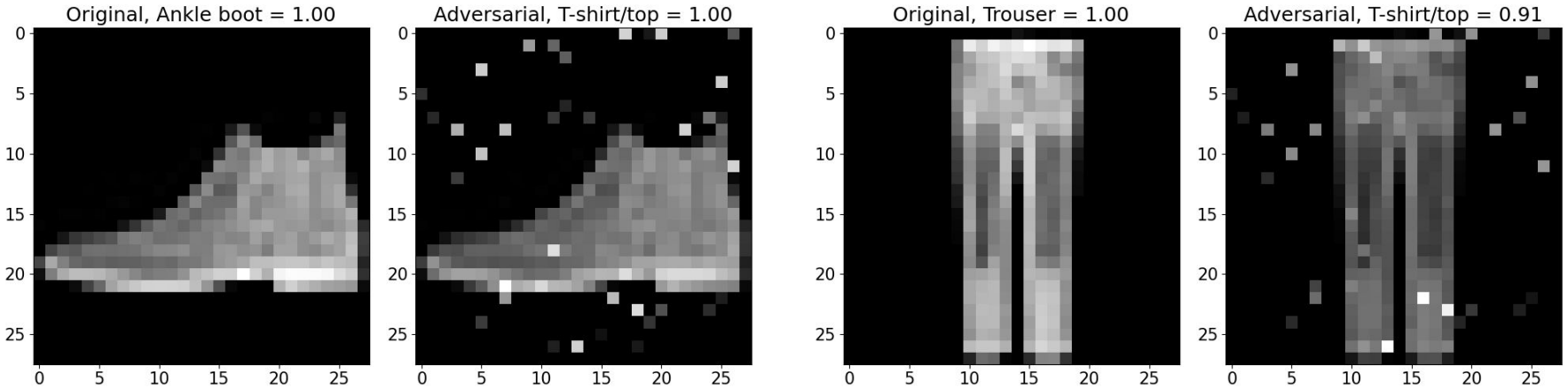
Evoluční návrh konfliktních příkladů pro neuronové sítě

Aplikované evoluční algoritmy

Bc. Ladislav Ondris (xondri07)

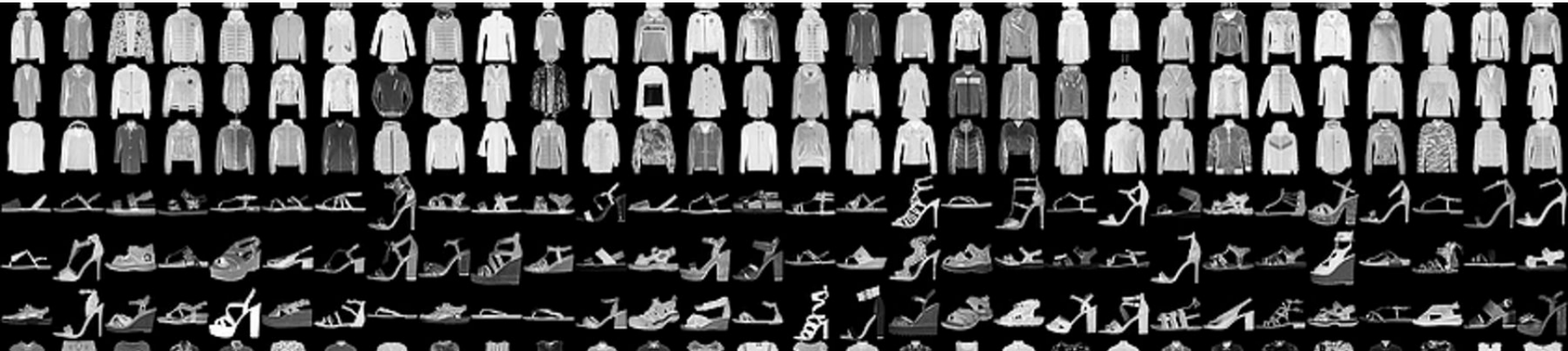
Popis

- Cílem je donutit konvoluční NN klasifikovat objekt do konkrétní třídy
- Hledání změn v obraze provádí genetický algoritmus



