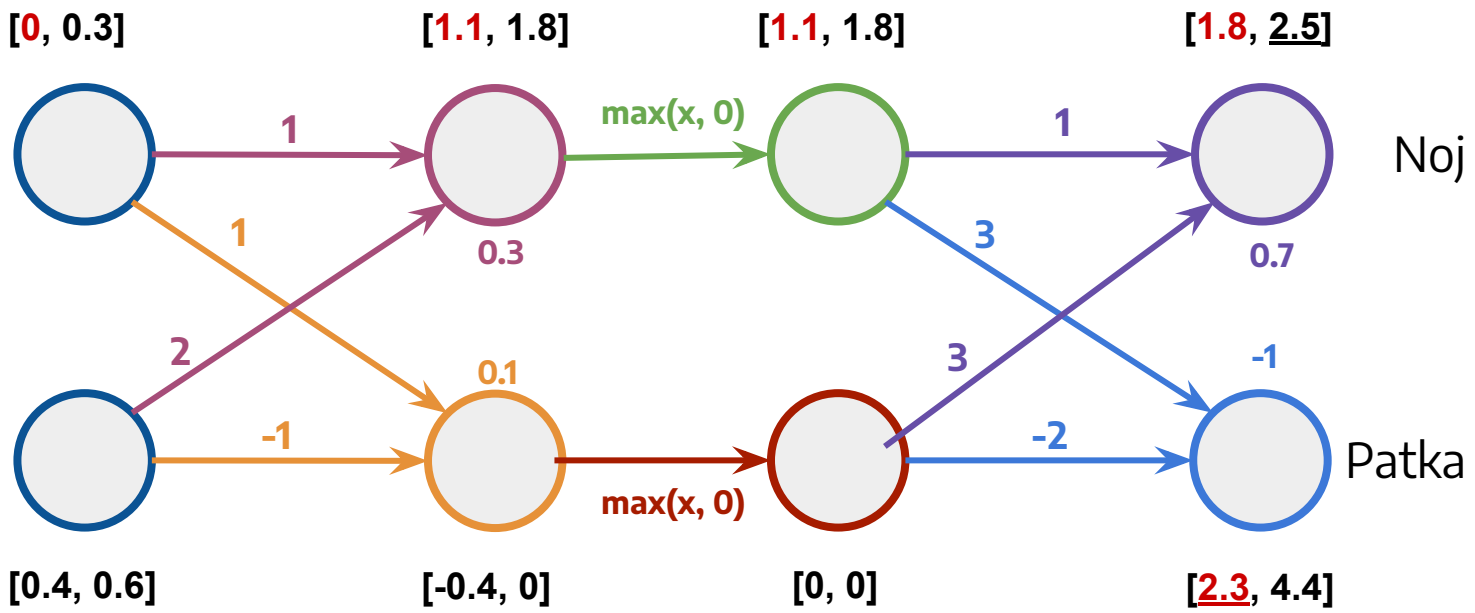
- Koristićemo primer od ranije: **x** = [0.2, 0.5], i mreža daje tačan odgovor „Patka”
- Želimo da verifikujemo da za sve vrednosti u ℓ_∞ okruženju oko **x** sa ε=0.1, neuralna mreža daje predviđanje „Patka”



- Kako je **2.6>2.5**, za svako **x’** u okruženju **x** važi da je **skor(Patka) > skor(Noj)**
  - ⇒ uspešno smo verifikovali da je mreža robusna!

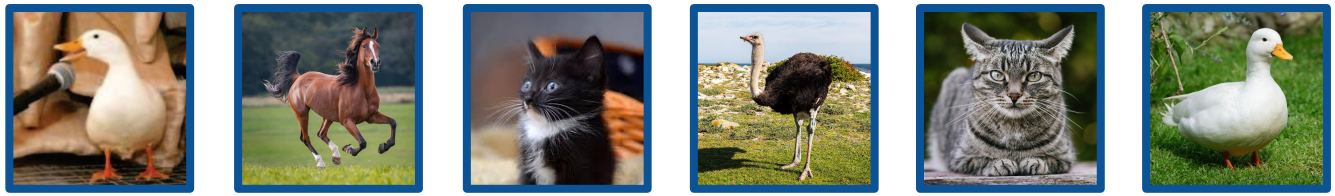
Box: verifikacija robusnosti

- Šta ako bismo imali **[0, 0.3]** kao opseg za prvi neuron ulaza umesto **[0.1, 0.3]**?



