MIT MLU

MIT Machine Learning Unit

MIT MLU je lokální výpočetní uzel poskytující podporu edge computingu pro umělou inteligenci a strojové učení.

MIT MLU se skládá z hardwarové a softwarové části, softwarové části jsou zkráceně označovány jako MUI – moduly umělé inteligence. Účelem MIT MLU je poskytnutí strojového učení především cílovým aplikacím pro správu informací.

MIT MLU přináší inovativní koncept strojového učení procesování různých typů podnikových informací. Zákazníkům tak přináší vysokou přidanou hodnotu v podobě redukce nákladů na interní procesy, jejich významného zefektivnění a v neposlední řadě v relativně rychlé implementaci s nízkým zapojením interních zdrojů zákazníka. Řešení je využitelné jak ve státní správě, tak pro soukromé subjekty.

MIT MLU v kombinaci s BPMN engine poskytuje možnost tvorby procesů, které jsou schopné samostatně pracovat s datovými vstupy až na té úrovni, že jsou schopné nahradit lidskou práci. Jedná se tak o univerzální řešení založené na vzájemné komunikaci autonomních procesů s cílovým systémem. Výsledkem jsou procesy, které jsou schopny samostatně rozpoznat a zpracovávat všechny druhy dat a následně rozhodnout, jaké kroky by měl cílový systém provést.

Klíčové vlastnosti

- Lokální výpočetní uzel poskytující podporu edge computingu pro umělou inteligenci a strojové učení
- Podpora AI algoritmů:
  - Decision Trees
  - Neural Networks (Computer Vision, Natural Language Processing, Tabular)
  - Support Vector Machines
- Rychlá implementace s nízkým zapojením interních zdrojů zákazníka
- Snadná integrace do existujícího prostředí
- Redukce nákladů na interní procesy a jejich zefektivnění
- Podpora pro Průmysl 4.0
- Rychlý vývoj a trénování modelů
- Snadné nasazení a správa modelů
- Špičkový výkon díky nejnovějším grafickým procesorům NVIDIA
- Robustní a centralizovaný monitoring a logování
- Podpora trénování modelů na více GPU zároveň
- Vysoká škálovatelnost výkonu s důrazem jak na trénink, tak i inferenci modelů
- Až 7 paralelně běžících učení
- GPU s výkonem až 3592,2 TFLOPS
- Maximální velikost modelu až 72 GB

EDGE COMPUTING

MIT MLU poskytuje hardwarové řešení lokálního výpočetní uzlu pro podporu edge computingu. Hardwarový modul s umělou inteligencí je možné snadno připojit do existujících nebo nově vznikajících infrastruktur, a to jak v podobě samostatného modulu, tak i v podobě propojených modulů do clusteru, pro dosažení vyšší výkonosti.

Výhodou edge computingu je tak především umístění výpočetní síly blíže ke koncovým zařízením a tím tedy snížení odezvy při komunikaci a objemu přenášených dat v porovnání se standardně využívaným cloudovým výpočetním centrem.

REDUKCE NÁKLADŮ NA INTERNÍ PROCESY A JEJICH ZEFEKTIVNĚNÍ

Strojové učení je součást Umělé inteligence, kdy se počítač učí jako člověk a sám provádí úkony. Algoritmy strojového učení identifikují vzory v datech a vytváří matematické modely, pomocí kterých dokážou vytvářet předpovědi pro zatím nepozorovaná data.

Strojové učení se dá použít takřka ve všech oblastech lidské činnosti, mezi odvětví aktivně využívající strojové učení spadají mimo jiné doprava, bankovní a finanční sektor, obchodní prodej, zdravotnictví ale i zemědělství.

