Využití programovatelných hradlových polí pro výpočet neuronových sítí

Ing. Jiří Čech <jiri.cech@tul.cz>, Ing. Karel Paleček, Ph.D.

Příspěvek shrnuje výsledky snahy optimalizace rychlosti výpočtu konvolučních neuronových sítí na dostupných výkonných platformách. Pro řešení náročného vyhodnocení obsáhlých obrazových dat upřednostňujeme programovatelná hradlová pole, která nabízí vysoký výpočetní výkon a nízkou energetickou spotřebu.

Klíčová slova: CNN, FPGA, VITIS AI, Pruning, DPU

Úvod

Příspěvek se zabývá problematikou vyhodnocení obrazu z hyperspektrálních kamer. Ty umožňují záznam scény v desítkách až stovkách barevných kanálů a poskytují enormní množství informací, které je potřeba co nejrychleji a nejefektivněji zpracovat. Na jejich vyhodnocení se využívají hluboké neuronové sítě. Pro zefektivnění vyhodnocovacího procesu se minimalizuje velikost natrénovaných sítí, které se pak ověřují na výkonné platformě. Pro rychlý výpočet je důležitá vysoká datová propustnost a velká míra paralelizace výpočtu. To jsou ideální podmínky pro využití programovatelných hradlových polí (FPGA). Dosažené realizace na FPGA porovnáváme s výpočtem na procesoru (CPU) a na grafické kartě (GPU).

Metodika

Vyhodnocujeme naměřená data z Hyperspektrální kamery, tzv. Hyperspektrální kostku, která obsahuje až několik Giga bajtů dat. Účelem vyhodnocení je na základě průběhu obrazového pixelu napříč nasnímanými „barevnými“ kanály, tzv. spektrální charakteristiky, odlišit a rozpoznat materiály ve scéně.

Pro dosažení přesných výsledků se do výpočtu jednoho pixelu zahrnuje i jeho okolí [1], čímž exponenciálně roste vyhodnocovaný vstupní vektor.

Zrychlení procesu výpočtu vyžaduje optimalizaci natrénované neuronové sítě a redukci její velikosti. Pro námi využívané konvoluční neuronové sítě (CNN) jsou možnosti jejich redukce následující:

a) Úprava struktury sítě
   Jednotlivé vrstvy neuronové sítě umožňují různá nastavení. Kromě základního parametru počtu neuronů či velikosti konvolučního jádra, lze upravit i velikost posuvu výpočtu konvoluce.
b) Redukce paměťové náročnosti [2]
   Základní výpočty a trénování sítí se provádí v datovém formátu vah sítě Float32 nebo Float16. Pokud váhy sítě převedeme na formát Int8, dosáhneme až čtyřnásobné redukce paměťové náročnosti a zrychlení výpočtu. Snížíme tím ale přesnost sítě v řádu desetin procent, výjimečně i jednotky procent.
c) Redukce vazeb
   Tato technika nahrazuje bezvýznamné váhy sítě, které mají hodnotu blízkou nule, skutečnou nulou. Tím dojde k ignorování výpočtu s danou vahou a ke snížení celkového počtu potřebných operací k výpočtu sítě.
d) Redukce neuronů [3]
   Pokud zredukujeme významný počet vah mezi neurony, mohou se nám objevit neurony, které jsou lineární kombinací jiných neuronů ve stejné vrstvě. Takové to duplicity neuronů je možné odstranit a zmenšit tak velikost dané vrstvy bez výrazného snížení přesnosti.

Takto připravená sít lze převést pomocí nástroje Vitis AI od firmy XIlinx pro použití v FPGA z řady Znyq-7000 SoCs a vyšší. Nástroj využívá „Deep Processing Unit“ (DPU) [4] pro akceleraci výpočtu na FPGA. Systém běží pod operačním systémem PetaLinux a znalost cílové platformy je potřebná pro správný převod sítě. Výsledek převodu je spustitelný soubor, který provede výpočet nad zadanými testovacími daty.

