Průmyslová umělá inteligence

Generativní umělá inteligence se většinou používá k řešení virtuálních úloh: generování textů, obrázků apod. Méně časté je její vyžití k řešení fyzických problémů, např. pro řízení a optimalizaci procesů v průmyslové praxi. Článek se zamýšlí nad vývojem generativní AI a nastiňuje její možnosti a hranice. Dále se věnuje metodě posilovaného učení, kterou autor pokládá za velmi perspektivní právě pro řešení úloh reálného světa.

Generativní umělá inteligence

Umělá inteligence (dále AI – *Artificial Intelligence*) zažívá v posledních několika letech nebývalý rozmach. Tento vědní obor se stále častěji prosazuje ve všech oblastech lidské činnosti, průmysl nevyjímaje. Mezi širokou veřejností je dnes pravděpodobně nejznámější podobor generativní AI, jenž celkem úspěšně ovládl psaní textů a tvorbu vizuálního obsahu. Každý již určitě viděl spoustu dechberoucích příkladů, jak generativní AI umí být tvůrčí a jak snadno řeší některé problémy, o kterých jsme si ještě nedávno mysleli, že jsou absolutní doménou člověka. Je tomu ale skutečně tak? Znamená to, že když stroje zvládnou tyto intelektuálně a talentově náročné činnosti, je pro ně zdánlivě snadná činnost, jako např. řízení auta, kompletace výrobků nebo kontrola kvality výrobní linky, též hračkou?

Je paradoxem, že přestože robotika a umělá inteligence si vždy kladly za cíl sejmout z lidstva odvěkou dřinu, mnoho úloh, jež získaly popularitu, se odehrává v oblastech, které spíš než úlevu od práce nahrazují kreativitu, umění a hravost lidské mysli.

Vzpomeňme např. porážku šampiona hry go Lee Se-dola v roce 2016 nebo dominanci nad ostatními šachovými programy na přelomu let 2017 a 2018 pomocí AI z dílny Google DeepMind metodou Monte-Carlo tree search, kombinovanou s umělými neuronovými sítěmi. Dalším příkladem jsou nyní populární generované obrázky pomocí DALL-E, automatické překlady a generování textu velkými jazykovými modely (LLM – *Large Language Model*), kde velký statistický model rozvíjí vstupní text tou nejpravděpodobnější cestou. Je fascinující, že daný text je v mnoha případech nejen správně gramaticky, ale odpovídá i požadavkům na sentiment, a to v mnoha jazycích tak, že se může zdát, že stroj, který jej generuje, skutečně „myslí“.

Generativní AI již v průmyslu své místo má a nejde pouze o pomoc programátorům PLC, kdy lze velké jazykové modely používat namísto hledání v obsáhlých manuálech, popř. je použít k ozřejmění cizího, komplikovaného programu PLC. Generativní AI umí vytvářet procesní diagramy UML (*Unified Modeling Language*) na základě poznámek z rozhovorů s manažery těchto procesů nebo profilovat uživatele a doporučovat jim specifický obsah.

Učení s učitelem (*supervised learning*)

Generativní AI staví na základech učení dopředných neuronových sítí s učitelem: neuronové síti předkládáme známé příklady dvojic jejích vstupů a požadovaných výstupů a metodou gradientního sestupu upravujeme váhy (často velké množství vah) tak, aby síť na daný vstup odpovídala požadovaným výstupem z této tzv. učicí množiny. Během učení sledujeme chybu na známé testovací množině dat (kterou však do učicí množiny nezahrnujeme) a síť považujeme za naučenou, když přestane tato chyba klesat. Tehdy ji lze implementovat do provozního prostředí. Dopředné neuronové sítě v každém kroku zpracovávají jen jeden vstupní vzor, na rozdíl od rekurentních neuronových sítí, které díky zpětné vazbě uvnitř sítě udržují stav, a jsou tedy vhodné pro zpracování posloupností vstupních vzorů. Generativní AI kombinuje tyto architektury neuronových sítí a metody jejich učení a rozšiřuje je např. o modelování pozornosti tak, aby bylo možné výslednou strukturu trénovat na obřích datech, a přitom aby získaný model co nejlépe statisticky reprezentoval vstupní data. Tento model je pak využit ke generování požadovaného obsahu.

