

V Plzni učí počítače rozumět lidské řeči. Využívají při tom služby CESNET

18.7.2023 - | Západočeská univerzita v Plzni

Nedávno jsme oslavili 30 let sítě CESNET. Spolehlivé rychlé připojení je pro odborná pracoviště v akademickém prostředí nezbytným základem. Ke studiu a výzkumu ale potřebují i nástroje pro spolupráci a komunikaci, ukládání dat, snadnou identifikaci a přístupy a také pro náročné výpočty. To vše jim CESNET na své síti dokáže nabídnout.

Jedněmi z aktivních uživatelů těchto služeb jsou vědci a studenti Fakulty aplikovaných věd Západočeské univerzity v Plzni. V oddělení umělé inteligence katedry kybernetiky se zabývají výzkumem v oblasti metod strojového učení pro inteligentní rozhodování, klasifikaci a predikci. Zaměřují se na hlasovou komunikaci člověka s počítačem v přirozeném jazyce, počítačové vidění neboli computer vision a rovněž na aplikace zkoumaných metod v dalších oborech.

„Katedra kybernetiky je rozkročená hodně do šířky. V rámci oddělení umělé inteligence pracujeme na řečových technologiích, máme tým, který se soustředí na computer vision, zabýváme se také obecnou umělou inteligencí a strojovým učním. Na katedře máme kolegy, kteří dělají robotiku a teorii řízení, ale i odborníky na biokybernetiku,“ říká Jan Švec, vědec v oboru umělé inteligence.

Bez MetaCentra by to nešlo

Pro počítání neuronových sítí, pro trénování řečových modelů, ale i pro řadu dalších úkolů potřebují na katedře velký výpočetní výkon. Vhodné podmínky jim zajišťuje prostředí MetaCentra, které provozuje CESNET. *„Poskytujeme propojené výpočetní a datové zdroje pro řešení velmi náročných úloh, které typicky přesahují možnosti samostatných akademických pracovišť,“* uvádí Jan Hoidekr z oddělení distribuovaných výpočtů CESNET.

Na plzeňské katedře kybernetiky využívají služby MetaCentra od roku 2007, kdy si pořídili první cluster a zapojili ho do společné infrastruktury. Spolupráci si chválí. *„Zapojení výpočetních a datových zdrojů do jednoho clusteru nám dává velkou flexibilitu. Pro naše účely potřebujeme zdroje MetaCentra jen část roku, takže by se nám nevyplatilo vše kupovat a provozovat vlastními silami. V tom vidíme podstatnou výhodu sdílení těchto kapacit,“* konstatuje Jan Švec.

Právě optimalizace napříč sítí CESNET je jedním z hlavních smyslů MetaCentra. Velkou výhodou je i jednotné pracovní prostředí s jednoduchou správou uživatelů a bezpečným přístupem ke službám. Zájemci se přihlašují jedním uživatelským účtem. *„Dokážeme tak velmi snadno přidávat nové uživatele i z řad studentů, spravovat je a zároveň lehce vytváříme sdílené projektové prostory pro spolupráci,“* dodává Jan Švec.

Podle Jana Hoidekra je spolupráce s katedrou kybernetiky vynikající zkušeností, oboustranně velmi užitečnou: *„Jsme přesvědčeni, že podobná partnerství najdou uplatnění napříč celou akademickou obcí bez ohledu na konkrétní obor.“*

Titulky pro Českou televizi

Katedra kybernetiky využívá MetaCentrum při úkolech na poli řečové technologie, rozpoznávání řeči, syntézy řeči nebo zpracování a porozumění obsahu komunikace. Tyto technologie byly využity například pro pilotní výzkumnou fázi projektu titulkování živého vysílání České televize či v projektu audiovizuálního archivu holocaustu. Vývoji aplikací pro oba projekty předcházela náročná výzkum a výpočty.

Nejdřív bylo zapotřebí zpracovat obrovské množství zdrojových dat metodou strojového učení, díky kterým se modely učí porozumět lidské řeči. Jak takové učení probíhá, popisuje Jan Lehečka, vědecko-výzkumný pracovník: *„Model trénujeme pomocí audionahrávek, které poslouchá a zpracovává. Za velmi krátkou dobu je schopen vstřebat obrovské množství informací. V další fázi model učíme, jak daný problém řešit. Ukazujeme mu, co konkrétně chceme, aby dělal, a on se to z malého množství dat snadno doučí, protože už má hlubokou znalost jazyka nabranou v první fázi.“*

Jde o výpočetně náročné úlohy s velkým objemem dat. Trénink je proto možný pouze díky superpočítačům, které jsou součástí MetaCentra. Jen pro představu: jeden dataset má desítky terabajtů. *„Náš konkrétní model se trénoval asi dva týdny na čtyřech výpočetních grafických kartách NVIDIA A100,“* doplňuje Jan Lehečka.

Po natrénování modelu se připraví aplikace pro finální uživatele. V případě živých titulků byly poznatky z pilotní výzkumné fáze projektu úspěšně aplikovány do komerční sféry a v současné době vyvíjí a provozuje službu automatického titulkování pořadů České televize společnost SpeechTech, s.r.o., na své vlastní infrastruktuře.

