Aproximace neuronových sítí

Biologií inspirované počítače (BIN) 2020/2021

Autor: Adam Láníček

10.5. 2021

Zadání projektu

- Seznamte se s knihovnou TensorFlow a s aproximačními vrstvami (https://github.com/ehw-fit/tf-approximate).

- Navrhněte genetický algoritmus (např. na bázi NSGA-II) pro přiřazování jednotlivých aproximačních násobiček k vrstvám pro vámi vybranou neuronovou síť. Důkladně vyhodnoťte výsledky a vliv různých parametrů na kvalitu aproximace.

Implementace

Trénink KNN AlexNet (5 konv.vrstev) na GPU
- CIFAR-10 dataset, 130 epoch

Std. evaluace
- 70,7% přesnost na neviděných datech

 # násobení v každé konv. vrstvě

X

NSGA-II algoritmus
- využití **pymoo** Python knihovny
- více-cílová optimalizace:
  - **max**imalizace přesnosti
  - **min**imalizace energie
- **křížení**: simulované binární křížení (celočíselné v rozsahu **0-34**, parametry eta, pravd.)
- **mutace**: **polynomiální celočíselná** (celočíselná v rozsahu **0-34** eta, pravd.)

mults[x][0] **jedinec** mults[x][0]

```
Testing approximate model with
ConvLayer 1: mul8u_125K.bin
ConvLayer 2: mul8u_12N4.bin
ConvLayer 3: mul8u_2P7.bin
ConvLayer 4: mul8u_96D.bin
ConvLayer 5: mul8u_NGR.bin
```

mults[x][0] mults[x][0] mults[x] [0]

8b násobičky (35x) a jejich energetická náročnost (EvoApproxLib)

```
mults = [
    ['125K', 0.384],
    ['12N4', 0.142],
    ['13QR', 0.0085],
```