Dataset

- **Fashion MNIST**
- 60k trénovacích a 10k testovacích obrázků
- 10 tříd objektů
- Rozlišení obrázku je 28x28

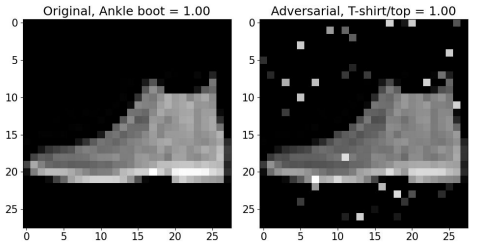


Model

- Jednoduchá konvoluční neuronová síť
- Natrénovaná na datasetu Fashion MNIST na 90% úspěšnost
- 622 tisíc trénovatelných parametrů
- Je použit pouze jako černá skříňka

Algoritmus

- Inspirován článkem **POBA-GA** [1]
- Implementace genetické algoritmu použita z knihovny **PyGad** [2]
- K hledání cílového vzoru je použito podmnožina **100 obrázků** datasetu kvůli časové náročnosti
- Populace se skládá z množiny 2D obrázků obsahující změny, které jsou přičteny k trénovacím obrázkům ve fitness funkci

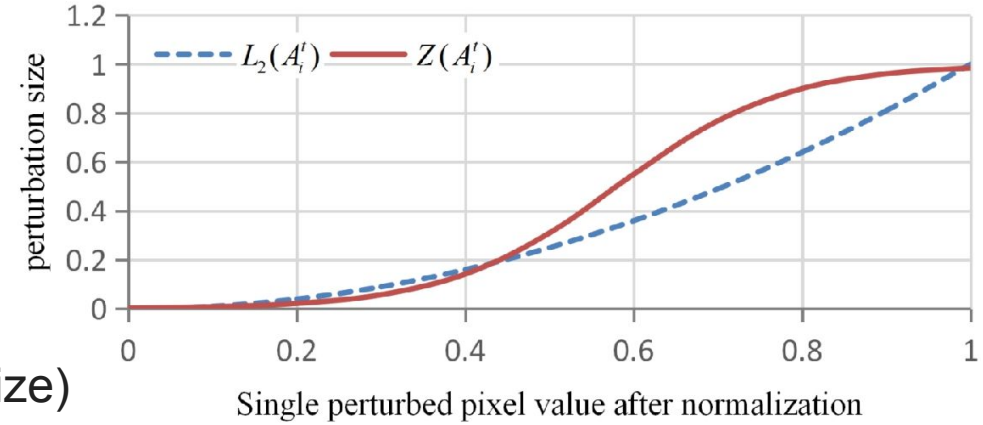


[1] J. Chen, M. Su, S. Shen, H. Xiong, and H. Zheng. POBA-GA: perturbation optimized black-box adversarial attacks via genetic algorithm. CoRR, abs/1906.03181, 2019. URL http://arxiv.org/abs/1906.03181.
[2] https://pygad.readthedocs.io/en/latest/

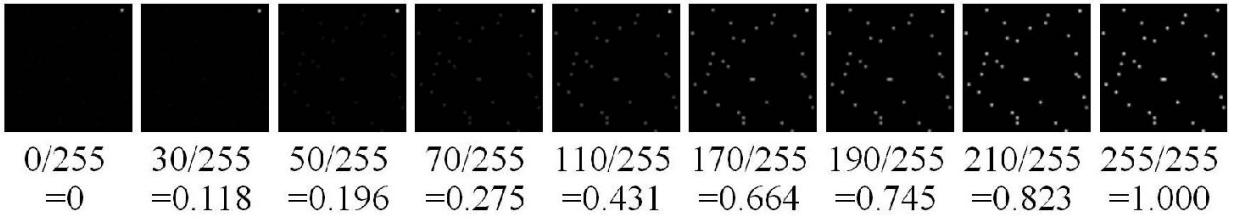
Funkce fitness

Ohodnocuje dvě věci:

- jak velká je pravděpodobnost třídy, která je cílem útoku, a
- jak znatelné jsou změny v obraze.