Ještě před současným boomem generativní AI došlo kolem roku 2012 k průlomu v oblasti strojového zpracování obrazu (počítačového vidění), kdy implementace výkonných herních grafických procesorů (GPU, z angl. *Graphics Processing Unit*) umožnila efektivně trénovat neuronové sítě právě pomocí gradientního sestupu. Rozhodující operací, kterou tyto sítě pro zpracování obrazu používají, je konvoluce. Zásadní část výpočtu konvoluce obnáší násobení reálných čísel, pro která jsou GPU optimalizovány. Jejich použití k výpočtům potřebným při trénování sítí na velkém množství zejména obrazových dat umožnilo automatickým laděním konvolučních filtrů pro zpracování obrazu, dosud nastavovaných ručně, dosáhnout výsledků lepších než lidských. Tyto průlomy de facto nastartovaly opětovný zájem o neuronové sítě, které mnoho odborníků na AI doposud považovalo jen za hračku.

Tato situace není v historii AI nová. Zájem o AI přicházel a opadal ve vlnách. Již klasickým příkladem je skepse k neuronovým sítím v knize Perceptrons Marvina Minského a Seymoura Paperta z roku 1969, založená na tom, že jednoduchý simulovaný neuron nedo-

Neuronové sítě trénované na velkém množství dat na výkonných grafických procesorech, integrovaných na běžně dostupných herních grafických kartách, začaly na některých úlohách z počítačového vidění překonávat lidské schopnosti. V srpnu 2011 zvítězil algoritmus konvoluční neuronové sítě ze švýcarské laboratoře IDSIA (*Istituto Dalle Molle di Studi sull'Intelligenza Artificiale*, Lugano) v soutěži v rozpoznávání dopravních značek s chybou dvakrát menší než druhý tým složený z lidí a provádějících klasifikaci. Na konci září 2012 konvoluční neuronová síť z University of Toronto dosáhla chyby 15,3 % při rozpoznávání obrazů databáze ImageNet (www.image-net.org), čímž téměř dvakrát překonala do té doby nejlepší klasifikátor založený na jiné technologii než konvoluční neuronové síti. Od té doby se s výsledky rozpoznávání objektů v obrazech pomocí konvolučních neuronových sítí (kterým se začalo říkat *deep learning*, hluboké učení) „roztrhl pytel".

káže implementovat ani funkci XOR (protože jeden neuron umí provádět pouze lineární separaci dvou množin dat), tolik běžnou v normálních počítačích, tedy z těchto neuronů nebude možné sestavit univerzální výpočetní nástroj. Tato interpretace uvrhla neuronové sítě do nelibosti na téměř dvě dekády, přestože použití většího množství umělých neuronů zapojených ve vrstvách dokáže teoreticky modelovat libovolnou matematickou funkci s požadovanou přesností.

Posilované učení (*reinforcement learning*)

Přestože oblast řešení úloh, zejména v rozpoznávání obrazů neuronovými sítěmi, a na to navázaná generativní AI je zajímavá i pro průmyslové aplikace, zásadní pro průmyslovou výrobu jsou problémy optimalizace (hledání konfigurace parametrů, např. průmyslového procesu) a řízení (reakce řídicími povely na stav průmyslového procesu). Těmto zadáním odpovídají podoblasti AI, které zahrnují celou škálu optimalizačních metod a tzv. posilované učení (anglicky *reinforcement learning*, dále jen RL – v literatuře tento termín zpravidla odkazuje na typ problému, ale i na třídu metod, které jej řeší). Jde o reálné problémy, které jsou stále většinou doménou lidských operátorů. Člověk si sice umí představit intuitivní řešení mnohadimenzionálního optimalizačního problému, ale často je spokojen s řešením, které je od optima velmi vzdálené. To ale není zdaleka tak dobré, jako kdybychom jeho řešení přenechali specializovanému algoritmu. Ještě před dvaceti lety byly problémy spojité optimalizace kolem tisíce dimenzí (hledání optimálního nastavení např. tisíce konfiguračních parametrů) považovány za akademické a s dimenzí řádově vyšší považovány za obtížně řešitelné. Abychom byli schopni efektivně řešit problémy řízení pomocí RL, je zvládnutí mnohadimenzionální optimalizace zpravidla nutností.

