AI MODELY TRÉNOVANÉ V METACENTRU

Seminář gridového počítání 2023 - MetaCentrum, 13. dubna 2023

Jan Lehečka
jlehecka@kky.zcu.cz
Katedra kybernetiky, ZČU v Plzni
NTIS, Fakulta aplikovaných věd, ZČU v Plzni

KATEDRA KYBERNETIKY

FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

AI NA KKY ZČU

Jazykové modely

- transformery (BERT, GPT, T5, …)
- klasifikace dokumentů, analýza sentimentu, porozumění textu
- dialogové systémy, chatboty

Řečové technologie

- rozpoznávání řeči (ASR), syntéza řeči (TTS)
- změna řečníka, konverze řeči, odstranění šumu, hlasový asistent

Počítačové vidění

- identifikace tváří, pózy ruky, znaková řeč
- ekologie a druhová ochrana, identifikace a detekce biologických druhů
- rozpoznávání ručně psaného písma, autonomní řízení
- medicína – sledování operačních nástrojů, analýza chování zvířat, 3D segmentace jater

KKY ZČU V METACENTRU

Cluster Konos

- 32x GPU NVIDIA GeForce GTX 1080 Ti, fronta **iti** (až 30 dnů)
- používáme často i jako JupyterLab pro náhradu osobních výpočetních stanic s GPU

Diskové pole

- vlastní diskové pole **plzen4-ntis** integrované do MetaCentra
- sdílená data KKY, veřejná data, osobní home

Používané technologie

- GPU clustery (Konos, Galdor, Zia, Adan)
- diskové pole brno12-cerit pro velké datasety (desítky TB)
- Python (TensorFlow, PyTorch, scipy, numpy, pandas, scikit-learn)
- vizualizace trénování – TensorBoard, Weights & Biases
- multi-node multi-GPU trénování (MPI)

JupiterLab + Singularity

- alternativa interaktivních jobů s GPU
- používáme vlastní image s balíky a nástroji v požadovaných verzích, obvykle:
  - PyTorch, TensorFlow, Keras
  - Transformers, Fairseq, Timm, HuggingFace
  - pandas, scikit-learn, matplotlib, librosa, opencv
  - ffmpeg, sox, Kaldi, Vim
- po spuštění jobu přijde e-mail s odkazem na přihlášení do JupyterLabu
  - vše běží v prostředí singularity jako standardní MetaCentrum úloha
  - GPU jsou dostupné a rovnou připravené k použití bez dalších modulů a balíků
  - všechny storage jsou dostupné
  - lokálně lze doinstalovat další balíky
- vycházíme vždy z připravených NGC image (`/cvmfs/singularity.metacentrum.cz/NGC`)
- snadný on-boarding nových členů týmu (studenti)

AI MODELY NATRÉNOVANÉ V METACENTRU

01

JAZYKOVÉ MODELY

BERT, GPT, T5, …

02

ŘEČOVÉ TECHNOLOGIE

Wav2Vec 2.0, SpeechT5

03

POČÍTAČOVÉ VIDĚNÍ

ViT, Swin, ResNeXt, …

04

APLIKACE

01

JAZYKOVÉ MODELY

BERT, GPT, T5, ...

NEJČASTĚJŠÍ POUŽITÍ LM

V NLP ÚLOHÁCH

Jako dodatečná informace o cílovém jazyce (úlohy ASR, OCR, strojový překlad, …)

NAŠEPTÁVAČE

Nápověda dalšího slova, pravopisná kontrola, …

GENERÁTORY TEXTU

Sumarizace textu, Q&A, chatboty, generátory článků, …

Trénování v Metacentru

Trénovací data

- webové stránky z projektu CommonCrawl
- pro češtinu 17 miliard slov
  - 120GB čistého deduplikovaného textu
  - 67 milionů normostran (police knih dlouhá 7km)

KKY modely

- **FERNET** (BERT 100M parametrů, 1 měsíc na 2xA100)
  - úlohy NLP, klasifikace dokumentů, analýza sentimentu, …
  - model zveřejněn pro nekomerční účely – https://huggingface.co/fav-kky/FERNET-C5
- **GPT-2** (100M parametrů, 2 měsíce na 2xA100)
  - generování textu, chatboty
- **T5** (200M parametrů, 2,5 měsíce na 2xA100)
  - opravy textu, chatboty

VELKÉ JAZYKOVÉ MODELY (LLM)

Velké transformery
> miliardy parametrů

Hodně trénovacích textů
> stovky miliard slov

Úloha trénování – predikce dalšího slova (tokenu)

Potřebný hardware – high-end GPU clustery

LLM v MetaCentru

Spuštění LLM

- **OpenAI** – **ChatGPT** & **GPT-3.5** & **GPT-4** (176B parametrů)
  - ***není možné spustit vlastní kopii modelu***
  - modely nejsou open-source, dostupné jen přes placené API

- **BigScience** – **BLOOM, BLOOMZ** (176B parametrů)
  - potřebuje 400GB GPU paměti
  - distribuovaný model přes 10xA100 nebo 9xA40)

- **Stanford** – **Llama**, **Alpaca** (7B parametrů)
  - potřebuje "jen" 16GB (1xA100, 1xA40)

- mnoho dalších – https://huggingface.co/models?pipeline_tag=text-generation

Natrénování vlastního LLM v MetaCentru?