Fitness = 1 / (loss + α * perturbation_size)



Experimenty obecně

- Vždy 30 opakování
- K běhu použito Metacentrum kvůli časové náročnosti

Experimenty se zaměřovaly na různé parametry GA:

- Velikost populace
- Pravděpodobnost křížení
- Pravděpodobnost mutace
- Počet mutujících genů
- Hodnota fitness funkce na počet generací

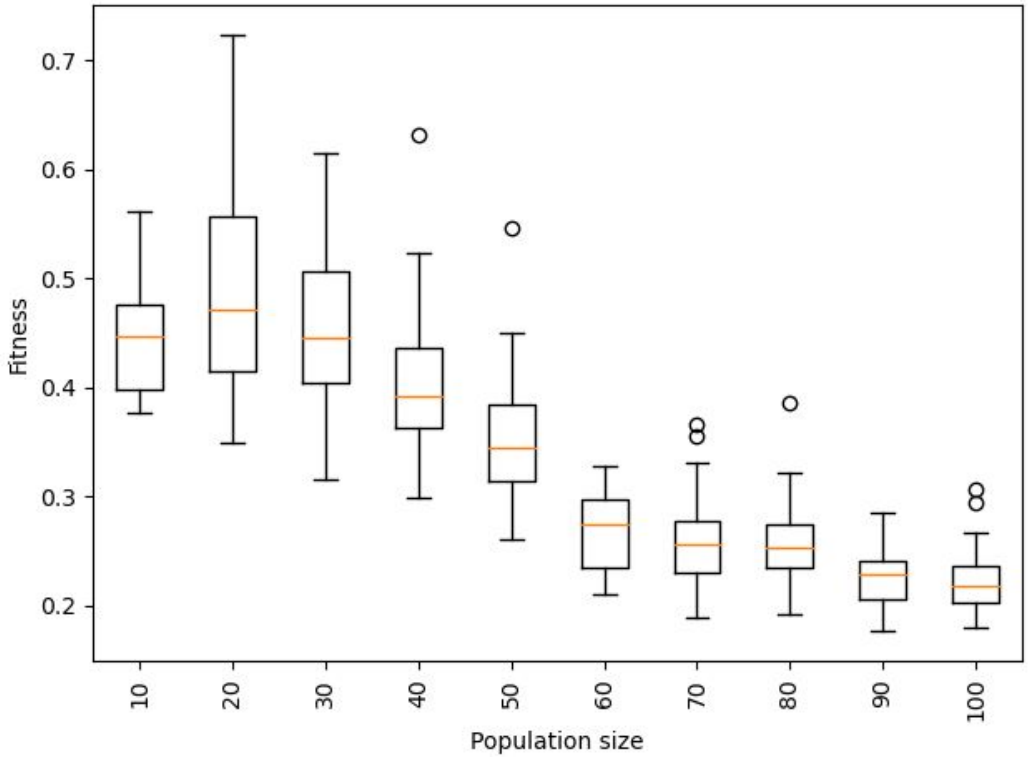
Experiment - Velikost populace

Férový experiment:

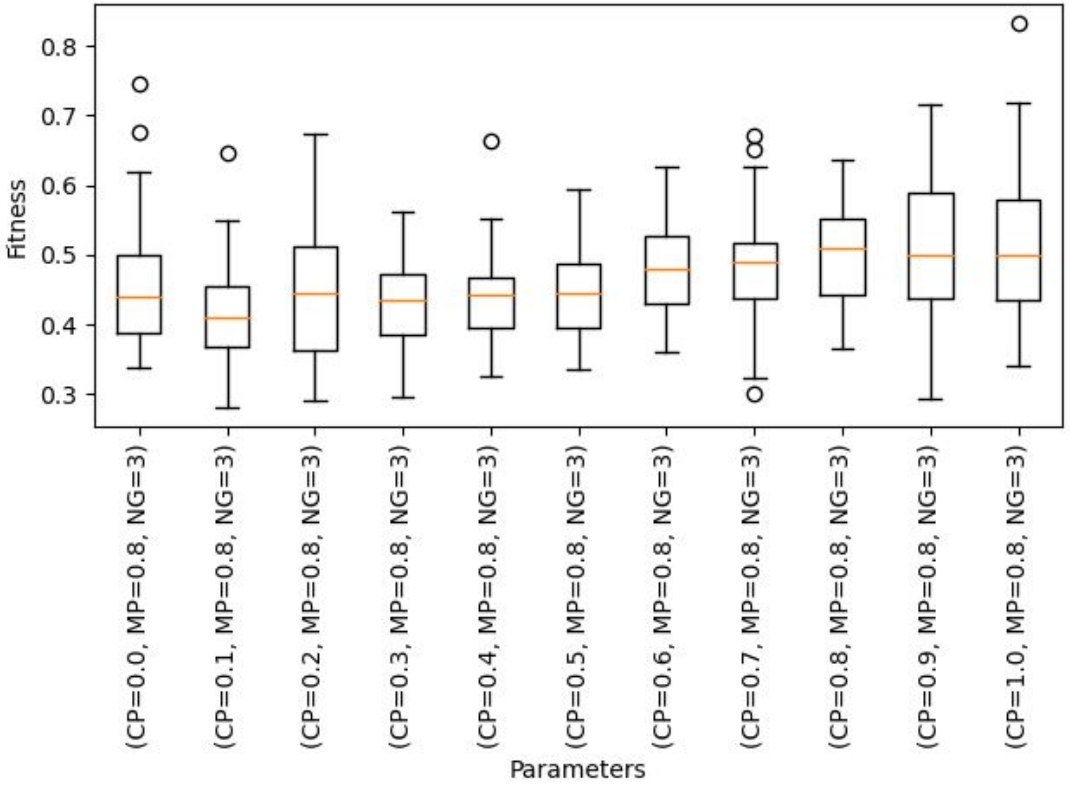
- Jedinců 10, generací 40
- Jedinců 20, generací 20
- Jedinců 100, generací 4

Parametry:

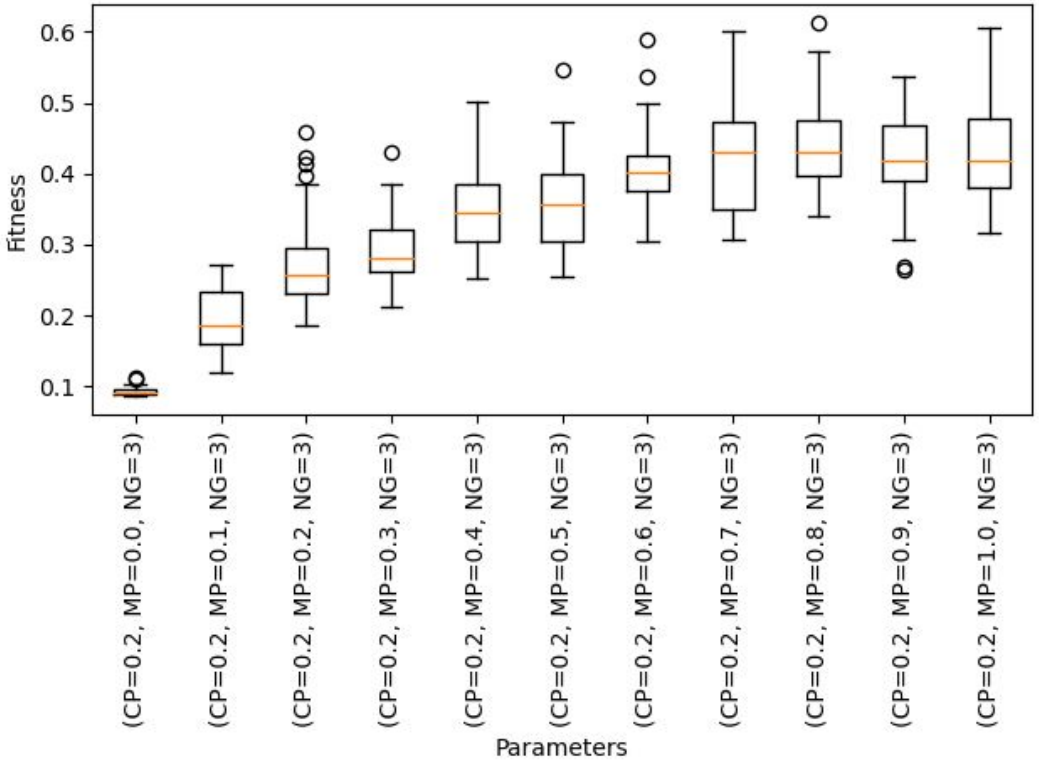
- Prav. křížení: 0.2
- Prav. mutace: 0.8
- Počet mut. genů: 3