RL počítá s existencí určitého prostředí (např. desková hra, dopravní systém, továrna), které je pozorováno tzv. agentem, jenž provádí kroky (tzv. akce) v čase tak, aby byl maximalizován součet očekávaných odměn (angl. *reward*) za určitou dobu (obvykle časově omezenou epizodu). Tato formulace problému RL je zajímavá nikoliv jen tím, že v principu popisuje mnoho reálných úloh, a to nejen v průmyslu, ale svým způsobem i veškeré lidské činnosti. Od učení s učitelem se ale liší zejména tím, že pro akce zpravidla neexistuje množina správných vstupů a výstupů, ze kterých se lze učit pomocí gradientního sestupu, a že agent svými akcemi prostředí ovlivňuje a zároveň musí reagovat na dynamicky se měnící stav prostředí, které pozoruje.

Z odměny přímo neplyne, co je správně, nicméně je to jediný dostupný signál, ze kterého se agent musí naučit, jaké přesně akce v daných krocích činit; jak přesně natočit volant v dané zatáčce, kterou právě vidí, o kolik snížit nebo zvýšit teplotu tavby skla, je-li v něm příliš mnoho bublin, o kolik zvýšit tlak v hydraulickém válci automatického bagru, který právě nakládá pásový dopravník kamenem. Ve většině případů, a to zejména na začátku učení, provádí agent objektivně špatné akce. Například bourá autem do zdi nebo šlape na brzdu při pokusu se rozjet. RL velmi dobře připomíná pokusy batolat prozkoumávat svět, přičemž se učí metodou pokus–omyl.

Zde přicházejí ke slovu tzv. black-box optimalizační algoritmy (pro hledání extrémů funkcí, u kterých neznáme jejich analytickou formulaci, a tedy ani nemůžeme odvodit její gradient pro nalezení extrému), které optimalizují parametry agentu a hodnotu odměny používají jako optimalizační kritérium. K reprezentaci agentu se velmi často používají opět umělé neuronové sítě, které zde slouží k rozpoznávání vzorů v pozorováních – k modelování prostředí. Pro optimalizaci parametrů (vah) agentu lze pak použít např. simulované evoluční algoritmy. Tento přístup má výhodu ve své jednoduchosti, kdy potřebujeme vlastně jen přístup k prostředí, reprezentaci agentu a vlastní evoluční algoritmus.

Tím, že akce generované agentem přímo ovlivňují data, která prostředí poskytuje formou pozorování, množina dat, která by bylo třeba generovat pro učení s učitelem, neúměrně roste. Tento problém je ještě komplikovanější v situaci, kdy akce nejsou jednoduché (např. vlevo, vpravo), ale spojité povely (např. vektor povelů pro motory průmyslového robotu).

Podstata problému posilovaného učení však přímo vybízí k překonání těchto teoretických překážek a k jeho využití na interakci s reálným průmyslovým hardwarem. Naučený agent může řídit roboty, nastavovat za běhu parametry výrobní linky, manipulovat s materiálem nebo ovládat celé průmyslové provozy. Je to podmnožina umělé inteligence, která si umí tzv. ušpinit ruce. Tím se RL zásadně liší od generativní AI. Už nejde o čistě softwarové řešení IT problémů, ale o skutečnou umělou inteligenci, která ovlivňuje okolní prostředí interakcí s ním. Tato AI má potenciál nahradit dělníka nebo operátora průmyslového provozu.

Průmysl 4.0, a co dál?

Zhruba před pěti lety jsme zaznamenali obrovský boom tzv. průmyslu 4.0, kdy se velké i malé podniky předháněly v tom, kdo rychleji digitalizuje, kdo dostane data ze svých soustruhů a vysokozdvižných vozíků rychleji do cloudu. Nyní, díky RL, přichází doba, kdy toto někdy až frenetické úsilí o digitalizaci průmyslu může začít nést ovoce. Není totiž snadné, mnohdy ani možné, učit agenty RL přímo v provozu. Představte si agenty učící se metodou pokus–omyl řídit vozidla v ulicích plných lidí nebo agent, který v každé epizodě svého učení zdemoluje pár oceláren. Agenty RL je tedy stále nutné učit v simulaci, na simulovaných prostředích. Zde s výhodou využijeme existenci tzv. digitálního dvojčete (z angl. *digital twin*), simulace prostředí, do kterého vypustíme simulovaný agent pouze pro potřeby učení. Stovky tisíc nabouraných aut za sekundu v simulaci nejsou problémem. Kromě toho může učení probíhat násobně rychleji (v roce 2016 jsme simulovali 180 let řízení auta na běžné herní grafické kartě za 24 hodin). Jestliže chápeme průmysl 4.0 jako prostředek k implementaci RL do průmyslu, má digitalizace velký smysl.

