

Když AI vylepšuje sama sebe

8.11.2024 - | Grantová agentura ČR

Výpočetně náročné metody umělé inteligence a zejména hluboké neuronové sítě jsou dnes klíčové pro zpracování obrazu, řeči, přirozeného jazyka nebo medicínských dat. Pro rozpoznání psa, auta, stromu či jiného objektu na obrázku musí například středně složitá neuronová síť typu ResNet-50, která má 25 milionů parametrů, provést 3,9 miliardy operací „vynásob a sečti“.

V dnešní době pronikají tyto metody z výkonných serverových klastrů do mobilních telefonů a jiných malých zařízení napájených bateriemi, a proto je nutné vyřešit jejich efektivní implementaci s ohledem na velmi omezené výpočetní zdroje (jako je místo na čipu nebo kapacita baterie).

Software, který umožňuje automatizovaně navrhovat neuronové sítě, a také hardwarové akcelerátory, ve kterých výpočet neuronových sítí probíhá, je intenzivně zkoumán a aplikován jak v technologických IT společnostech (včetně těch gigantických), tak i v univerzitním prostředí.

Cílem plně automatizovaného návrhu neuronové sítě je na základě zadaných dat (např. v podobě databáze, která obsahuje desetitisíce obrázků a informaci o jejich třídě) a dalších požadavků (např. maximální povolená doba klasifikace a maximální chybovost klasifikace) vygenerovat neuronovou síť a její softwarovou implementaci, která dokáže obrázky spolehlivě klasifikovat.

Tým prof. Sekaniny z Fakulty informačních technologií Vysokého učení technického v Brně nahlížel na metody, které automatizovaný návrh neuronových sítí realizují, jako na metody pro vícekritériální optimalizaci. Konkrétně se zabýval neuroevolučními algoritmy, které s využitím metafor biologické evoluce prochází prostor možných neuronových sítí a hledají ty, jež nejlépe splňují zadané požadavky. Ty jsou však často protichůdné. Neuroevoluční algoritmus v prvním kroku generuje náhodné kandidátní implementace neuronových sítí a zjišťuje jejich schopnost řešit zadaný problém. Ve druhém kroku pak pomocí selekce vybírá rodičovské neuronové sítě, ze kterých za pomoci algoritmů „křížení“ a „mutace“ vytváří nové kandidátní neuronové sítě. Tento postup se opakuje, dokud není nalezena vhodná neuronová síť.

„Očekáváme, že neuroevoluční přístup nám umožní pro danou úlohu nalézt nejvhodnější architekturu neuronové sítě, kterou by člověk-expert nevymyslel,“ říká o motivaci pro tento výzkum prof. Sekanina. „Zadání následně děláme těžší tím, že hledáme nejen dobře fungující neuronovou síť, ale také vysoce optimalizovaný hardwarový akcelerátor, který zajistí, aby klasifikace trvala co nejkratší dobu a byla energeticky co nejméně náročná,“ doplňuje.

V projektu byly intenzivně využívány principy aproximativního počítání. *„Pokud se drobné chyby, které vzniknou například kvůli zjednodušenému (aproximativnímu) násobení, téměř neprojeví na kvalitě výstupu neuronové sítě, tak není třeba provádět miliardy přesných a energeticky náročných násobení. Takto můžeme při používání neuronové sítě v závislosti na řešené úloze uspořit až nižší desítky procent energie,“* popisuje výsledky projektu prof. Sekanina.

Jedním z výstupů projektu je unikátní metoda založená na neuroevolučních algoritmech, která umožňuje automatizovaně vytvářet klasifikátory obrázků a současně optimalizovat typ použitého násobení ve zvolených vrstvách sítě, a tím redukovat spotřebu energie.

Kromě automatizovaného návrhu neuronových sítí výzkumníci také rozvíjeli metody, díky nimž budou moci automatizovat proces navrhování zjednodušených (aproximativních) elektronických obvodů,

které jsou stavebními bloky energeticky úsporných hardwarových akceleratorů neuronových sítí. Principy automatizovaného návrhu využili i v dalších úlohách, např. při klasifikaci signálů z elektroencefalografie nebo příznaků indukované dyskineze, která se projevuje mimovolnými pohyby po podání léku např. u pacientů s Parkinsonovou chorobou.

Výsledky projektu otevřely nové cesty k návrhu vysoce optimalizovaných akceleratorů neuronových sítí, které by se v budoucnu mohly objevit v komerčních zařízeních. Vznikla také řada technik, jako je např. rychlý výpočet konvoluční vrstvy s aproximativním násobením na grafické kartě, umožňujících zredukovat vysokou výpočetní náročnost použitých algoritmů.

Kromě prof. Sekaniny k výsledkům zásadně přispěli doc. Ing. Zdeněk Vašíček, Ph.D., a Ing. Vojtěch Mrázek, Ph.D. Na projektu se dále podíleli 2 postdoktorandi a 12 studentů, převážně doktorandů. Metody návrhu aproximativních komponent byly vytvářeny ve spolupráci s New York University (Abu Dhabi). Případová studie v oblasti klasifikace příznaků indukované dyskineze vznikla ve spolupráci s University of York, UK.

<https://gacr.cz/kdyz-ai-vylepsuje-sama-sebe>