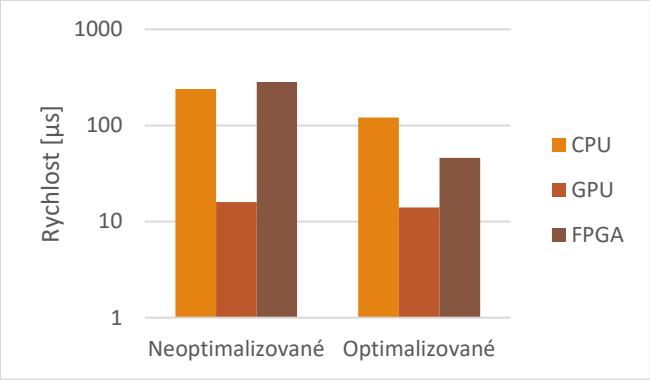
Výsledky a diskuze

Na dostupném testovacím hyperspektrálním vzorku „Pavia University“, zaznamenaném satelitním

senzorem ROSIS, jsme natrénovali konvoluční neuronovou síť schopnou rozeznat materiály ve vzorku s přesností 96 %. Výsledná síť obsahuje 1,8M vah a na její vypočet je potřeba provést 3,6 M operací.

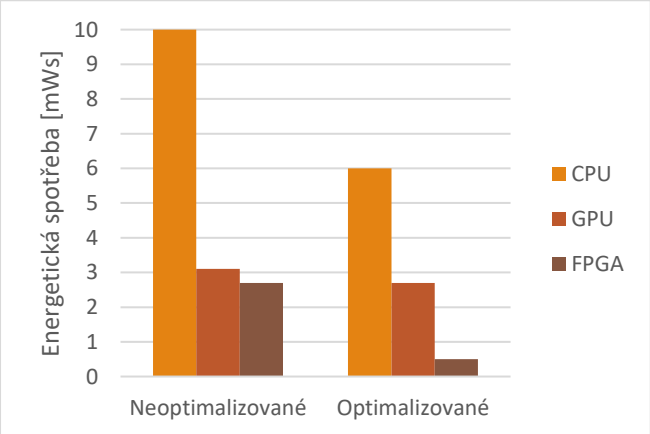
Poté jsme provedli optimalizaci sítě, při které se podařilo udržet původní úspěšnost vyhodnocení 96 % a snížit celkový počet vah na 0,1 M a počet potřebných operací na 0,2 M.

Realizaci obou sítí jsme provedli pomocí prostředí Tensorflow na CPU (Intel Core i7-7820X, který má 8 jader na 4,3GHz), na GPU (Nvidia Geforce RTX 2070) a na FPGA (jedno jádro DPU 4096 o velikosti 70k LUT a 704 DSP). Výsledky rychlosti výpočtu jsou na obrázku 1. Vypočítán byl průměr z dvaceti výpočtů sítě po 33 tisících vzorcích.



Obrázek 1: Porovnání rychlosti výpočtu CNN

Dalším zajímavým aspektem k porovnání je spotřeba dané výpočetní platformy. Ta je rozhodujícím kritériem pro použití v bateriových systémech, na dronech či v satelitech.



Obrázek 2: Porovnání energetické spotřeby při výpočtu CNN

Závěr

Otestovali jsme využitelnost FPGA platformy pro výpočet konvolučních neuronových sítí vyhodnocující hyperspektrální data. Pro neoptimalizovanou síť byly dosažené výsledky podobné CPU platformě, ale díky provedeným optimalizacím se podařilo zrychlit výpočet a přiblížit se tak více času GPU platformy.

Zajímavější je výsledek energetické spotřeby, kde FPGA platforma je ve výsledku sice 3× pomalejší než GPU, ale spotřeba je 10× nižší.

Dosáhli jsme výkonu až 50 TOPS, což je srovnatelný výkon dosažený pro jiné neuronové sítě [5]. Plánujeme v budoucnu využít až čtyři DPU jednotek místo jedné a tím celý proces ještě více zrychlit.

Poděkování

Tato práce byla podpořena z projektu Studentské grantové soutěže (SGS) na Technické univerzitě v Liberci v roce 2020.

Reference

[1] Y. Chen, H. Jiang, C. Li, X. Jia and P. Ghamisi, "Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks," in IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 10, pp. 6232-6251, Oct. 2016.

[2] Yang, Jiwei, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang and Xian-Sheng Hua. “Quantization Networks.” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 7300-7308. arXiv:1911.09464

[3] R. Reed, "Pruning algorithms-a survey," in IEEE Transactions on Neural Networks, vol. 4, no. 5, pp. 740-747, Sept. 1993.

[4] Xilinx, Vitis AI, 2020. [Online]. Available: https://www.xilinx.com/support/documentation/sw_manuals/vitis_ai/1_0/ug1414-vitis-ai.pdf

[5] K. Tajiri and T. Maruyama, “FPGA Acceleration of a Supervised Learning Method for Hyperspectral Image Classification,” 2018 International Conference on Field-Programmable Technology (FPT), Naha, Okinawa, Japan, 2018, pp. 270-273.