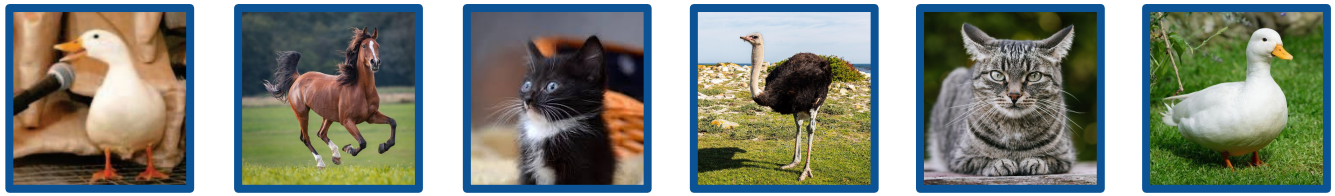
- Kako **2.3>2.5** ne važi, nismo uspeli da verifikujemo da je mreža robusna!
  - (možda zaista nije, a možda intervalni račun unosi preveliku grešku)

Nova evaluacija: tačnost, robusnost, verifikabilnost

$$x = \{$$  $$\}$$

ACC = 83.3%

Nova evaluacija: tačnost, robustnost, verifikabilnost

$$x = \{$$  $$\}$$

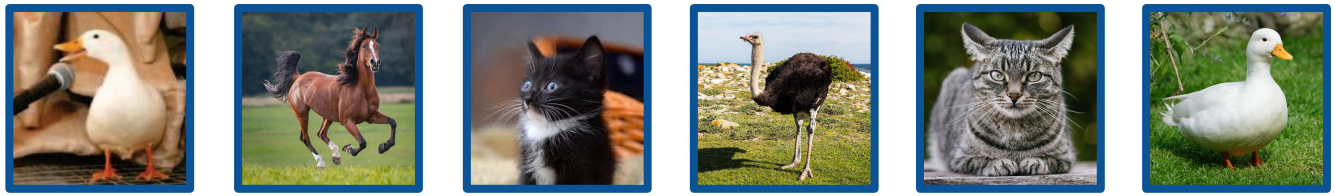
Standard: ✓ ✗ ✓ ✓ ✓ ✓

PGD: ✗ ✗ ✗ ✗ ✓ ✗

ACC = 81.9%

ADV = 16.7%

Nova evaluacija: tačnost, robusnost, verifikabilnost

$$x = \{$$  $$\}$$

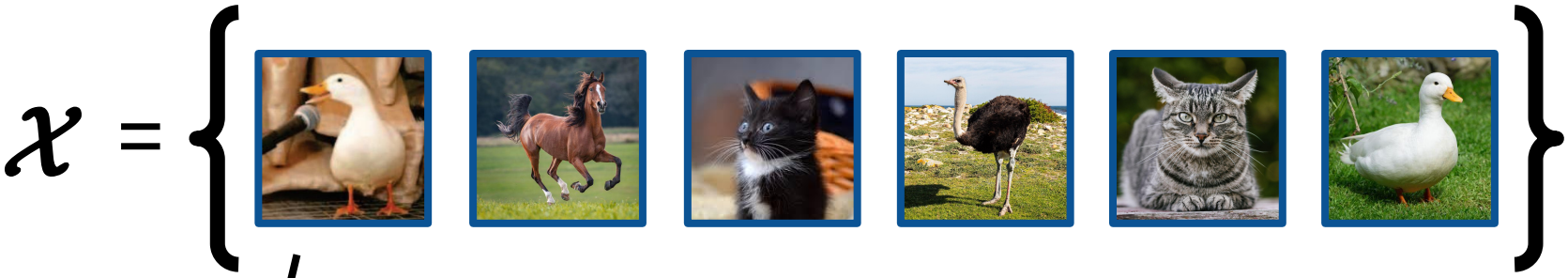
Standard: ✓ ✗ ✓ ✓ ✓ ✓

PGD: ✓ ✗ ✓ ✗ ✓ ✓

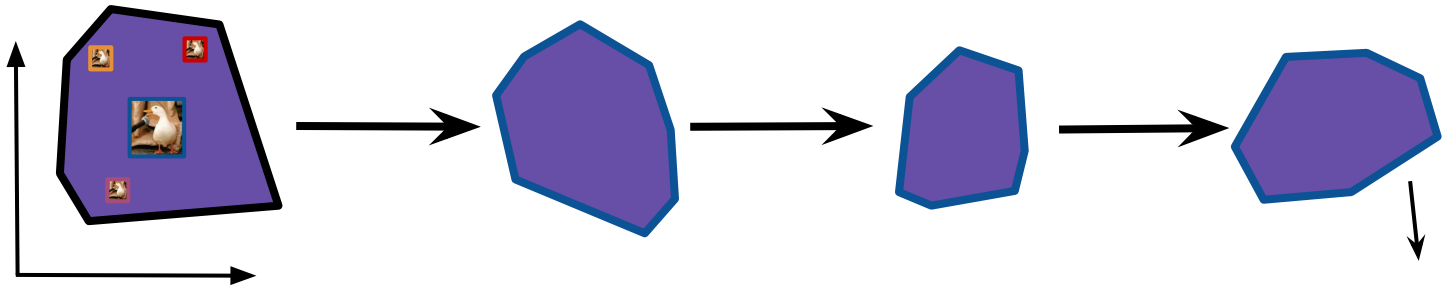
ACC = 81.9%

ADV = 76.3%

Nova evaluacija: tačnost, robusnost, verifikabilnost

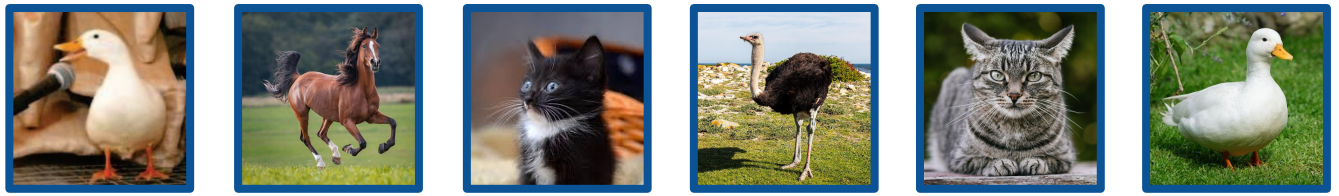


(verifikacija putem konveksnih aproksimacija unutar ℓ_∞ sa ε=0.3)



Da li je za sve vektore skorova u ovom regionu skor za klasu „patka” najveći?

Nova evaluacija: tačnost, robusnost, verifikabilnost

$x =$ 

Standard:	✓	✗	✓	✓	✓	✓
PGD:	✓	✗	✓	✗	✓	✓
Verified:	✗	✗	✗	✗	✗	✗

ACC = 81.9%

ADV = 76.3%

VER = 0%

- Slično kao ranije, **VER** je izuzetno loš za većinu mreža...
- Poboljšali smo kriterijume ponovo, ali kako ih ponovo dostići?

Dokaziva adversarijalna odbrana

- Uključivanje **propagacije konveksnih aproksimacija** u proces treniranja
- Rezultat:

ACC = 80.4%

ADV = 75.3%

VER = 62.1%

- Značajno bolja verifikabilnost: za 62.1% ulaznih primera možemo da **garantujemo** da mreža daje tačan odgovor za sve dozvoljene perturbacije
- **Glavni izazov:** skalabilnost tesnijih aproksimacija

Rezime

- Moderne neuralne mreže **nisu pouzdane**, i podložne su raznim vrstama **adversarijalnih napada** **ACC↗ ADV↘**