Má to ale několik úskalí. Zejména, posilované učení a optimalizace ještě nejsou dostatečnou „komoditou" jako *deep learning* nebo generativní AI. To činí návrh produktů, které vnášejí optimalizaci a řízení do průmyslu, výrazně složitější než aplikace komoditních oblastí AI. To je zřejmě důvod, proč se v posledních několika letech vyrojilo mnoho start-upů, které nabízejí promyšleně zabalený chatbot založený na generativní AI. Kromě těchto firem však již existují specializované společnosti, které nabízejí produkty pro řešení průmyslových problémů rozpoznávání obrazu, v nichž AI hraje klíčovou roli a zároveň jde o hotové výrobky přímo integrovatelné do průmyslové výroby. Cílový průmyslový zákazník tak nemusí stavět svůj tým pro AI, kupovat specializované komponenty, jako např. kamery, a z nich si samostatně navrhovat řešení. Místo toho si může objednat hotový produkt, který pokud možno co nejrychleji začlení do svého provozu.

Vzhledem k tomu, že při optimalizaci či posilovaném učení se zde vždy interaguje s prostředím, je složitější je implementovat, na rozdíl třeba od jazykových modelů, kdy jde o čistě softwarovou záležitost. Implemen-

tace v mnoha případech naráží na bezpečnostní požadavky (AI bude přece jen něco řídit) a případně velmi konzervativní seznamy povolených systémových komponent pro daný provoz. Nicméně doufáme, že se i zde bude využití průmyslové AI rozrůstat, neboť příležitostí pro řízení a optimalizaci je v průmyslu velmi mnoho.

Jednou se tak pravděpodobně dočkáme toho, že AI se bude používat všude tam, kde se uplatní její neúnavnost, konzistence a přesnost. Že v průmyslových provozech nahradí úmornou lidskou práci tak, jak slibuje robotika už od dob jejích počátků, a lidé se budou moci věnovat příjemnějším činnostem tak, jak si posteskla spisovatelka Joanna Maciejewska: „Chci, aby umělá inteligence prala prádlo a umývala nádobí, abych se mohla věnovat umění a psaní, a ne aby za mě umělá inteligence malovala obrazy a psala texty, abych se mohla věnovat praní prádla a umývání nádobí.“.

Vzhledem k tomu, že vlnu generativní AI máme, nebo budeme mít velmi brzy za sebou, může se dostavit jistá skepse, která v nejhorším případě vyústí v další „AI zimu“. Podobě jako v 70. letech minulého století nastalo zklamání z perceptronu může i dnes vyrůst mnoho zklamaných lidí, nejen investorů, ale i průmyslníků, neboť obecně chápaná generativní AI nedodá řešení všech problémů, zvláště ne těch skutečných průmyslových. Je to buď proto, že na AI podvědomě klademe nerealistické nároky (slíbili nám obecnou AI, která všechno řeší, a přitom to jen generuje texty a obrázky, zatímco v továrně musí dělníci pracovat u pásu dál ručně), nebo prostě nenasazujeme správný druh AI na provozy, kde se přímo nabízí použití jiné než generativní AI.

Je tedy třeba mít oči otevřené. Správně identifikovat problémy, se kterými nám AI může pomoci, a slepě nepodlehnout iluzi, že AI je vyřešený problém a teď už jen zbývá ji použít. Každý průmyslový problém je pro AI výzvou, když ne na poli algoritmů, tak při jejich využití v nehostinném prostředí průmyslového provozu, kdy např. v těžkém a horkém prostředí nelze instalovat žádnou běžnou elektroniku, natož citlivé počítačové vybavení. Přestože technika počítačového vidění a generování obsahu pomocí AI za poslední dekádu významně pokročily, je na průmyslu, aby ostatní zbylé problémy správně identifikoval a použil metody umělé inteligence, které jsou pro ně vhodné, zejména při optimalizaci a řízení prostřednictvím posilovaného učení.

Ing. Jan Koutník, Ph.D.

