BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2026
http://excel.fit.vutbr.cz

Veľké jazykové modely v návrhu aproximatívnych obvodov

Bc. Martin Tomašovič*

Abstrakt

Aproximatívne obvody predstavujú alternatívu k presným digitálnym obvodom, kde sa za cenu kontrolovanej chyby dosahujú nižšia spotreba, latencia a plocha. V tejto práci sa zameriavame na návrh aproximatívnych 8-bitových násobičiek pomocou kombinácie evolučných algoritmov a veľkých jazykových modelov (LLM). Navrhujeme evolučný prístup, v ktorom sa vyvíja populácia obvodov a populácia šablón promptov riadiacich LLM. LLM je využitý ako operátor mutácie, ktorý generuje nové návrhy obvodov.
Experimenty ukazujú, že navrhovaná metóda dokáže nájsť riešenia s lepším kompromisom medzi chybou a plochou v porovnaní s existujúcimi návrhmi z knižnice EvoApproxLib.

xtomas36@vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Úvod

Táto práca skúma využitie kombinácie veľkých jazykových modelov (LLM) a evolučných algoritmov pri aproximácii obvodov. Konkrétne sme sa zamerali na 8-bitové násobičky. LLM funguje ako operátor mutácie a rekombinácie. Myšlienka spočíva v tom, že LLM využíva svoje znalosti získané predtrénovaním na navrhovanie inteligentných operácií. Metóda tiež využíva LLM na evolúciu šablón promptov, z ktorých následne generuje nové obvody. V každej generácii algoritmus vyberá šablóny promptov a populáciu obvodov pomocou stratégie prežitia NSGA-II s metrikou „Pruning Crowding Distance“ pre určenie rozptylu (crowding).

V tejto oblasti existuje niekoľko prác, ktoré používajú podobnú metodológiu. GPTAC [1] je najbližšia práca, kde dotrénovaný model GPT-2 generuje aproximatívne obvody. GPTAC využíva algoritmy, ktoré upravujú výstup počas inferencie s cieľom ho zlepšiť a zvýšiť generalizáciu. Preukázali zlepšenie oproti metóde Versatile Efficiency-Accuracy Configurable Batch Error Estimation [2] a 9,9 % zlepšenie plochy oproti staršiemu Evoapprox8b; avšak so 7,5 % vyššou latenciou pri rovnakých obmedzeniach MSE. Ďalšou súvisejúcou metódou je viackriteriálne kartézske genetické programovanie [3]. Je to state-of-the-art prístup, ktorý využíva dvojrozmerné pole programovateľných uzlov pre evolúciu aproximatívnych obvodov z presných a spravuje populácie pomocou algoritmu NSGA-II.

V porovnaní s týmito metódami sa náš prístup odlišuje tým, že využíva evolúciu a voľne dostupný LLM bez ďalšieho dotrénovania ako operátor mutácie a rekombinácie.

2. Navrhnutá metóda

Obvody sú reprezentované pomocou matematických rovníc na úrovni logických hradiel. Hradlá AND, OR, XOR a NOT sú reprezentované notáciou prevzatou z programovacieho jazyka C. Tento formát je pre človeka čitateľný a pochopiteľný. Na základe predošlých experimentov sme zistili, že podobnosť s programovacími jazykmi dovoluje LLM využiť znalosti z predtrénovania. Pre generovanie návrhov obchodov bol vybraný model dotrénovaný pre riešenie matematických úloh a generovanie kódu.

Na začiatku algoritmu (Obrázok 1) sa inicializuje populácia obvodov (z datasetu), tiež sa pripravia šablóny promptov z externých JSON súborov a nastaví sa cieľový obvod definovaný chybou a plochou. Algoritmus je ukončený po dosiahnutí vopred definovaného počtu iterácií. Počas každej iterácie sa mutáciou vytvárajú nové obvody z aktuálnej populácie v každej generácii (Obrázok 2). Fitness promptov sa aktual-

izujú na základe meraní chýb a plochy príslušných obvodov. Nová populácia obvodov sa vytvára na základe fitness obvodov. Fitness obvodov ovplyvňuje fitness promptov. Pre obe populácie používame selekciu $(\mu+\lambda)$. Pre výber obvodov sú fitness funkcie stredná kvadratická chyba (MSE) a plocha. Meranie chyby je vykonané konverziou do programovacieho jazyka C. Plocha sa odhaduje na základe počtu jednotlivých typov hradiel. Veľkosti hradiel zodpovedajú 45 nm technológii. Výber šablón promptov používa odlišné fitness funkcie. Konkrétne prvou fitness funkciou je priemerný rozdiel medzi aktuálnou MSE a cieľovou MSE a druhou je pomer validných a nevalidných obvodov generovaných pomocou danej šablóny. Na základe evolučného cieľa nahrádza MSE vo fitness funkciách maximálnu možnú chybu (WCE).