  - **Adversarijalne odbrane** (uključivanje napadača u proces treninga) često rade dobro u praksi, ali ne daju garancije **ACC↗ ADV↗**

- Korišćenjem **konveksnih aproksimacija** možemo **verifikovati** da data mreža ne greši ni za jedan ulaz u okruženju nekog ulaza **x**
- Mreže trenirane standardno ili korišćenjem adversarijalne odbrane je izuzetno teško verifikovati **ACC↗ ADV↗ VER ↘**
  - **Dokazive adversarijalne odbrane** (uključivanje konveksnih relaksacija i verifikacije u proces treninga) dovode do **dokazivo robusnih mreža** **ACC↗ ADV↗ VER ↗**

Resursi

- Predavanje Petra Veličkovića na temu neuralnih mreža (NI v3.0): ni.mg.edu.rs/static/resources/v3.0/sre1_neuralne_pv.pdf
- ETH kurs iz sigurne i pouzdane veštačke inteligencije: sri.inf.ethz.ch/teaching/riai2020
  - Snimci svih predavanja iz 2020. su javno dostupni na Youtube: youtu.be/playlist?list=PLWjm4hHpaNg6c-W7JjNYDEC_kJK9oSp0Y
  - Najnoviji radovi i predavanja SRI grupe sa ETH na ovu temu: safeai.ethz.ch
- DALL·E (Pikaču u odelu koji šeta psa): openai.com/blog/dall-e/
  „Tri talasa veštačke inteligencije": youtu.be/watch?v=-O01G3tSYpU
- Prvi radovi koji su primetili adversarijalne napade: arxiv.org/abs/1312.6199 i arxiv.org/abs/1412.6572
- Primeri adversarijalnih napada
  - Saobraćajni znaci: arxiv.org/abs/1707.08945
  - Modni detalji: cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf
  - One pixel attack: arxiv.org/abs/1710.08864
  - Audio napad na Mozilla DeepSpeech prepoznavanje govora: arxiv.org/abs/1801.01944
  - Napad na NLP question answering sistem: arxiv.org/abs/1707.07328
  - Reinforcement learning (Pong): arxiv.org/abs/1702.02284
  - Sve je toster: arxiv.org/abs/1712.09665
  - Majica protiv detekcije: arxiv.org/abs/1910.11099

Hvala na pažnji!

Pitanja?