Jan Koutník je spoluzakladatel EVOPTIMA s. r. o., (www.evoptima.com) a odborník na umělou inteligenci v průmyslové inteligentní automatizaci. Po získání titulu Ph.D. na ČVUT působil jako výzkumník v IDSIA a byl spoluzakladatelem firmy NNAISENSE SA sídlící v Luganu ve Švýcarsku.

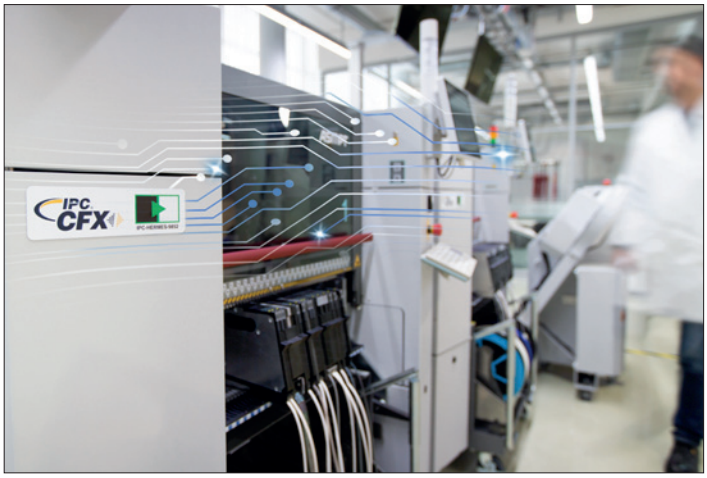
Automatická změna programu strojů řízená výrobkem

Na výrobních linkách desek elektroniky SMT s hardwarem a softwarem ASMPT lze nyní změny programu provádět plně automaticky, bez čtečky čárových kódů. Data z příslušné desky s plošnými spoji jsou předávána ze stroje na stroj prostřednictvím standardizovaného rozhraní IPC-HERMES-9852.

Průmyslový standard IPC-Hermes-9852 umožňuje elektronický přenos údajů o deskách plošných spojů SMT mezi stroji celé výrobní linky. Na základě toho mohou tiskárny pájecích past řady DEK TQ, inspekční systémy SPI Process Lens a osazovací stroje Siplace od ASMPT automaticky načítat příslušný výrobní program bez nutnosti číst čárový kód na každém stroji. Otevřené standardizované rozhraní také umožňuje automatické přizpůsobení přepravy šířce přepravovaných desek plošných spojů.

Již na začátku linky, při vykládání ze zásobníku, jsou data plošného spoje přenášena do tiskárny pájecí pasty buď prostřednictvím čtečky čárových kódů, nebo rozhraním IPC-Hermes-9852. Systém potom tato data použije k výběru programu pro výrobu. Je-li nutná změna, automaticky se načte nový výrobní program. V případě potřeby je povolán příslušný specialista, aby provedl přechod přes Works Operations. Works Operations je aplikace, která pomáhá řídit zaměstnance na výrobní lince a přidělovat jim odpovídající úkoly v souladu s jejich dostupností a dovednostmi.

Po vytištění je deska s plošnými spoji předána stroji Process Lens, kde je kontrolována pájecí pasta. Současně s fyzickým předáním desek systém SPI (*Solder Paste Inspection*) zkontroluje v řídicím systému linky údaje o deskách plošných spojů a zvolí nebo změní program kontroly. Poté jsou na řadě osazovací stroje, které stejně jako tiskárna pájecí pasty zkontrolují v řídicím systému údaje o desce, je-li třeba, automaticky upraví program nebo přivolají obsluhu přes Works Operations.



Obr. 1. Údaje o desce plošných spojů lze bez problémů předávat ze stroje na stroj prostřednictvím rozhraní IPC-Hermes-9852

„Doposud nebylo možné předávat údaje o desce s plošnými spoji ze stroje na stroj, ale každý stroj si musel tato data vyžádat individuálně od systému vyšší úrovně podle načteného čárového kódu,“ vysvětluje Volker Sindel, Senior Product Manager ve společnosti ASMPT. „Nyní lze data přenášet mnohem elegantněji prostřednictvím rozhraní IPC-Hermes-9852. Již není třeba čtečka čárových kódů na každém stroji. Díky standardizovanému rozhraní může být automatická změna programu podporována i systémy třetích stran v rámci inteligentní výrobní linky.“

[ASMPT Ltd.: *Die Leiterplatte steuert die SMT-Linie*. Březen 2024.]

(*Bk*)