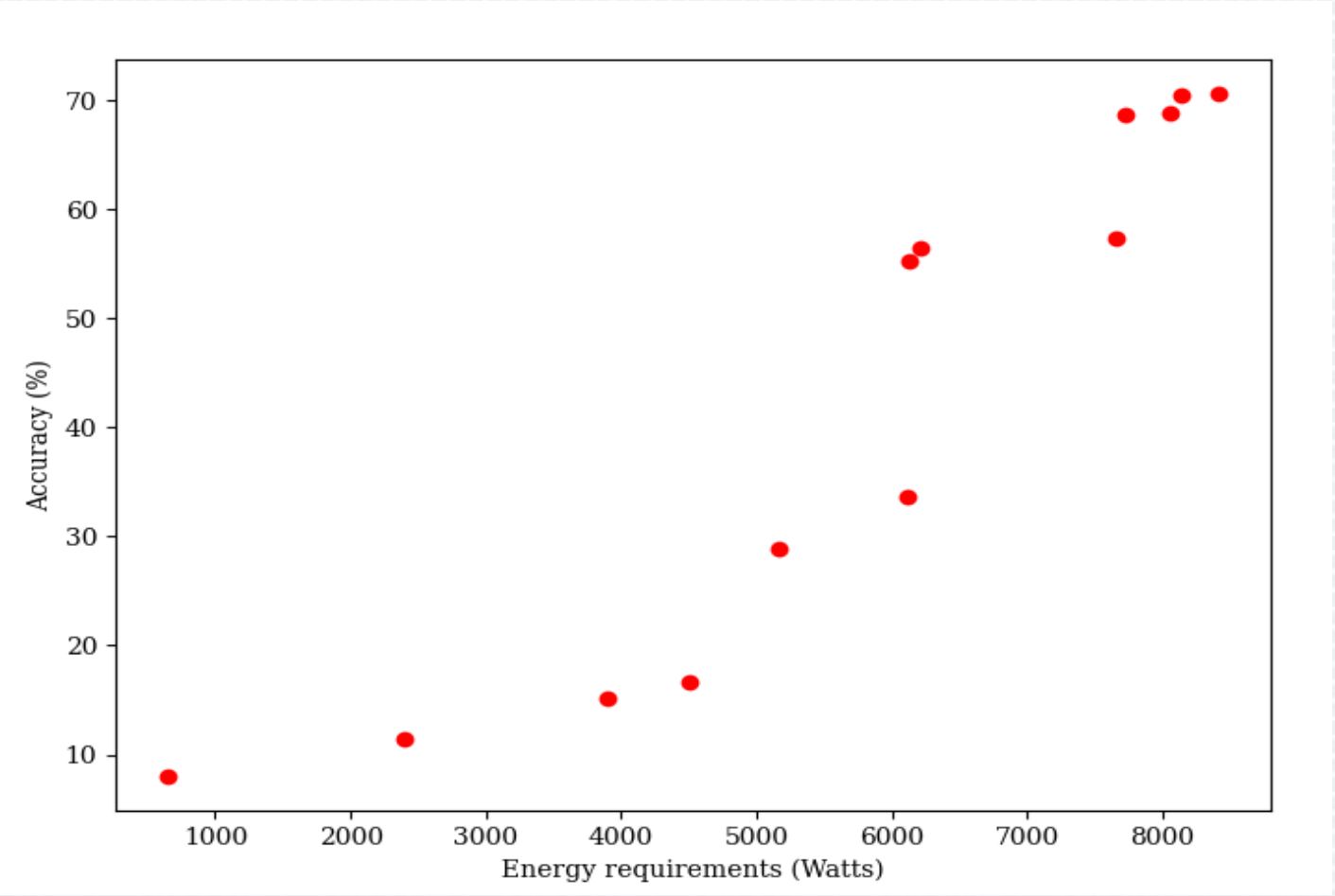
Vyhodnocení přesnosti inference

Převzato a upraveno

Vlastní implementace

Experimenty 1/4