System rozpoznává mluvenou řeč a přepisuje ji v reálném čase do textu. Cílem je přesný přepis, ale i přesné načasování, aby se text na obrazovce objevil ve chvíli, kdy je vyřčen. Problém nastává u pořadů, v nichž jsou různé šumy na pozadí nebo kde hovoří více lidí najednou. Tyto pořady je pak obtížné správně automaticky přepsat. Typicky jde o diskusní relace a sportovní přenosy. Tady musí nastoupit takzvaní stínoví řečníci, kteří zvukovou stopu přemluví tak, aby byla pro automat srozumitelná, a ten ji potom bez problémů přepíše.

Projekty z orální historie

Technologie rozpoznávání řeči a indexování byla použita i ve významném projektu z oblasti orální historie. Jde o audiovizuální archiv holocaustu, který zachycuje svědectví přeživších nahaná nadací USC Shoah Foundation. *„Hlavním úkolem bylo rozpoznání a přepis řeči ve specifické situaci a ve specifickém prostředí. Lidé zde často mluví v afektu, s emocemi, angličtina obvykle není jejich rodným jazykem, mají akcent... Z našeho pohledu jde tedy o náročnou úlohu,“* vysvětluje Aleš Pražák, specialista na rozpoznávání řeči.

Důležitá je přesnost při přepisu vlastních jmen a míst, protože podle nich je pak možné vyhledávat v archivu. Kdyby jména byla špatně přepsána, tak by i výstupy z archivu byly zkreslené. Cenná svědectví uložená v uceleném archivu s možností vyhledávání slouží badatelům a studentům ke zpracovávání konkrétních vědeckých témat. Prostřednictvím výpovědí jednotlivých lidí dávají dohromady příběhy a propojují informace.

V návaznosti na tento projekt byli výzkumníci z katedry kybernetiky osloveni Ústavem pro studium totalitních režimů (ÚSTR), který disponuje vlastním orálním archivem. K jednotlivým rozhovorům má ústav sady naskenovaných dokumentů, které doplňují nahrané výpovědi: vyšetřovací spisy, cestovní dokumenty, protokoly...

„Propojili jsme proto zpracování řeči a zpracování obrazu. Skeny jsme přepsali do textu, což umožňuje vyhledávání a zpracovávání informací,“ říká Jan Švec. Spolupráce s kolegy z ÚSTR jde nad rámec archivů orální historie. V minulém roce byl dokončen archiv skenovaných dokumentů NKVD/KGB, které byly získány z ukrajinských archivů vedených na československé občany.

Němci nám závidí

Na Fakultě aplikovaných věd jsou se službami CESNET nadmíru spokojeni. *„Jsem nadšen z toho, jak pracuje e-infrastruktura CESNET. V Evropě taková řešení úplně obvyklá nejsou,“* uvádí Jan Švec, *„například kolegů v Německu jsem se ptal, jakým způsobem počítají na clusterech. Odpověď byla, že mají někde nějaký cluster a když jej chtějí využít, tak musí dojít za svým profesorem, ten jim otevře*

účet a pak mohou počítat, ovšem pouze v určenou dobu. Diví se, když jim říkáme, že u nás máme národní infrastrukturu a jak to na ní funguje. My řekneme studentům: Tady si vyplňte přihlášku. Přidáme je do systému a za hodinu mohou ukládat data a počítat na výkonných serverech, prostě využívat veškerou infrastrukturu. Navíc je vše sdílené přes celou akademickou sféru. Když máme nějaký problém, tak se obrátíme na uživatelskou podporu a hned to vyřešíme.“

Kromě e-infrastruktury CESNET je fakulta zapojena i do dalších výzkumných infrastruktur, například LINDAT/CLARIAH-CZ, což je síť pro výzkumníky v lingvistice a příbuzných oborech. Výzkumná data tady ukládají do repozitáře zcela v souladu s principy otevřené vědy, které prosazuje projekt EOSC. *„Jakmile máme dataset, který stojí za to uveřejnit, tak ho publikujeme. Přidáváme i detailní metadata. Tuto cestu už máme relativně prošlápnutou a víme, jak s tím naložit. Naše zkušenosti budeme předávat i v pracovních skupinách projektu EOSC-CZ, do kterých jsme zapojeni,“* dodává Jan Švec.

Jak porozumět DNA

Na Fakultě aplikovaných věd Západočeské univerzity si neumí představit, jak by mohli pracovat bez infrastruktury CESNET a jejích služeb. Nic takového ale nehrozí. Mohou tak vymýšlet, jak nadstandardní kapacity využít pro nové úkoly.

„Umělá inteligence přesahuje už teď i do dalších oborů na fakultě. Jedním z nich je biokybernetika. Ani toto odvětví se neobejde bez velkého výpočetního výkonu a odpovídajícího úložného prostoru,“ upozorňuje Jan Hoidekr z CESNET.

Výzkumný asistent Fakulty aplikovaných věd ZČU Lukáš Kuhajda na něj navazuje: *„Zdroje dostupné v infrastruktuře CESNET používáme převážně na trénování neuronových sítí, konkrétně v projektu Regulace v DNA. Výstupy modelu můžeme okamžitě ověřit v laboratoři.“*

Potvrzuje tak, že vyvíjené technologie umí fakulta využít nejen pro porozumění textu, ale dokonce i pro porozumění DNA sekvencí či proteinových sekvencí. A to je teprve začátek...

<https://info.zcu.cz/clanek.jsp?id=5579>