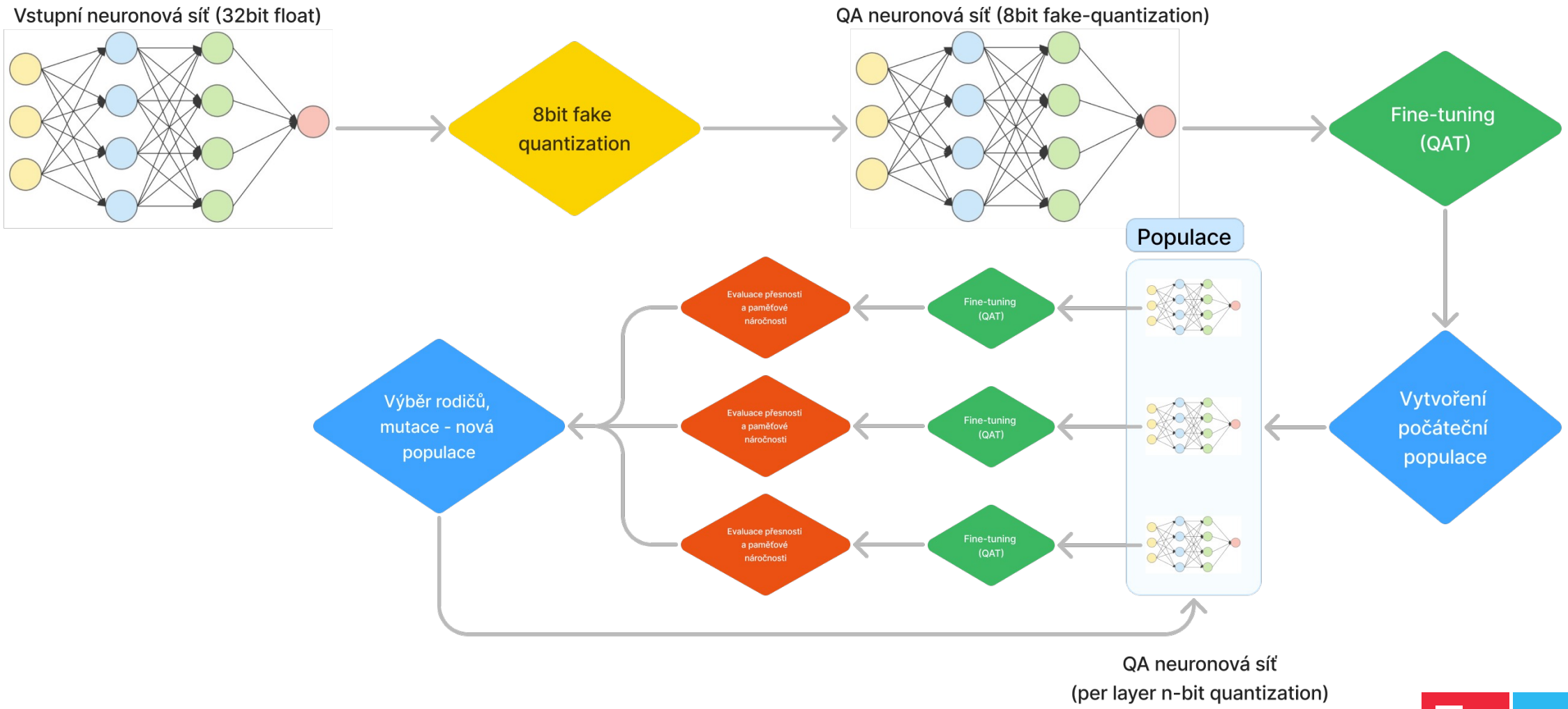
Automatická kvantizace neuronových sítí

Miroslav Šafář
xsafar23
Brno 2023

Cíl práce

- Navrhnout systém pro automatické určování úrovně kvantizace jednotlivých vrstev vstupní neuronové sítě
- Systém implementovat
- Vyhodnotit kvalitu navržených řešení (přesnost x paměťová náročnost)

Návrh řešení



Stav řešení

- Prozkoumány možnosti kvantizace v TensorFlow a PyTorch
- Implementován systém kvantizace jednotlivých vrstev na různé přesnosti pro jednoduché neuronové sítě pomocí TensorFlow
- Ověřování funkcionality TensorFlow pro kvantizaci složitějších neuronových sítí, u kterých dochází před inferencí k „batchnorm fold“

Přečtená literatura

- Bjorck, J., Gomes, C., Selman, B. a Weinberger, K. Q. Understanding Batch Normalization. arXiv, 2018. DOI: 10.48550/ARXIV.1806.02375
- Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W. et al. A Survey of Quantization Methods for Efficient Neural Network Inference. arXiv, 2021. DOI: 10.48550/ARXIV.2103.13630.
- Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M. et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. arXiv, 2017. DOI: 10.48550/ARXIV.1712.0587
- Li, Y., Shen, M., Ma, J., Ren, Y., Zhao, M. et al. MQBench: Towards Reproducible and Deployable Model Quantization Benchmark. arXiv, 2021. DOI: 10.48550/ARXIV.2111.03759
- Li, Z. a Gu, Q. I-ViT: Integer-only Quantization for Efficient Vision Transformer Inference. arXiv, 2022. DOI: 10.48550/ARXIV.2207.01405.
- Sekanina, L. Neural Architecture Search and Hardware Accelerator Co-Search: A Survey. IEEE Access. 2021, sv. 9, s. 151337–15136
- Wang, K., Liu, Z., Lin, Y., Lin, J. a Han, S. HAQ: Hardware-Aware Automated Quantization with Mixed Precision. arXiv, 2018. DOI: 10.48550/ARXIV.1811.08886
- Yao, Z., Dong, Z., Zheng, Z., Gholami, A., Yu, J. et al. HAWQV3: Dyadic Neural Network Quantization. arXiv. 2020. DOI: 10.48550/ARXIV.2011.10680
- Deb, K., Pratap, A., Agarwal, S. a Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation. 2002, sv. 6, ˇc. 2, s. 182–197. DOI: 10.1109/4235.996017.

Co dál

- Implementace EA (NSGA-II)
- Provádění experimentů s navrženým systémem (MobileNet, EfficientNet)
- Zhodnocení výsledků navrženého systému

Děkuji za pozornost