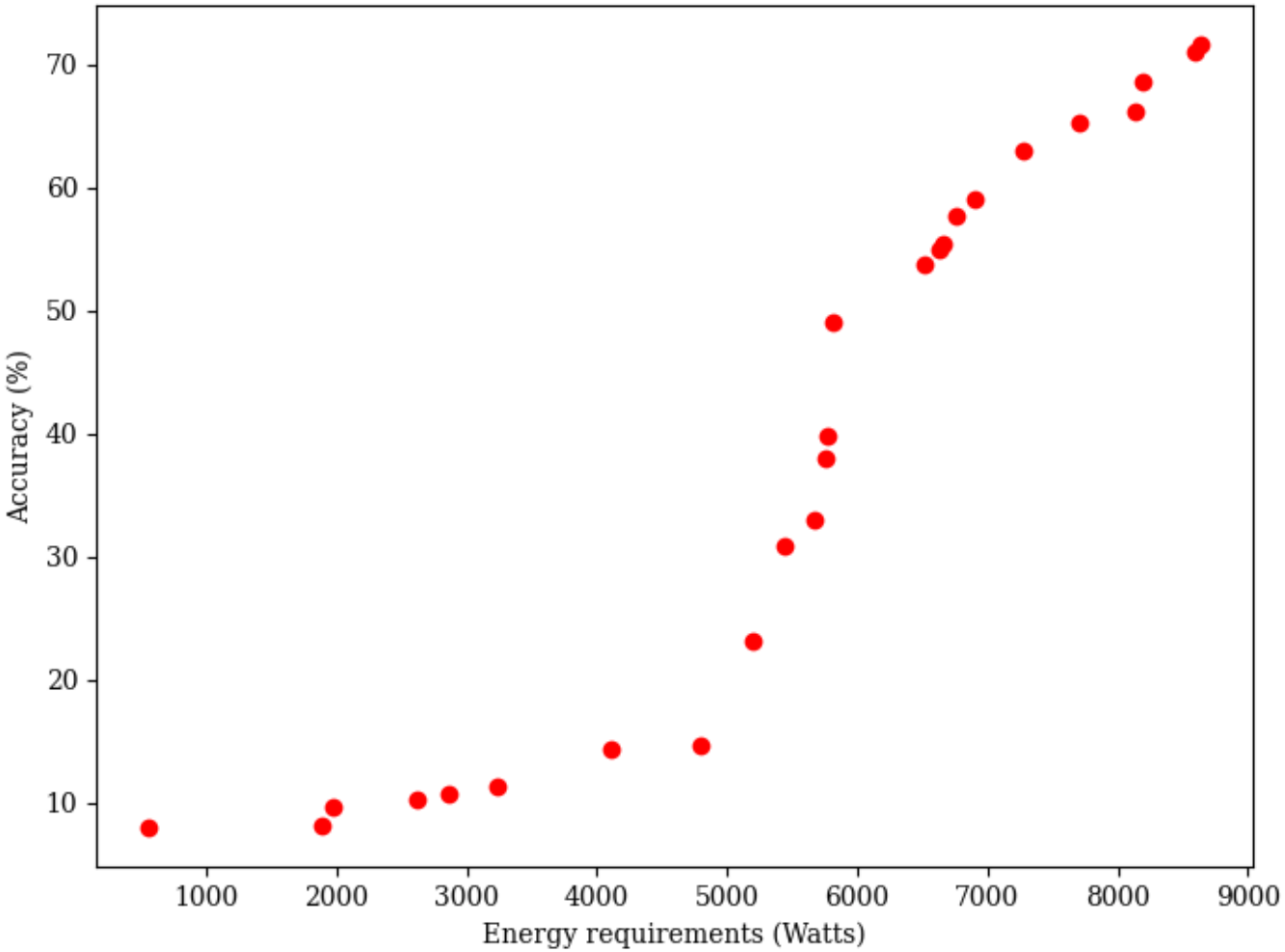
- **20** členů
- **20** generací
- křížení **100%**
- mutace 50%
- eta mutace = 3
- eta křížení = 3
- běh cca 1,5h