Mezi hlavní přínosy tak můžeme zařadit mimo jiné:

- Zlepšení uživatelské přívětivosti – cílený obsah, inteligentní návrhy činností založené na předchozích akcích uživatele, inteligentní nápovědy a virtuální asistenti.
- Optimalizace procesů – inteligentní agenti (počítačové programy) s doménovou znalostí blížící se lidské dokážou se strojovou přesností analyzovat a optimalizovat výkony průmyslových výrobních strojů z jinak lidsky neanalyzovatelných dat.
- Dolování znalostí – strojové učení dokáže identifikovat vzory a struktury v nestrukturovaných datech a použít je pro užitečné přehledy a analýzy.
- Nižší náklady – inteligentní automatizace rutinních procesů dokáže uvolnit čas, prostředky i lidské zdroje.

MIT MLU je tedy navrženo tak, aby podporovalo všechny tyto oblasti využití a mnohé další.

Další vlastnosti

- Až 512 GB RAM DDR4 3200MHz
- CPU 3.6 GHz – 64 jader – 128 vláken
- 4 GPU s výkonem až 3592,2 TFLOPS
- Redundantní napájení pro vysokou dostupnost
- Až 7 paralelně běžících učení
- Velikost modelu až 72 GB

Flexibilita a škálovatelnost

- Možnost využití jako jednotlivého HW modulu, ale také jako spolupracujících jednotek (clusteru)
- Škálovatelnost výkonu s důrazem jak na trénink, tak i samotnou inferenci

Robustní a centralizovaný monitoring a logování

- Průběh inference nad modely
- Průběh výpočtu inference
- Rychlost výpočtu inference
- Komplexní monitoring využití HW

Dynamická registrace služeb

- Snadná integrace existujících systémů na MLU

AUTOMATIZACE VÝVOJOVÝCH PROCESŮ

Základním prvkem MIT MLU je softwarový modul strojového učení a umělé inteligence (MUI). Ten cílí na zjednodušení rutinních činností v oblasti správy informací pro koncové uživatele, kterým umožňuje efektivní práci v dodaných systémech. V současné době se jedná o dvě základní oblasti, těmi jsou automatizovaná klasifikace informací a automatizace procesů v aplikaci.

Oblast informačních systémů ovšem není jediným místem, kde je možné umělou inteligenci využít. Často skloňovaným pojmem posledních let se stal Průmysl 4.0, též označovaný jako čtvrtá průmyslová revoluce. Tímto pojmem se skrývá koncept chytrých továren, které převezmou opakující se a jednoduché činnosti, které byly doposud vykonávány lidmi. Je v něm využito metod strojového vnímání, autokonfigurace a autodiagnostiky a počítačového spojení strojů a dílů.

MODULARITA

MIT MLU nabízí moduly, které jsou schopné samostatně pracovat s datovými vstupy na té úrovni, že nahradí lidskou práci. Nejedná se tak o přímou automatizaci zpracování dat na základě definovaných parametrů, ale o univerzální řešení založené na vzájemné komunikaci autonomního modulu s cílovým systémem. Po prvotním zaučení jsou tak moduly MIT MLU schopny samostatně rozpoznat a zpracovávat všechny druhy dat a rozhodovat se, jaké následné kroky by měly být v systému provedeny a tuto informaci mu předávat.

Moduly umělé inteligence jsou tedy softwarové komponenty, které mají za úkol provádět operace strojového učení. Poskytují rozhraní pro jejich napojení do existujících i nově vznikajících softwarových řešení a pro svou práci využívají hardwarové vrstvy, z pohledu AI především TPU a GPU.

Implementace do stávajícího prostředí

MIT MLU nabízí velmi snadnou integraci v porovnání s výměnou celého softwarového řešení. Díky standardizovanému RESTful rozhraní je možné modul integrovat do téměř jakéhokoli existujícího řešení.

Mimo to také nabízí předem generované klienty v nejrozšířenějších programovacích jazycích, kteří umějí přímo komunikovat s klasifikačním modulem MIT MLU a stávající software tak jen využije jimi poskytnuté rozhraní.

Škálovatelnost

Systém umožňuje vyřešit velkým i menším podnikům problémy se zpracováním velkého objemu dat prostřednictvím škálovatelnosti na vyžádání.