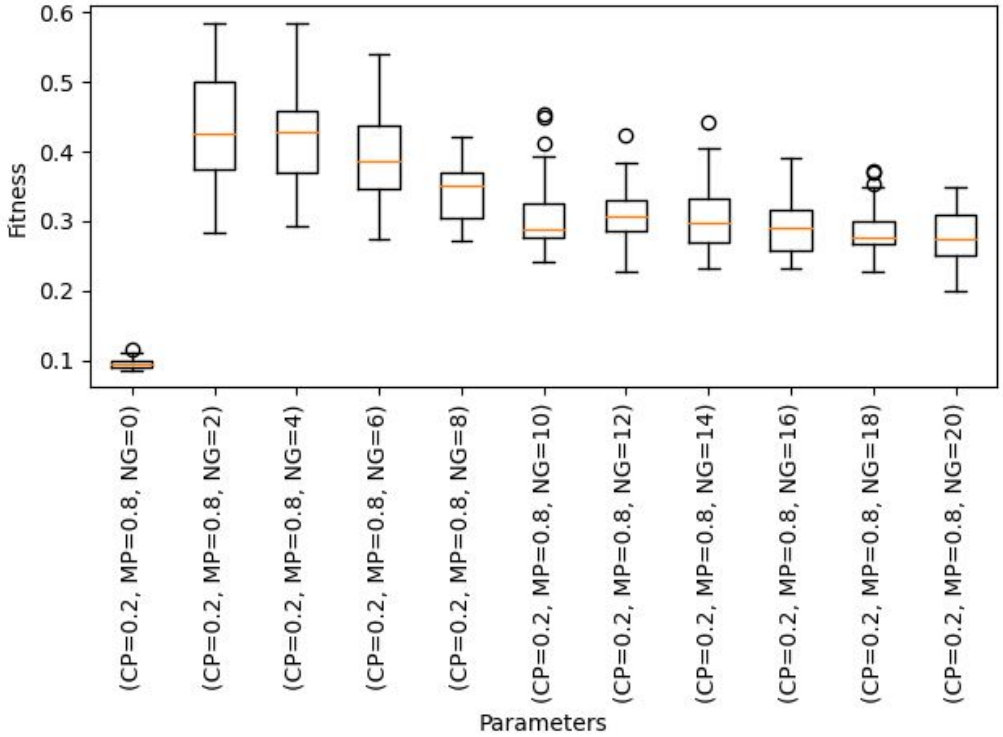
Experiment - Pravděpodobnost křížení



Experiment - Pravděpodobnost mutace



Experimenty - Počet mutujících genů



Experimenty - Počet generací

Parametry:

- Populace: 30
- Prav. křížení: 0.2
- Prav. mutace: 0.8
- Počet mut. genů: 3