- ověřená hypotéza – menší populace poskytuje omezenější možnost křížení/mutací a tedy v rámci běhu generuje méně rozsáhlou Pareto-optimální množinu potenciálních řešení (duplikovaná řešení v grafu zobrazena nejsou)

Experimenty 2/4

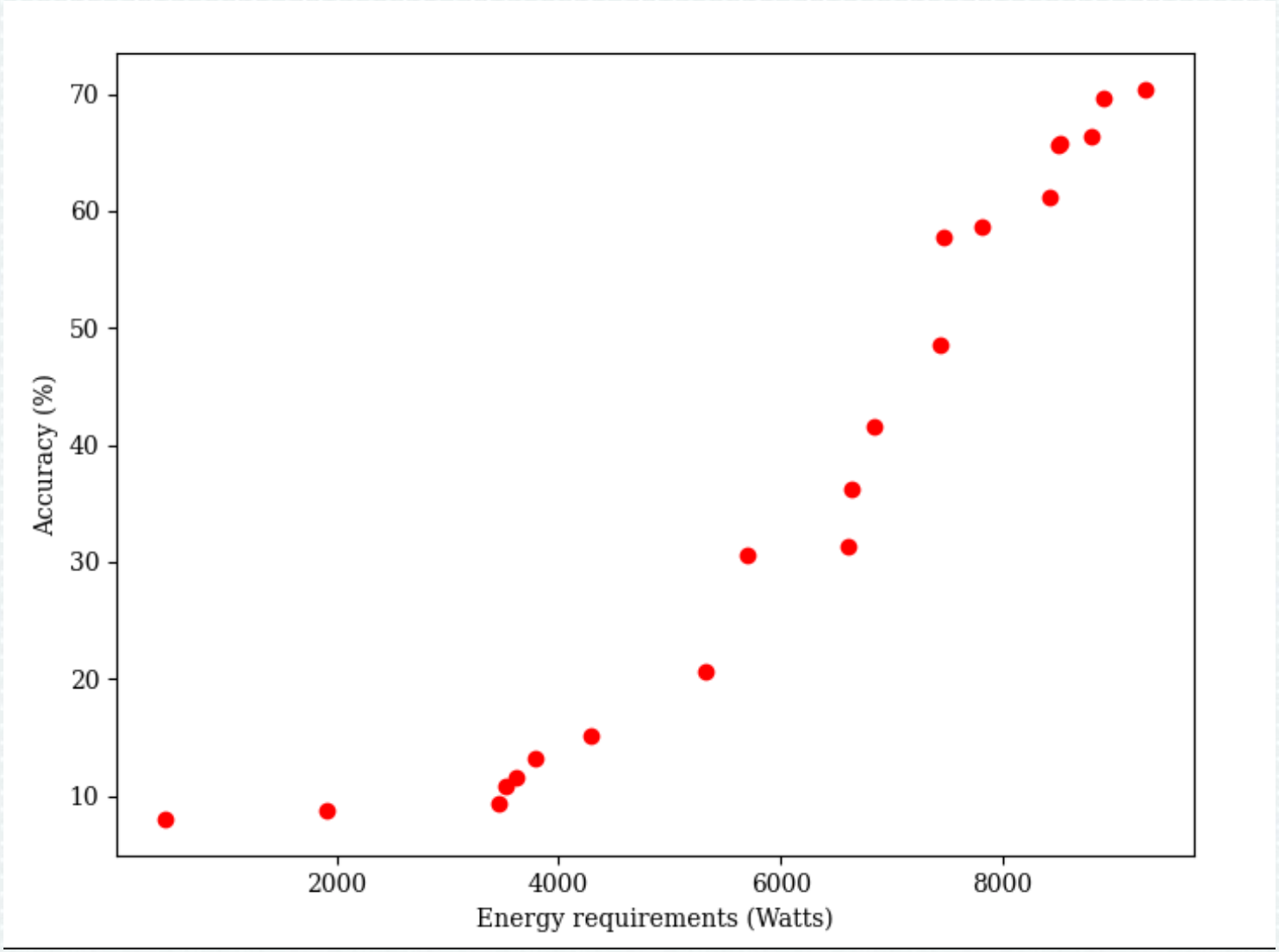
- **40** členů
- **20** generací
- křížení **100%**
- mutace 50%
- eta mutace=3
- eta křížení=3
- běh cca 3h



- ověřená hypotéza – více členů v populaci, více řešení (viz předešlý snímek)