Díky použitému řešení se zásadním způsobem zvyšuje flexibilita, výkonnost a kvalita způsobu práce informačních systémů s daty, což jsou aspekty potřebné pro úspěšné fungování podnikového prostředí.

MLU Type	Min. configuration	Max. configuration	Option 1	Option 2	Option 3
CPU	AMD Ryzen™ Threadripper™ PRO 5955WX - 16 Core/32 Threads - 64 MB vyrovnávací paměti L3	AMD Ryzen™ Threadripper™ PRO 5995WX - 64 Core / 128 Threads - 256 MB vyrovnávací paměti L3	AMD Ryzen™ Threadripper™ PRO 5965WX - 24 Core / 48 Threads - 128 MB vyrovnávací paměti L3	AMD Ryzen™ Threadripper™ PRO 5975WX - 32 Core / 64 Threads - 128 MB vyrovnávací paměti L3	AMD Ryzen™ Threadripper™ PRO 5975WX - 32 Core / 64 Threads - 128 MB vyrovnávací paměti L3
GPU	2x NVIDIA RTX A4000 - NVIDIA Ampere Architecture - 16 GB GDDR6 - 6144 CUDA Cores - 192 Tensor Cores - Single-precision performance 19.2 TFLOPS - Tensor performance až 153.4 TFLOPS	3x NVIDIA A40 - NVIDIA Ampere Lovelace Architecture - 48 GB GDDR6 - propustnost paměti 696 GB/s - 10752 CUDA Cores - 336 Tensor Cores - Single-precision performance 74,8 TFLOPS - Tensor performance až 1197,4 TFLOPS	7x NVIDIA RTX 4000 SFF - NVIDIA Ada Lovelace architecture - 20 GB GDDR6 - propustnost paměti 280 GB/s - 6144 CUDA Cores - 192 Tensor Cores - Single-precision performance 19.2 TFLOPS - Tensor performance až 306.8 TFLOPS	2x NVIDIA RTX A5000 NVIDIA Ampere architecture - 24 GB GDDR6 - propustnost paměti 768 GB/s - 8192 CUDA Core - 256 Tensor Cores - Single-precision performance 27.8 TFLOPS - Tensor performance až 222.2 TFLOPS 3x NVIDIA RTX A4000 - NVIDIA Ampere Architecture - 16 GB GDDR6 - propustnost paměti 448 GB/s - 6144 CUDA Cores - 192 Tensor Cores - Single-precision performance 19.2 TFLOPS - Tensor performance až 153.4 TFLOPS	1x NVIDIA RTX A6000 - NVIDIA Ampere architecture - 48 GB GDDR6 - propustnost paměti 768 GB/s - 10752 CUDA Cores - 336 Tensor Cores - Single-precision performance 38,7 TFLOPS - Tensor performance až 309,7 TFLOPS 5x NVIDIA RTX A4000 - NVIDIA Ampere Architecture - 16 GB GDDR6 - 6144 CUDA Cores - 192 Tensor Cores - Single-precision performance 19.2 TFLOPS - Tensor performance až 153.4 TFLOPS
RAM	64 GB DDR4 ECC	512 GB DDR4 ECC	256 GB DDR4 ECC	256 GB DDR4 ECC	48 GB DDR4 ECC
Storage	2TB RAID1	10TB zrcadlené (RAID1) úložiště - 2x SATA 10TB HDD	4TB zrcadlené (RAID1) úložiště - 2x SATA 4TB HDD	4TB zrcadlené (RAID1) úložiště - 2x SATA 4TB HDD	4TB zrcadlené (RAID1) úložiště - 2x SATA 4TB HDD
Power supply	Redundantní 1000 W PSU	Redundantní 2000 W PSU	Redundantní 1600 W PSU	Redundantní 1600 W PSU	Redundantní 1600 W PSU
Maximální počet paralelních učení na GPU	2	3	7	5	6
Maximální velikost modelu v GPU	16 GB	72 GB	16 GB	24 GB	48 GB
Celková dostupná kapacita pro inferenci v GPU	32 GB	256 GB	128 GB	128 GB	128 GB
Celková dostupná kapacita pro inferenci v CPU	32 GB	256 GB	128 GB	128 GB	24 GB
Cuda cores	12288	32256	43008	34816	41472
Tensor Cores	384	1008	1344	1088	1296
Single-precision	38,4 TFLOPS	224,4 TFLOPS	134,4 TFLOPS	113,2 TFLOPS	134,7 TFLOPS
Tensor performance	306,8 TFLOPS	3592,2 TFLOPS	2147,6 TFLOPS	904,6 TFLOPS	1076,7 TFLOPS
Vhodné pro	Minimální vhodná konfigurace pro vyrovnaný výkon jak v rámci trénování, tak v rámci inference modelů	Maximální vhodná konfigurace pro vyrovnaný výkon jak v rámci trénování, tak v rámci inference modelů	Konfigurace optimalizovaná pro efektivní trénování a inferenci méně komplexních modelů	Vyrovnaná konfigurace optimalizovaná jak pro efektivní trénink, tak inferenci většiny mainstreamově používaných modelů	Konfigurace pro trénování komplexních modelů, s vysokým výkonem pro inferenci modelu