- možné, ale výpočetně extrémně náročné
- odhadovaná cena trénování GPT-3: 12 milionů dolarů (energie 1,3GWh)

02

ŘEČOVÉ TECHNOLOGIE

Wav2Vec 2.0, SpeechT5

Wav2Vec 2.0

- state-of-the-art rozpoznávač řeči; end-to-end neuronová síť
- před-trénink (self-supervised)
  - data: 80 tisíc hodin neanotovaného audia (~9 let non-stop poslechu)
    - 10TB na disku
    - nutno dostat co nejblíž ke GPU
    - nelze před výpočtem stěhovat do /scratch
  - doba trénování: base model 14 dnů na 4xA100, large model 2 měsíce
  - cluster zia, fronta **gpu_max** (až 30 dnů), NFS scratch disk 360TB
- dotrénování – fine-tuning (supervised)
  - úloha ASR (12 hodin na 4xA100)
  - další úlohy: VAD, identifikace řečníka, …

SpeechT5

- aktuální trend v řečových AI technologiích
- univerzální multi-modální model audio + text
  - úlohy ASR, TTS, konverze audia, změna řečníka, …
- trénovací data: audio (130 tisíc hodin) + text (17 miliard slov)
- aktuálně běží před-trénování (odhad: 50 dnů na 4xA100)

03

POČÍTAČOVÉ VIDĚNÍ

ViT, Swin, ResNeXt, ...

POČÍTAČOVÉ VIDĚNÍ

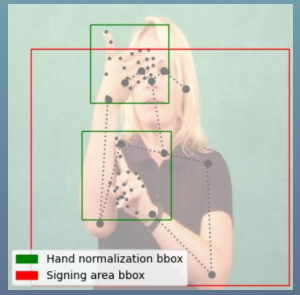
Potřebný hardware ⇒ high-end GPU clustery

dlouhá doba **Trénování**

převážně **Supervised** ⇒ Velké datové sady

Velikost a typ vstupu ⇒ Paměťová náročnost

PROJEKTY

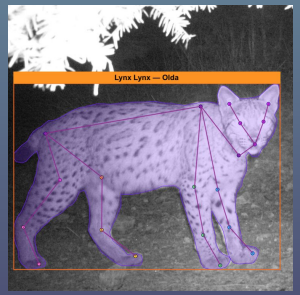


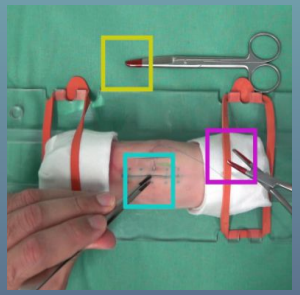
Odhad Pózy, Znakový Jazyk

Konvoluční sítě, Transformery a Zpracování sekvencí

Ekologie a druhová ochrana

CNN- a Transformer-based modely pro detekci, klasifikaci a odhad pózy



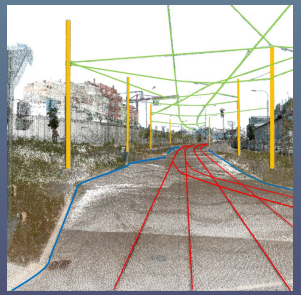


Medicína

Detekce a Sledování objektů, 3D Segmentace z CT

Autonomní řízení

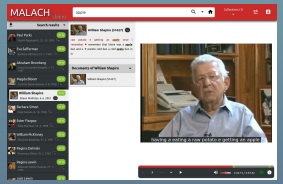
Zpracování lidarových a kamerových dat. Antikolizní systém.



04
APLIKACE

Aplikace modelů

- titulkování živých televizních přenosů ČT
- audiovizuální archiv MALACH
  - state-of-the-art výsledky, čeština, angličtina, němčina, slovenština
  - https://malach.kky.zcu.cz/



- audiovizuální archiv ÚSTR
  - https://naki-ustr.zcu.cz/
- archiv KGB ÚSTR
  - https://archivkgb.zcu.cz/

- rozpoznávání izolovaných znaků
  - https://huggingface.co/spaces/matyasbohacek/spoter-demo-test
- syntéza řeči a konzervace hlasu pro pacienty s laryngektomií

DÍKY VŠEM, ZEJMÉNA:

meta@cesnet.cz
Lukáš Hejtmánek
Jan Hoidekr
tým KKY

DĚKUJI
ZA POZORNOST

Jan Lehečka jlehecka@kky.zcu.cz

Lukáš Picek picekl@kky.zcu.cz

KATEDRA KYBERNETIKY

FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

NTIS