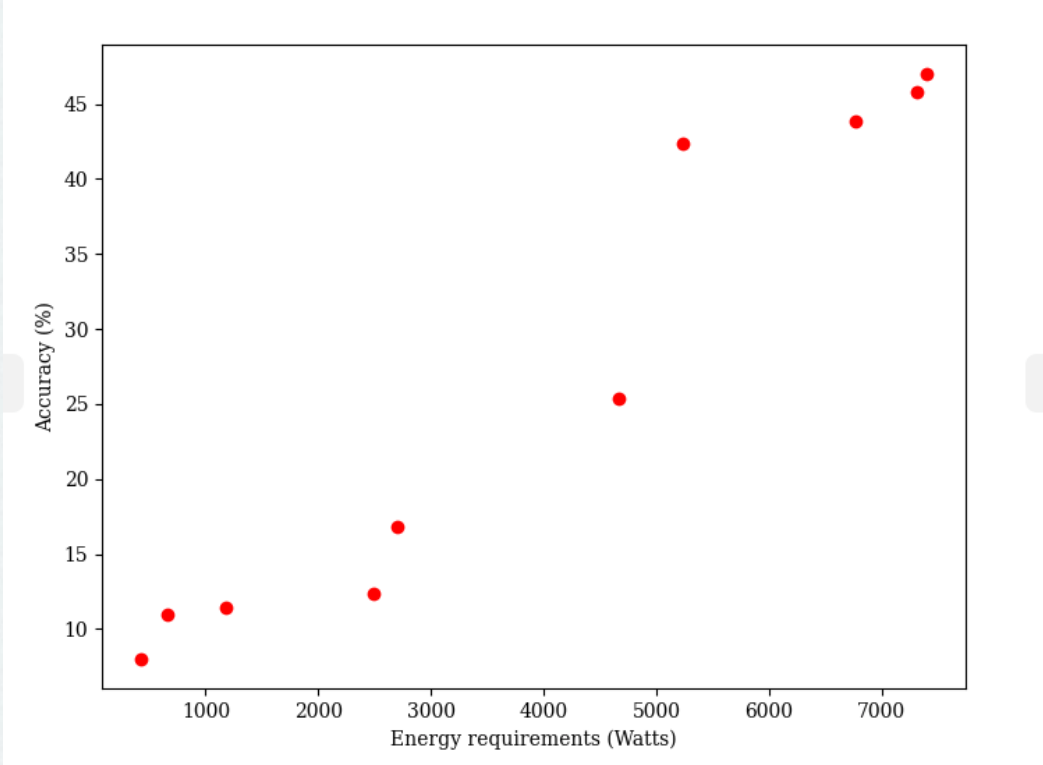
Experimenty 3/4

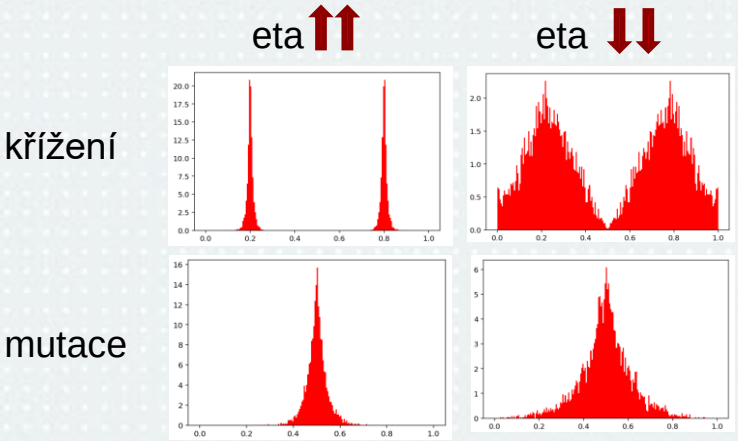
- **100** členů
- **5** generací
- křížení **50%**
- mutace 50%
- eta mutace=3
- eta křížení=3



Experimenty 4/4

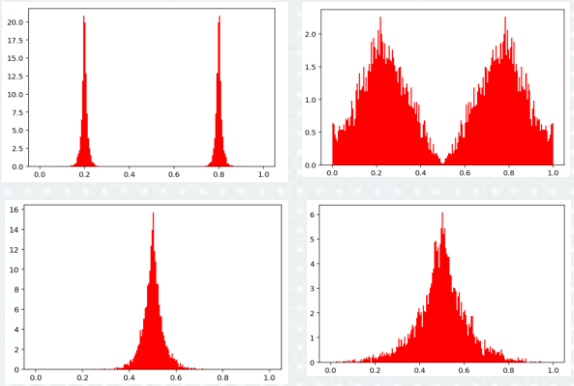
- **70** členů
- **10** generací
- křížení **50%**
- mutace 50%
- **eta mutace=30**
- **eta křížení=30**





Závěr

- článek **ALWANN: Automatic Layer-Wise Approximation of Deep Neural Network Accelerators without Retraining** (viz zadání) uvádí, že při použití 8b násobiček místo konvolučních vrstev dojde k ušetření 30% energie při snížení přesnosti o **0,6%**
  - moje experimenty tento závěr o přesnosti potvrdily, kdy vybrané kombinace násobiček téměř v každém běhu dosáhly o **0,5%** nižší přesnosti ve srovnání s neuronovými sítěmi s klasickými konvolučními vrstvami
- největší vliv na kvalitu aproximací, konkrétněji **na množství vygenerovaných potenciálně použitelných kombinací**, měla:
  - velikost populace
  - **eta** parametr u křížení a mutace



Děkuji za pozornost!