Škálovatelnost CPU AMD Ryzen Threadripper PRO			
3945WX	3955WX	3975WX	3995WX
5955WX	5965WX	5975WX	5995WX

Škálovatelnost RAM DDR4			
3200 MHz	3000 MHz	2933 MHz	2800 MHz
2666 MHz	2400 MHz	2133 MHz	
ECC Support	ECC	Non-ECC	
Buffering support	Buffered	Un-buffer	
Možné kapacity a velikost paměťových modulů (celková kapacita/velikost jednoho modulu)			
32 GB / 8x 4 GB	64 GB / 8x 8 GB	128 GB / 8x 16 GB	256 GB / 8x 32 GB
512 GB / 8x 64 GB	1024 GB / 8x 128 GB	2048 GB / 8x 256 GB	

Škálovatelnost GPU

Grafické karty NVIDIA
A = NVIDIA Ampere Architecture
L = NVIDIA Ada Lovelace Architecture

1 slot, aktivní chlazení 1 Slot ac					
Model	**Architektura**	**RAM**	**CUDA Core**	**Tensor Core**	**NVLink**
RTX A4000	A	16 GB	6144	192	Ne
2 sloty, aktivní chlazení 2 Slot ac					
Model	**Architektura**	**RAM**	**CUDA Core**	**Tensor Core**	**NVLink**
RTX 4000 SSF	L	20 GB	6144	192	Ne
RTX A4500	A	20 GB	7168	224	Ano
RTX A5000	A	24 GB	8192	256	Ano
RTX A5500	L	24 GB	10240	320	Ano
RTX A6000	A	48 GB	10752	336	Ano
RTX 6000 Ada Gen.	L	48 GB	18176 TFLOPS	568	Ne
2 sloty, pasivní chlazení 2 Slot pc					
Model	**Architektura**	**RAM**	**CUDA Core**	**Tensor Core**	**NVLink**
A40	A	48 GB	10752	336	Ano
L40	L	48 GB	18176	568	Ne

Přípustné kombinace

1 slot **2 slot**	**0x**	**1x**	**2x**	**3x**	**4x**	**5x**	**6x**	**7x**
0x		X	X	X	X	X	X	X
1x	X	X	X	X	X	X		
2x	X	X	X	X				
3x	X	X						

Kontaktujte nás

M.I.T. Consulting

M.I.T. Consulting s.r.o., Baarova 1542/48, 140 00 Praha 4
Pobočka Olomouc: Třída Svornosti 192/23, 779 00 Olomouc
Telefon: +420 603 212 079
www.mit-consulting.cz

EVROPSKÁ UNIE
Evropský fond pro regionální rozvoj
Operační program Podnikání
a inovace pro konkurenceschopnost

Projekt byl podpořen ze strukturálních fondů EU v rámci Operačního programu Podnikání a inovace pro konkurenceschopnost (Aplikace VIII. výzva)

MIT MLU