Ďalej nastavujeme obmedzenie strednej absolútnej chyby (MAE) pre násobenie nulou. To je nevyhnutné pre použitie obvodov v akcelerátoroch hlbokých neurónových sietí [4]. Je to preto, že mnohé algoritmy (dropout, masky pozornosti) používajú násobenie nulou ako masku a mierne odlišná hodnota by spôsobila logickú chybu.

Posledným krokom v cykle je kríženie a mutácia (Obrázok 3). Obe operácie vykonáva veľký jazykový model (LLM). Na tento účel používame šablóny meta-promptov. Existujúce šablóny promptov sa vložia do meta-šablón a LLM vytvorí nové šablóny promptov pre mutovanie obvodov.

Každú odpoveď od LLM tiež spracúvame volaním LLM s históriou, aby sme z nej extrahovali čistý obvod alebo šablónu promptu.

3. Experimenty

V experimentoch sme sa zamerali na 8-bitové násobičky bez znamienka. Ako model vykonávajúci kríženie a mutáciu šablón promptov sme nastavili GPT-OSS-120B a pre mutáciu obvodov model Qwen3-Coder480B. Optimalizujeme pre plochu a jednu z chýb, buď WCE, alebo MSE.

Pre strednú absolútnu chybu (MAE) optimalizujeme s požiadavkou, aby násobenie nulou dávalo výsledok nula, z dôvodov uvedených v sekcii 2.

Nastavenia experimentov sú nasledovné: počet preživších príkazov = 8, počet preživších obvodov = 30, pravdepodobnosť kríženia = 0,9, pravdepodobnosť mutácie = 0,9, počet generácií = 27, LLM teplota = 0,4.

Priemerne trvá volanie väčšieho LLM 10 s, menšieho 5,3 s a meranie obvodu 0,5 s. Počet novo objavených násobičiek pre WCE–plochu je 24 a MSE–plochu je 23.

Výsledné evolučne získané obvody porovnávame so zdrojovým datasetom EvoApproxLib[1]. Z tohto datasetu pochádza počiatočná populácia. Pareto fronty zobrazujeme na obrázkoch 5 a 6. Okrem dlhých behov algoritmu sme na kratších behoch štatisticky vyhodnotili algoritmus oproti zjednodušenej variante, a to spôsobom odvodeným od Hill-climbing algoritmu. Meranie bolo vykonané pre rovnaké počty volaní LLM, ktoré predstavujú výpočetne najdrahšiu operáciu. Po vyhodnotení pomocou Mann-Whitneyho U testu na hyperobjemoch Pareto front sme zistili, že sa štatisticky významne líšia, pričom Hill-climbing vykazuje horšie výsledky.

4. Zhrnutie

V tejto práci sme navrhli metódu aproximácie obvodov. Hlavným prínosom je navrhnutý algoritmus a skutočnosť, že pre optimalizáciu obvodov nepotrebujeme dolaďovať (fine-tune) veľké jazykové modely.

Poďakovanie

Chcel by som sa poďakovať svojmu školiteľovi prof. Ing. Lukášovi Sekaninovi, Ph.D. za jeho cenné poznatky a rady.

Literatúra

[1] Sipei Yi, Weichuan Zuo, Hongyi Wu, Ruicheng Dai, Weikang Qian, and Jienan Chen. Gptac: Domain-specific generative pre-trained model for approximate circuit design exploration. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 15(2):349–360, 2025.

[2] Sanbao Su, Chang Meng, Fan Yang, Xiaolong Shen, Leibin Ni, Wei Wu, Zhihang Wu, Junfeng Zhao, and Weikang Qian. Vecbee: A versatile efficiency–accuracy configurable batch error estimation method for greedy approximate logic synthesis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(11):5085–5099, 2022.

[3] Radek Hrbacek, Vojtech Mrazek, and Zdenek Vasicek. Automatic design of approximate circuits by means of multi-objective evolutionary algorithms. In *2016 International Conference on Design and Technology of Integrated Systems in Nanoscale Era (DTIS)*, pages 1–6, 2016.

[1] https://ehw.fit.vutbr.cz/evoapproxlib/

[4] Vojtech Mrazek, Lukas Sekanina, and Zdenek Vasicek. Libraries of approximate circuits: Automated design and application in cnn accelerators. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 10(4):406–418, 2020.